BAYESIAN HIERARCHICAL MODELING OF HIGH-THROUGHPUT GENOMIC DATA WITH APPLICATIONS TO CANCER BIOINFORMATICS AND STEM CELL DIFFERENTIATION

by

Keegan D. Korthauer

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

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2015

Date of final oral examination: 05/04/15

The dissertation is approved by the following members of the Final Oral Committee:
Christina Kendziorski, Professor, Biostatistics and Medical Informatics
Michael A. Newton, Professor, Statistics
Sunduz Keleş, Professor, Biostatistics and Medical Informatics
Sijian Wang, Associate Professor, Biostatistics and Medical Informatics
Michael N. Gould, Professor, Oncology

in memory of my grandparents
Ma and Pa
FL Grandma and John

ACKNOWLEDGMENTS

First and foremost, I am deeply grateful to my thesis advisor Christina Kendziorski for her invaluable advice, enthusiastic support, and unending patience throughout my time at UW-Madison. She has provided sound wisdom on everything from methodological principles to the intricacies of academic research. I especially appreciate that she has always encouraged me to eke out my own path and I attribute a great deal of credit to her for the successes I have achieved thus far. I also owe special thanks to my committee member Professor Michael Newton, who guided me through one of my first collaborative research experiences and has continued to provide key advice on my thesis research. I am also indebted to the other members of my thesis committee, Professor Sunduz Keleş, Professor Sijian Wang, and Professor Michael Gould, whose valuable comments, questions, and suggestions have greatly improved this dissertation. As an aspiring academic statistician and collaborative researcher, I could not ask for better role models.

Many thanks go out to the former and current members of the Kendziorski group. I truly appreciate the helpful critiques and suggestions on my work, as well as the positive and supportive atmosphere we created. I also thank my fellow classmates and Qualifying Exam study group for their camaraderie and scholarly advice.

I am very thankful for the opportunity to work with many collaborators on engaging projects throughout my PhD studies. In particular I thank Dr. Janet Rader and Dr. William Bradley of the Medical College of Wisconsin, and Dr. Li-Fang Chu and Dr. Ron Stewart of the James Thomson Lab. These experiences have cemented in me a great appreciation of the value of interdisciplinary collaborations.

Finally, I would like to express my utmost appreciation to all of my family and friends for their infinite support. I would not be where I am today without those who have always believed in me, even if I tried to convince them not to. Thank you to my partner in life Matthew for his unbounded optimism and encouragement, my parents for showing me the value of hard work and always building me up, my late grandparents for their wise words and warm hugs, my siblings for always being there to listen to me, and my friends for inspiring me.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

BAYESIAN HIERARCHICAL MODELING OF HIGH-THROUGHPUT GENOMIC DATA WITH APPLICATIONS TO CANCER BIOINFORMATICS AND STEM CELL DIFFERENTIATION

Keegan D. Korthauer

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

Abstract

Advances in the ability to obtain genomic measurements have continually outpaced advances in the ability to interpret them in a statistically rigorous manner. In this dissertation, I develop, evaluate, and apply statistical and computational methods to uncover novel insights in cancer bioinformatics as well as explore and characterize stem cell expression heterogeneity. A unifying theme of this work is the utilization of Bayesian hierarchical models to represent biological systems and processes. The first Bayesian hierarchical modeling framework integrates diverse sets of genomic information to identify cancer patient subgroups. The second framework identifies cancer driver genes based on somatic mutation profiles. Finally, the third framework identifies genes that exhibit differential regulation of expression across cell populations.

The recently developed survival-supervised latent Dirichlet allocation (survLDA) model, when applied to genomic data, is a method for identifying underlying groups ('topics') of co-occurring clinical and genomic features that are predictive of a time-to-event outcome. This framework is able to capture patient heterogeneity as well as incorporate many diverse data types, but the potential in utilizing the model for predictive inference has yet to be explored. This is evaluated empirically using clinical data from The Cancer Genome Atlas (TCGA) project. Using simulation studies, we examine the conditions under which the model shows potential for accurate patient-specific prediction. We show that in order to accurately identify patient subgroups, the necessary sample size depends on the size of the model being used (number of topics), the size of each patient's document, and the number of patients considered.

The second framework is a Model-based Approach for identifying Driver Genes in Cancer (MADGiC), which infers causal genes in cancer based somatic mutation profiles. The model takes advantage of external sources of information regarding background mutation rates as well as the potential for specific mutations to result in functional consequences. In addition, it recognizes that driver genes tend to have characteristic mutation patterns and leverages such information from the Catalogue Of Somatic Mutations In Cancer (COSMIC) database. As such, MADGiC encodes valuable prior information in a novel manner and incorporates several key sources of information that were previously only considered in isolation. This results in improved inference of driver genes, as demonstrated in simulation studies. Further advantages are illustrated in an analysis of ovarian and lung cancer data from TCGA.

The third framework provides a generative model for studying cellular heterogeneity in single-cell RNA-seq experiments. Specifically, it is known that gene expression often occurs in a stochastic, bursty manner. When profiling across many cells, these bursty gene expression patterns may be exhibited by multimodal distributions. We develop a Bayesian nonparametric mixture modeling approach that explicitly accounts for these multimodal patterns. Identifying these bursty expression patterns as well as detecting differences across biological conditions, which may represent differential regulation, is an important first step in many single-cell experiments. Through simulation we demonstrate that the modeling framework is able to detect regulation patterns of interest under a wide range of settings. Consistent with prior biological knowledge about the heterogeneity of a set of stem cell lines, we show that the framework detects fewer differentially regulated genes among undifferentiated lines compared to those involving differentiated cell types.

1 INTRODUCTION

Since the first microarray studies were published more than fifteen years ago (DeRisi et al., 1997), advances in the ability to obtain diverse measurements from the genome have continued to occur at a rapid pace. Technological improvements have resulted in new methodologies for and increased efficiency of sequencing, phenotyping, and genotyping. These developments continue to increase the ease (and decrease the cost) of probing the genome and phenome of an individual.

As genomic measurements become faster, cheaper, and more high-throughput, it is clear that larger and larger quantities of rich data are being generated. The sheer volume and complexity of presently available data contains a wealth of genetic information awaiting discovery through novel insights. One example is the emergence of projects such as The Cancer Genome Atlas (TCGA), which provides a wide array of sequencing, transcriptome and epigenome profiling, as well as phenotyping for large cohorts of patients of several different cancer types. This is the motivation behind two statistical frameworks presented here that focus on integrating a wide variety of well-established data types to further understand the molecular basis of cancer. Additionally, rapid technological advances introduce the continual need for rigorous statistical methods development as each new technology introduces new possibilities and unique challenges. This motivates a statistical framework to identify differentially regulated genes using a newly introduced protocol that measures RNA-seq at the single-cell level.

The statistical frameworks presented in this dissertation all make use of Bayesian

hierarchical modeling. This type of statistical methodology is particularly well-suited to the high-dimensional genomic setting in many ways. First, the flexibility of these models to accommodate many variables, via joint modeling or additional hierarchies, provides the opportunity for the integration of diverse data types. In addition, the incorporation of existing biological knowledge is often straightforward with the use of prior distributions or hyperparameters. Further, while the curse of dimensionality can be a barrier to statistical inference, empirical Bayesian techniques can be used to leverage commonality across genes or samples. Finally, evaluation of Bayesian hierarchical modeling is often straightforward by simulating out of a generative model. These advantages are exploited in the following three frameworks.

The first framework integrates any number of diverse data types with the aim of discovering meaningful cancer subtypes. First presented in Dawson and Kendziorski (2012), the survival-supervised latent Dirichlet allocation (survLDA) model is a method for identifying underlying groups ('topics') of co-occurring clinical and genomic features that are predictive of a time-to-event outcome, such as survival time. These clinical and genomic features are created by translating diverse data types into 'words', where each patient can be thought of as a collection of these words. Features that are predictive of survival outcome may be of great clinical utility in the area of personalized genomic medicine, which aims to make informed decisions for a particular patient's well-being in the presence of disease heterogeneity.

While the survLDA framework is designed to capture patient heterogeneity as well as incorporate many diverse data types, the potential in utilizing the model for predictive inference has yet to be explored. Toward this end, we carry out a

simulation study to examine the conditions under which the model shows potential for accurate patient-specific prediction. We show that in order to accurately identify patient subgroups, the necessary sample size depends on the size of the model being used (number of topics), the size of each patient's document, and the number of patients considered. Patient-specific prediction is also evaluated empirically on data from the TCGA ovarian cancer cohort. We also discuss further considerations, such as the possibility of alternative word creation strategies and the implications of word replication (the number of times a given word appears in a document). More details are provided in Chapter 2.

We introduce a second statistical method that focuses on data integration in the realm of cancer bioinformatics. Also motivated by the enormous amount of genetic data made publicly available by databases such as The Cancer Genome Atlas (TCGA) project and the Catalogue of Somatic Mutations In Cancer (COSMIC), we aim to identify mutated genes that contribute to the disease process (driver genes). The main challenge in identifying driver genes is that cancer genomes rapidly accumulate benign mutations (passengers) after tumor initiation, and thus the passenger mutations greatly outnumber the drivers. Further complicating matters, passenger mutations do not occur at a uniform rate throughout the genome.

Early statistical methods for identifying driver genes in cancer relied primarily on frequency-based criteria (i.e. identifying driver genes as those showing higher mutation rates than expected by chance). However, more recent studies have identified many other properties of drivers such as increased functional impact, enrichment for specific types of mutations, and highly structured spatial patterns. Though tools exist for

probing some of these factors one at a time, we have developed a unified framework to identify driver genes that incorporates all of these criteria. This is done by jointly modeling the mutational event with its functional impact, as well as incorporating the mutational enrichment and patterns as prior information. The method is called MADGiC, a Model-based Approach for identifying Driver Genes in Cancer, and shows substantially increased power (with a well controlled false discovery rate) compared to competing methods in simulation studies. Further advantages are demonstrated in case studies using data from the TCGA ovarian and lung cancer cohorts. For more details on this method, see Chapter 3.

The emergence of new technologies has also been a motivating factor in the development of statistical methods. Just in the past couple of years it has become feasible to probe the entire transcriptome of a single cell in a high-throughput manner. With this new single-cell RNA-seq technology, heterogeneity within a population of cells can be thoroughly characterized. This is in contrast to traditional RNA-seq experiments, which allow for the quantification of average transcript abundance on large collections of cells. This is of great interest since it is clear that transcription often occurs in a stochastic manner, which can result in multimodal distributions within gene (where some individual cells are on at a low level, and others are on at a high level, for example).

Identifying such multimodal genes and using them to characterize subgroups within and across biological conditions is an important first step in many single-cell RNA-seq experiments. Toward this end, we present a nonparametric Bayesian mixture model-based clustering approach. The approach facilitates the identification

of multimodal genes as well as genes that are differentially regulated (with differential expression, differential modalities, differential proportions within modal groups, or a combination thereof) across multiple biological conditions. Using simulated data we demonstrate that the modeling framework is able to detect regulation patterns of interest under a wide range of settings. Applied to a dataset of human embryonic stem cells, we show that the amount of multimodal genes varies by cell type and that fewer differentially regulated genes are detected in cell types that are more similar to each other in terms of differentiation state. Details of the approach are given in Chapter 4.

2 PREDICTING CANCER SUBTYPES USING SURVIVAL-SUPERVISED LATENT DIRICHLET ALLOCATION MODELS

2.1 Background

The goal of personalized genomic medicine is to utilize rich, diverse data types, such as high-throughput genomic information from multiple platforms, in the presence of disease heterogeneity to make informed decisions for a particular patient's well-being. To date, however, little has been accomplished in the way of utilizing this rich source of data to make individualized decisions in the clinical setting. While gene expression signatures have proven extremely useful in predicting outcomes in patients (for example breast cancer recurrence (Mook et al., 2007; Sparano and Paik, 2008) and colon cancer recurrence (Clark-Langone et al., 2010)), these approaches tend to categorize patients into a few groups and rely on a single source of genomic information. Personalized medicine, by definition, will require even more refined and specific categories, which will be more effective and informative if multiple sources of data are utilized.

Personalized genomic medicine seeks to fully characterize how genome and phenome heterogeneity relate to an outcome of clinical importance, such as response to treatment. Characterizing genome and phenome heterogeneity is of particular importance in cancer since the same disease can result from many different genomic events or abnormalities, and specific subgroups may have different treatment responses. If

we could catalog, for every patient, the specific genomic and downstream events that gave rise to cancer cells, this could be used to identify cancer subtypes. If we also knew how each of the subtypes responded to different treatments, it would be possible to make personalized treatment recommendations based on subtype identification.

Most existing methods for characterizing new cancer subtypes on the basis of genomic data do not integrate multiple sources of information. In general, subtype characterization techniques attempt to discover patterns in a genetic measurement that group subjects into distinct clusters using a single data type (e.g. gene expression). For example, unsupervised methods, such as hierarchical clustering (Hu et al., 2009; Mackay et al., 2011), K-means clustering (Tothill et al., 2008), and non-negative matrix factorization consensus clustering (Frigyesi and Höglund, 2008) have been used to group samples into subtypes from expression microarray data. Clustering algorithms often require some form of dimension reduction based on choosing those genes with sufficient magnitude and/or variation across the samples (Hu et al., 2009; Tothill et al., 2008; Shen et al., 2012), often result in molecular subtypes that are seemingly unrelated to established histological subtypes, and suffer from being unstable (i.e. slightly different settings, initial conditions, or selection criteria can produce different results). Hierarchical clustering suffers additionally due to the inherently subjective choice of where to cut a dendrogram (Mackay et al., 2011).

Supervised learning techniques are particularly suited to the task of classification once subtypes have been characterized. Classification can be based on distance to centroids (e.g. mean expression profiles of each subtype) (Mackay et al., 2011), or a host of other machine learning strategies such as support vector machines, decision

trees, or neural nets (Eddy et al., 2010). While some of these methods may lend themselves more naturally to integration of multiple data types, they often result in complex decision rules that are difficult to interpret. An alternative approach is relative expression analysis (RXA), which identifies a set of genes whose ranking of expression level best predicts subtype (Eddy et al., 2010). RXA is potentially more interpretable and reproducible than supervised machine learning methods with complex decision rules, but it is not clear how the method could integrate additional sources of genomic information, particularly if they are not ordinal.

A novel approach to the integration of multiple sources of high- throughput data is motivated by the success of a particular class of models used in text mining. The latent Dirichlet allocation (LDA) model as introduced by Blei et al. (2003) is a framework for characterizing documents as weighted combinations of topics, where topics are themselves made up of weighted combinations of words. Notably, high weight is given to frequently co-occurring words within a topic. An LDA model fit on a corpus can be applied to new documents to determine their document specific distributions over topics.

Consider the following toy example. Suppose a sales representative of a textbook publishing company is interested in selling books to university faculty and students, and has access to all e-mail correspondence from university accounts. Further suppose that this representative is statistically savvy and wants to analyze the e-mails using an LDA model. Three hypothetical topics are: topic 1 with high weight on words such as ‘publications’, ‘grants’, ‘conferences’, and ‘tenure’; topic 2 with high weight on words such as ‘advising’, ‘postdoc’, ‘conferences’, and ‘manuscript’; topic 3 with high

weight on words such as ‘homework’, ‘exams’, ‘bars’, ‘frisbee’. These topics provide the seller with some information regarding the types of books that may be interesting to this group of potential customers. Furthermore, the topics can be used to provide information on an e-mail’s sender. For example, a seller might infer that an e-mail with most of the words taken from the third topic came from a student, whereas an e-mail with most of the words coming from the second topic might be from a senior professor. An e-mail that had half of its words coming from topic three and half from topic one might be from an assistant professor who just joined an ultimate frisbee team that frequents drinking establishments. Given this type of information, particular advertisements can be better targeted to those most likely to respond to them. We are not interested in selling books, or in particular in identifying groups of words that co-occur together in e-mails (topics) and then classifying each e-mailer by those groups so that advertisements can be more specific and effective. We are, however, interested in a problem that is structurally very similar; namely, identifying groups of clinical and genomic features that co-occur in patients so that important patient subgroups can be identified and treatments may be better targeted toward the sub-groups most likely to respond to them.

Ideally, in our application, the patient ‘e-mails’ would contain comprehensive genomic and clinical information. This sort of catalog is not available, but Dawson and Kendziorski (2012) detail how one can be constructed from multiple sources of clinical and genomic data. They then develop and apply an extension of LDA that allows for supervision by time-to-event outcomes (survial-supervised LDA, or survLDA). Although existing statistical methods could be used to classify patients into subgroups

and identify predictive genomic aberrations once a patient's 'e-mail' is constructed, survLDA harbors two critical advantages. First, unlike classical models incorporating dimension reduction (Li and Luan, 2003; Ghosh and Yuan, 2010; Pang et al., 2010; Chen and Wang, 2009; Li and Li, 2004; Ma et al., 2007; Chen et al., 2010b), LDA is flexible enough to account for heterogeneity within the patient population. Rather than simply identifying predictive features common to the entire patient population, LDA characterizes complex interactions in these features, some of which may only apply to a subset of patients. Second, supervised LDA, unlike supervised clustering approaches (Dettling and Buhlmann, 2002; Li and Gui, 2004), identifies topics that are directly interpretable (e.g. this topic is associated with poor survival, so the collection of genomic aberrations contained within it are also associated with poor survival).

In addition to flexibility and interpretability, the LDA model also presents an opportunity for data integration. Constructing documents (patient 'e-mails') based on the presence of genomic aberrations transforms diverse genomic measurements to the same scale (words). This transformation allows for the combination of numeric data of varying scale (such as expression and methylation) as well as categorical data (such as SNPs). In addition, since we are not constrained to one particular level of experimental unit, such as a gene, non-genomic clinical information can easily be incorporated (e.g. treatments, stage, grade, and so on). Beyond existing genomic data and clinical covariates, the framework also makes room for new data types, given a word-generation scheme that outlines criteria for abnormal measurements.

In Dawson and Kendziorski (2012), survLDA was used to estimate groups of

co-occurring aberrations that displayed differential survival within the cohort, but the potential for patient-specific prediction was not investigated. In order to understand the circumstances under which patient-specific prediction is feasible for survival-supervised LDA models, utility for patient-specific prediction is here evaluated empirically on the TCGA ovarian cohort, and simulated data is fit with the survLDA model to assess classification accuracy in varying conditions. The survLDA model will be detailed in Section 2.2. Empirical results from applying survLDA, described in detail in Dawson and Kendziorski (2012), are summarized in Section 2.3. Section 2.4 derives predictors for the survLDA model, and predictive utility in the TCGA cohort is assessed in Section 2.5. A simulation study is outlined in Section 2.6, and limitations and issues for further study are discussed in Section 2.7.

2.2 The survLDA model

Here we detail the survival-supervised LDA model as proposed in Dawson and Kendziorski (2012). Consider a collection of D documents which contain N_i words, where $i = 1, ..., D$. Let the documents arise from a vocabulary set of V words, which contains at least the set of unique words in the corpus. Assume that the words are assigned to documents conditional on a set of K latent 'topics'. A topic is represented by a discrete distribution over the vocabulary and is parameterized by a length V vector τ_k, where $k = 1, ..., K$. Further, assume that each document i is represented as a mixture of the K latent topics, parameterized by the K-vector θ_i. Finally, let each document be associated with a time-to-event outcome T_i and censoring indicator δ_i which are associated with topic proportions through a survival regression model (here

we use the Cox proportional hazards model).

The system-wide parameters for the survLDA model are the topic V-vectors $\tau_{1:K}$, the K-vector Dirichlet parameter, the K-vector of survival regression coefficients β, and baseline hazard $h_0(\cdot)$. Given these values for document i, the N_i words and survival response T_i arise from the following generative process:

1. Draw topic proportions $\theta_i \sim Dirichlet(\alpha)$
2. For each of the N_i words, indexed by j
   a) Draw a topic assignment $Z_{ij}|\theta_i \sim Multinomial(1, \theta_i)$
   (Where $Z_{ij} \in \{1, \ldots, K\}$)
   b) Draw a word $W_{ij}|Z_{ij}, \tau_{1:K} \sim Multinomial(1, \tau_{Z_{ij}})$
   (Where $W_{ij} \in \{1, \ldots, V\}$)
3. Compute the K-vector $\bar{Z}_i$ s.t. $\bar{Z}_{ik} = \#\{Z_{ij} = k\}/N_i$
4. Draw a survival response $T_i|\bar{Z}_i, \beta, h_0$ from the survival function corresponding to a Cox proportional hazards model with hazard function $h(t|\bar{Z}_i) = h_0(t)\exp\{\beta'\bar{Z}_i\}$

Note that the regression coefficient β_k represents the effect of topic k on survival, where a negative parameter corresponds to an increase in survival and conversely a positive parameter corresponds to a decrease in survival. A variational expectation-maximization (EM) algorithm may be used to estimate the system-wide hyperparameters α, $\tau_{1:K}$, β and h_0 as detailed in Dawson and Kendziorski (2012). Specifically, the

fully factorized variational distribution of the latent variables $q_i(\theta_i, Z_{i,1:N_i})$ is chosen, such that

$$q_i(\theta_i, Z_{i,1:N_i}|\gamma_i, \phi_{i,1:N_i}) = q_i(\theta_i|\gamma_i) \prod_{j=1}^{N_i} q_i(Z_{ij}|\phi_{ij}) \tag{2.1}$$

where $\theta_i|\gamma_i \sim Dir(\gamma_i)$ and $Z_{ij}|\phi_{ij} \sim Discrete(\phi_{ij})$. A lower bound for the marginal log-likelihood (evidence lower bound) in terms of (2.1) is:

$$E_{q_i}[\log p(\theta_i, Z_{i,1:N_i}, W_{i,1:N_i}, T_i, \delta_i|\alpha, \tau_{1:K}, \beta, h_0)] - E_{q_i}[\log q_i(\theta_i, Z_{i,1:N_i})] \tag{2.2}$$

In the E-step, variational parameters $\{\gamma_i, \phi_{ij}\}$ are chosen to maximize (2.2) so as to approximate the marginal log-likelihood. The updates are given by

$$\gamma_i^{new} = \alpha + \sum_{j=1}^{N_i} \phi_{ij} \qquad \text{and}$$

$$\begin{aligned} \phi_{ijk}^{new} \quad \propto \quad & \exp\left[\Psi(\gamma_{ik}) - \Psi(\sum_{g=1}^{k} \gamma_{ig}) + \sum_{v=1}^{V} I(W_{ij} = v) \log \tau_{kv} + \delta_i \frac{\beta_k}{N_i}\right. \\ & \left. -H_0(T_i) \left(\prod_{m=j} \exp\left(\frac{\beta}{N_i}\right)' \phi_{im}\right) \exp\left(\frac{\beta_k}{N_i}\right)\right], \end{aligned}$$

where $\Psi(\cdot)$ denotes the digamma function and ϕ_{ijk} is normalized such that $\sum_{k=1}^{K} \phi_{ijk} = 1$. In the M-step, the system-wide hyper-parameters $\{\tau_{1:K}, \beta, h_0\}$ are chosen to maximize (2.2) summed over the entire corpus at the current values of the document-specific variational parameters (here we adopt the simplifying assumption as in Blei

and McAuliffe (2008) that α is fixed at $(\alpha_0/K, ..., \alpha_0/K)$ where α_0 is specified *a priori*). These two steps are iterated until convergence of all parameters is achieved. For details on derivation of these terms and specific update formulas for $\{\tau_{1:K}, \beta, h_0\}$, see the Appendix of Dawson and Kendziorski (2012).

2.3 Empirical results from TCGA application

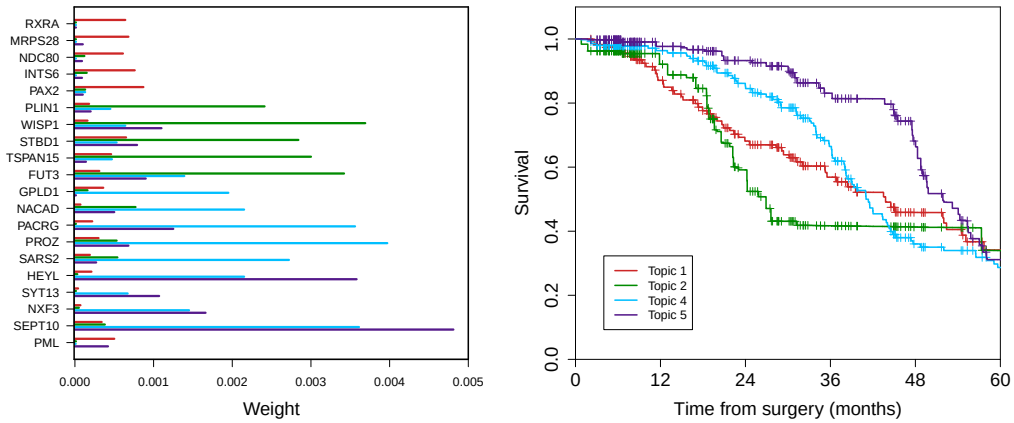
To illustrate how the survLDA model can be applied to clinical and genomic data, we briefly review the analysis in Dawson and Kendziorski (2012). The goal of this analysis is to discover subgroups of the patient population with co-occurring sets of genomic aberrations that are correlated with overall survival. With these sets in hand, we may gain valuable insight into the disease process by examining the individual aberrations within each set, as well as the potential for patient-specific prediction of survival given the most likely subgroup membership (see Section 2.4). The application of LDA-type models to genomics requires some discussion. When applying LDA to text corpora, words and documents are already defined and only minimal pre-processing (e.g. removal of stop words) is necessary. In contrast, it is not immediately clear how to move from numeric measurements of the genome to 'words'; we aim to create words wherever there is evidence for an aberration. In this way, we hope to characterize the topics based on meaningful sets of co-occurring genomic events.

Dawson and Kendziorski (2012) constructed documents from expression, methylation, and clinical covariates for 448 women in the ovarian cancer cohort of The Cancer Genome Atlas (TCGA) project (National Cancer Institute and National Hu-

man Genome Research Institute, 2011). Expression and methylation measurements were translated into 'gene' words in the following manner. For a given patient and expression or methylation measurement, if that measurement is deemed multimodal for the patient population (using the *Mclust* Model-Based Clustering method (Fraley and Raftery, 2002, 2006)) and the patient's measurement lies within the minor mode, a word is generated for the gene corresponding to that measurement. The clinical covariate of interest, adjuvant drug therapy, was also used to generate words. For relatively common drugs (given to at least 10% of the women), a word was added to each patient's document for each of the drugs administered. As a final step in document construction, words were filtered based on univariate association with survival, which resulted in a vocabulary size of 1264, and mean document size of 347 words.

The survLDA model was applied to these documents and their associated survival times (from all-cause mortality) using a Weibull model for the baseline hazard and assuming the existence of 6 free topics plus a background. A background topic was incorporated that contained high weight on only two drug words that were considered a priori uninteresting as they constituted inclusion criteria for the TCGA study (i.e. each patient in the TCGA received the two drugs). Of interest is whether patients who differ with respect to their highest weighted topics (collections of co-occurring aberrations and adjuvant therapy) also show marked differences in survival.

For each topic k, a topic-specific survival curve was estimated by weighting each patient's survival information by their estimated proportion of words coming from that topic (θ_{ik}). Comparing the curves from each of the 6 non-background topics, it



Left: A sample of words with high topic partiality among four selected topics, with each word's corresponding topic-specific weight on the x-axis. Colors correspond to topics displayed in the right panel. Right: Weighted Kaplan-Meier curves for 4 of the estimated topics.

Figure 2.1: Kaplan-Meier curves by topic and words with high topic partiality

was found that two had poorer than baseline survival and two had better. The two most extreme topics had a rather large difference in survival; the estimated 2-year survival for patients predicted to be composed exclusively of the 'best' topic was 92%, compared with 65% for patients predicted to be composed exclusively of the 'worst' topic. Figure 2.1 displays some of the characteristics of the fitted model. Examples of words that are partial to certain topics are highlighted in Figure 2.1 (left), where for example gene PLIN1 is much more highly weighted in topic 2 compared to topics 1, 4 and 5. From Figure 2.1 (right) we see that topic 2 is associated with decreased 2 year survival, suggesting that aberrations in this gene tend to occur more often in patients with shorter survival times. While these results identify patient subgroups

based on co-occurrences of groups of genomic aberrations and provide insight into which groups of aberrations commonly co-occur, clinical utility of the survLDA model ultimately depends on the potential for patient-specific prediction.

2.4 Prediction in the survLDA model

In this framework we are interested in estimating the topic distribution over words θ^* for a new patient's document $w_{1:N}$ given fitted parameters $\{\alpha, \tau_{1:K}\}$ for a training set. We assume that the new patient's document is generated from the same vocabulary as the training set. As with model fitting, the posterior mean of $\theta^*|w_{1:N}, \alpha, \tau_{1:K}$ is intractable and must be approximated via variational inference. This is similar to the procedure outlined in Section 2.2 except that here the variational parameters are updated as in unsupervised LDA (that is, all survival-related terms are dropped). The variational parameter updates are given by

$$\gamma^{new} = \alpha + \sum_{j=1}^{N} \phi_j \qquad \text{and}$$

$$\phi_{jk}^{new} \propto \exp\left[\Psi(\gamma_k) - \Psi(\sum_{g=1}^{k} \gamma_g) + \sum_{v=1}^{V} I(W_j = v) \log \tau_{kv}\right]$$

where again ϕ_{jk} is normalized such that $\sum_{k=1}^{K} \phi_{jk} = 1$. The posterior mean of $\theta^*|w_{1:N}, \alpha, \tau_{1:K}$ is then approximated by the expectation with respect to the variational distribution $E_q[\theta^*] = \frac{\gamma}{\sum_{g=1}^{K} \gamma_g}$. Note that we may also obtain an estimate of the posterior mean of the realized topic proportions $\bar{Z}^*$ in a similar manner, where $E_q[\bar{Z}^*] = \frac{1}{N} \sum_{j=1}^{N} \phi_j$.

Given these posterior estimates, measures related to topic membership can be predicted for the new patient. This may be done qualitatively with θ^* (e.g., "This patient is predicted to belong primarily to the second topic and survival for that topic is poor, hence prognosis is bad.") or quantitatively with $\bar{Z}^*$ (e.g., predicting median survival time using the parametric survival model). Specifically, the predicted p^{th} percentile of lifetime can be obtained by solving the following equation for $\hat{t}_p$

$$\exp\ \left[-H_0(\hat{t}_p)\ \exp\ (\beta' \bar{Z}^*)\right] = \frac{p}{100}.$$

2.5 Evaluation of prediction in TCGA cohort

To evaluate the utility of survLDA for patient-specific prediction, we split the TCGA cohort into a training and test set (75% and 25% of the cohort, respectively). Document creation and survLDA model fitting procedures as previously described were applied to the training set, again assuming a model with 6 free topics plus a background. The fitted survLDA model was used to predict topic membership for the test set, using the prediction approach described in the previous section. Results were evaluated on association between predicted membership in topics estimated to have a negative effect on overall survival for those women in the test set. That is, we compared the overall survival for those who were predicted to have majority weight on the topics estimated to be 'bad' with the overall survival for those who were predicted to have majority weight on the topics estimated to be 'good'.

Although the survLDA model identifies two groups of patients that differ significantly in survival based on estimated topic membership in the training set (log-rank

$p = 0.0005$), there is no difference in survival by predicted topic membership in the test set (log-rank $p = 0.428$). These initial investigations of predictive capability of this model for the TCGA cohort suggest that in this setting predictive capacity is limited. This problem could result from low sample size (number of patients), small document size (number of words), or the manner in which the words were constructed. The first two issues will be addressed in the next section, and the third will be revisited in Section 2.7.

2.6 Simulation study to assess predictive performance

LDA models are most commonly applied to text corpora, and although the size of the documents created for the TCGA application is rather typical, it is usually the case that a larger number of documents is used. For example, the original LDA paper (Blei et al., 2003) as well as the supervised LDA paper (Blei and McAuliffe, 2008) considers corpora sizes in the thousands. Though not exclusively used for large corpora, it is not clear how well model parameters can be estimated with fewer documents, as in the setting of the TCGA ovarian cohort. Thus it remains to be determined how many documents, and of what size, are needed to carry out patient-specific prediction. In terms of patient-specific prediction, we will focus on the task of classifying patients into the 'bad' topic(s) (topic k has a negative effect on survival when $\beta_k > 0$).

A simulation study to evaluate power to classify documents (patients) into the 'bad' topics was carried out using a K=7-topic survLDA model with effect sizes β =(-1, -0.6, -0.3, 0, 0.3, 0.6, 1), where 0 corresponds to a background (no effect)

topic. The number of topics and their effect sizes were chosen to be similar to the TCGA analysis in Section 2.3. The vocabulary was fixed at V=1000 unique words, which is also on the order of the vocabulary size in the TCGA application. The number of documents was varied among 100, 500, and 1000, with each containing either 250 or 500 words. In total, the design contains 6 different combinations of sample size and document size parameters. Words were generated according to the process detailed in Section 2.2, with hyperparameter α equal to a K-length vector where each entry equals $1/K$, and hyperparameter $\tau_{1:K}$ drawn from a Dirichlet(γ) distribution with γ as the V-length vector with each entry equal to 2. The γ prior was chosen in such a way that the observed $\tau_{1:K}$ values were consistent with empirical results from the TCGA application (this corresponded to approximately 25% of the total topic weight being applied to the top 100 words). Finally, uncensored survival times were generated from the Weibull survival model with shape and scale parameters of 2 and 0.04, respectively (Bender et al., 2005). These values were also chosen to represent the observed distribution of survival times in the TCGA cohort.

To evaluate predictive capability, the fitted topic-specific distributions over words $\tau_{1:K}$ were used to estimate document-specific distributions over those topics $\hat{\theta}$ on a set of 1000 independently generated documents, with size N corresponding to the respective training set. Summary statistics including misclassification rate, SSE of θ, and p-value of the log-rank test for difference in survival of two predicted groups were examined. The misclassification rate was computed for the task of classifying each patient into one of two groups based on the estimated porportion of words coming from 'good' or 'bad' topics. Specifically, the task is to predict whether or not a patient

has majority of weight on topics 4, 5 and 6 (if $\sum_{k=4}^{6} \theta_k * 1[\beta_k > 0] > 0.5$), since in this setting, topics 4, 5, and 6 all have positive coefficients (negative effect on survival). The sum of squared errors (SSE) of θ was defined as $\sum_{i=1}^{D^*} \sum_{k=1}^{K} (\theta_{ik} - \hat{\theta}_{ik})^2$, where $\hat{\theta}_{ik}$ is the estimated proportion of words for document i coming from topic k and D^* is the number of documents in the test set (1000). These quantities are averaged over a total of 200 replications.

Table 2.1: Mean (SD) for simulated 7-topic model

	D=100		D=500		D=1000	
	N=250	N=500	N=250	N=500	N=250	N=500
SSE θ	486.27	517.82	573.68	674.22	619.98	709.60
	(47.02)	(108.42)	(140.70)	(189.69)	(172.92)	(226.31)
Misclass.	0.324	0.286	0.236	0.221	0.201	0.214
	(0.067)	(0.102)	(0.091)	(0.109)	(0.089)	(0.119)

Results for the 7-topic model, displayed in Table 2.1, indicate some improvement in classification accuracy for increased sample size. With 500 documents, accuracy is around 75% for the task of classifying documents into two groups. This accuracy is as high as 80% for the case of 1000 documents with 250 words each, but drops to near 67% for the case of 100 documents with 250 words each. The SSE of the document-specific topic proportions θ_i show a different pattern, however. Within the ranges of parameters explored, SSE θ increased with an increasing number of documents in the training set and words per document. The p-value for the log-rank test for the difference in survival of the predicted 2-group classification was significant

in almost every run for each case (non-significant for 0-5% of runs, depending on case), indicating a significant difference in the overall survival for the two groups, even with about 20-30% of documents misclassified. The survival of the two groups in an example case is shown in Figure 2.2 (left), where there is a modest but consistent difference in survival for those with majority weight on the 'good' versus 'bad' topics.

To investigate whether these patterns persist with the use of fewer topics, and whether predictive performance is improved in this setting, an additional simulation study was carried out. This time, the number of topics was restricted to 3, with effect sizes $\beta = (-1, 0, 1)$. All other parameters were unchanged. Each model was again fit a total of 200 times and evaluated on a test set of 1000 documents of the same size as the documents in the respective training set. Here the misclassification rate was computed for the task of classifying each patient into either the 'worst' topic, or not the 'worst' topic, since there is only one 'bad' topic instead of three as in the previous case.

Table 2.2: Mean (SD) for simulated 3-topic model

	D=100		D=500		D=1000	
	N=250	N=500	N=250	N=500	N=250	N=500
SSE θ	188.35	191.72	63.54	58.85	36.03	24.43
	(147.91)	(282.94)	(144.83)	(172.44)	(50.48)	(6.73)
Misclass.	0.169	0.146	0.064	0.050	0.043	0.028
	(0.166)	(0.206)	(0.105)	(0.104)	(0.034)	(0.006)

The simulation results displayed in Table 2.2 show that under these conditions the

mean misclassification rate of 'worst' topic membership ranged from 2.8 to 16.9%, and an increased number of documents resulted in decreased mean and variation. Further, increased number of words per document resulted in decreased mean misclassification rate. A similar pattern was observed for the SSE of the document-specific topic proportions θ_i, except that with 100 documents, the SSE θ actually increased with more words per document, and so did its variation. Finally, the p-value of the log-rank test of the Cox proportional hazards model for predicted 'worst' topic membership was always significant for all cases except the 100 document, 500 word case where it was significant on average. The survival of the two groups in an example case is shown in Figure 2.2 (right), where there is a clear difference in survival for those with majority weight on the 'good' versus 'bad' topic.

These results suggest that under these hyperparameter settings and specific Weibull survival model, for only 500 to 1000 documents, there is reasonably good potential for patient-specific predictive capabilities under a 7-topic survLDA model ($\sim 80\%$). Additionally, a 3-topic survLDA model may be very predictive with at least 500 documents ($\sim 90\%$), with accuracy increasing with sample size and number of words.

2.7 Discussion

The task of characterizing genomic features that are predictive of clinical response is important for gaining insight into the disease processes, and the task of classifying patients into subgroups based on their particular genomic aberrations could have substantial impact on the field of personalized medicine. Few methods exist that can

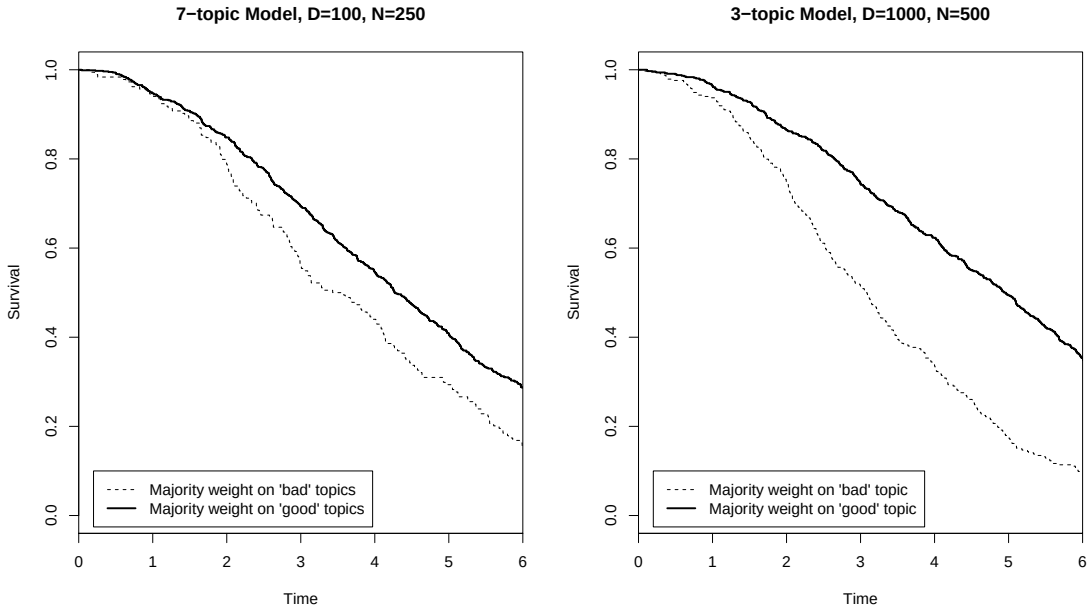


Figure 2.2: Kaplan-Meier curves for one simulated test set comparing those with majority weight on 'good' versus 'bad' topics. Left: the 7-topic case, with 100 documents and 250 words per document. Right: the 3-topic case, with 1000 documents and 500 words per document.

tackle both of these issues simultaneously in the presence of population heterogeneity. Latent Dirichlet allocation and its supervised extensions model documents as mixtures of underlying groups of features (topics). The underlying topics are themselves meaningful in that high-weight items tend to co-occur, and they also provide a mechanism for classification of new documents based on the estimated proportion of items drawn from each topic.

The settings under which accurate inference can be performed on new documents using survival-supervised LDA has not been well-explored. Simulation results presented here show that under some settings (e.g. a 3 topic model), accurate

patient-specific classification can be made with a modest sample size. Further, they provide some information about how prediction accuracy improves with increased document size and number of documents. They also show that with a larger number of topics, predictive performance declines, likely due to the increase in the number of parameters estimated.

As with all simulation studies, this one reflects a somewhat idealized scenario, thus a few limitations should be noted. First, the procedure assumes that we know the true number of underlying topics. This is not the case in practice, and there is no clear best method for choosing K. One technique that has been utilized is to pick a K for which the log-likelihood of the data given K is maximized (Griffiths and Steyvers, 2004). Another is to minimize the perplexity of a held-out set using cross-validation (Blei et al., 2003). Further study is necessary to fully understand the consequences of misspecification of topic number in this setting. Next, the effect sizes of topics on survival were chosen to be similar in magnitude to those empirically observed in the TCGA cohort. More investigation is needed into whether this choice reflects other situations, perhaps with different cancer types or different time-to-event outcomes. Finally, the simulation study does not deal with the complication of document creation.

Document creation in this setting involves translating a set of clinical and genomic data of various types into words. The word generation strategy employed in the TCGA application involved two filtering steps in order to weed out genes that looked similar for all patients as well as to select those with the most promise in terms of univariate association. We emphasize that the details of the chosen procedure

represent only one manner in which documents could be created, and briefly discuss some alternatives.

Another potential dimension reduction strategy could involve filtering genes based on variance level, keeping only those above a certain threshold level of variability across patients. This would accomplish a screening for genes that look most similar across patients, but would not necessarily be very successful at identifying subgroups among the patients. For example, variability could be high simply due to a few extreme outliers. In addition to alternate filtering strategies, there are numerous possibilities for generating words to mark extreme expression or methylation measurements. For example, cut-off points could be defined for high and low values beyond which merit a gene word. However, it is not clear how to choose a meaningful cut-off point for each gene. Choosing a percentile, say 10%, assumes that each gene in the set has two subgroups of identical size (one with high extreme values and the other low).

Two important factors regarding document composition that merit further study are word replication and topic 'strength'. Word replication, or the number of times a given word appears in a document or corpus, may affect inference in that more unique words tend to result in higher topic specificity. For both simulation studies there was little within-document replication. For cases with document size of 250 (500), there were approximately 214 (371) unique words on average. Across documents, each word in the vocabulary appeared on average in 25% (50%) of the documents for the N =250 (N =500) case. These settings could be varied by changing the number of words in the vocabulary relative to document size; with more words to draw from, a higher proportion of each document will be unique words and each word will appear

less often. In order to improve document creation strategies, it will be necessary to characterize the impact of word replication on predictive performance.

The 'strength' of a hypothetical topic lies in the set of words that are highly weighted for only that topic. Compared to the situation in which a set of words is highly weighted in more than one topic, presence of a high-strength set of words will increase specificity of topic assignment. While intuitively there exist many opportunities for 'strong' topics in text mining applications such as inferring latent topics of magazine articles, it is not necessarily so for documents constructed from the genome. For example, words such as 'Cabernet' and 'vineyard' may be highly weighted in topics about wine tasting, but unlikely to appear in many other topics if a corpus under study contains a mix of wine, sports, health, travel, and science articles. For gene words, as constructed in the TCGA application, we have no such prior that a certain gene is likely to appear in only one or two latent topics. This may result in lower topic specificity in documents constructed from genomic data in this manner.

In summary, survival-supervised latent Dirichlet allocation is a framework for characterizing clinically relevant cancer subtypes on the basis of diverse sources of clinical and genomic data. It allows integration of highly diverse data types and supervision by time-to-event outcomes of interest. In this study we have evaluated survLDA for its potential for patient-specific inference. Simulation results suggest that prediction accuracy can be greatly improved with increased sample size. Thus with a larger sample of patients and with additional insight into word generation procedures, survLDA has potential to translate diverse sets of 'omics' data into

meaningful and powerful clinical outcome predictions.

Supplementary notes

The results presented here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov/. The manuscript based on this work is published as a chapter in the edited volume Advances in Statistical Bioinformatics (Korthauer et al., 2013).

3 A MODEL-BASED APPROACH FOR IDENTIFYING DRIVER GENES IN CANCER

3.1 Background

Cancer is thought to result from the accumulation of causal somatic mutations throughout the lifetime of an individual. These cancer-driving mutations function by altering one of three broad classes of genes: oncogenes, which activate neoplastic activity; tumor-suppressor genes, which decrease a cell's ability to inhibit abnormal cell proliferation; and stability genes, which affect a cell's damage repair mechanisms (Kinzler and Vogelstein, 1997; Vogelstein and Kinzler, 2004). A first causal mutation in one of these classes of genes (or a rate-limiting combination thereof) leads to tumorigenesis, and subsequent causal mutational events drive tumor progression by providing a selective advantage to the cancer cells through positive selection (Vogelstein and Kinzler, 2004; Wood et al., 2007; Bozic et al., 2010; Vogelstein et al., 2013).

A major area of cancer research revolves around identifying these causal mutations, as doing so may provide new insights into gene function as well as potential targets for drug development. Methods for distinguishing genes with causal mutations ('driver genes') from those containing only background mutations ('passenger genes') which are irrelevant to cancer growth are also vital in making sense of the vast amounts of information being gathered from tumor sequencing studies such as The Cancer Genome Atlas project (`http://cancergenome.nih.gov/`) and the Cancer Genome Project (`http://www.sanger.ac.uk/research/projects/cancergenome/`).

A common approach to this problem is to identify genes that harbor significantly more somatic mutations than expected by chance. Methods using this approach, termed 'frequency-based' methods, rely on an estimate of a background mutation rate which represents the rate of random passenger mutations. Early frequency-based methods assumed a single background rate, constant across the genome and common to all samples (Ding et al., 2008). However, a number of features are known to affect mutation rate: mutation type (transition versus transversion), nucleotide context (which base is at the mutation site), dinucleotide context (which bases are located at neighboring sites to the mutation), replication timing of the region, and expression level of the gene. Further details are provided in Section 3.2.3.

In an effort to get a more accurate estimate of the background mutation rate, subsequent frequency-based methods have been developed that adjust for one or more of these factors. Sjoblom et al. (2006) account for nucleotide and dinucleotide context in searching for drivers of breast and colorectal cancer. MuSiC (Dees et al., 2012) accounts for mutation type and allows for sample-specific mutation rates; and in addition to these factors, Lawrence et al. (2013) (MutSigCV) also allow for the inclusion of gene-specific factors such as expression level and replication timing.

Although useful, a main limitation of methods based solely on mutation frequency is the inherent assumption that driver genes have relatively high mutation rates. This is often not the case. Indeed, with a few notable exceptions such as TP53 and KRAS, which show consistently high mutation rates in many cancers, most driver genes harbor surprisingly few mutations (Wood et al., 2007; Vogelstein et al., 2013). Consequently, additional criteria need to be incorporated into the search beyond

frequency if reliable driver gene identifications are to be made.

Recent developments provide at least two new sources for such information. The first are methods such as SIFT (first reported by Ng and Henikoff (2001), later updated by Kumar et al. (2009)), Polyphen (Adzhubei et al., 2010), and MutationAssessor (Reva et al., 2011) that incorporate information from sequence context, position, and protein characteristics to assess a mutation's functional impact. Recognizing the advantage of prioritizing genes by functional impact information, Gonzalez-Perez and Lopez-Bigas (2012) exploited bias in these scores as evidence of driver activity in their method OncodriveFM.

To account for both frequency and function, Youn and Simon (2011) (referred to hereinafter as YS) model mutation type, account for sample-specific mutation rates, and incorporate BLOSUM80 (BLOcks Substitution Matrix) alignment scores (Henikoff and Henikoff, 1992) into their approach. BLOSUM80 alignment scores reflect empirical probabilities associated with amino acid substitutions; and YS use these scores as a measure of functional impact. The idea is that if an amino acid substitution is rarely seen, it is likely detrimental. Although useful, power and specificity is gained by using methods such as those mentioned above that directly assess functional impact specific to the gene and mutation of interest (Ng and Henikoff, 2001).

In addition to advances regarding our ability to assess functional impact at the single nucleotide level, major advances have also been made with respect to our understanding of the spatial pattern of mutations within driver genes. Indeed, Vogelstein et al. (2013) recently noted that the best way to identify driver genes

is not through their mutation frequency as has often been done in the past, but rather through their spatial patterns of mutation. Vogelstein's claim is based on the recognition that oncogenes are often mutated recurrently at the same amino acid positions while tumor suppressor genes tend to have an over-abundance of truncating mutations (frameshift indels, nonsense mutations, or mutations at the normal stop codon). These characteristic patterns were not known just a few years ago, since they only become apparent with very large sample sizes. For example, even when looking at a dataset with close to 500 samples such as the TCGA ovarian, spatial patterning of mutations is not obvious (see Figure 3.1).

Recognition of such spatial patterns has been facilitated largely by a project to Catalogue Somatic Mutations in Cancer (Forbes et al., 2011). The so-called COSMIC project was initiated by the NIH in 2004 and is ongoing, with new datasets being added several times per year. The database currently contains mutation information for close to one million samples in over 40 tissue types, including data from several thousand whole exomes. Recent results from an integrative analysis across multiple cancers in COSMIC identified "highly characteristic and non-random" patterns of mutation that were not apparent when studying cancers by type in isolation (Vogelstein et al., 2013). In particular, results demonstrated that many known oncogenes consistently harbor mutations at relatively few specific amino acid positions, suggesting that oncogenic activity does not result from random mutation(s) in an oncogene, but rather requires a mutation in one of a few locations. OncodriveCLUST (Tamborero et al., 2013) was designed to exploit such evidence of positional clustering to identify oncogenes (but like OncodriveFM does not utilize other sources of information such

as frequency of mutation and functional impact). Non-random mutational patterns are also observed in known tumor suppressor genes, which tend to exhibit an over-abundance of protein-truncating alterations. Figure 3.1 provides a few examples. As we demonstrate, accounting for these non-random spatial patterns and abundance of truncating mutations improves both the sensitivity and specificity with which driver genes may be identified.

In addition to these recently characterized patterns of mutation, it is well known that alteration of DNA repair genes such as BRCA1 or BRCA2 leads to an increased accumulation of mutations (Birkbak et al., 2013). For those samples with mutations in known DNA repair genes, any global increase in mutation rate will be accommodated by our model's sample-specific background rate estimation (see Section 3.2.3.3).

Table 3.1: Summary of features of methods to identify driver genes

Methods	Mutation Type	Frequency	Gene-specific Background	Functional Impact	Spatial Patterning
MADGiC	✓	✓	✓	✓	✓
MuSiC	✓	✓			
YS	✓	✓		✓	
MutSigCV	✓	✓	✓		
OncodriveFM	✓			✓	
OncodriveCLUST	✓				✓

In summary, the most important sources of information to consider when identifying driver genes include: mutation frequency, mutation type, gene-specific features such as replication timing and expression level that are known to affect background rates of mutation, mutation-specific scores that assess functional impact, and the spatial patterning of mutations that only becomes apparent when thousands of samples

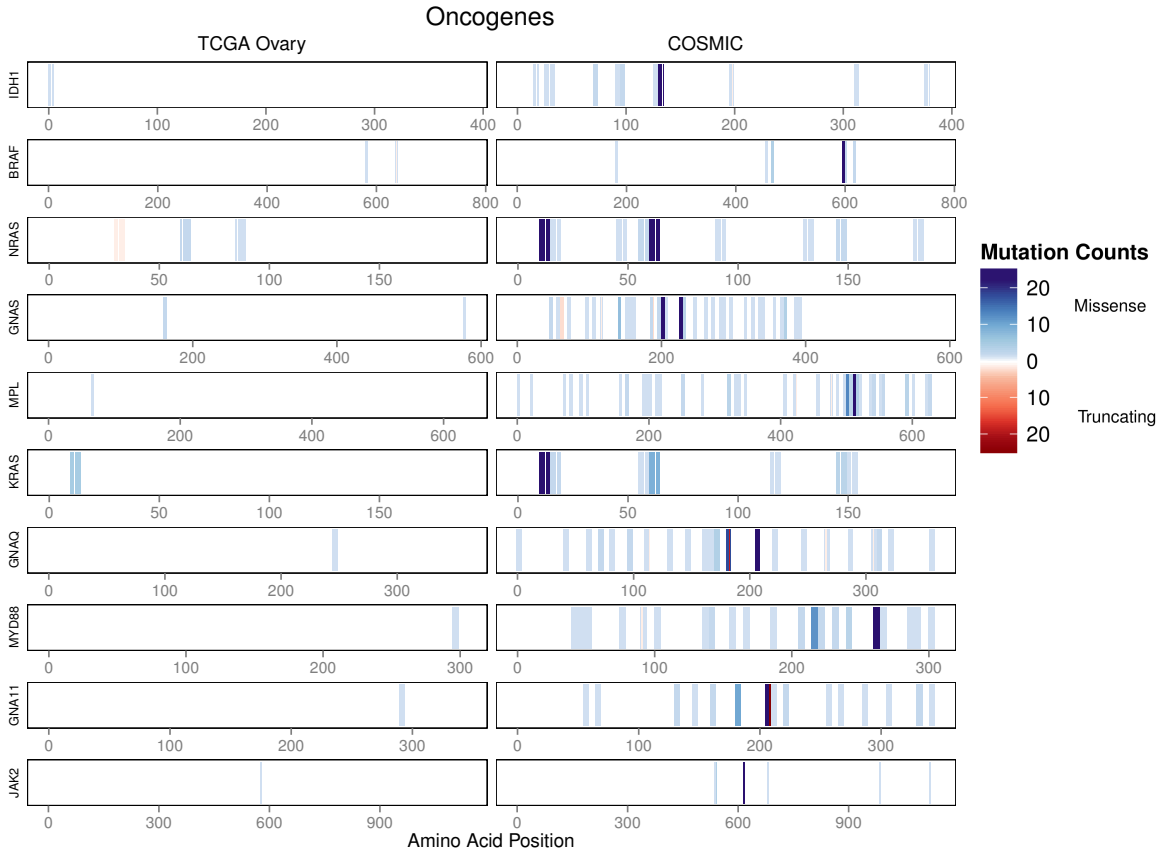


Figure 3.1: Oncogenic mutational patterns

Counts of samples with mutation by position and type for TCGA ovarian and COSMIC datasets. Displayed are ten genes with the lowest entropy in COSMIC (putative oncogenes) that have at least one mutation in TCGA ovarian. Blue represents missense mutations and red represents a location with at least one truncating mutation. Each vertical bar spans five amino acids and darker colors correspond to more mutations. For genes with more than 500 mutations, a random sample of 500 was plotted, and positions with more than 25 mutations are given the same color intensity as those with 25 mutations.

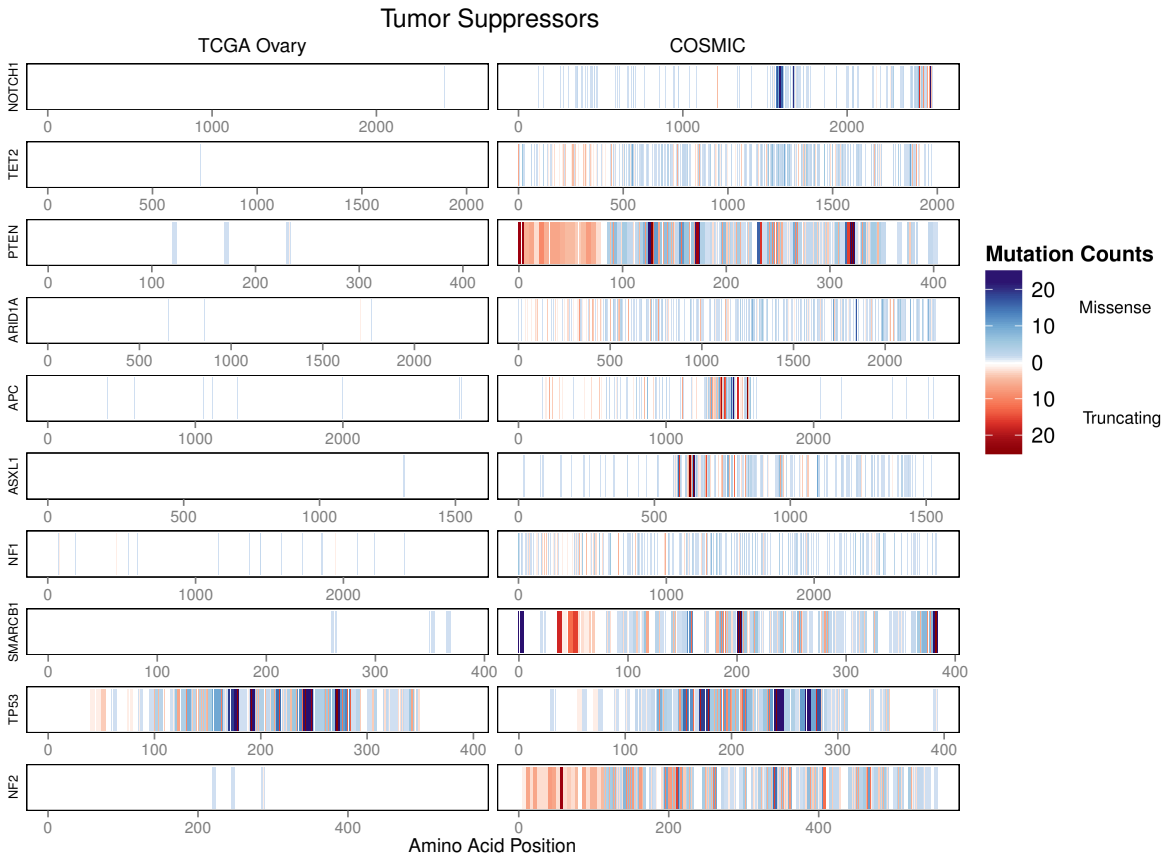


Figure 3.2: Tumor Suppressor mutational patterns

Counts of samples with mutation by position and type for TCGA ovarian and COSMIC datasets. Displayed are ten genes with the highest proportion of truncating mutations (putative tumor-suppressor genes) that have at least one mutation in TCGA ovarian. Blue represents missense mutations and red represents a location with at least one truncating mutation. Each vertical bar spans five amino acids and darker colors correspond to more mutations. For genes with more than 500 mutations, a random sample of 500 was plotted, and positions with more than 25 mutations are given the same color intensity as those with 25 mutations.

are considered. Previously developed methods incorporate many of these features (see Table 3.1 for an overview), but not all at once. In this chapter, we provide a unified empirical Bayesian Model-based Approach for identifying Driver Genes in Cancer (MADGiC) that utilizes each of these features. The Bayesian framework provides a natural way to leverage the mutational patterns observed in COSMIC as prior information and provides gene-specific posterior probabilities of driver gene activity. The posterior probabilities are informed by mutation frequency relative to a background model that incorporates mutation type and the gene-specific features mentioned above as well as position specific functional impact scores. Results from a simulation study in Section 3.3.1 suggest improved performance over currently available methods. Further advantages are demonstrated in an analysis of data from the Cancer Genome Atlas (TCGA) project (Section 3.3.2).

3.2 Methods

3.2 TCGA somatic mutation data

The TCGA somatic mutation datasets consist of exome somatic mutation calls between tumor tissue samples and normal samples (from either matched tissue or blood) of cancer patients and are freely available for download from the TCGA data download portal (available at https://tcga-data.nci.nih.gov/tcga/). Each somatic mutation is annotated for the sample(s) in which it occurs, its chromosome and position, the gene in which it is located, the allele found in the reference genome, the specific nucleotide change(s), and the type of mutation (silent, missense, nonsense, frameshift indel, in frame indel). The analysis presented here includes all available

ovarian and squamous cell lung cancer samples as of October 1, 2013.

3.2.1.1 Ovarian cancer

In the collection of 463 ovarian cancer samples there are 5,849 silent mutations (mutations that do not alter the amino acid sequence) located in 4,369 genes and 21,800 nonsilent mutations (mutations that cause a change in the amino acid sequence) located in 10,164 genes. The median (range) total number of mutations per sample is 60 (1-209). For silent mutations, the median (range) is 11 (0-51), and 41 (0-175) for nonsilent mutations. There is very little positional overlap of mutations across samples and only 62 genes have a nonsilent mutation in more than 10 samples.

3.2.1.2 Squamous cell lung carcinoma

In the collection of 178 squamous cell lung cancer samples there are 15,883 silent mutations located in 8,191 genes and 49,418 nonsilent mutations located in 13,238 genes. The median (range) total number of mutations per sample is 299.5 (4-3922). For silent mutations, the median (range) is 71.5 (0-1374), and 229 (3-2548) for nonsilent mutations. There is very little positional overlap of mutations across samples, but 649 genes have a nonsilent mutation in more than 10 samples.

The vast majority of squamous cell lung cancer cases are attributed to cigarette smoking (Kenfield et al., 2008). Since cigarette smoking is a known mutagen that results in an increased mutation rate as well as characteristic mutation signatures (Pleasance et al., 2009), it is plausible that the driver genes may differ between smokers and nonsmokers because they are subject to different mutational processes.

This is problematic since most methods assume there exists a common set of driver genes. To minimize the possibility of including non smoking-related cancer cases in the analysis, samples with a mutation rate below the 5th percentile that were also recorded as current or lifelong non-smokers at the time of data collection were excluded. This resulted in the removal of 10 samples.

3.2.1.3 Simulated data

To facilitate comparisons with existing methods, two types of simulations were considered. For SIM I simulations, 100 sets of random passenger mutations were obtained by shuffling the observed mutations from the TCGA ovarian dataset while preserving nucleotide context and mutation type, but ignoring gene-specific factors that affect mutation rate such as replication timing and expression level (see Figure 3.5 and Section 3.2.3 for details of these gene-specific factors). Each mutation in a given sample was randomly assigned a new position, drawn from all possible positions with the same reference nucleotide and mutation type. Next, 100 sets of thirty driver genes were randomly selected from the Cancer Gene Census (a set of nearly 500 genes that have been implicated in some form of cancer, manually curated by Futreal et al. (2004)) and nonsilent mutations were randomly added at three levels: either 3, 5, or 10 mutations (total across all samples; 10 genes at each level). The choice of 30 driver genes was made to be on the order of the median number of genes identified as drivers in the case studies. This resulted in a total of 100 unique simulated datasets. One hundred sets of random passenger mutations were obtained for SIM II in a similar way, but accommodating the dependence of mutation rate on replication

timing and expression level. Specifically, in this case each mutation in a given sample was randomly assigned a new position, drawn from all possible positions with the same reference nucleotide and mutation type in the same replication timing and gene expression categories. As in SIM I, 100 sets of thirty driver genes were randomly selected from the Cancer Gene Census and nonsilent mutations were randomly added at three levels.

This same process was repeated to generate 100 SIM I and 100 SIM II datasets using the TCGA lung data since it was suspected that some sample characteristics may influence the ability to detect driver genes. In particular, the lung dataset differs from the ovarian in that it has less than half the number of samples but more than twice the number of somatic mutations. In addition, the sample-specific mutation rates are much more heterogeneous in the lung dataset compared to the ovarian. This can be seen in the ranges of detected mutations per sample reported above. Note that the absolute number of mutations in the true driver genes is the same for both simulation sets, and consequently the relative mutation rate for driver genes in the simulated ovarian data is higher than that in the simulated lung data.

3.2 Driver gene model framework

Our primary aim is to prioritize genes that have been somatically mutated in cancer based on the likelihood that they are driver genes. A driver gene is defined as a gene harboring a mutation that provides a selective advantage to the cancer cell. The empirical Bayesian hierarchical mixture model framework we develop considers three sources of evidence for driver activity: (1) increased frequency of mutation

compared to a gene-specific background mutation model, (2) evidence of functional impact, and (3) a non-random spatial pattern of mutations. We detail the generative model framework in Section 3.2.2.1 and the calculation of the posterior probabilities in Section 3.2.2.2. Parameter estimation is discussed in Section 3.2.2.3 and the use of spatial pattern data to inform the prior probability of oncogenic activity is described in Section 3.2.2.4.

3.2.2.1 Generative model

Consider a single gene indexed by g, from a total of G genes having at least one nonsilent somatic mutation. Note that nonsilent mutations include missense mutations, frameshift indels, and in frame indels. Further consider an independent sample of size J, indexed by j, each with at least one nonsilent somatic mutation in one or more of the G genes. Let $\vec{X}_g = X_{1g}, ..., X_{Jg}$ be the vector of observed nonsilent mutation states of gene g for all samples (where $X_{jg} \in \{0, 1\}$ and $1 =$ one or more nonsilent mutations anywhere in the gene; $0 =$ no mutations in the gene). Next, let $\vec{S}_g = S_{1g}, ..., S_{Jg}$ be the vector of functional impact scores for mutations in gene g for all samples. Finally let $Z_g \in \{0, 1\}$ be the indicator that gene g exhibits driver activity.

We are interested in the posterior probability that gene g is a driver gene given the mutation status and impact score for that gene across J independent samples:

$$P(Z_g = 1|\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x}) = \frac{P(Z_g = 1)\prod_{j=1}^{J} P(S_{jg} = s_j, X_{jg} = x_j|Z_g = 1)}{\sum_{k\in\{0,1\}} P(Z_g = k)\prod_{j=1}^{J} P(S_{jg} = s_j, X_{jg} = x_j|Z_g = k)} \tag{3.1}$$

We assume that the presence of mutations in gene g and sample j depends on driver status. Specifically, $X_{jg}|\, Z_g = z \sim Bern((1-z)b_{jg} + zd_g)$ where $b_{jg} \in (0,1)$ is the background (passenger) mutation probability for sample j, gene g, and $d_g \in (0,1)$ is the driver mutation probability for gene g. To enforce that the driver mutation probability is at least as high as the average passenger mutation probability (i.e. that $d_g > \bar{b}_{\cdot g}$), we let $d_g \sim Beta(\alpha, \beta)$ truncated below at $\bar{b}_{\cdot g}$.

Likewise, we assume that the distribution of functional impact scores across all genes and all samples depends on driver status. Specifically, $S_{jg}|\, X_{jg} = 1, Z_g = z \sim (1-z)f^p + zf^d$, where f^p is the distribution of functional impact scores for passenger genes and f^d is the distribution of functional impact scores for driver genes. Note that we are assuming a common functional impact score profile for all driver mutations, and another for all passenger mutations, independent of mutation frequency.

3.2.2.2 Likelihood and posterior calculations

For J independent samples with observed mutation states $\vec{x}$ and scores $\vec{s}$, the data likelihood for gene g given driver status Z_g, driver mutation probability d_g, and estimates $\hat{b}_{jg}$, $\hat{f}^p$, $\hat{f}^d$ is

$$
\begin{aligned}
&P(\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x} | Z_g = z, d_g = \delta) \\
&\quad = \prod_{j=1}^{J} P(S_{jg} = s_j | X_{jg} = x_j, Z_g = z) P(X_{jg} = x_j | Z_g = z, d_g = \delta) \\
&\quad = \delta^{z \sum_{j=1}^{J} x_j} (1-\delta)^{z(J - \sum_{j=1}^{J} x_j)} \prod_{j=1}^{J} \hat{f}^d(s_j)^{x_j z} (\hat{b}_{jg} \hat{f}^p(s_j))^{x_j(1-z)} (1-\hat{b}_{jg})^{(1-z)(1-x_j)}
\end{aligned}
$$

Note that this probability depends on d_g, which is unknown. Thus, we calculate the prior predictive distributions $P(\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x} | Z_g = 1)$ and $P(\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x} | Z_g = 0)$ by averaging over the prior distribution of d_g. Then,

$$
P(\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x} | Z_g = 1) = \frac{B(\alpha^*, \beta^*)[1 - F_{(\alpha^*, \beta^*)}(\bar{b}_{.g})]}{B(\alpha, \beta)[1 - F_{(\alpha, \beta)}(\bar{b}_{.g})]} \prod_{j=1}^{J} \hat{f}^d(s_j)^{x_j}
$$

$$
P(\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x} | Z_g = 0) = \prod_{j=1}^{J} (\hat{f}^p(s_j) \hat{b}_{jg})^{x_j} (1 - \hat{b}_{jg})^{1 - x_j}
$$

where $F_{(\alpha,\beta)}$ is the cumulative distribution function of the beta distribution with shape parameters (α, β); B is the Beta function; $\alpha^* = \sum_{j=1}^{J} x_j + \alpha$; and $\beta^* = J - \sum_{j=1}^{J} x_j + \beta$. Then the final form of the posterior probability is easily obtained from Equation 3.1.

3.2.2.3 Parameter estimation

We use the background mutation model that will be described in Section 3.2.3 to get an empirical Bayes estimate of b_{jg}. Recall that the global hyperparameters α and β govern the prior probability that a driver gene is mutated and consequently they are estimated using the method of moments from tissue-specific mutation data of known cancer genes (from the Cancer Gene Census (Futreal et al., 2004)) in COSMIC. To avoid overfitting the model, any samples included in a dataset of interest should be removed prior to estimation of the hyperparameters. Here, for example, TCGA ovarian and lung samples were removed and the fitted values were ($\hat{\alpha} = 0.15$, $\hat{\beta} = 6.60$) for ovarian and ($\hat{\alpha} = 0.27$, $\hat{\beta} = 5.83$) for lung. From the plot of the fitted distributions in Figure 3.3, we see that the distribution of probabilities for the lung dataset are shifted slightly to the right. This is consistent with overall higher somatic mutation rates in squamous cell lung cancer samples compared to ovarian cancer samples.

Note that we also investigated the use of genes annotated as 'High Confidence Drivers' by Tamborero et al. (2013) instead of those in the Cancer Gene Census to obtain the estimated hyperparameters. The resulting fitted values for the ovarian cohort were ($\hat{\alpha} = 0.17$, $\hat{\beta} = 6.73$). The difference is small in part due to considerable overlap in the genes considered among the two sets. Here, there were 144 genes from the High Confidence Driver set that were mutated in ovarian tissue samples in COSMIC (after TCGA samples were removed); 83 of those 144 are also included in the Cancer Gene Census. The slight change in hyperparameter estimates did not alter the set of genes with posterior probability greater than 0.95. For a more extensive evaluation of the sensitivity of the model to hyperparameter specification, see Section

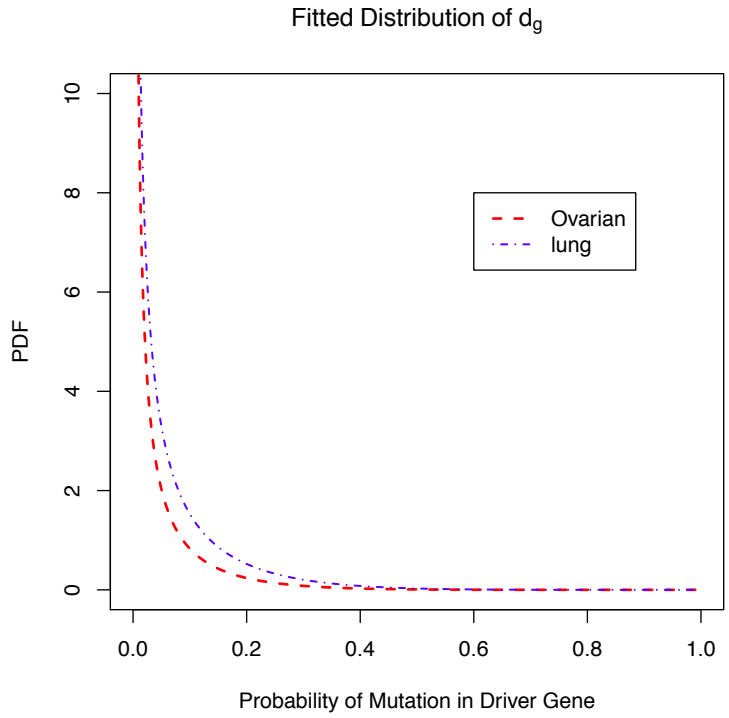


Figure 3.3: Fitted prior distributions of d_g for TCGA ovarian (red) and TCGA lung (blue)

3.3.4.

To assess functional impact, we use SIFT (Sorting Intolerant From Tolerant) scores from Liu et al. (2011), which range from zero to one, transformed such that scores closer to one represent high impact (Kumar et al., 2009). If there is more than one nonsilent mutation in gene g for sample j, we let S_{jg} take the value of the maximum functional impact score for all mutations in the gene. If there are no nonsilent mutations in gene g for sample j, we let $S_{jg} = -1$. To estimate $f^d(\cdot)$, the distribution of functional impact scores for driver genes, we first obtain SIFT scores for a random sample of nonsilent mutations, generated by shuffling the observed

mutations subject to the constraints of the background mutation model. We then estimate $f^d(\cdot)$ using nonparametric spline regression on the ratio of the simulated null to the observed full distribution $f(\cdot)$ of scores across bins of the score range, a technique used by Efron et al. (2001) to estimate the non-null distribution of z-scores in the analysis of gene expression microarray experiments. Specifically, 50 equally spaced bins and a natural spline with 5 degrees of freedom were used to estimate the probability that an observation in a given score bin comes form the null distribution (entire set of random passenger mutations) versus the full distribution (observed dataset) of scores. Though our functional impact score of choice here is SIFT, this non-parametric approach accommodates other available functional impact scoring schemes. The estimated full, null, and non-null curves for the TCGA ovarian and lung datasets are shown in Figure 3.4.

3.2.2.4 Quantifying gene-specific mutation patterns

Motivated by the fact that genes showing a random pattern of mutations across cancers in COSMIC are less likely to be drivers than those showing concentrated mutations (more likely to be oncogenes) or those showing an overabundance of protein-truncating mutations (more likely to be tumor suppressors) (Vogelstein et al., 2013), for every gene g we calculate a prior probability of driver activity ($P(Z_g = 1)$) using all mutations observed for that gene in COSMIC (excluding TCGA ovarian and lung cancer cases and only including data from whole-gene screens).

To quantify evidence of concentrated mutations, for each gene we calculate its positional entropy compared to a random distribution of mutations across all

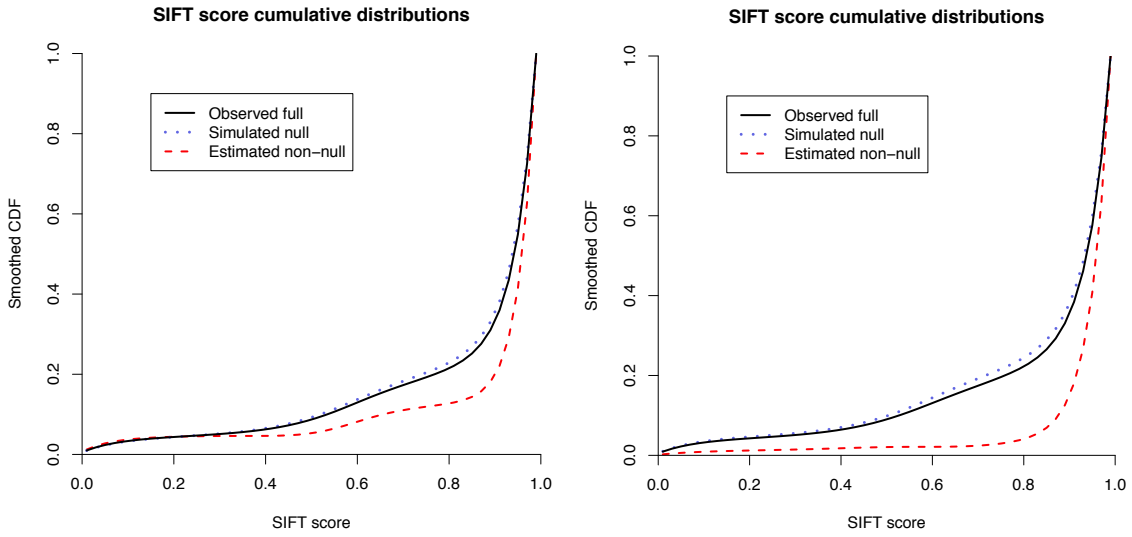


Figure 3.4: Fitted SIFT score distributions for TCGA ovarian (left) and TCGA lung (right)

amino acids. Genes with low entropy are ones with highly concentrated mutations. Specifically, we define a gene as showing evidence of oncogenic activity if it has less than 75% of maximum positional entropy, defined as the observed entropy divided by maximum entropy, where the maximum entropy is defined as the entropy when all mutations are evenly distributed across all amino acids. Observed entropy is the Bayesian plug-in estimator of the Shannon entropy, calculated from the observed counts of mutations at each position in the gene using the Dirichlet-multinomial pseudocount model (Hausser and Strimmer, 2009).

The entropy calculations are performed as follows: Let H be the Shannon entropy:

$$H = -\sum_{k=1}^{p} \theta_k log_2(\theta_k),$$

where p represents the number of amino acid positions in a given gene, and θ_k represents the proportion of mutations in the gene that occur at position k. Estimation of θ_k is performed under the Dirichlet-multinomial pseudocount model such that

$$\hat{\theta}_k = \frac{m_k + a_k}{\sum_{k=1}^{p}(m_k + a_k)},$$

where m_k represents the number of mutations occurring at position k and a_k represents a pseudocount (here we use the pseudocount corresponding to the minimax prior so that $a_k = \sqrt{\sum_{k=1}^{p} m_k}/p)$. In the calculation of the observed entropy H_{obs}, the observed number of mutations at each position m_k is used. In the calculation of the maximum entropy H_{max}, the number of mutations at each position when the mutations (same total as observed) are randomly spread over all p positions m'_k is used. Finally, the proportion of maximum entropy is taken as

$$E = \frac{H_{obs}}{H_{max}}.$$

Smaller values of E represent increasing oncogenic activity; here we consider values of E less than 0.75 as evidence of low entropy and we assign π_0 to some threshold value T when $E < 0.75$. The threshold value used in MADGiC ($T = 0.5$) was chosen based on a sensitivity analysis that demonstrated little variability in the number of driver genes identified for a reasonable range of values (refer to Section 3.3.4).

Similarly, we test each gene to see if it has a significantly higher proportion of truncating mutations than the proportion of truncating mutations observed over all genes. Genes with significantly low entropy or a significantly high proportion of

truncating mutations ($p < 0.05$) are assigned a higher prior probability of oncogenic activity ($P(Z_g = 1) = 0.5$); otherwise $P(Z_g = 1) = 0.01$. Examples of genes with oncogenic or tumor suppressor mutation patterns are shown in Figures 3.1 and 3.1, respectively. Again, the specific values of 0.5 and 0.01 here are arbitrary, but empirical sensitivity analyses (see Section 3.3.4) demonstrated little variability in results for values between 0.25 and 0.75 and between 0.005 and 0.05, respectively.

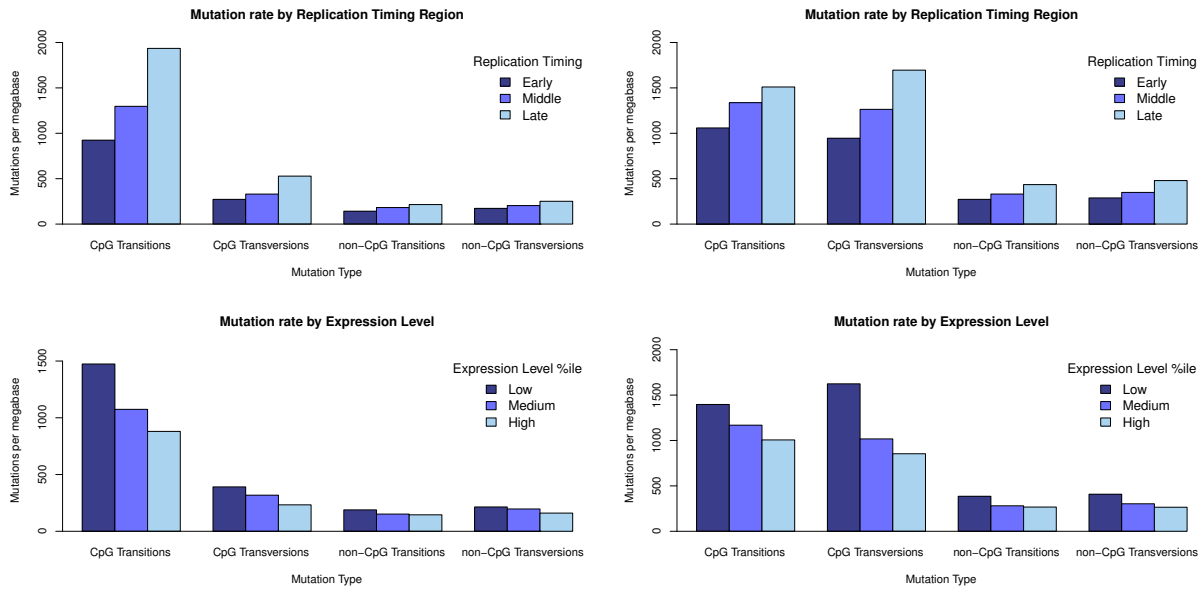
We note that that there may be a bias in assigning an increased prior probability of driver activity to genes that have been sequenced more times than others (since the COSMIC database contains targeted sequencing projects wherein only a certain set of genes were screened for mutations). However, this problem will diminish as more and more information is added to COSMIC. It is important to point out, also, that more sequencing information does not automatically mean that a gene will have more weight in the prior; the additional information has to show evidence of either positional patterns or high proportion of truncating mutations, both of which are adjusted for the total number of mutations observed.

3.2 Background mutation model

We build on the YS background model and extend it to incorporate external information that has been shown to affect mutation rates, namely replication timing and expression level. Substantial variation in somatic mutation rates, up to 33% in normal and 60% in cancer cells, has been attributed to variation in the replication timing of DNA (Koren et al., 2012; Woo and Li, 2012). In short, regions that replicate later have higher mutation rates due to the decreased amount of time the replication

machinery has to perform repairs compared to earlier replicating regions (Pleasance et al., 2009).

Figure 3.5: Mutation rate by mutation type and gene-specific factors in TCGA ovarian (left) and lung (right)



Mutation rate is shown to depend significantly on replication timing region and expression level. Specifically, mutation rate is shown for three replication timing regions (top) and for three levels of expression (bottom) for four types of mutations in TCGA ovarian (left) and lung (right). Within each mutation type, Chi-Square tests of mutation counts stratified by replication timing or expression level categories were found to be significant ($p < 0.05$)

Figure 3.5 (top) shows this effect in the TCGA ovarian (left) and lung (right) datasets. As shown, the pattern persists when looking only at a specific mutation type (transitions vs transversions) and nucleotide context (CpG vs non-CpG dinucleotide). Thus, it is not likely that the pattern can be explained by differences in rates of specific types of mutations across the regions. If this factor were to be ignored,

then the background rate for late-replication regions would likely be underestimated, whereas the background rate for early-replicating regions would be overestimated.

Variation in mutation rate has also been observed with gene expression level. Specifically, Chapman et al. (2011) discovered that there are fewer mutations observed in genes that are expressed at a higher level on average in cancer cells. It is thought that transcription-coupled repair mechanisms are responsible for this effect. As in the case of replication timing, the differences in mutation rate by expression level remain largely consistent within mutation type and nucleotide context. This can be seen in Figure 3.5 (bottom), where the mutation rate is plotted for three gene expression level categories. As with replication timing, if this factor is ignored, the background rate for lowly expressed genes will be underestimated.

These two gene-specific factors explain additional variation in mutation rate beyond that contributed by the position-specific factors of mutation type and nucleotide context. However, some genes still have an inexplicably high mutation rate even after accounting for all the previously mentioned factors. Notably, the class of genes known as olfactory receptors has a near 2-fold increase in mutation rate compared to genes with similar replication timing and expression levels in the two TCGA datasets examined (see Figure 3.6). This effect is also evident in the rate parameter estimates for both the lung and ovarian datasets displayed in Tables 3.3 and 3.4. Here we classify genes as olfactory receptors using gene symbols to obtain a set of size 323 genes. While it is unclear why these genes have elevated rates of somatic mutation, they are known to exhibit substantial genetic diversity in terms of both single nucleotide polymorphisms and copy number variation (Hasin et al., 2008). Consequently, the

background model adjusts for the expected increase in the number of background mutations for this class of genes.

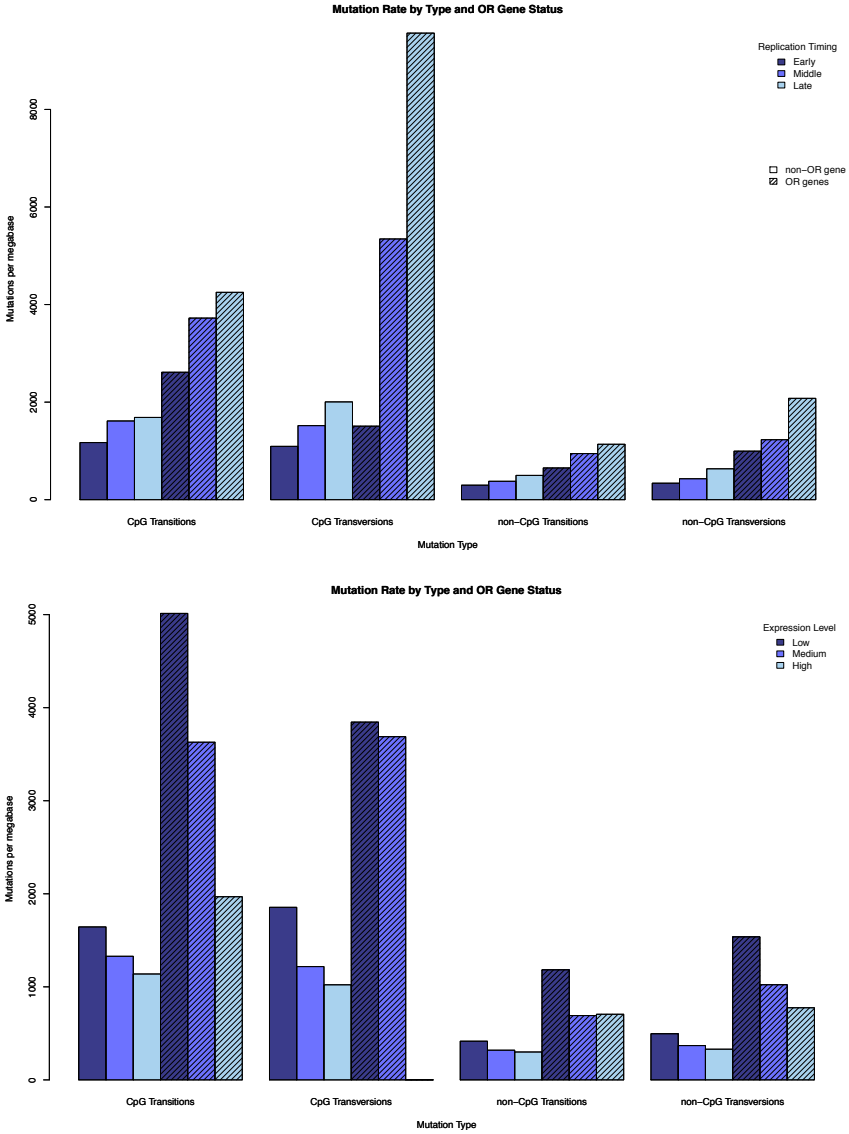


Figure 3.6: Mutation rate by mutation type, Olfactory Receptor (OR) gene status, and replication timing region (top) or expression level category (bottom) in TCGA lung

3.2.3.1 Adjusting for gene-specific factors

In order to incorporate the gene-specific factors of replication timing and expression level into the background mutation model, external estimates of replication timing were first obtained from Chen et al. (2010a), who sequenced the DNA from HeLa cell lines at various stages of the synthesis phase of the cell cycle and provided timing estimates over 100kb windows across the entire genome. As a robust proxy for replication timing, we divided the genome into three equal parts: (1) Early, (2) Middle and (3) Late replicating regions by splitting on the tertiles of the observed distribution. This is desirable since replication timing is not perfectly correlated across cell types, and we do not have ovarian cell line data. However, we note that the implementation of MADGiC is flexible enough to accommodate other sources of replication timing data.

Next, average expression levels of each gene were obtained from the 91 cell lines in the Cancer Cell Line Encyclopedia (CCLE) database with RNA-seq data (Barretina et al., 2012), and genes were divided into tertiles of expression. Averages across many tissue types in the CCLE were used rather than matched expression measurements from TCGA since the pattern of decreased mutations with increased expression was more stable within mutation type, and because this same set of expression data could be used in studies of a different cancer or in studies where expression data was not available. However, as with replication timing data, if other sources of expression data are available, they may be specified in the MADGiC package.

Let λ_n, $n \in \{1, 2, 3\}$ be the relative rates of mutation for a position in replication timing category n, and let ϵ_h, $h \in \{1, 2, 3\}$ be the relative rates of mutation for a

Table 3.2: Background mutation rate parameters define the background mutation rate for position k of sample j

Mutation Type	CpG context	Nucleotide Change	Probability
Transition	-	A:T $\rightarrow$ G:C	$p_1\lambda_{u_k}\epsilon_{\nu_k}\delta^{\omega_k}q_j$
Transversion	-	A:T $\rightarrow$ C:G or T:A	$p_2\lambda_{u_k}\epsilon_{\nu_k}\delta^{\omega_k}q_j$
Transition	non CpG	C:G $\rightarrow$ T:A	$p_3\lambda_{u_k}\epsilon_{\nu_k}\delta^{\omega_k}q_j$
Transversion	non CpG	C:G $\rightarrow$ A:T or G:C	$p_4\lambda_{u_k}\epsilon_{\nu_k}\delta_{\omega_k}q_j$
Transition	CpG	C:G $\rightarrow$ T:A	$p_5\lambda_{u_k}\epsilon_{\nu_k}\delta^{\omega_k}q_j$
Transversion	CpG	C:G $\rightarrow$ A:T or G:C	$p_6\lambda_{u_k}\epsilon_{\nu_k}\delta^{\omega_k}q_j$
Other (Indel)	-	In Frame	$p_7\lambda_{u_k}\epsilon_{\nu_k}\delta^{\omega_k}q_j$
	-	Frameshift	$p_8\lambda_{u_k}\epsilon_{\nu_k}\delta^{\omega_k}q_j$

position in expression level category h. In addition, let δ be the relative rate of mutation for olfactory genes compared to all others. These parameters are incorporated into the background model of YS as additional multiplicative factors. They are in addition to the mutation-type and nucleotide context-specific rate parameters p_m, $m = 1, ..., 8$ defined in the YS model.

3.2.3.2 Description of relative rate parameters

To translate the relative rate parameters into probabilities at the base pair position level, we obtain, for every position k in the exome, the number of possible changes resulting in: (1) silent transitions $c_k \in \{0, 1\}$, (2) silent transversions $d_k \in \{0, 1, 2\}$, (3) nonsilent transitions $e_k \in \{0, 1\}$, and (4) nonsilent transversions $f_k \in \{0, 1, 2\}$.

In addition we obtain the mutation type $t_k \in \{1, 3, 5\}$ of a transition occurring at position k and the mutation type $v_k \in \{2, 4, 6\}$ of a transversion occurring at position k, both of which depend on the reference nucleotide and CpG dinucleotide context.

The replication timing category indicator $u_k \in \{1, 2, 3\}$ and expression level category indicator $v_k \in \{1, 2, 3\}$ for position k are easily obtained from the gene-specific factor data. Finally, $\omega_k \in \{0, 1\}$ is the indicator that position k is in an olfactory receptor gene.

From these constants and rate parameters, we can calculate the probability of a specific nucleotide change at any position in the exome as in Table 3.2. To obtain probabilities for a mutation type we multiply by the number of such changes that are possible. For example, the probability of a nonsilent transversion at position k in sample j is $f_k q_j p_4 \lambda_1 \epsilon_3$ if the position has a G or C reference nucleotide (non-CpG) and is in an early-replicating and high-expressing gene that is not an olfactory receptor.

The relative rate parameter estimates $\hat{p}_m$, $\hat{\lambda}_n$, $\hat{\epsilon}_h$, and $\hat{\delta}$ determine background mutation rate probabilities and thus, ideally, they should be obtained by fitting the model only to genes containing silent mutations. Using all genes would mean that driver genes are included, which would violate our assumption that driver genes do not follow the background mutation model. However, because indels are nonsilent, we also include genes that have at most one nonsilent mutation. This introduces potential selection bias in the sample-specific mutation rates q_j so we follow YS and introduce another parameter r to account for the bias.

3.2.3.3 Estimation of relative rate parameters

As in Youn and Simon (2011), we use the method of moments to estimate the relative rate parameters for mutation type p_m and selection bias r, as well as the additional parameters λ_n, ϵ_h, and δ. Estimates for p_m and r in the gene-specific factors model are

of a similar form as in the original model, but are averaged over the three replication timing regions and expression level categories.

Let K denote the set of nucleotide positions used for silent mutation detection, and L be the subset of nucleotide positions for genes with at most one nonsilent mutation. Additionally, let K^{OR} and L^{OR} denote the subsets of K and L that are located in olfactory receptor genes. Let $Y_{jk} \in \{\text{ts}, \text{tv}\}$ be an indicator variable for the type of mutation in sample j at position k (ts = transition, tv = transversion). Given estimates $\hat{p}$ and $\hat{r}$ and fixing the reference categories $s_2 = \epsilon_2 = 1$, estimates for λ_n and ϵ_h, where n, $h \in \{1, 3\}$, are obtained by taking the expectation of the number of transitions or transversions

$$\hat{\lambda}_n = \frac{1}{3}\sum_{h=1}^{3} \frac{\sum_{\substack{j \\ u_k=n \\ t_k=m \\ \nu_k=h}} \left[\sum_{m=1,3,5} I(Y_{jk} = \text{ts}) G_{m,2,h} + \sum_{m=2,4,6} I(Y_{jk} = \text{tv}) H_{m,2,h} \right]}{\sum_{\substack{j \\ u_k=2 \\ t_k=m \\ \nu_k=h}} \left[\sum_{m=1,3,5} I(Y_{jk} = \text{ts}) G_{m,n,h} + \sum_{m=2,4,6} I(Y_{jk} = \text{tv}) H_{m,n,h} \right]}$$

$$\hat{\epsilon}_h = \frac{\sum_{n=1}^{3} \sum_{\substack{j \\ u_k=n \\ t_k=m \\ \nu_k=h}} \hat{\lambda}_n \left[\sum_{m=1,3,5} I(Y_{jk}=\text{ts}) G_{m,n,2} + \sum_{m=2,4,6} I(Y_{jk}=\text{tv}) H_{m,n,2} \right]}{\sum_{n=1}^{3} \sum_{\substack{j \\ u_k=2 \\ t_k=m \\ \nu_k=h}} \hat{\lambda}_n \left[\sum_{m=1,3,5} I(Y_{jk}=\text{ts}) G_{m,n,h} + \sum_{m=2,4,6} I(Y_{jk}=\text{tv}) H_{m,n,h} \right]}$$

$$\hat{\delta} = \frac{\sum_{h=1}^{3}\sum_{n=1}^{3}\sum_{\substack{j\\ u_k=n\\ t_k=m\\ \nu_k=h}} \hat{\epsilon}_h\hat{\lambda}_n\left[\sum_{m=1,3,5} I(Y_{jk}{=}\text{ts})G^{OR}_{m,n,h} + \sum_{m=2,4,6} I(Y_{jk}{=}\text{tv})H^{OR}_{m,n,h}\right]}{\sum_{h=1}^{3}\sum_{n=1}^{3}\sum_{\substack{j\\ u_k=2\\ t_k=m\\ \nu_k=h}} \hat{\epsilon}_h\hat{\lambda}_n\left[\sum_{m=1,3,5} I(Y_{jk}{=}\text{ts})G^{OR^c}_{m,n,h} + \sum_{m=2,4,6} I(Y_{jk}{=}\text{tv})H^{OR^c}_{m,n,h}\right]}$$

where the constants are

$$G_{m,n,h} = \hat{p}_m\left(\sum_{\substack{k\in K\\ u_k=n\\ t_k=m\\ \nu_k=h}} c_k,\ +\hat{r}\sum_{\substack{k\in L\\ u_k=n\\ t_k=m\\ \nu_k=h}} e_k\right)$$

$$H_{m,n,h} = \hat{p}_m\left(\sum_{\substack{k\in K\\ u_k=n\\ v_k=m\\ \nu_k=h}} d_k,\ +\hat{r}\sum_{\substack{k\in L\\ u_k=n\\ v_k=m\\ \nu_k=h}} f_k\right)$$

and likewise,

$$G^{OR}_{m,n,h} = \hat{p}_m\left(\sum_{\substack{k\in K^{OR}\\ u_k=n\\ t_k=m\\ \nu_k=h}} c_k,\ +\hat{r}\sum_{\substack{k\in L^{OR}\\ u_k=n\\ t_k=m\\ \nu_k=h}} e_k\right) \text{ and } G^{OR^c}_{m,n,h} = \hat{p}_m\left(\sum_{\substack{k\in K\setminus K^{OR}\\ u_k=n\\ t_k=m\\ \nu_k=h}} c_k,\ +\hat{r}\sum_{\substack{k\in L\setminus L^{OR}\\ u_k=n\\ t_k=m\\ \nu_k=h}} e_k\right).$$

The quantities $H^{OR}_{m,n,h}$ and $H^{OR^c}_{m,n,h}$ follow similarly.

The results of fitting the YS background model (M_{YS}) and our background model adjusted for gene-specific factors (M_{GS}) are shown in Tables 3.3 and 3.4 for the ovarian and lung case studies, respectively. The mutation type rate parameters p_m are very similar in the two models for both datasets, and the direction of the replication

timing and expression level rate parameters are consistent with the patterns observed in Figure 3.5, namely that later replication timing region or lower expression level results in an increase in the relative mutation rate. We also note that the parameters characterizing relative rate of mutation at GC nucleotides (p_3, p_4, p_5 and p_6) are larger than the rest. This is consistent with the relationship between GC content and mutation rate in previous studies (Chapman et al., 2011).

We obtain empirical Bayes estimates of the sample-specific overall mutation rates q_j (by assigning the prior distribution of q_j to be Uniform(a, b) and estimating the posterior mean). The hyperparameters $(\hat{a}, \hat{b})$ are found via maximum likelihood estimation given relative rate parameters ($\hat{r}, \hat{\vec{p}}$, $\hat{\vec{\lambda}}$, $\hat{\vec{\epsilon}}$, and $\hat{\delta}$). In this way, the posterior distribution of q_j depends on the observed mutations in sample j, as well as the data-wide parameter estimates of the relative rates of the different types of mutations. Finally, the background probability b_{jg} that a gene g is mutated in sample j under the background model (i.e. given g is a passenger) is approximated by summing the probability of a background mutation across all base pairs in the gene. Note that this procedure is different from YS, who calculate b_{jg} as the expectation with respect to the posterior distribution of q_j; the resulting estimates of b_{jg} using YS are also empirical Bayes estimates and are very similar to the estimates obtained by our procedure, except that the former requires $J * G$ numerical integrations and the latter only J which provides considerable improvement in computation time.

Table 3.3: Fitted parameters for the YS background model and our model accounting for gene-specific (GS) factors in TCGA ovarian

	Selection	Mutation Type and Nucleotide Context							Timing		Expression		OR
	Bias	Transitions		Transversions			Indels		Early	Late	Low	High	Gene
	$\hat{r}$	$\hat{p}_3$	$\hat{p}_5$	$\hat{p}_2$	$\hat{p}_4$	$\hat{p}_6$	$\hat{p}_7$	$\hat{p}_8$	$\hat{\lambda}_1$	$\hat{\lambda}_3$	$\hat{\epsilon}_1$	$\hat{\epsilon}_3$	$\hat{\delta}$
M_{YS}	0.51	1.79	9.44	0.59	1.29	1.33	0.03	0.11	-	-	-	-	-
M_{GS}	0.52	1.87	11.07	0.66	1.30	1.35	0.04	0.13	0.77	1.24	1.06	0.89	1.86

Table 3.4: Fitted parameters for the YS background model and our model accounting for gene-specific (GS) factors in TCGA lung

	Selection	Mutation Type and Nucleotide Context							Timing		Expression		OR
	Bias	Transitions		Transversions			Indels		Early	Late	Low	High	Gene
	$\hat{r}$	$\hat{p}_3$	$\hat{p}_5$	$\hat{p}_2$	$\hat{p}_4$	$\hat{p}_6$	$\hat{p}_7$	$\hat{p}_8$	$\hat{\lambda}_1$	$\hat{\lambda}_3$	$\hat{\epsilon}_1$	$\hat{\epsilon}_3$	$\hat{\delta}$
M_{YS}	0.27	2.57	6.82	0.61	2.22	3.59	0.01	0.07	-	-	-	-	-
M_{GS}	0.31	2.66	7.52	0.68	2.34	4.19	0.01	0.06	0.71	1.56	1.18	0.90	2.40

3.2 Implementation and evaluation

In order to evaluate the utility of incorporating functional impact scores in the model, as well as assess what could be gained with a score that was better able to distinguish between passenger and driver mutations, MADGiC was evaluated under three different functional impact profiles: (a) ignoring functional impact, (b) realistic impact, and (c) high impact.

For the realistic impact setting, score profiles are assigned to the spiked-in driver mutations that correspond to the 95^{th} percentile of the sum of SIFT scores for all genes with that number of observed mutations. For example, the observed 95^{th} percentile for the sum of SIFT scores for genes with 3 observed mutations in the ovarian dataset was 2.99 which corresponded to a profile of scores for those three mutations of (1, 1, 0.99). This profile represents two mutations with a SIFT score value of 1 and one with a score of 0.99 (recall that higher score indicates more predicted functional impact). If more than one observed profile corresponded to the same 95th percentile of the sum, one of them was chosen randomly for each spiked-in driver gene. FI scores for the shuffled passenger mutations were exactly the SIFT scores of those mutations under this setting.

For the high impact setting, passenger scores are drawn from Beta(1,1.5) and driver scores are set equal to one. The last setting represents some idealistic functional impact (FI) scoring system in which the distributions of driver and passenger scores are well-separated (i.e. passenger mutations tend to have low functional impact and driver mutations always have high functional impact) and is designed to assess the upper bound for the amount of improvement than can be achieved by incorporating

FI. The background model was fit as described in Section 3.2.3 and the posterior probabilities of each gene being a driver were computed as described in Section 3.2.2. Genes with posterior probability greater than 0.95 were classified as drivers.

For comparison, the frequency-based methods YS and MutSigCV were also evaluated (YS evaluated for only 50 simulations due to computation time). Genes were classified as drivers by YS or MutSigCV if the Benjamini-Hochberg adjusted p-value was less than 0.05. MuSiC was not evaluated since it requires a post-processing step to filter the output, for which general guidelines are not provided by Dees et al. (2012); and OncodriveFM and OncodriveCLUST were only evaluated for the case study data since it is not possible to specify simulated SIFT scores for these approaches. For the Oncodrive methods, we considered genes with q-values less than 0.05.

While we can comment on the characteristic differences among the driver genes identified in the case studies, it should be noted that we do not have a list of 'true positive' driver genes for the ovarian or lung cancer data. As a proxy, we use the list of 125 genes identified as drivers by Vogelstein et al. (2013). Note that although some hyperparameters in MADGIC were estimated using COSMIC data (see Section 3.2.2.3), the TCGA ovarian and lung datasets were removed prior to estimation, and no information regarding the list of drivers in Vogelstein et al. (2013) was used. Further, a sensitivity analysis was conducted to examine the effect of the weight placed on COSMIC in assigning prior probabilities that a gene is a driver (see Section 3.3.4).

3.2 Software and database versions

R code to implement this method is available at `http://www.biostat.wisc.edu/~kendzior/MADGiC/`. Unless otherwise noted, analyses are carried out using R (R Core Team, 2014) version 2.14.2. The method of Youn and Simon (2011) was implemented using the code provided at the following website: `http://linus.nci.nih.gov/Data/YounA/software.zip`. MutSigCV version 1.3 was used (Lawrence et al., 2013). OncodriveFM and OncodriveCLUST were implemented using the web version of Intogen Mutations software suite version 2.4.1-maintenance at www.intogen.org with OncodriveCLUST genes threshold = 3 for the ovarian case study and 5 for the lung case study (all other parameters set as default). No filtering of genes based on gene expression was carried out. The COSMIC database version 66 was used in informing the prior probability of oncogenic activity as well as for estimating the hyperparameters of the prior probability of mutation in a driver gene. Only information obtained from whole-gene screens was utilized since mutations at targeted positions may be biased toward positional clustering. All genome coordinates were mapped to the hg18 assembly (NCBI build 36.1).

3.3 Results

3.3 Application to simulated data

To facilitate comparisons with existing methods, the simulation study considers two types of simulations: SIM I simulations that ignore the dependence of mutation rate on replication timing and expression level and SIM II simulations that do not.

Table 3.5: Simulation results

					MADGiC		
			MutSigCV	YS	No FI	SIFT	Ideal FI
SIM I	Ovary	Power	0.05	0.30	0.42	0.51	0.86
		FDR	0.04	0.04	0.04	0.04	0.02
	Lung	Power	0.01	0.16	0.27	0.31	0.75
		FDR	0.07	0.08	0.07	0.06	0.03
SIM II	Ovary	Power	0.06	0.33	0.45	0.55	0.86
		FDR	0.02	0.32	0.08	0.09	0.04
	Lung	Power	0.02	0.36	0.30	0.34	0.77
		FDR	0.58	0.97	0.32	0.30	0.05

Power and FDR averaged over 100 SIM I datasets, where dependence of mutation rate on replication timing and expression level is ignored and 100 SIM II datasets, where this dependence is preserved. The first set of simulations was designed to mimic TCGA ovarian data, which has a relatively large sample size, an average number of mutations, and relatively little variability among sample-specific mutation rates; the second set is based on TCGA lung data, with smaller sample size, larger number of mutations, and greater heterogeneity in sample-specific mutation rates.

Within each simulation setup, we evaluate the ability of MADGiC and competing methods to identify true driver genes in a scenario that mimics TCGA ovarian (with a relatively large sample size and average number of mutations) as well as one that mimics TCGA lung (relatively small sample size and large number of mutations). In addition, MADGiC was evaluated under three different functional impact settings in order to assess to what degree the inclusion of an FI score may result in increased power. As expected, performance depends on each of these characteristics.

As shown in Table 3.5, FDR is well controlled for all methods when mutation rate is assumed constant across replication timing region and expression level. In the more realistic SIM II setting, FDR increases for methods that do not accommodate this

dependence. For the simulated lung data, false discovery rates (FDRs) are generally higher and power is generally lower for all methods. This is likely due in part to the higher passenger mutation rate relative to the true driver mutation rate as well as greater heterogeneity in sample-specific mutation rates.

When functional impact scores are able to separate driver mutations from passengers (the ideal FI case), MADGiC is very well powered to detect true driver genes and has a well-controlled FDR. In contrast, when no FI information is used, the power of MADGiC is decreased but is still highest among approaches using the ovarian-based simulations, with only moderate increases in FDR. In the lung-based simulations, YS has higher power than MADGiC, but the FDR is considerably inflated. Under the more intermediate FI setting that is based on observed SIFT score profiles, MADGiC has more power than when FI information is ignored, with comparable FDR. Thus, MADGiC performs best when FI scores are set to be near ideal, however, it still shows favorable performance when SIFT scores are used (and also when no FI is used).

3.3 Application to TCGA somatic mutation data

3.3.2.1 Ovarian cancer

MADGiC identified 19 genes with a posterior probability of being a driver greater than 0.95. Table 3.6 (top) displays the number of genes found by each method, along with the proportion of those found that were also on the list of putative drivers from Vogelstein et al. (2013) (the 'Putative Driver Rate'). Table 3.7 displays the 19 genes from the ovarian case study that were found by MADGiC to have a posterior

Table 3.6: Case study results

		MADGiC	YS	MutSigCV	Oncodrive FM	Oncodrive CLUST
Ovary	Total Found	19	70	5	21	20
	Put. Driver Fraction	0.579	0.129	0.400	0.381	0.250
Lung	Total Found	47	585	7	85	55
	Put. Driver Fraction	0.213	0.019	0.571	0.153	0.145

For each method applied to the two case studies (TCGA ovarian and lung), we report the total number of driver genes identified, along with the proportion of those found that are putative drivers (i.e. they are on the list identified by Vogelstein et al. (2013)).

probability of being a driver of at least 0.95, along with whether they were significant by the other models (YS, MutSigCV, OncodriveFM, and OncodriveCLUST), and whether they are in the set of drivers identified by Vogelstein et al. (2013).

YS identified 70 significant genes after adjustment for multiple comparisons, 14 of which were also found by our model. MutSigCV identified five significant genes after adjustment for multiple comparisons, two of which were identified by our model and YS. Six of the 21 genes identified by OncodriveFM and three of the 20 genes identified by OncodriveCLUST were also found by MADGiC. We note that 57.9% of the drivers identified by our model are contained in the list from Vogelstein et al. (2013), while the same figure is 12.9% for YS, 40.0% for MutSigCV, 38.1% for OncodriveFM, and 25.0% for OncodriveCLUST. Of the five genes identified by MADGiC but not YS, four have five or fewer mutated samples and two of those are putative drivers.

Figure 3.7 displays the proportion of genes found by each method in each replication timing and expression level category. Here we see that MADGiC is not biased toward

finding genes in the high background mutation categories (late replication timing or low expression) compared to the distribution of all genes. In contrast, YS finds the highest proportion of genes in the high mutation rate categories.

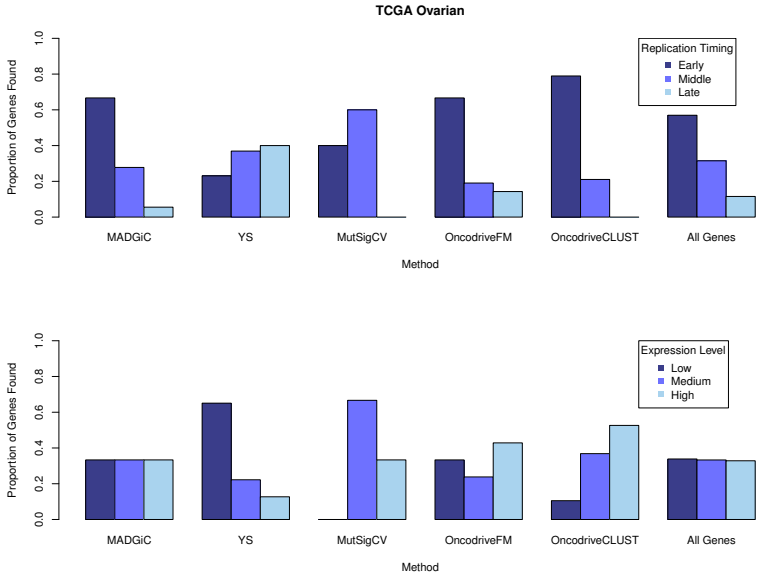


Figure 3.7: The proportion of driver genes identified by each method in each replication timing (top) and expression level (bottom) category for the TCGA ovarian case study.

3.3.2.2 Squamous cell lung carcinoma

Although the lung data set is structurally different than ovarian with a smaller sample size and much higher average mutation rate, the qualitative results from each method are similar. MADGiC identified 47 genes with a posterior probability of being a driver greater than 0.95. Table 3.6 (bottom) displays the number of genes found by each method, along with the proportion of those found that were also on the list of putative drivers from Vogelstein et al. (2013). Table 3.8 displays the 47 genes from

Table 3.7: Genes with posterior probability > 0.95 in the ovarian dataset

Gene	Post	Samples	younsimon	MutSigCV	fm	clust	Vogel	Rep	Exp
TP53	1.000	383	✓	✓	✓	✓	✓	Early	High
NF1	1.000	23	✓		✓		✓	Early	Medium
BRCA1	1.000	19	✓		✓		✓	Early	High
RB1	1.000	15	✓	✓	✓		✓	Middle	Medium
CDK12	1.000	14	✓		✓			Early	High
CREBBP	1.000	11					✓	Early	Medium
KRAS	1.000	5	✓		✓	✓	✓	Middle	Low
NRAS	0.999	4	✓			✓	✓	Early	Medium
EFEMP1	0.998	7	✓					Early	Medium
PTEN	0.997	5	✓				✓	Early	Low
NLRP4	0.995	8	✓					Middle	High
CSMD3	0.993	26	✓					Late	Low
FBXW7	0.991	5					✓	Early	Low
MAP3K19	0.985	10	✓					Middle	-
KIT	0.983	8	✓				✓	Middle	Low
NF2	0.982	4					✓	Early	Medium
GPS2	0.980	3						Early	High
RALY	0.959	3						Early	High
ACTRT1	0.958	6	✓					-	Low

the ovarian case study that were found by MADGiC to have a posterior probability of being a driver of at least 0.95, along with whether they were significant by the other models (YS, MutSigCV, OncodriveFM, and OncodriveCLUST), and whether they are in the set of drivers identified by Vogelstein et al. (2013).

YS identified 585 significant genes after adjustment for multiple comparisons, 45 of which were also found by our model. MutSigCV identified seven significant genes after adjustment for multiple comparisons, six of which were identified by MADGiC and YS. Eight of the 85 genes identified by OncodriveFM and seven of the 55 genes identified by OncodriveCLUST were also found by MADGiC. We note that 21.3% of the drivers identified by our model are contained in the list from Vogelstein et al. (2013), while the same figure is 1.9% for YS, 57.1% for MutSigCV, 15.3% for

OncodriveFM, and 14.5% for OncodriveCLUST. As in the ovarian case study, YS is biased toward identifying genes in the high background mutation rate categories (see Figure 3.8). Specifically, of the 448 genes significant only by YS that also have complete replication timing and expression information, 400 (89%) are in either the late replicating region, the low expression category, or both. In addition, only one of these additional genes was also identified by Vogelstein et al. (2013). Of the two genes identified by MADGiC but not YS, one has five or fewer mutated samples and the other is a putative driver.

Note that the results presented here for MutSigCV are slightly different than those observed in Lawrence et al. (2013) since we have removed 10 samples, used a q-value threshold of 0.05 instead of 0.10, and used the most updated version of MutSigCV (see Sections 3.2.1.2 and 3.2.4 for details).

3.3 Gene length bias

We (and others, including Lawrence et al. (2013)) have observed that frequency-based methods suffer from a bias toward identifying long genes as significant. In fact, in the OV case study, there is a significant bias toward longer genes by all three methods we considered. This can be seen in the enrichment p-values in Table 3.9, calculated by comparing the mean length of identified driver genes with mean coding sequence lengths from 100,000 random subsets of the same number of mutated genes identified by each approach (e.g. MADGiC found 19 drivers in the OV case study, so we compare the mean length of these 19 with the mean coding sequence length of 100,000 random subsets of 19 mutated genes). In the LUSC case study, there is a significant

Table 3.8: Genes with posterior probability > 0.95 the lung dataset

Gene	Post. Prob.	Samples Mutated	YS	MutSig CV	Oncodrive FM	Oncodrive CLUST	Putative Driver	Replication Category	Expression Category
TP53	1.000	141	✓	✓	✓	✓	✓	Early	High
CSMD3	1.000	81	✓					Late	Low
RYR2	1.000	76	✓					Late	Low
ZFHX4	1.000	65	✓					Late	Low
PCDHA6	1.000	53	✓					Middle	High
KMT2D	1.000	43	✓				✓	Early	High
FAM135B	1.000	37	✓					Late	Low
ERICH3	1.000	34	✓					Late	Low
CDH10	1.000	31	✓					Late	Low
ZNF804B	1.000	30	✓					Late	Low
PCDH11X	1.000	28	✓					-	Low
PIK3CA	1.000	27	✓		✓	✓	✓	Late	Medium
SPHKAP	1.000	27	✓					Middle	Low
NFE2L2	1.000	27	✓	✓			✓	Early	-
CDKN2A	1.000	26	✓	✓	✓	✓	✓	Early	Low
MROH2B	1.000	26	✓					Middle	-
TPTE	1.000	24	✓					Late	Low
CDH12	1.000	23	✓					Middle	Low
KEAP1	1.000	22	✓	✓	✓	✓		Early	High
SLCO1B3	1.000	20	✓					Middle	Low
CDH9	1.000	19	✓					Late	Low
KLHL1	1.000	17	✓					Early	Low
PTEN	1.000	15	✓	✓	✓	✓	✓	Early	Low
POTEA	1.000	13	✓					Early	Low
RB1	1.000	12	✓		✓		✓	Middle	Medium
KRTAP5-5	1.000	9	✓	✓				Early	-
MYH2	0.999	28	✓					Middle	Low
LILRA1	0.999	16	✓					Middle	High
KRTAP1-1	0.999	9	✓					Early	-
PSG2	0.998	10	✓					Early	-
HRAS	0.998	5	✓		✓	✓	✓	Early	High
HCN1	0.996	31	✓					Late	Low
TGIF2LX	0.996	9	✓					-	Low
SGIP1	0.995	12	✓					Middle	Low
RP1	0.994	28	✓					Middle	Low
ZIC1	0.993	16	✓					Late	Low
CNTNAP5	0.991	28	✓					Late	Low
GPS2	0.990	5						Early	High
FBXW7	0.984	11	✓			✓	✓	Early	Low
SLC39A12	0.983	10	✓					Early	-
TECRL	0.982	9	✓					Late	Low
ADAMTS12	0.978	28	✓					Middle	Medium
MMP16	0.971	15	✓					Late	Low
ZFP42	0.966	8	✓					Middle	Low
ZNF208	0.961	24	✓					Late	Low
NOTCH1	0.958	14			✓		✓	Early	High
HLA-A	0.957	7	✓					Middle	-

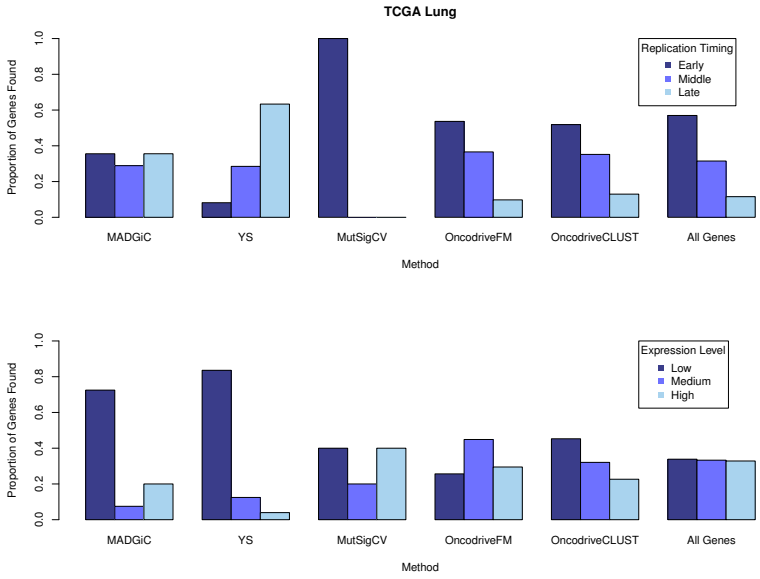


Figure 3.8: The proportion of driver genes identified by all methods stratified by replication timing (top) and expression level (bottom) categories for the TCGA lung case study.

bias toward longer genes by MADGiC and YS, but not MutSigCV. Here MutSigCV does appear to handle the length bias more appropriately, however it is much more conservative than the other two methods and only identifies 7 genes. Adding just three more genes (i.e. taking the top 10 genes ranked by p-value) results in adding two longer genes (lengths $\sim$ 15,000bp and $\sim$ 5,000bp) increasing the mean length from 1165 to 2970 and suggesting a length bias as with the other approaches ($p = 0.06171$).

Table 3.9: Length bias observed in the TCGA lung and ovarian case studies

Case Study	Method	Mean Length CDS	Enrichment P-value
OV	MADGiC	3390	0.02611
	YS	5418	<0.00001
	MutSigCV	4649	0.02298
LUSC	MADGiC	3714	0.00498
	YS	2978	<0.00001
	MutSigCV	1165	0.96703

3.3 Empirical sensitivity analysis of prior specification

In our driver mutation model, we incorporate prior information in two places. First, we inform the prior probability of driver activity by the spatial pattern of mutations in COSMIC (see Section 3.2.2.4). We also estimate the hyperparameters of the distribution of mutation status of driver genes using mutation frequencies in COSMIC (see Section 3.2.2.3). Presented here is a summary of the empirical sensitivity analyses that were carried out to assess the degree in variation in the results over a range of these prior settings.

3.3.4.1 Range of parameters explored

We set the prior probability of a given gene being a driver to some threshold level T if that gene has at least ten observed mutations in COSMIC and shows evidence of either tumor suppressor (biased toward truncating mutations) or oncogenic activity (mutations tend to overlap at the same amino acid position) as defined above. If it does not show evidence, we assign a fixed baseline prior probability π_0. Here we examine several combinations of the threshold T and fixed baseline prior π_0 to see how sensitive the model is to these parameters when fitting to the TCGA ovarian

data. Specifically, we look at the number of genes with posterior probability greater than 0.95 (defined as drivers), as well as how many are on the list of putative driver genes identified by Vogelstein et al. (2013). We explore the threshold values T of 0.05, 0.25, 0.50, 0.75, and none (flat prior) along with fixed baseline prior π_0 values of 0.995, 0.99, 0.975, and 0.95.

We also examine the results under two settings of the prior distribution of mutation probability in driver genes: that where the hyperparameters are elicited from the distribution of mutational frequencies in COSMIC for non-TCGA ovarian cancer samples (α= 0.15, β = 6.6))and that where we adopt an uninformative, flat prior (α =β = 1).

3.3.4.2 Parameters chosen for use in MADGiC

Table 3.10 displays the number of driver genes identified, along with how many of those are also in Vogelstein's set in parentheses, for the 20 different settings of π_0 and T and two different settings for the hyperparameters of d_g. The top panel of Table 3.10 contains the results for using the hyperparameters elicited from COSMIC (as described above) and the bottom contains the results using the flat uniform prior for d_g.

Based on these results, we choose a fixed baseline π_0 of 0.99 and a threshold value T of 0.5 (genes showing evidence of tumor suppressor or oncogenic activity based on positional data in COSMIC are assigned a prior probability of being a passenger of $\pi_0 = 0.5$). Overall, we see a marked effect of using COSMIC mutation frequency information for the prior of d_g compared to using a flat prior. In addition, there

Table 3.10: Number of driver genes identified (along with how many of those are also in Vogelstein's set)

Using COSMIC-derived hyperparameters for d_g

	Fixed Baseline π_0			
Threshold T	0.995	0.99	0.975	0.95
None	8 (5)	13 (7)	20 (7)	27 (8)
0.75	11 (8)	16 (10)	23 (10)	29 (10)
0.50	12 (9)	19 (11)	26 (11)	32 (11)
0.25	15 (11)	25 (14)	32 (14)	38 (14)
0.05	24 (16)	41 (22)	48 (22)	54 (22)

Using non-informative hyperparameters for d_g

	Fixed Baseline π_0			
Threshold T	0.995	0.99	0.975	0.95
None	6 (5)	8 (5)	9 (5)	10 (5)
0.75	6 (5)	10 (7)	11 (7)	12 (7)
0.50	7 (6)	11 (8)	12 (8)	13 (8)
0.25	8 (7)	11 (8)	12 (8)	13 (8)
0.05	10 (9)	15 (11)	16 (11)	17 (11)

was little variation in the number of driver genes or Vogelstein genes for threshold values between 0.25 and 0.75. The number of Vogelstein genes was also robust to changes in baseline π_0. We choose to use the COSMIC-derived hyperparameters for the probability of mutation in driver genes since it was able to detect more driver genes, including more Vogelstein genes, for all settings of the baseline π_0 and threshold T. The particular values of 0.99 and 0.5 are arbitrary, but allow for the model to incorporate evidence from both prior sources without letting it dominate the model (as in the case of threshold value of 0.05, where the number of driver and Vogelstein genes is double to triple compared to the case of ignoring positional information).

3.4 Discussion

MADGiC is an integrative model that provides posterior probabilities for improved inference for driver gene identification. The empirical Bayesian framework provides a natural way to incorporate several critical features together that were previously only considered in isolation. In addition to modeling key features of the observed mutation data, MADGiC also leverages the non-random mutational patterns observed across many cancer types in the COSMIC database to inform the prior probability of driver activity. Until recently, these spatial patterns were only evident in well-studied cancer genes that were the focus of targeted sequencing studies. Over the past few years, however, the COSMIC database has accumulated data from thousands of whole genomes and whole exomes, enabling a systematic search over all genes. The use of a database that collects mutation position data from multiple studies for each cancer type is vital, as the characteristic spatial patterns observed across thousands of cancers are not discernible when analyzing data from a single cancer in isolation.

The performance of MADGiC shows promise both in simulations and case studies. The simulation studies suggest that MADGiC has favorable operating characteristics relative to existing methods, and further highlights specifically the amount of advantage gained by incorporating functional impact scores. As the quality of these scores improves, so too should the power of MADGiC. In addition, the simulation study demonstrates that the operating characteristics of all approaches can vary widely with sample size, mutation frequency, and heterogeneity in sample-specific mutation rates. It also demonstrates that MADGiC's integration of data across multiple sources

facilitates the identification of putative driver genes showing relatively few mutations, a result also observed in the case studies. Specifically, as seen in Tables 3.7 and 3.8, there are several genes with only three to five samples mutated that are identified as drivers by MADGiC but not other approaches. The fact that many of these are also on the putative driver list of Vogelstein et al. (2013) suggests that they are not false positives.

A limitation of all methods investigated stems from our assumption that the somatic mutation calls are complete and accurate. While it has been observed that properties of tumor samples (e.g. low allelic fraction) are responsible for introducing systematic sequencing bias, methods for improving the sensitivity of mutation callers have been developed (Yost et al., 2013). As these methods continue to improve, so too will results from MADGiC. A further limitation of frequency-based methods that was noted in Lawrence et al. (2013) is the bias toward longer genes. Although MADGiC, YS, and MutSigCV each account for gene length, the driver genes are still enriched for longer genes in all three methods in both case studies except for MutSigCV in the lung cancer study (see Section 3.3.3 for details). However, MutSigCV is more conservative than the other two methods and the bias reappears as the gene list size increases. The bias is likely a result of additional, perhaps unknown factors that affect the rate of mutation of these longer genes. The fact that none of the methods are able to completely overcome this bias demonstrates that this is an ongoing challenge for frequency-based methods.

So far, we have only considered modeling one gene at a time. Thus, when computing the posterior probability that a given gene is a driver, no information pertaining

to any other genes is considered, beyond that used to estimate the parameters in the background mutation model. However, a nonsilent mutation in any one of a group of coordinately regulated genes (for example A, B, and C) could cause the same selective advantage to a cancer cell. In this situation, evidence of driver activity of gene A would increase given nonsilent mutations in genes B and C in other samples. A number of methods are available for identifying pathways containing driver genes (Vaske et al., 2010; Vandin et al., 2012; Ciriello et al., 2012). Extensions of MADGiC to accommodate pathway structure should further improve our ability to identify drivers of cancer.

Supplementary notes

The results presented here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov/. This work was supported by NIH GM102756. The manuscript based on this work is published in *Bioinformatics* (Korthauer and Kendziorski, 2015).

4 BAYESIAN NONPARAMETRIC MIXTURE MODELING OF EXPRESSION DYNAMICS IN SINGLE-CELL RNA-SEQ EXPERIMENTS

4.1 Background

Traditional RNA-seq experiments, referred to hereinafter as 'bulk' RNA-seq, allow for the quantification of transcript abundance on collections of thousands to millions of cells (Shapiro et al., 2013). Though useful in many settings, bulk RNA-seq quantifies the average signal seen in the population of cells under study. In contrast, technologies are rapidly improving for measuring mRNA transcript abundance within a collection of single cells. Specifically, microfluidics-based single-cell RNA-seq (scRNA-seq) measurements in aggregate have shown high accuracy in recapitulating bulk measurements (Wu et al., 2014; Shalek et al., 2014). Though the data structure of single-cell RNA-seq experiments is theoretically identical to that of bulk RNA-seq, the analysis of single-cell RNA-seq data introduces unique challenges and opportunities which necessitate the development of new statistical and computational tools (Stegle et al., 2015; Shapiro et al., 2013).

Single-cell RNA-seq provides the opportunity to answer emerging scientific questions that were elusive with only averages. Clearly, measuring a single cell instead of averaging a pool of cells introduces additional biological signal if there is heterogeneity in the amount of transcript in each cell. This heterogeneity is of great interest, for example, in studies of the differentiation potential of individual cells (Ohnishi et al.,

2014), the identification of subpopulations of cells (Shalek et al., 2013; Treutlein et al., 2014; Buettner et al., 2015), and the study of expression kinetics (Sanchez and Golding, 2013; Kim and Marioni, 2013; Marinov et al., 2014). Beyond the anticipated heterogeneity among cells from different biological conditions (e.g. differentiation lineages), there is still more variation that can be attributed at least in part to the stochasticity of transcription (Kim and Marioni, 2013). For example, even in genetically homogeneous cell populations, the total mRNA content within individual cells can vary by a factor of more than five (Marinov et al., 2014).

Our primary interest lies in comparing expression stochasticity and heterogeneity across samples and biological conditions. Stochasticity in gene expression levels within a single cell is due in part to the randomness involved in the gene regulation process (Kærn et al., 2005; Shahrezaei and Swain, 2008; Sanchez and Golding, 2013). Specifically, the production of mRNA relies on the interaction of multiple regulatory proteins as well as many biochemical reactions. Both of these steps are inherently stochastic at the single molecule level and consequently introduce a degree of randomness (Sanchez and Golding, 2013; Kim and Marioni, 2013).

These random processes have been studied with two-state Markov processes that model rates of transition between 'on' and 'off' states as well as mRNA transcription and decay rates in both time-lapse microscopy (Munsky et al., 2012) and scRNA-seq (Kim and Marioni, 2013) experiments. While time-lapse microscopy enables measurement of a single gene in a single cell over time, only a handful of genes can be considered in a typical experiment. This limitation is avoided in scRNA-seq, however a true time-series is not possible with this technology since each sampled

cell can only be measured once due to cell lysis (i.e. we can only observe the cells at a snapshot in time). The method of Kim and Marioni (2013) instead regards a scRNA-seq dataset as a sample from the underlying stationary distribution and estimates rate parameters under the strong assumption that the decay rate is equal to one. However, it is not possible to estimate rate parameters in units of time from snapshot data, which precludes the comparison of rate parameters across samples or conditions (Stegle et al., 2015).

In order to study differences in expression stochasticity across samples and conditions, we avoid fitting specific kinetic models that rely on measurements over time. Instead we focus on a key observation that certain systematic variations in regulatory mechanisms can result in multimodal distributions of expression across samples. For example, multiple modes may represent the existence of multiple underlying cellular states (Birtwistle et al., 2012; Singer et al., 2014), as depicted in Figure 4.1. Specifically, multiple modes have been shown to represent different promoter integration sites that exhibit different signal strengths (Larson, 2011). In addition, multiple modes may result from strong feedback signals in a positive feedback loop or slow promoter transitions (Kærn et al., 2005).

Bimodality has been described and studied for scRNA-seq in terms of a nonzero mode and a zero mode (Shalek et al., 2013). The zero mode consists of so-called dropout events which occur when a gene has nonzero measurements in some cells, but is not detected in others. It is thought that this phenomenon results from transcripts being missed during the reverse-transcriptase step during amplification (Stegle et al., 2015). Due to the small amounts of starting mRNA material in individual cells,

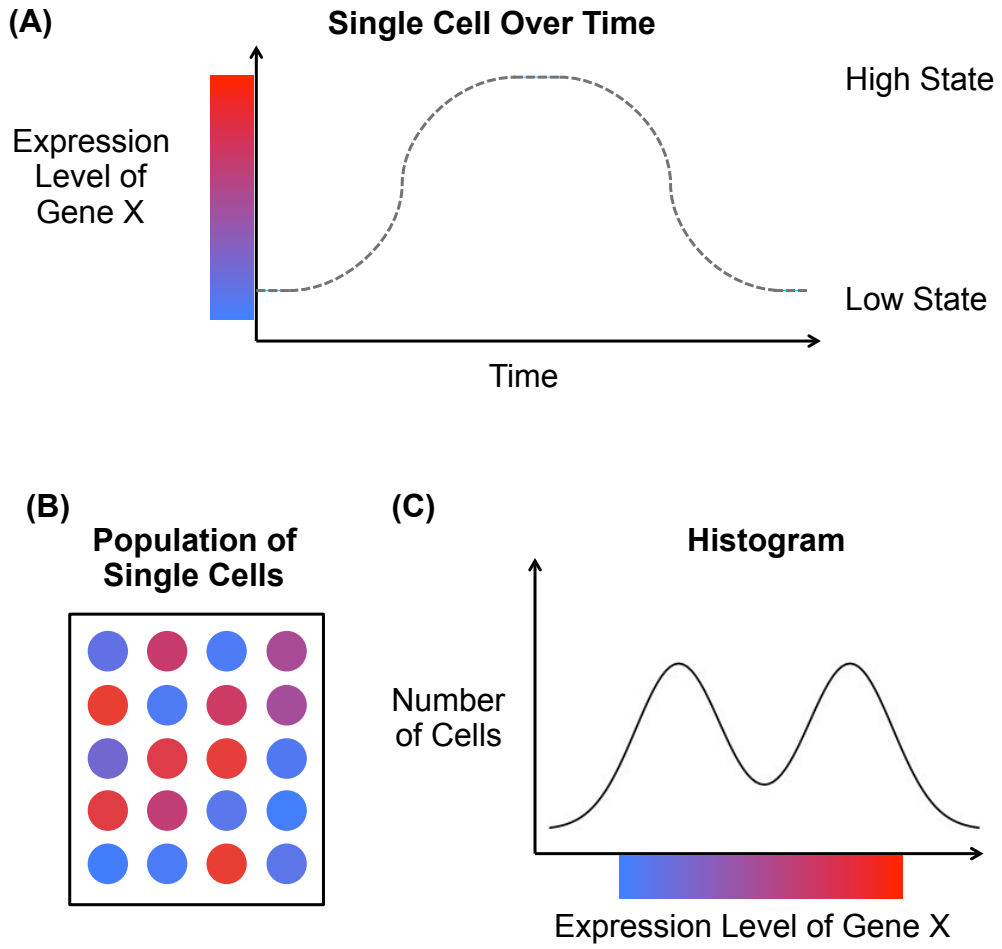


Figure 4.1: Schematic of single-cell expression dynamics and how they can lead to heterogeneity within cell populations. (A) Time series of the expression of gene X in a single cell, which switches back and forth between a high and low state. (B) Population of individual cells shaded by level of expression of gene X at a snapshot in time. (C) Histogram of the expression of gene X for the cell population in (B).

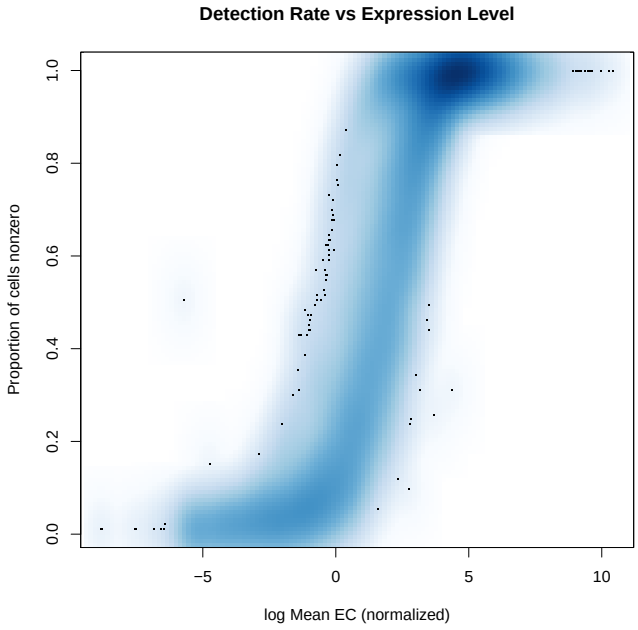


Figure 4.2: Density scatterplot of the detection rate (proportion of cells with nonzero measurements in that gene) versus the log average expression level in cell type DEC.

single-cell RNA-seq protocols involve an unprecedented degree of amplification, which may explain why dropouts are not present to the same degree in bulk. The dropout rate is strongly related to the average expression of the transcript (Shalek et al., 2014; Kharchenko et al., 2014), which suggests that transcripts present at higher levels have a better chance of being amplified (see Figure 4.2). We note that it is possible that some dropout events represent cases where a gene is turned off in an individual cell rather than 'missed', however the rate at which they occur is consistent with the rate observed for control transcripts that are spiked in to every cell (see Appendix D, Figure D.1). Further, we observe considerable presence of multimodality among

the nonzero measurements, which has not been accounted for in scRNA-seq analysis methods to date.

By characterizing these multimodal patterns, we will gain a better understanding of expression stochasticity and heterogeneity of regulatory mechanisms. Here we propose a nonparametric Bayesian modeling framework to infer which genes exhibit these multimodal patterns, and also detect which genes exhibit different patterns across two biological conditions, which we define as differential regulation (DR). In contrast to traditional differential expression (DE), this problem is not as straightforward as detecting a mean shift across different distributions, as there are many ways that two multimodal distributions could differ (see Section 4.8 for examples of different DR patterns). The rest of this chapter is outlined as follows: in Section 4.2 we describe the cell line data for the case study, Sections 4.3-4.4 introduce the generative model, Sections 4.5-4.8 describe the process of fitting the model and inferring which genes fall into each DR pattern category, Sections 4.9.1-4.9.2 present simulation studies to assess sensitivity of the model to prior specification and ability to correctly identify and classify DR genes, and Section 4.10 presents results from application to a case study of human stem cell lines.

4.2 Thomson Lab human stem cell line data

Single-cell RNA-seq data was obtained from the James Thomson Lab at the Morgridge Institute for Research (Leng et al., 2015). Here we present data from two undifferentiated human embryonic stem cell lines: the male H1 line (60 cells) and the female H9 line (87 cells). In addition, we include data from two differentiated

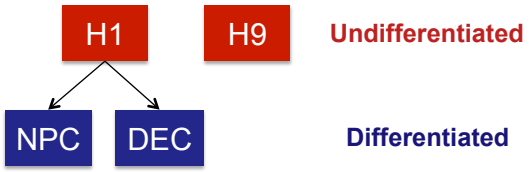


Figure 4.3: Relationship of cell types used in case study

cell types that are both derived from H1: DEC (64 cells) and NPC (86 cells). The relationship between these four lines is summarized by the diagram in Figure 4.3. It is of interest to characterize the differences in regulation of gene expression among and within these four cell types to obtain insight into the differentiation process.

4.3 Dirichlet Process Mixture of normals

Let $Y_g^c = (y_{g1}^c, ..., y_{gJ_c}^c)$ be the log-transformed nonzero expression measurements of gene g for a collection of J_c cells in condition c out of 2 total conditions. For simplicity of presentation, we drop the dependency on g for now, and let the total number of cells with nonzero measurements be J. We assume that under the null hypothesis of equivalent dynamics (i.e. no condition effect), $Y = \{Y^c\}_{c=1,2}$ can be modeled by a conjugate Dirichlet process mixture (DPM) of normals given by

$$\begin{aligned} y_j^c &\sim N(\mu_j, \sigma^2) \\ \mu_j &\sim G \\ G &\sim DP(\alpha, G_0) \\ G_0 &= N(\mu_0, \sigma_0^2) \end{aligned} \tag{4.1}$$

where DP is the Dirichlet process with base distribution G_0 and precision parameter α, and $N(a, b)$ is the Normal distribution parameterized with mean a and variance b. Here, the parameters μ_0 and σ_0^2 are the prior mean and variance for the base distribution of the component-specific means μ_j, and σ^2 is the known component-specific variance (fixed across all components). Let K denote the number of components (unique values among $(\mu, \tau) = \{\mu_j, \tau_j\}_{j=1}^{J}$). Note that two observations indexed by j and j' are from the same cluster if and only if (μ_j, τ_j)=$(\mu_{j'}, \tau_{j'})$. Under this formulation the expected number of clusters is given by Antoniak (1974) as

$$E[K] = \sum_{j=1}^{J} \frac{\alpha}{(\alpha + j - 1)}. \tag{4.2}$$

Thus, as the value of the hyperparameter α increases, the expected number of components obtained from fitting the DPM of normals increases monotonically for fixed J. The value of α needs to be chosen carefully, as the model is very sensitive to its specification (Escobar and West, 1995). Shotwell and Slate (2011) suggest fixing α in a principled manner to incorporate prior beliefs about the number of components. As such, we perform a sensitivity analysis to justify our choice of α, which is reported in Section 4.9.1.

4.4 Product Partition Models

The posterior distribution of μ is intractable even for moderate sample sizes. This is because the number of possible partitions (clusterings) of the data grows extremely rapidly as the sample size increases (according to the Bell number). However, if we let $Z_{=}(z_1, ..., z_J)$ be the vector of component memberships of gene g for all samples,

where the number of unique Z values is K, the likelihood of Y conditional on Z can be viewed as a product partition model (Shotwell and Slate, 2011; Hartigan, 1990). Thus it can be written as a product over all cluster-specific component likelihoods:

$$f(Y|Z,\mu) = \prod_{k=1}^{K} f(y^{(k)}|\mu_k) \tag{4.3}$$

where $y^{(k)}$ is the vector of observations belonging to component k.

Integrating over the parameters μ_k in Equation 4.3, we can obtain the conditional posterior distribution of the data given the clustering:

$$\begin{aligned} f(Y|Z) &= \int f(Y|Z,\mu)p(\mu|Z)d\mu = \int \Big[\prod_{k=1}^{K} f(y^{(k)}|\mu_k)p(\mu_k) \Big] d\mu \\ &= \prod_{k=1}^{K} \int f(y^{(k)}|\mu_k)p(\mu_k)d\mu_k \\ &= \prod_{k=1}^{K} f(y^{(k)}) \end{aligned} \tag{4.4}$$

Where $f(y^{(k)})$ is the component-specific distribution after integrating out the cluster-specific parameter μ_k. In the conjugate normal setting, this has a closed form given by

$$f(y^{(k)}) = \prod_{i=1}^{n^{(k)}} N(y_i^{(k)}|m_{ik}, s_{ik}) \tag{4.5}$$

where $n^{(k)}$ is the number of observations in cluster k. The posterior parameters of the normal distribution also have closed form due to the conjugacy of the model given by Equation 4.1. These parameters are given by

$$
\begin{aligned}
m_{ik} &= \frac{\sigma^2\mu_0 + \sigma_0^2 \sum_{l=1}^{i-1} y_l^{(k)}}{\sigma^2 + (i-1)\sigma_0^2} \\
s_{ik} &= \frac{\sigma^2\sigma_0^2}{\sigma^2 + (i-1)\sigma_0^2} + \sigma^2
\end{aligned}
\tag{4.6}
$$

An equivalent formula to Equation 4.5 that does not rely on recursive definitions for the posterior parameters in Equation 4.6 is given by

$$
\begin{aligned}
f(y^{(k)}) = \frac{\sigma}{(2\pi\sigma^2)^{n^{(k)}/2}\sqrt{n^{(k)}\sigma_0^2 + \sigma^2}} * exp\Big(\frac{-\sum_{i=1}^{n^{(k)}}(y_i^{(k)})^2}{2\sigma^2} - \frac{\mu_0^2}{2\sigma_0^2}\Big) \\
* exp\left(\frac{\frac{\sigma_0^2(\sum_{i=1}^{n^{(k)}} y_i^{(k)})^2}{\sigma^2} + \frac{\sigma^2\mu_0^2}{\sigma_0^2} + 2\mu_0\sum_{i=1}^{n^{(k)}} y_i^{(k)}}{2(n^{(k)}\sigma_0^2 + \sigma^2)}\right)
\end{aligned}
\tag{4.7}
$$

The product partition Dirichlet process mixture model can be simplified as follows

$$
\begin{aligned}
y_j \,|z_j = k, \mu_k &\sim N(\mu_k, \sigma^2) \\
\mu_k &\sim N(\mu_0, \sigma_0^2) \\
z &\sim \frac{\alpha^K\Gamma(\alpha)}{\Gamma(\alpha + J)} \prod_{k=1}^{K} \Gamma(n^{(k)})
\end{aligned}
\tag{4.8}
$$

Then we can obtain the joint posterior distribution of the data Y and clustering Z by incorporating Equation 4.8:

$$
\begin{aligned}
f(Y,Z) &= f(Y|Z)f(Z) \\
&= f(Z)\prod_{k=1}^{K} f(y^{(k)}) \\
&= \frac{\alpha^K\Gamma(\alpha)}{\Gamma(\alpha+J)}\prod_{k=1}^{K}\left[\frac{\sigma}{(2\pi\sigma^2)^{n^{(k)}/2}\sqrt{n^{(k)}\sigma_0^2+\sigma^2}} * exp\Big(\frac{-\sum_{i=1}^{n^{(k)}}(y_i^{(k)})^2}{2\sigma^2} - \frac{\mu_0^2}{2\sigma_0^2}\Big)\right. \\
&\quad \left. * exp\left(\frac{\frac{\sigma_0^2(\sum_{i=1}^{n^{(k)}} y_i^{(k)})^2}{\sigma^2} + \frac{\sigma^2\mu_0^2}{\sigma_0^2} + 2\mu_0\sum_{i=1}^{n^{(k)}} y_i^{(k)}}{2(n^{(k)}\sigma_0^2+\sigma^2)}\right)\Gamma(n^{(k)})\right] \\
&\propto f(Z|Y)
\end{aligned}
\tag{4.9}
$$

4.5 MAP partition estimation

The fitting of the model given in Equation 4.8 was carried out using the algorithm *modalclust* developed by Dahl (2009). This method obtains the maximum a posteriori (MAP) clustering of the data given a fixed component-specific variance estimate and prior parameters for the component-specific means. The MAP clustering is the partition that yields the highest posterior mass (see Equation 4.9).

The hyperparameters for the cluster-specific means were chosen so as to encode a heavy-tailed distribution over the parameters. Specifically, the parameters were set to $\mu_0 = 0$ and $\sigma_0^2 = 100$. The Dirichlet concentration parameter was set to $\alpha = 0.10$, a choice of which is shown in Section 4.9.1 to be robust to many different settings in a sensitivity analysis.

Note that we could also think of placing a prior distribution over the cluster-specific

variance parameter σ^2 as in Shotwell and Slate (2011), where similar closed form solutions for the posterior distribution of the data could also be obtained. However, if we can assume a fixed variance, the algorithms for obtaining estimates of Z_j under the conjugate normal product partition formulation are fast and deterministic (Dahl, 2009). In the unknown variance case, estimates of Z_j require sampling from the (closed-form) full conditional distributions of Z_j and involve Polya urn Gibbs sampling algorithms which are much more computationally intensive (MacEachern, 1994; Bush and MacEachern, 1996; MacEachern and Müller, 1998). In addition, as shown in Section 4.9.1, despite the computational time burden this method does not show improved performance over the algorithm from Dahl (2009) when the variance is well-specified.

Since the model-fitting procedure relies on a fixed cluster variance parameter and a method of estimating it is not provided by Dahl (2009), we implement a procedure using the mixture modeling framework from *Mclust* (Fraley et al., 2012). Briefly, the model with the lowest BIC under the equal-variance constraint is obtained, and the resulting maximum likelihood variance estimate is used as input for the *modalclust* (Dahl, 2009) algorithm. The sensitivity analysis in Section 4.9.1 shows that this procedure has favorable performance under a variety of settings.

4.6 Modeling multimodality within condition

Applying the model to scRNA-seq data within one biological condition, the primary interest lies in detecting which genes exhibit the characteristic multimodality that may be indicative of systematic variation in expression levels. Doing so can provide

insight into the level of heterogeneity among cells belonging to a particular condition, as well as indicate which genes specifically are not expressed at constant levels.

Genes are assessed for multimodality by examining the number of components in the MAP partition, obtained as described in Section 4.5, as well as some additional filtering criteria for robustness. In summary, if a gene meets the following criteria, then it is considered multimodal:

1. There is more than one component in the MAP partition with at least 5 cells
2. Components are separated by at least 3 standard deviations
3. Bayes factor-like score (defined below in Equation 4.10) indicates evidence for multimodality

The first filter criteria is designed to be robust to outlier cells. The second is based on the observation that the model shows poor performance to identify the correct number of components when they are not well-separated (see simulation studies in Section 4.9.2). The last criteria is included so that results are robust to variance specification. The Bayes factor-like score is defined as

$$Score_g = log\left(\frac{f(Y_g, Z_g = Z_{MAP})}{f(Y_g, Z_g = (1, ...1))}\right) \tag{4.10}$$

where $f(Y_g, Z_g = Z_{MAP})$ is the result of plugging in the MAP partition for Z in Equation 4.9, and $f(Y_g, Z_g = (1, ...1))$ is the result of that same equation assuming that only one component exists. Note that the cluster-specific variance estimate σ^2

for the denominator is obtained in a similar manner as described in Section 4.5, but restricting the number of clusters to one.

4.7 Approximate Bayes factor score for condition independence

Ultimately, we would like to calculate a Bayes Factor for the evidence that the data arises from two independent condition-specific models (differential regulation (DR)) versus one overall model that ignores condition (equivalent regulation (ER)). Let $\mathcal{M}_{DR}$ denote the differential regulation hypothesis, and $\mathcal{M}_{ER}$ denote the equivalent regulation hypothesis. A Bayes factor in this context for gene g would be:

$$BF_g = \frac{f(Y_g|\mathcal{M}_{DR})}{f(Y_g|\mathcal{M}_{ER})}$$

where $f(Y_g|\mathcal{M})$ denotes the posterior distribution of the observations from gene g under the given regulation hypothesis. Note, however, that there is no analytical solution for this distribution for the Dirichlet process mixture model. However, under the product partition model formulation, we can get a closed form solution for $f(Y_g, Z_g|\mathcal{M})$. Clearly, the partition Z_g cannot be integrated out. We propose to use an approximate Bayes Factor score:

$$Score_g = log\left(\frac{f(Y_g, Z_g|\mathcal{M}_{DR})}{f(Y_g, Z_g|\mathcal{M}_{ER})}\right) = log\left(\frac{f_{C1}(Y_g^{C1}, Z_g^{C1}) f_{C1}(Y_g^{C2}, Z_g^{C2})}{f_{C1,C2}(Y_g, Z_g)}\right)$$

where the score is evaluated at the MAP estimate $\hat{Z}_g$. A high value of this

score presents evidence that a given gene is differentially regulated. Significance of a positive score is assessed via a permutation test as follows: permute the condition labels, obtain the MAP estimate within the new 'conditions', and calculate the new score. This was done for 1,000 initial permutations for all genes, and then 10,000 more for those that had an unadjusted p-value of less than 0.05.

4.8 Classification of significant DR genes

For genes that are identified as Differentially Regulated (DR) based on a significant permutation p-value (less than 0.05 after adjustment for multiple comparisons using the method of Benjamini and Hochberg (1995)), a post-hoc procedure is applied to classify them into four categories that represent distinct DR patterns of interest as shown in Figure 4.4. Specifically, we are interested in determining whether a gene with a significant BF score is best described by a traditional differential expression (DE) pattern where the number of modes in each condition is the same, but with differing mean. Another interesting pattern that could arise is that of differential modality (DM) where there are a different number of modes in each condition. There could also be the same number of modes in each condition, but a differential proportion (DP) of cells within each mode dependent on condition. Finally, there could be differences in both the number of modes and their mean (DB).

To classify the DR genes into these patterns (DE, DM, DP, and DB), we implement a posthoc procedure. Let c_1 be the number of components in condition 1, c_2 the number of components in condition 2, and c_{OA} the number of components overall (when pooling condition 1 and 2). Only components containing at least 3 cells are

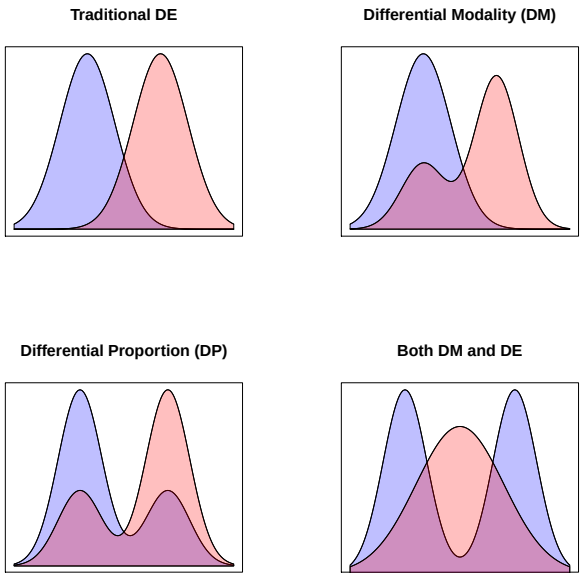


Figure 4.4: Diagram of plausible differential regulation patterns (histograms), including traditional differential expression (upper left), differential modality (upper right), differential proportion within each mode (lower left), and both differential modality and differential expression (lower right).

considered to minimize the impact of outliers. Note that for interpretability we only consider cases where the DR gene satisfies the following criterion

$$c_1 + c_2 \geq c_{OA} \geq min(c_1, c_2) \tag{4.11}$$

These bounds on the number of components overall represent the two extreme cases: condition 1 does not overlap with condition 2 at all, versus one condition completely overlaps with the other. Any cases outside of these boundaries are not readily interpretable in this context. Algorithm B.1 (pseudocode presented in

Appendix B) describes the action to take for all other possible combinations of c_1, c_2, and c_{OA}.

The genes that are not classified as either DE, DP, DM, or DB are considered 'no calls', abbreviated NC. These represent patterns that are not of primary interest, such as those with the same number of components within each condition and overall, but not significantly different cluster-specific means. Genes with this pattern that are significantly DR could arise if, for example, the cluster-specific variances differ across conditions. We do not infer differences of these types since it is possible that they could be explained by cell-specific differences in technical variation (Kharchenko et al., 2014).

An additional step to improve the power to detect genes in the DP category was also implemented. This step was motivated by the observation that power to detect DP genes is low when either the sample size is low or when components are not well-separated in the simulation studies. Thus, for genes that were not significantly DR by permutation but had the same number of components within condition as overall, with cluster-specific means that were not significantly different, Fisher's exact test was used to test for independence with biological condition. If the p-value for that test is less than 0.05, then the gene is added to the DP category. This additional step did not result in the addition of any false positives in the simulation study.

4.9 Simulation studies

4.9 Sensitivity analysis of hyperparameter α and variance specification

In this section we present the results of a simulation study to evaluate the ability of the product partition model to detect the existence of multiple clusters when they truly exist as well as its false positive rate in detecting multiple clusters when there is only one. Since the expected number of clusters depends on the value of α (the Dirichlet concentration parameter) as well as the sample size J (see Equation 4.2), we vary both of these numbers. In addition, we also evaluate performance under different cluster variance (σ^2) specifications: three fixed variance settings (underestimated, true, and overestimated) as well as the maximum likelihood estimates from *Mclust* Fraley et al. (2012), and a setting in which a prior is placed on the variance (using the R package *Profdpm* by Shotwell (2013)). We evaluate each choice of α (0.001, 0.01, 0.02, 0.05, 0.10, 1.0), J (20, 30, 50, 100, 200, 500), and σ^2 (0.7, 1, 1.3, *Mclust* MLE, *Profdpm* prior) under four scenarios:

1. **Null Scenario:** samples are drawn from a standard normal distribution

2. **Two close components:** samples are drawn from a mixture of two normals (equal weight), one standard normal and one with mean equal to two and variance equal to one

3. **Two moderately-separated components:** samples are drawn from a mixture of two normals (equal weight), one standard normal and one with mean

equal to four and variance equal to one

4. **Two well-separated components:** samples are drawn from a mixture of two normals (equal weight), one standard normal and one with mean equal to six and variance equal to one

The specific settings for the distance between the components was chosen to reflect that observed in bimodal genes from the case study (see Section 4.10.2). Specifically, average distance between two modes (standardized by cluster standard deviation) in the case study was close to 4, and ranged from about 3 to 20. Each combination of α, J, σ^2, and scenario is evaluated for 500 replications. The model was fit in the same way as described in Section 4.5 with vague cluster-specific prior parameters $\mu_0 = 0$ and $\sigma_0 = 100$. *Mclust* was constrained to the equal cluster variance setting and allowed a maximum of 5 components. *Profdpm* was implemented using the 'gibbs' method with default input settings for prior parameters and iterations. Unless otherwise specified, all analyses were carried out using R software version 3.1.1 (R Core Team, 2014).

The evaluation for each scenario includes the concordance of MAP clustering estimates with the true clustering, measured using the Rand index (proportion of concordant observation pairs (Hubert and Arabie, 1985)). The higher the Rand index, the better the clustering estimate. Table 4.1 displays the overall median Rand index across all sample sizes, alpha settings, and replications (along with the range of the median Rand index in each of the settings).

In addition, each scenario is also assessed for false discovery rate (how often the MAP estimate identifies something other than the true number of clusters). Results

are shown for the correctly specified variance case in Figure 4.5 and when the variance estimated via *Mclust* in Figure 4.6. The results for the additional variance settings are displayed in Appendix A (see Figure A.1 for the underestimated variance case, Figure A.2 for the overestimated variance case, and Figure A.3 for the case with variance modeled with a Gamma prior and fit via *Profdpm*).

	Scenario			
Variance	1 (Null)	2 (Distance 2)	3 (Distance 4)	4 (Distance 6)
0.7	1.00 -	0.69 (0.52-0.72)	0.95 (0.93-0.96)	1.00 (0.99-1.00)
1.0	1.00 -	0.49 (0.48-0.50)	0.95 (0.93-0.96)	1.00 -
1.3	1.00 -	0.49 (0.48-0.50)	0.95 (0.93-0.96)	1.00 -
Profdpm	1.00 (0.05-1.00)	0.49 (0.48-0.55)	0.92 (0.48-0.95)	1.00 (0.63-1.00)
Mclust	1.00 -	0.49 (0.48-0.50)	0.95 (0.93-0.96)	1.00 -

Table 4.1: Overall median Rand index (range of median over 500 replications) for simulation study

In the interest of balancing the misclassification rates under the various scenarios and variance specifications, we choose a value of α of 0.10. Except in the hardest case (Scenario 2 with very close clusters), this setting of α is fairly robust to variance misspecification. In Scenario 2, the misclassification rate is high and the Rand index is low even when the variance is correctly specified. These results suggest that the PPM is not reliable for detecting the existence of more than one component when the components are very close together. This result is in accordance with what is seen in the case studies, where very few bimodal genes are detected with standardized component distances less than 3. We also note that the *modalclust* algorithm with *Mclust* variance estimates outperforms the *Profdpm* method which places a prior distribution on the variance and estimates the MAP by sampling from the posterior.

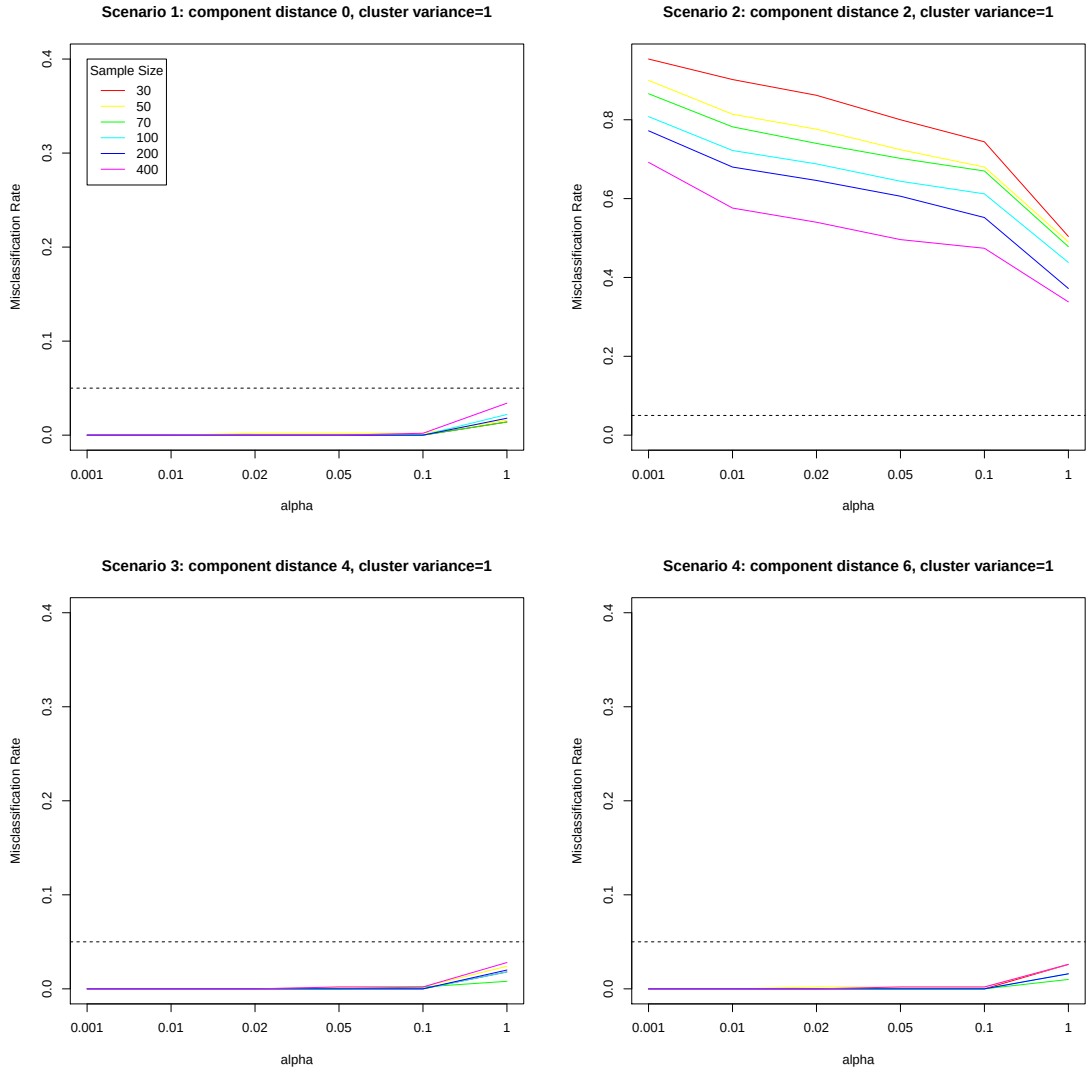


Figure 4.5: Proportion of replicates failing to identify the correct number of clusters in scenario 1 (upper left), 2 (upper right), 3 (lower left) and 4 (lower right) when correct cluster variance is specified

4.9 Identification of differentially regulated genes

Here we implement a simulation study to assess the performance of the Bayes Factor score to identify DR genes, as well as the post-hoc procedure to classify them into

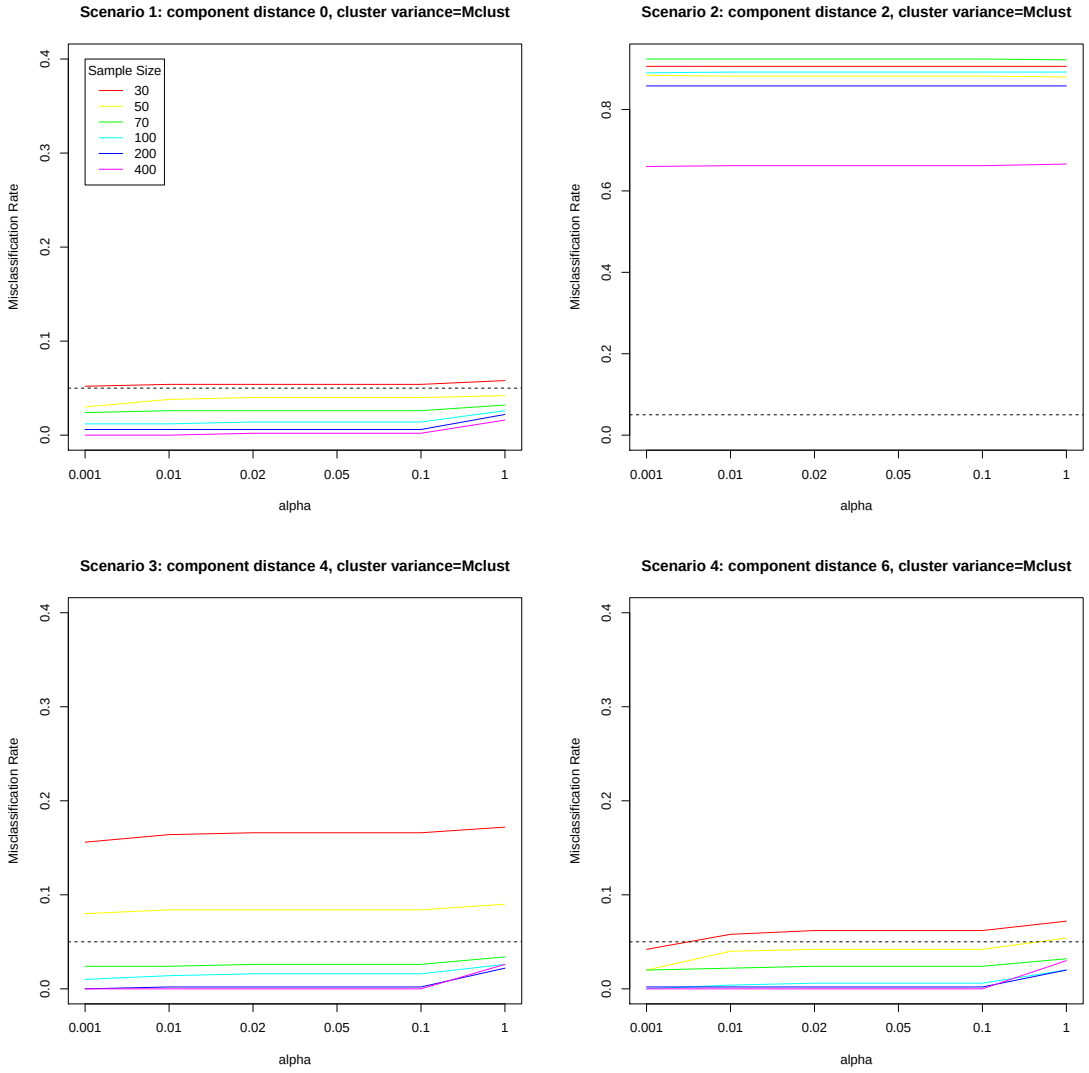


Figure 4.6: Proportion of replicates failing to identify the correct number of clusters in scenario 1 (upper left), 2 (upper right), 3 (lower left) and 4 (lower right) when cluster variance is estimated via *Mclust*

the four categories (DE, DP, DM, and DB) - see Section 4.8 for more details. A set of 10,000 genes was generated from normal mixture distributions for two conditions at two different sample size settings (50 cells in each condition and 100 cells in each

condition). The majority of the genes (9,000) were simulated out of the same model in each condition, and the other 1,000 represent genes with the four types of differential regulation (DR) outlined in Figure 4.4. Of the 9,000 null genes, 5,000 were generated from a single normal component (EE) and the other 4,000 from a two-component normal mixture (EP). Component-specific variance parameters were fixed at 1, and distances between component means Δ_μ were varied for the two-component case, with 800 genes at each setting of $\Delta_\mu \in \{2, 3, 4, 5, 6\}$. The 1,000 DR genes were split equally into the four categories:

- DE: single component with different mean in each condition
- DP: two components with equal component means across conditions, proportion in low mode is 0.33 for condition 1 and 0.66 for condition 2
- DM: single component in condition 1, two components in condition 2 with one overlapping component and 25% of condition 2 cells belong to the second mode
- DB: single component in condition 1, two components in condition 2 with no overlapping components and mean of condition 1 is half-way between the means in condition 2.

For all scenarios, where there are two components (or one in each condition) the distance between means was also varied by Δ_μ as in the EP case. When there are two components, if not otherwise specified, half of the cells belong to each component. The evaluation of model performance on this simulated set is assessed based on (1) ability to detect the correct number of components, (2) ability to detect significantly

DR genes, and (3) ability to classify DR genes into their correct categories. These three criteria are explored in the next three sections, respectively.

4.9.2.1 Modality within condition and overall

Similarly to the sensitivity analysis of Section 4.9.1, we first examine the ability of the framework to detect the correct number of components within each condition and overall. The results, separated by gene category, are shown in Table 4.2. The table also includes the proportion for which the correct number of components were identified within each condition and overall (row labeled 'All Three'). The classification rates improve with increased sample size. The settings with one component in each condition (EE and DE) have the highest classification rates, whereas lower rates are observed in the more difficult cases with more than one component in both conditions (EP and DP).

		True Gene Category					
Sample Size	Condition	EE	EP	DE	DP	DM	DB
50	1	0.965	0.688	0.976	0.648	0.964	0.972
	2	0.972	0.687	0.948	0.632	0.688	0.680
	Overall	0.989	0.749	0.740	0.676	0.760	0.996
	All Three	0.934	0.599	0.684	0.552	0.660	0.664
100	1	0.991	0.753	0.996	0.700	0.996	0.992
	2	0.988	0.752	0.988	0.676	0.756	0.740
	Overall	0.998	0.754	0.776	0.668	0.812	1.00
	All Three	0.978	0.694	0.768	0.600	0.752	0.736

Table 4.2: Proportion of genes in each category where the correct number of components were identified

4.9.2.2 Detection of DR genes

Next, we examine the ability of the modeling framework to identify the non-null genes as significantly DR using the permutation test on the Bayes factor scores. The power to detect each category gene as DR is shown in Table 4.3. The overall false discovery rate is 0.016 for sample size 50 and 0.024 for sample size 100. Note that the calculations here are taken before the classification step, so power is defined as the proportion of genes from each simulated category that are detected as DR. Power to detect DR genes is improved with increased sample size. The DE and DM cases exhibit higher power than the DP and DB cases.

	True Gene Category				
Sample Size	DE	DP	DM	DB	Overall
50	1.00	0.580	0.680	0.640	0.725
100	1.00	0.800	0.912	0.712	0.856

Table 4.3: Power to detect DR genes by true category

4.9.2.3 Classification of DR genes

Next, we examine the ability of the framework as a whole to detect *and* classify each DR gene into its corresponding category. In contrast to the previous section, here power is defined as the proportion of genes detected and classified correctly. Here we define two different types of FDR to differentiate between a false discovery at the DR score stage and a false discovery at the classification stage. Briefly, for a given category we let

$$FDR_{null} = \frac{\text{number of null genes assigned to that category}}{\text{total number of genes assigned to that category}}$$

$$FDR_{class} = \frac{\text{number of genes assigned that belong to a different category}}{\text{total number of genes assigned to that category}}$$

		Gene Category			
Sample Size	Statistic	DE	DP	DM	DB
50	Power	0.924	0.532	0.572	0.628
	FDR_{null}	0.022	0.000	0.000	0.000
	FDR_{class}	0.125	0.000	0.021	0.133
100	Power	0.984	0.556	0.724	0.712
	FDR_{null}	0.0290	0.000	0.000	0.020
	FDR_{class}	0.258	0.000	0.011	0.103

Table 4.4: Power to detect and classify DR genes by category

The results for the power and two different types of FDR are shown in Table 4.4. We see that the FDR_{class} is rather high in the DE and DB cases. This means that a substantial proportion of the DR genes classified as DE or DB truly belong to another category. To investigate whether these misclassified genes tend to belong to certain cluster mean distance settings, we next examine the power and FDR_{class} by Δ_μ. These results are displayed in Table 4.5, and demonstrate that the misclassification events happen most often in the scenarios where Δ_μ is small. Examining these events in more detail, it is the case that the majority of them occur when the correct number of components is not identified.

Finally, we examine the classification step in isolation. Table 4.6 displays the proportion of true positive DR genes that are assigned to the correct DR category. It is clear that the ability of the algorithm to correctly classify DR genes improves as

Sample Size	Gene Category	Cluster mean distance Δ_μ 2	3	4	5	6
50	DE	0.92 (0.21)	0.94 (0.28)	0.96 (0.04)	0.86 (0)	0.94 (0.04)
	DP	0 (0)	0.16 (0)	0.58 (0)	0.94 (0)	0.98 (0)
	DM	0.02 (0.50)	0.14 (0)	0.84 (0.05)	0.92 (0)	0.94 (0)
	DB	0 (1)	0.30 (0.17)	0.92 (0.10)	1.00 (0.17)	0.92 (0.06)
100	DE	0.96 (0.58)	1.00 (0.32)	1.00 (0)	1.00 (0)	0.96 (0)
	DP	0 (0)	0.20 (0)	0.64 (0)	0.96 (0)	0.98 (0)
	DM	0 (1)	0.68 (0)	0.96 (0)	1.00 (0)	0.98 (0.02)
	DB	0.02 (0.50)	0.54 (0.25)	1.00 (0.14)	1.00 (0)	1.00 (0.06)

Table 4.5: Power (FDR_{class}) to detect and classify genes in each category stratified by Δ_μ

the component mean distance increases.

Sample Size	Gene Category	Cluster mean distance Δ_μ 2	3	4	5	6
50	DE	0.92	0.94	0.96	0.86	0.94
	DP	0.00	0.67	0.85	1.00	0.98
	DM	0.08	0.26	0.98	0.94	0.98
	DB	0.00	0.44	0.92	1.00	0.92
100	DE	0.96	1.00	1.00	1.00	0.96
	DP	0.00	0.32	0.82	0.98	0.98
	DM	0.00	0.72	0.96	1.00	0.98
	DB	0.03	0.61	1.00	1.00	1.00

Table 4.6: Correct classification rates (proportion of true positive DR genes assigned to the correct category) stratified by Δ_μ

4.10 Case studies

4.10 Data normalization and preprocessing

For each of the cell types described in Section 4.2, expected counts were obtained from RSEM (Li and Dewey, 2011). In each condition there are a maximum of 96 cells, but all have fewer than 96 cells due to removal by quality control standards. Some cells were removed due to cell death or contamination, indicated by a very low percentage of mapped reads (Leng et al., 2015). DESeq median normalization (Anders and Huber, 2010) was carried out using the `MedianNorm` function in the *EBSeq* R package to obtain library sizes. The library sizes were applied to scale the count data. Note that ideally, counts would also be normalized for varying cell size with the use of spike-in data (Brennecke et al., 2013), however spike-ins were not used in the experiment. Further, genes with zero measurements in more than 75% of the cells within each condition were removed from consideration due to low sample size for fitting the DP mixture models.

4.10 Results

4.10.2.1 Multimodal genes within condition

The number of genes with 1, 2, 3, 4, or 5 components for each cell type is displayed in Table 4.7. For each, we see that one and two component genes are by far the most common, with only a handful of genes containing 5 components. We see that the H1 cell type has the largest proportion of genes with only one component (70.1%), whereas NPC has the largest proportion of genes with more than one component

(57.1%).

	Number of Components					Total
Cell Type	1	2	3	4	5	Considered
H1	7821	3082	224	25	3	11155
H9	5471	5079	426	30	6	11012
DEC	5085	4632	491	57	5	10270
NPC	4635	5599	503	54	4	10795

Table 4.7: Total number of genes by number of components identified

4.10.2.2 Differentially regulated genes across conditions

The number of significant DR genes for each cell type comparison is shown in Table 4.8. Note that the comparison of H1 and H9 detects the least amount of DR genes, a finding that is consistent with the fact that both of these are undifferentiated human stem cell lines and it is expected that they are the most similar among the comparisons. Histograms of the top 20 significantly differentially regulated genes for each of the four DR categories are plotted in Appendix C, Figures C.1-C.16.

	DR Category			
Comparison	DE	DP	DM	DB
H1 vs NPC	1079	428	1148	1772
H1 vs DEC	1576	338	1435	1973
NPC vs DEC	1159	512	800	1124
H1 vs H9	794	142	146	506

Table 4.8: Number of DR genes identified in Thomson Lab cell line data

4.11 Discussion

Single-cell RNA-seq experiments provide unprecedented ability to probe cellular heterogeneity in a transcriptome-wide manner, and while many established tools for the analysis of bulk RNA-seq data are widely available, it is imperative that the issues relating to increased biological and technical variability be carefully considered in any scRNA-seq analysis. A recent overview and set of recommendations for when bulk RNA-seq tools can safely be applied to scRNA-seq data is provided by Stegle et al. (2015). Bulk RNA-seq tools are not directly applicable in the study of expression stochasticity, since bulk experiments only provide measurements of average expression across a pool of cells.

To our knowledge, we have presented the first statistical method to detect differences in scRNA-seq experiments that explicitly accounts for potential multimodality of the distribution of expressed cells in each condition. This is of great interest, since these multimodal expression patterns may represent biological heterogeneity within otherwise homogeneous cell populations. When these patterns differ across two conditions, this could mean that a gene is regulated differently in the two groups. We have introduced a set of four interesting patterns to summarize the key features that can differ between two conditions. When applied to cells at different differentiation states, this information may provide insight into which genes are responsible for driving phenotypic changes.

We stress that our approach is inherently different from a method that detects traditional differential expression, such as that developed by Kharchenko et al. (2014),

which aims to detect a shift in the mean. In addition to identifying genes that have different mean expression levels across conditions, our modeling framework allows us to identify subpopulations within each condition that have differing levels of expression (i.e. which cells belong to which component). For such genes, the clustering automatically provides an estimate of the proportion of cells in each condition that belong to each subpopulation. We also do not require specification of the total number of components, which can vary for each gene. We note that Kharchenko et al. (2014) also present a method for identifying subpopulations of cells, but these are identified on the basis of similarity measures across all genes at once. Further, our method is able to detect and characterize more complex differences than a mean shift (e.g. difference in the number of subpopulations, or modes).

Based on extensive simulation studies, we conclude that the nonparametric Bayesian hierarchical modeling framework is able to reliably detect the true number of components when the sample size is large enough and the components are well-separated. In addition, the algorithm to classify DR genes into their true category is robust when components are well-separated and improves with increasing sample size. We observe that multimodality is present at considerable levels in two undifferentiated and two differentiated human stem cell lines. This further validates the need for flexible methods, such as the one presented here, that do not assume unimodality or fix the number of components. We also point out that the smallest number of differentially regulated genes is found in the comparison between the two undifferentiated cell lines, which are thought to be the most homogeneous of all the comparisons examined.

Improvements could be made to the framework in order to overcome limitations. First, as noted in Section 4.1, a large source of technical variation in scRNA-seq experiments is the amount of dropout observed. While the modeling framework proposed here is appropriate for data arising out of a mixture of gaussian components, it does not explicitly account for dropout events. This may become an important issue when there are global differences in the dropout rate between two conditions, as it could result in inference that the two conditions have differential regulation when they in fact do not. This can be seen in a hypothetical example gene where two conditions have the same number of components and the means are the same in each, but the dropout rate is twice as high in one of the conditions compared to the other. Since we assume (based on the relationship between average expression magnitude of a gene and its dropout rate among cells, described in Section 4.1) that dropout events are more likely to occur for low-magnitude expressors, the condition with the higher dropout rate may appear to have higher expression than the other if only the nonzero observations are considered.

To account for such differences, one may think of explicitly modeling the dropout events alongside the nonzero counts. Recent scRNA-seq analyses have done just that by modeling them as separate mixture components (Shalek et al., 2014; Kharchenko et al., 2014). In Shalek et al. (2014), the model assumes that all dropouts arise from a spike-mass component at zero, and the primary goal is to test the null hypothesis of whether the amount of zeroes is consistent with the expected number of zeroes due to technical noise. Modeling of the nonzero observations is done independently of the dropout events, and any inference performed on them (i.e. differential expression

analysis) assumes that the observations arise from a unimodal lognormal component. The authors found that for close to 10% of the genes under study, the unimodal model failed a goodness-of-fit, which motivates the need for more flexible models.

Kharchenko et al. (2014) include a low-magnitude Poisson component to accommodate the excess zeroes from dropout events. In contrast to Shalek et al. (2014) however, the zeroes could arise from either the Poisson dropout component, or the main negative binomial count distribution. The advantage of this zero-inflated modeling approach is the ability to directly estimate the probability that a given zero observation is generated from the dropout component versus the count distribution.

Our approach could be similarly extended to augment the mixture model to include a zero component. However, we are not primarily interested in estimating differences in proportions of zeroes (because the current belief is that the majority of dropouts arise due to technical error). Instead we believe there is much to gain by developing a normalization procedure that will not only adjust for differences in sequencing depth, but also for differences in the level of technical noise that causes dropout events. Efforts to do so are under way. We note that this task could likely be improved by utilizing spike-in data, as there are likely unidentified factors that affect detection rates beyond the average transcript expression and number of genes detected per cell.

One final limitation is that our framework does not make inference on differential variability across conditions. While our model does *allow* for the variance to differ between conditions, it does not identify or characterize patterns associated with it. Detecting this type of difference may be of interest, however the sources of cell-specific

technical variability are still not well-understood, so differences across condition could possibly be confounded with these unknown sources of variation.

A SENSITIVITY ANALYSIS FOR ADDITIONAL VARIANCE SETTINGS

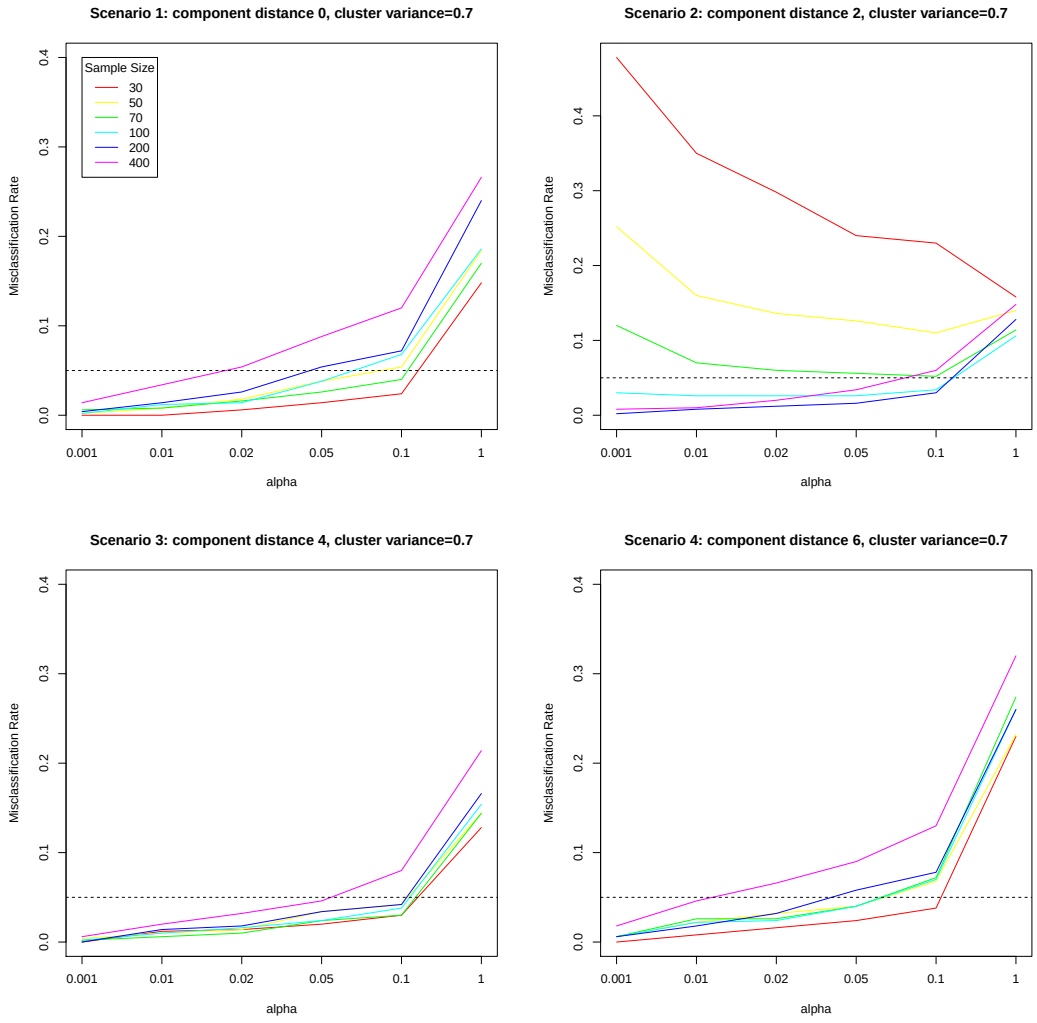


Figure A.1: Proportion of replicates failing to identify the correct number of clusters in scenario 1 (upper left), 2 (upper right), 3 (lower left) and 4 (lower right) when cluster variance is underestimated

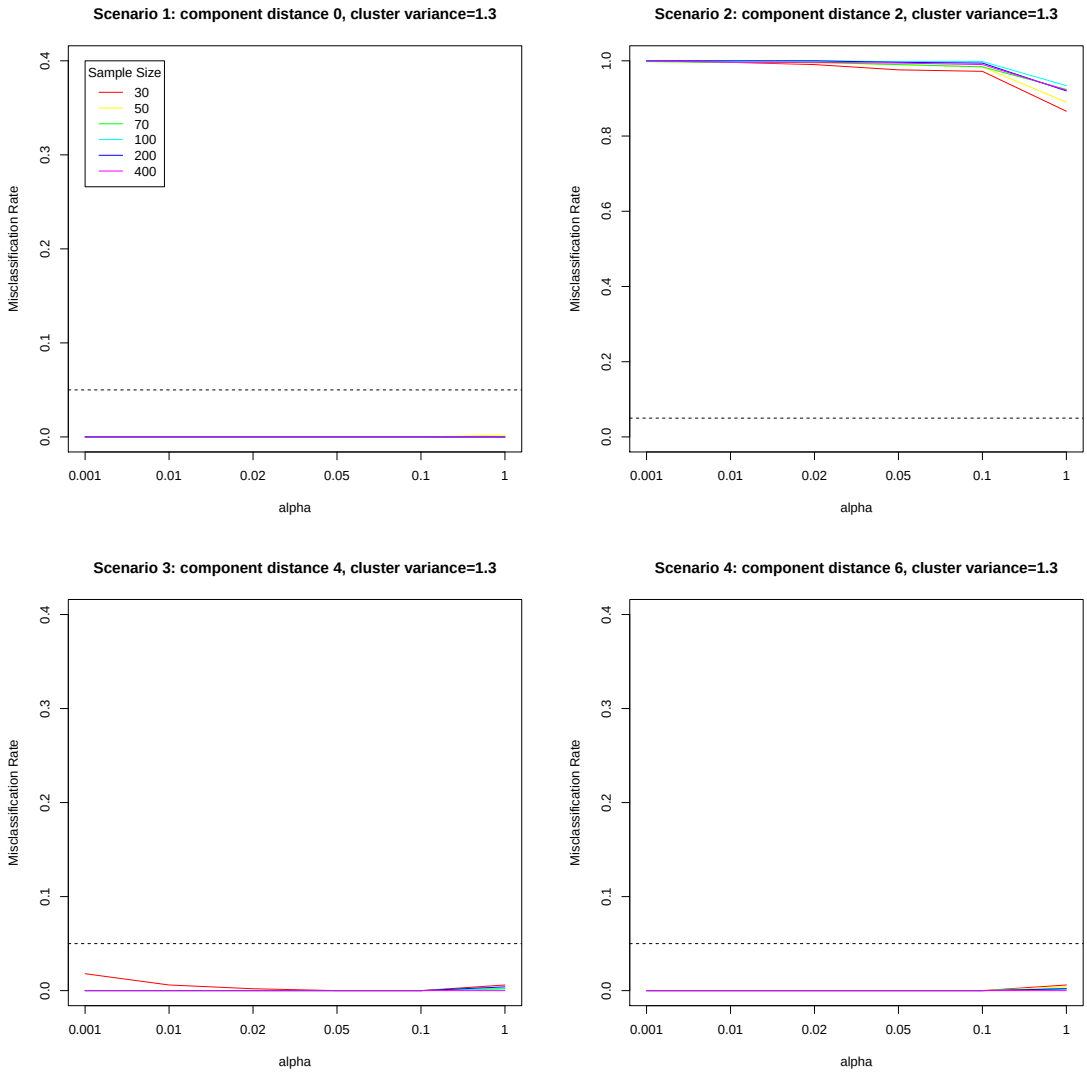


Figure A.2: Proportion of replicates failing to identify the correct number of clusters in scenario 1 (upper left), 2 (upper right), 3 (lower left) and 4 (lower right) when cluster variance is overestimated

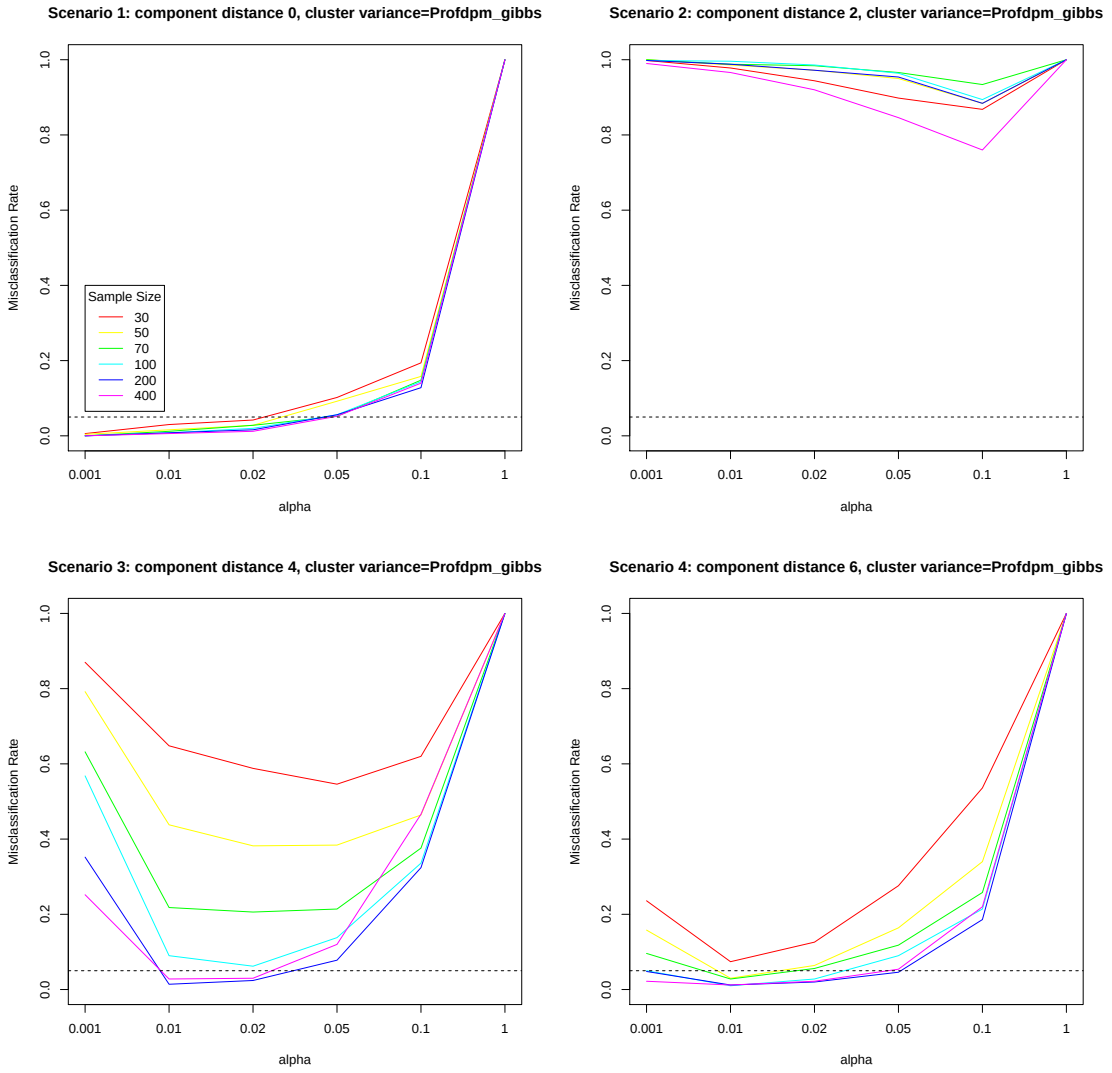


Figure A.3: Proportion of replicates failing to identify the correct number of clusters in scenario 1 (upper left), 2 (upper right), 3 (lower left) and 4 (lower right) when cluster variance is modeled with a Gamma prior and fit via Profdpm

B DR CLASSIFICATION ALGORTHM

Pseudocode for the classification of DR genes in to the categories DE, DP, DM, or DB. Note that the genes that are not classified as either DE, DP, DM, or DB are considered 'no calls', abbreviated NC. More detail is provided in the main text in Section 4.8.

Algorithm B.1 DR classification

if $c_1 = c_2$:

if $c_1 = c_2 = 1$:

if $c_{OA} = 1$: perform t-test of cluster means

if significant at 0.01 level ⇒ DE

else if not significant at 0.01 level ⇒ NC

if $c_{OA} = 2$ ⇒ DE

else if $c_1 = c_2 \geq 2$:

if if $c_1 = c_2 = c_{OA}$: Fisher's exact test for condition independence

if significant at 0.01 level ⇒ DP

else if not significant at 0.01 level ⇒ NC

else if $c_1 = c_2 < c_{OA}$: perform pairwise t-tests of cluster means

if at least one pair significant at 0.01 level ⇒ DE

else if none significant at 0.01 level ⇒ NC

else if $c_1 \neq c_2$: perform pairwise t-tests of cluster means

if at least one pair not significant at 0.01 level ⇒ DM

else if all pairs are significant at 0.01 level ⇒ DB

C TOP DR GENES IN THOMSON DATA

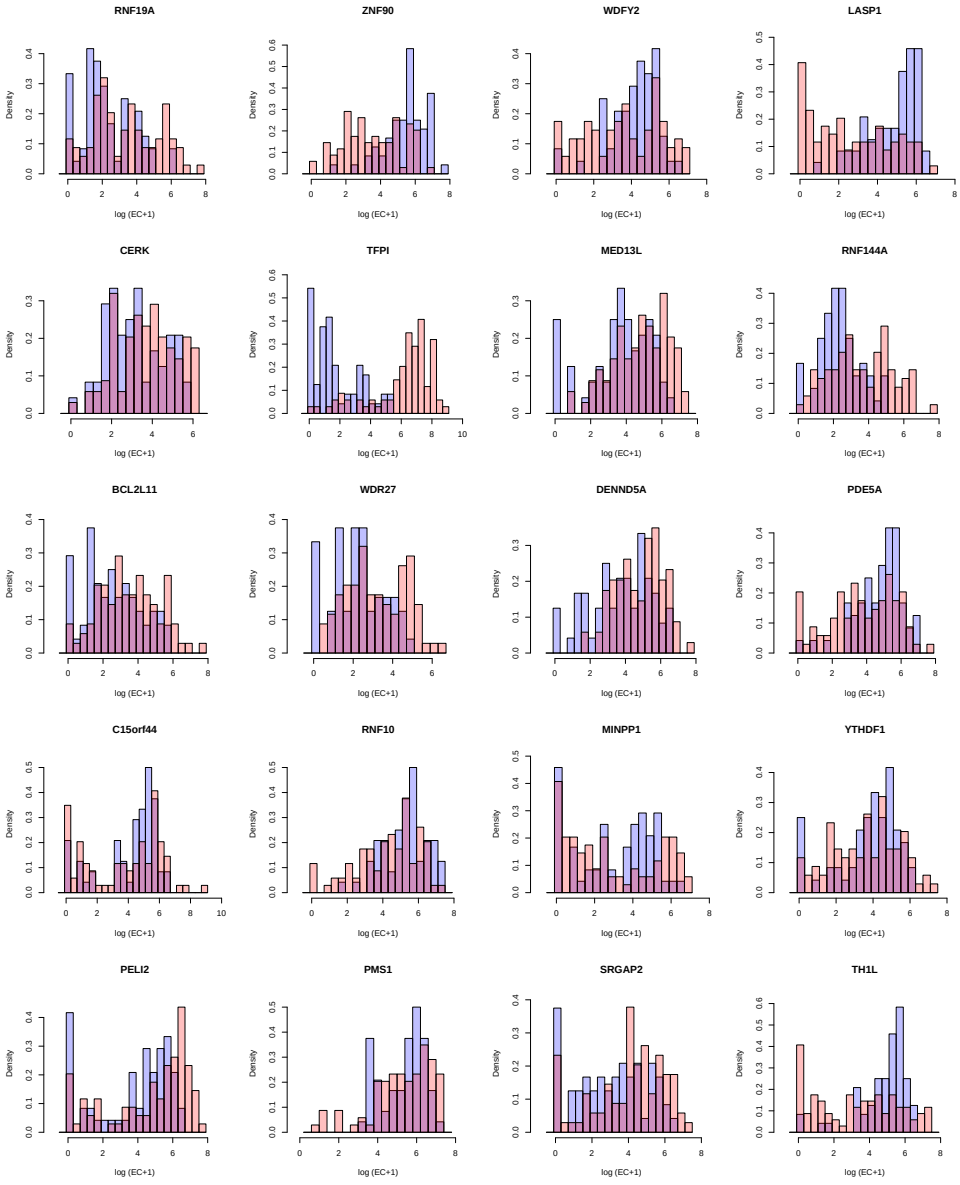


Figure C.1: Top 20 DE genes for H1 vs NPC ranked by Bayes factor score

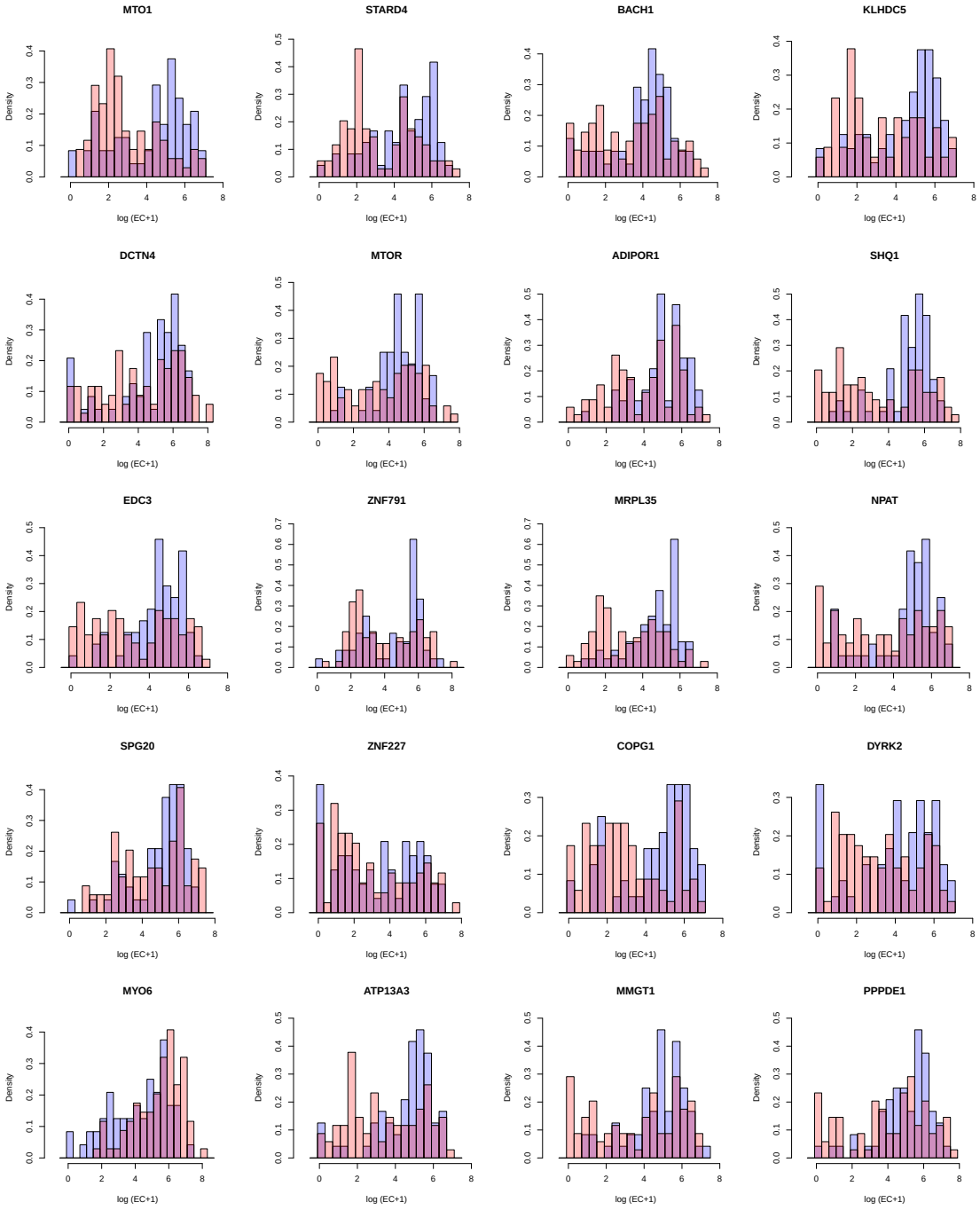


Figure C.2: Top 20 DP genes for H1 vs NPC ranked by Bayes factor score

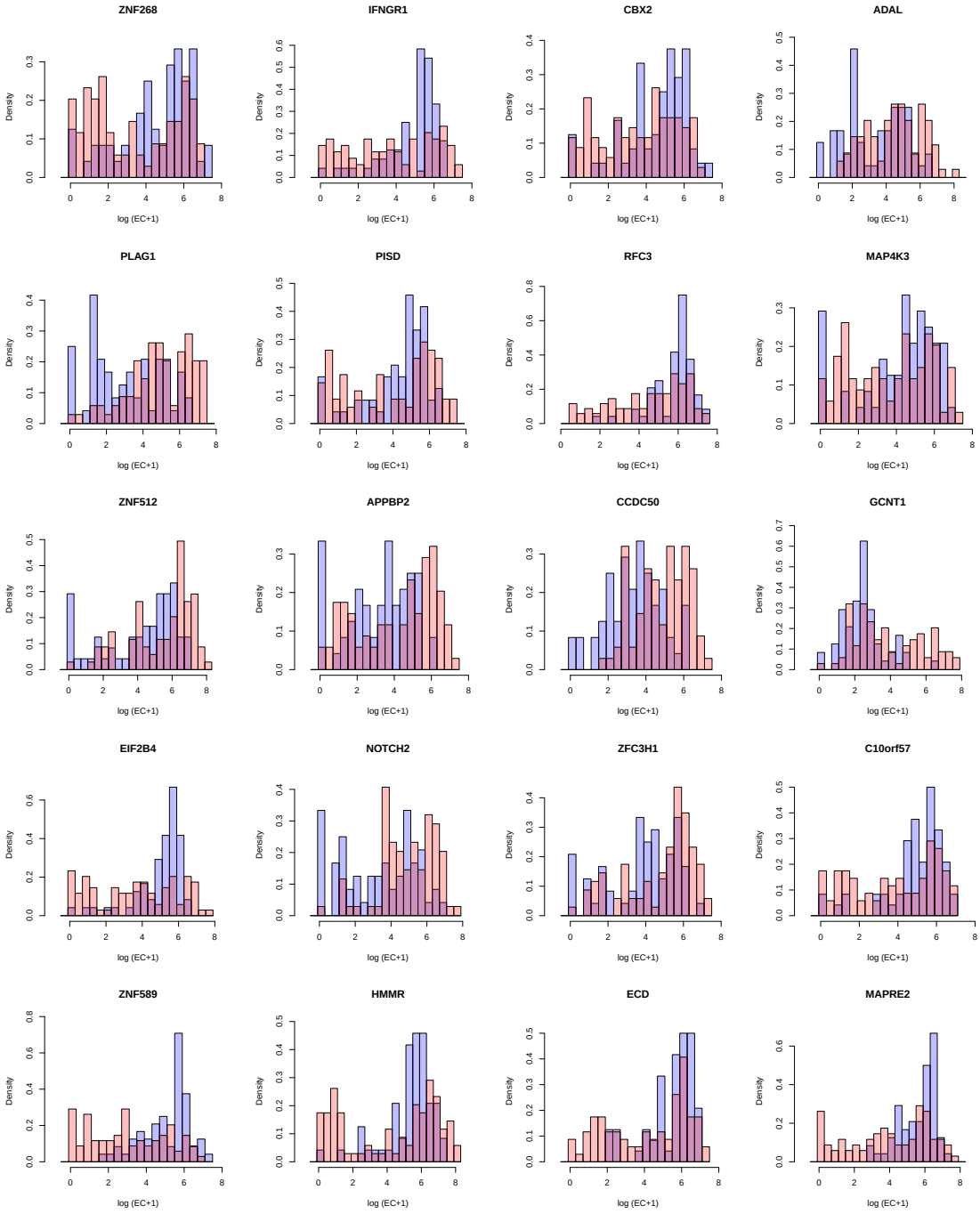


Figure C.3: Top 20 DM genes for H1 vs NPC ranked by Bayes factor score

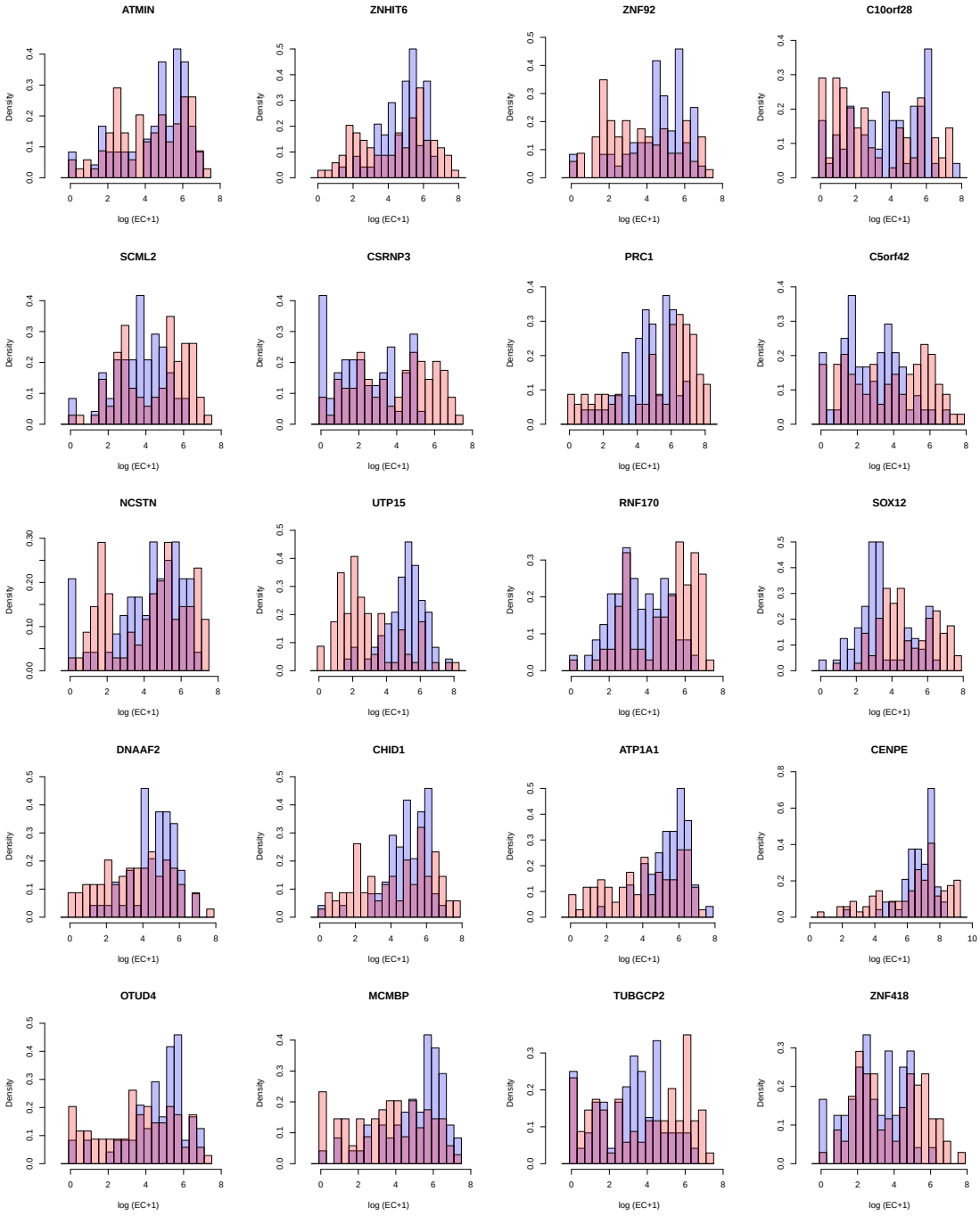


Figure C.4: Top 20 DB genes for H1 vs NPC ranked by Bayes factor score

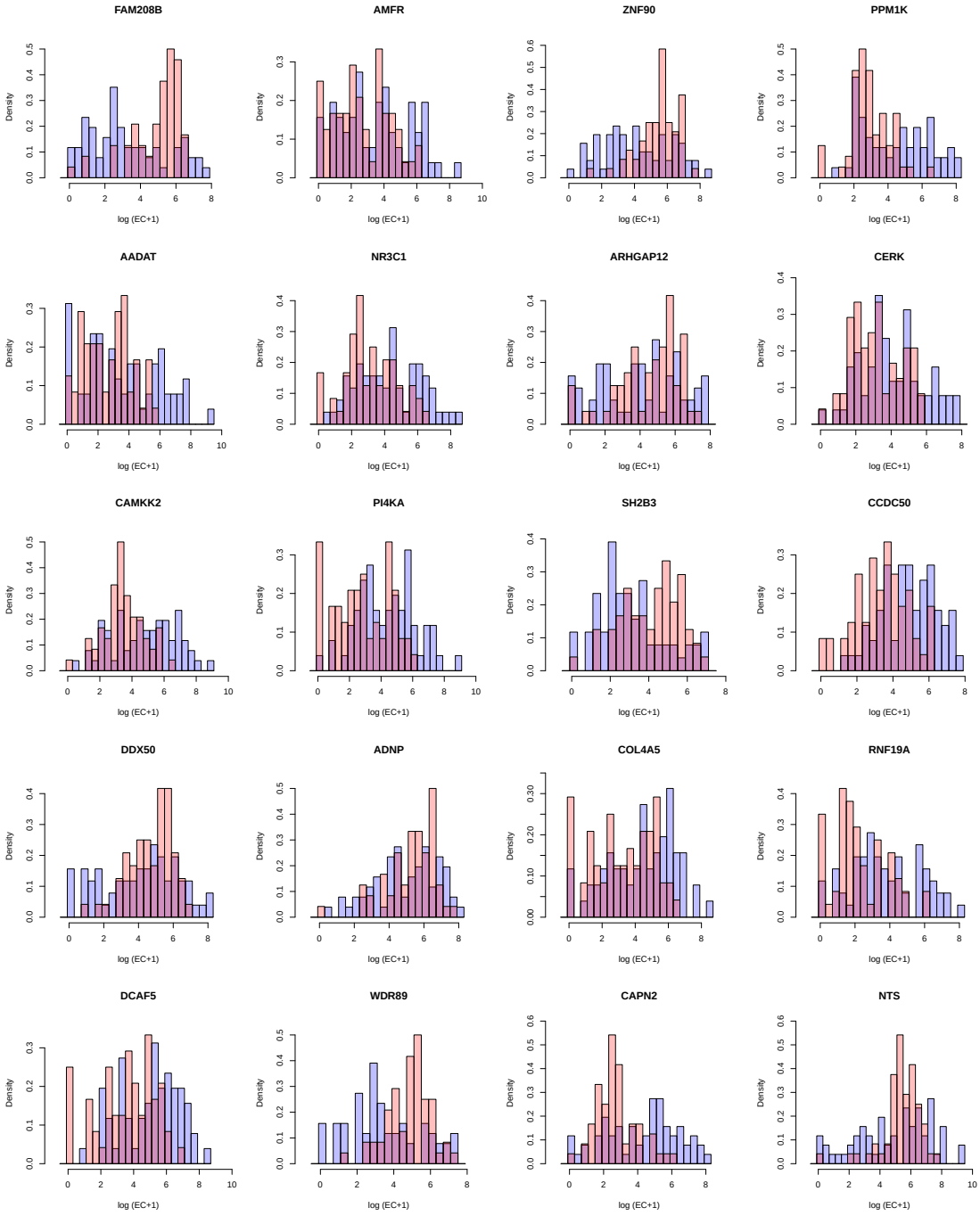


Figure C.5: Top 20 DE genes for H1 vs DEC ranked by Bayes factor score

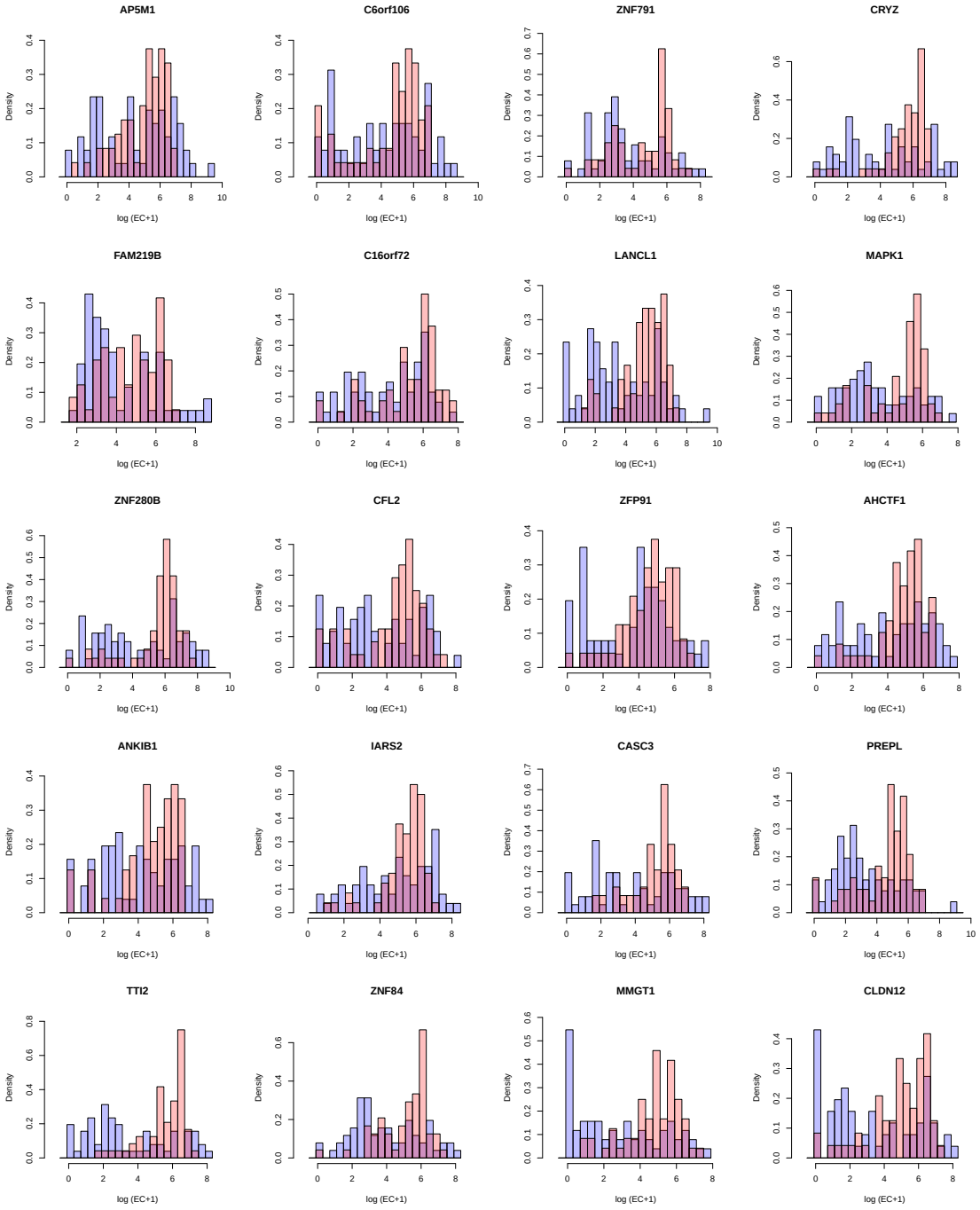


Figure C.6: Top 20 DP genes for H1 vs DEC ranked by Bayes factor score

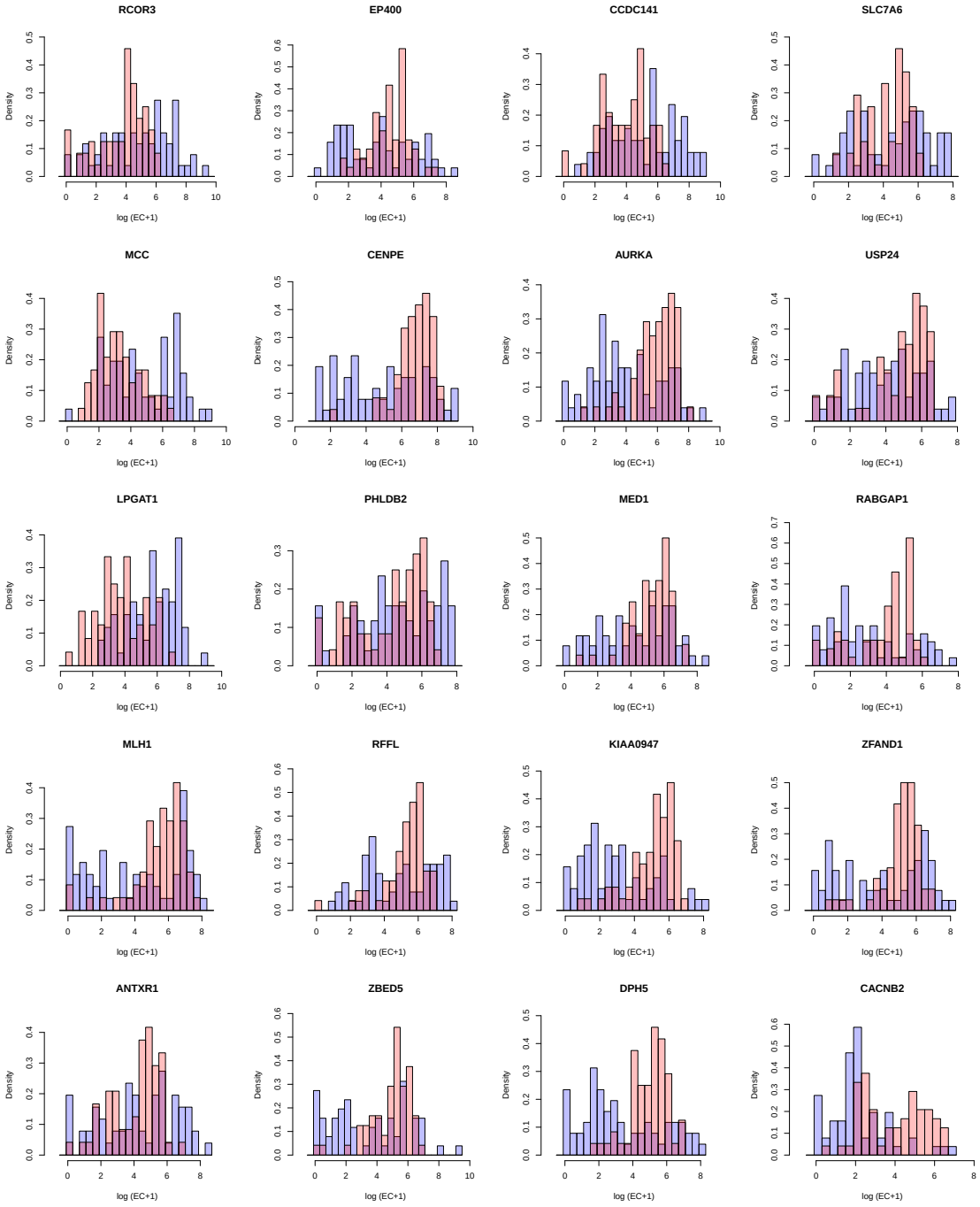


Figure C.7: Top 20 DM genes for H1 vs DEC ranked by Bayes factor score

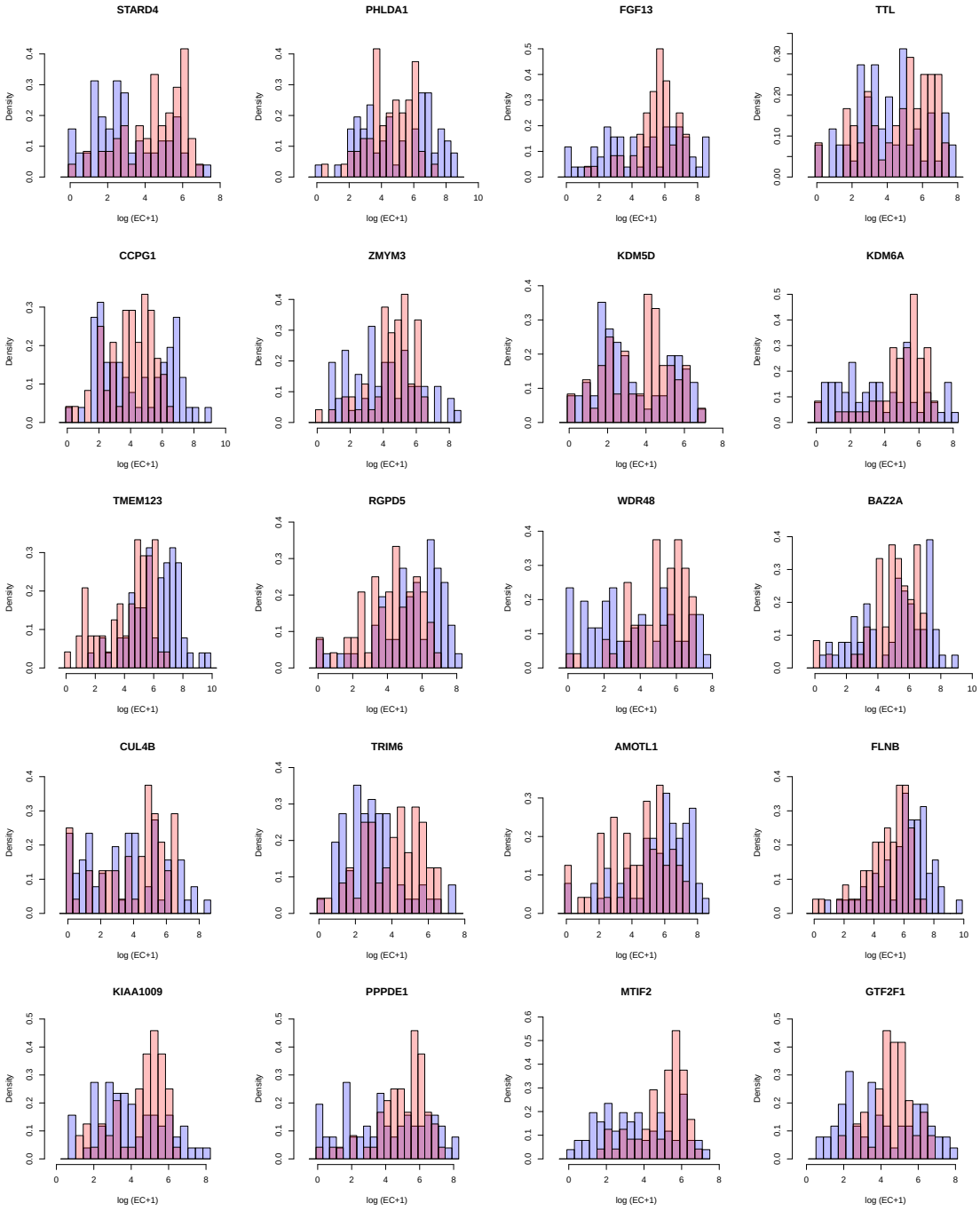


Figure C.8: Top 20 DB genes for H1 vs DEC ranked by Bayes factor score

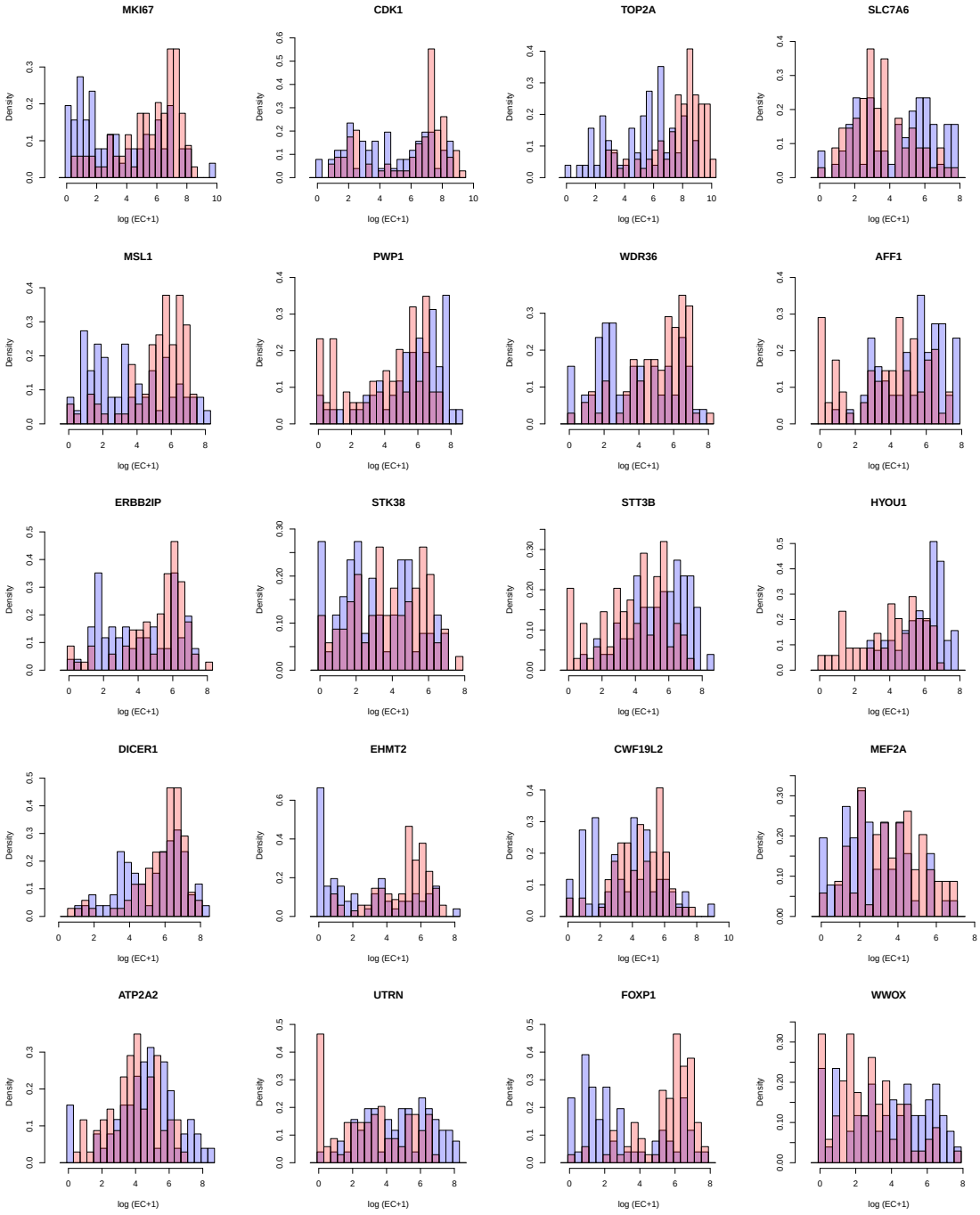


Figure C.9: Top 20 DE genes for DEC vs NPC ranked by Bayes factor score

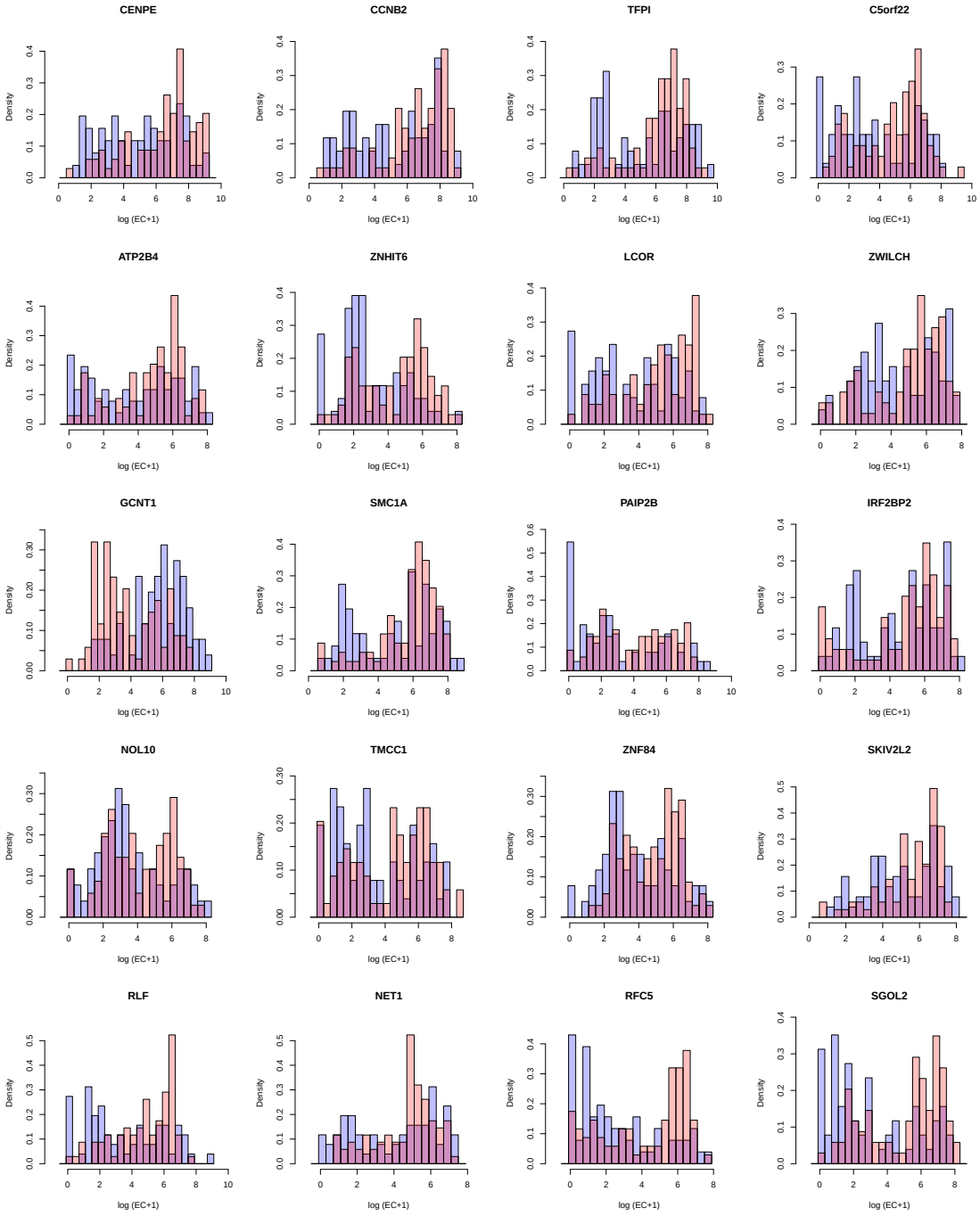


Figure C.10: Top 20 DP genes for DEC vs NPC ranked by Bayes factor score

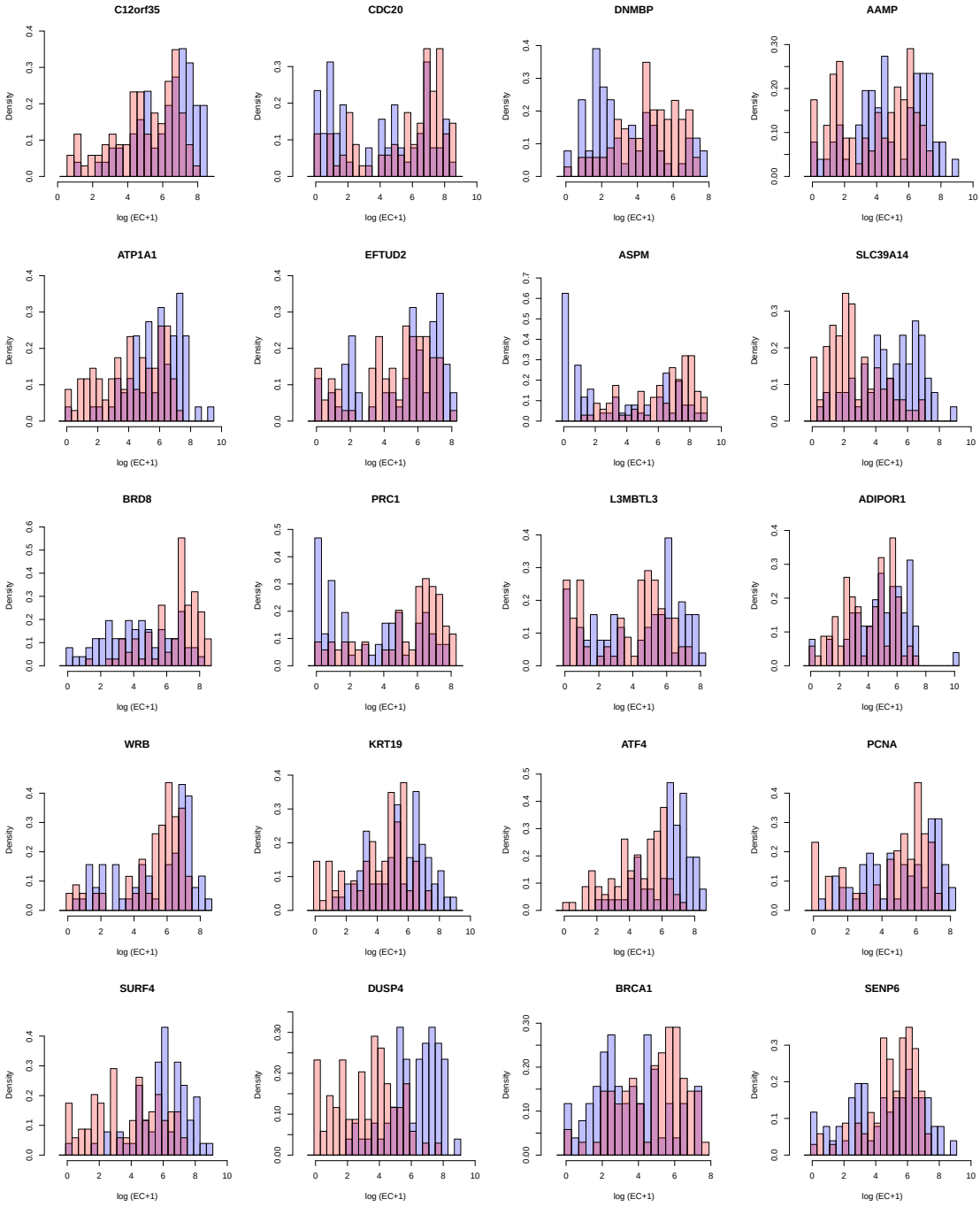


Figure C.11: Top 20 DM genes for DEC vs NPC ranked by Bayes factor score

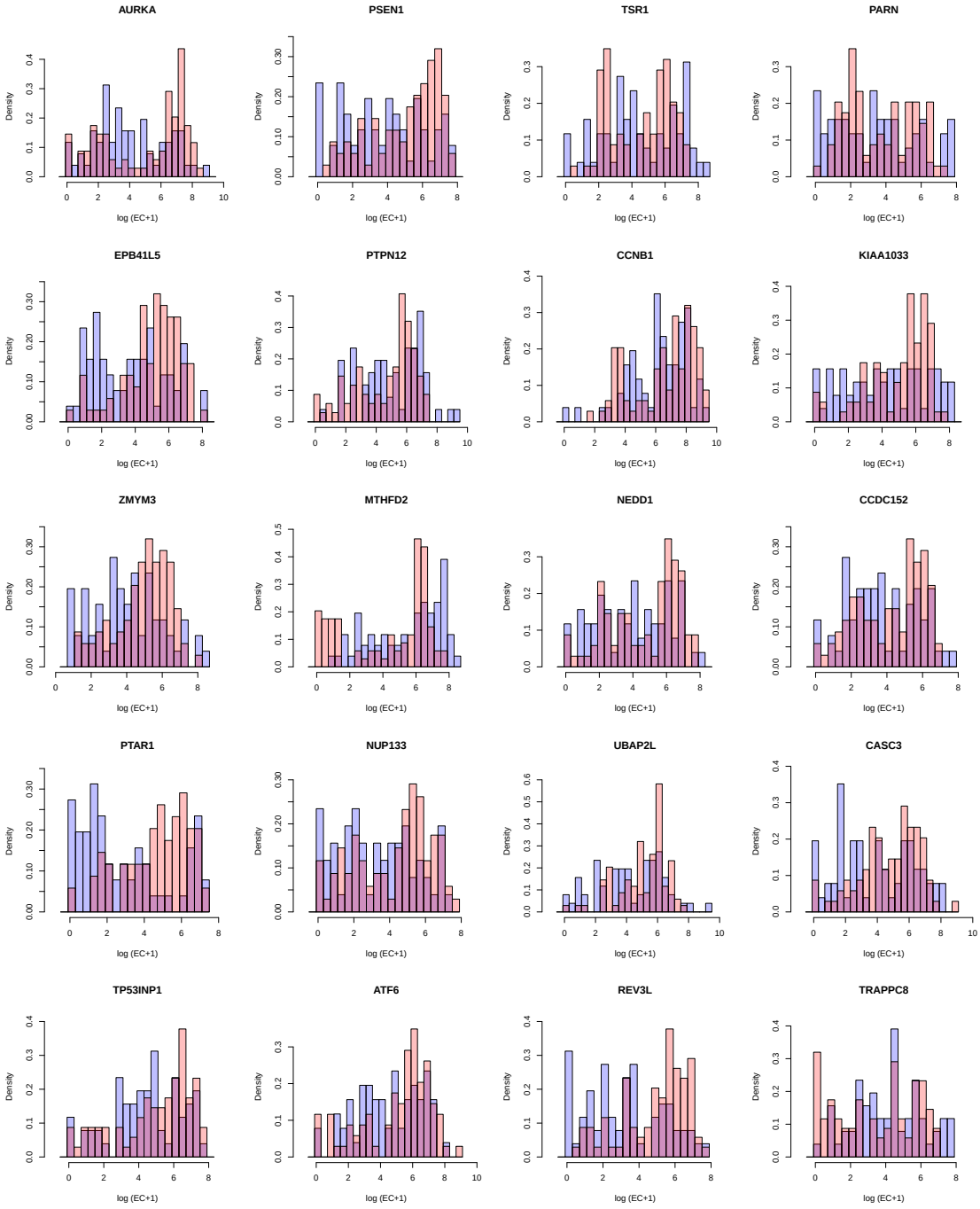


Figure C.12: Top 20 DB genes for DEC vs NPC ranked by Bayes factor score

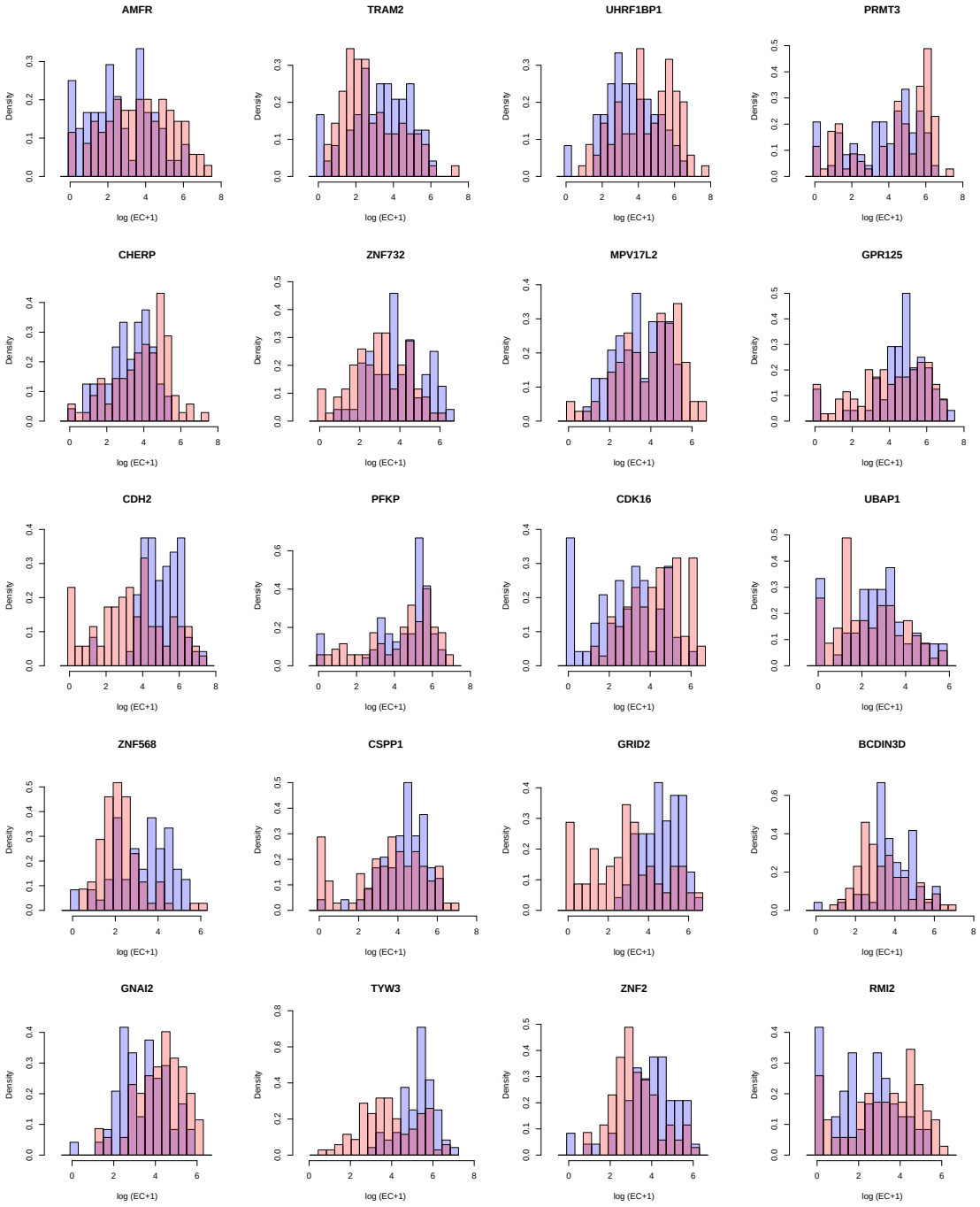


Figure C.13: Top 20 DE genes for H1 vs H9 ranked by Bayes factor score

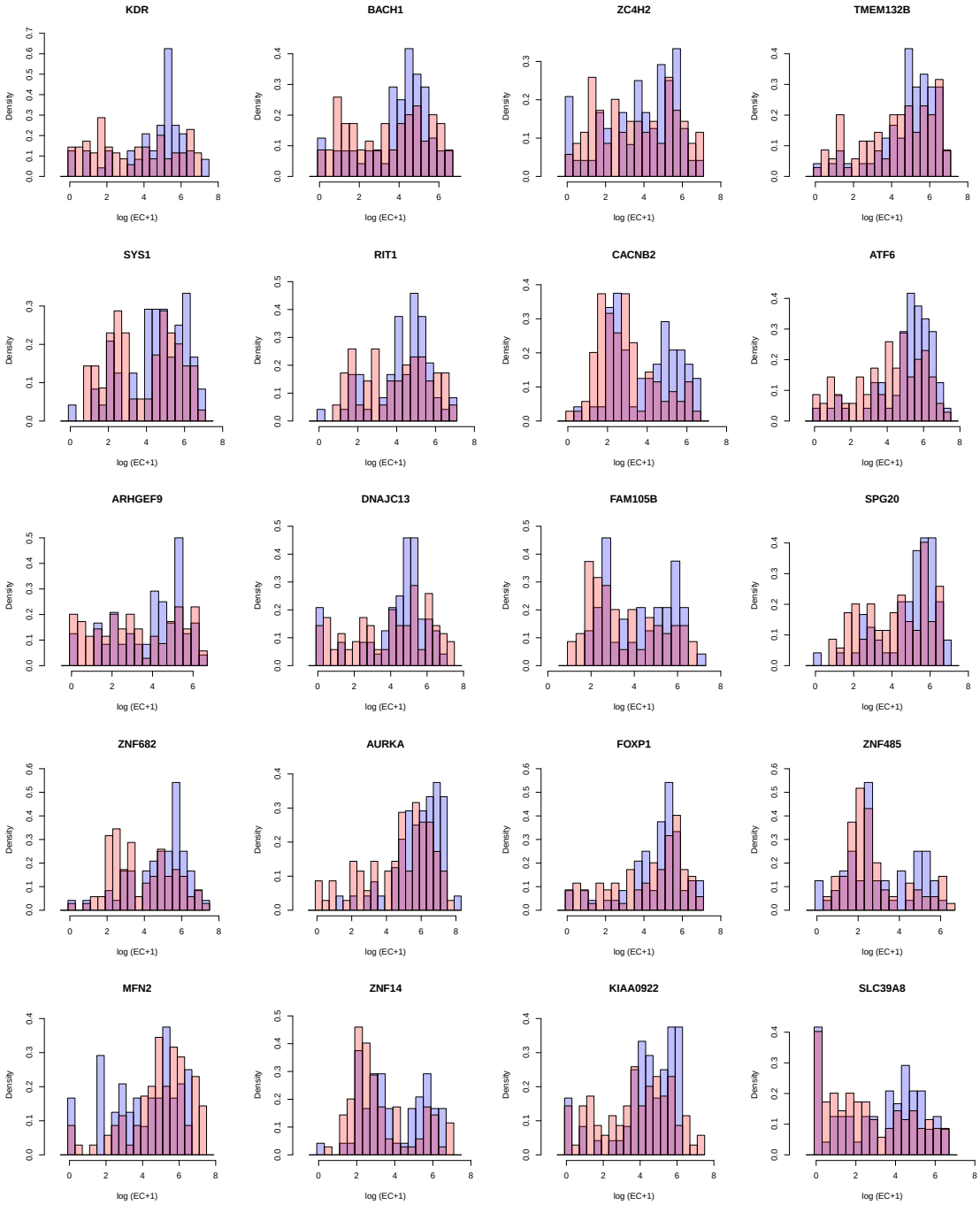


Figure C.14: Top 20 DP genes for H1 vs H9 ranked by Bayes factor score

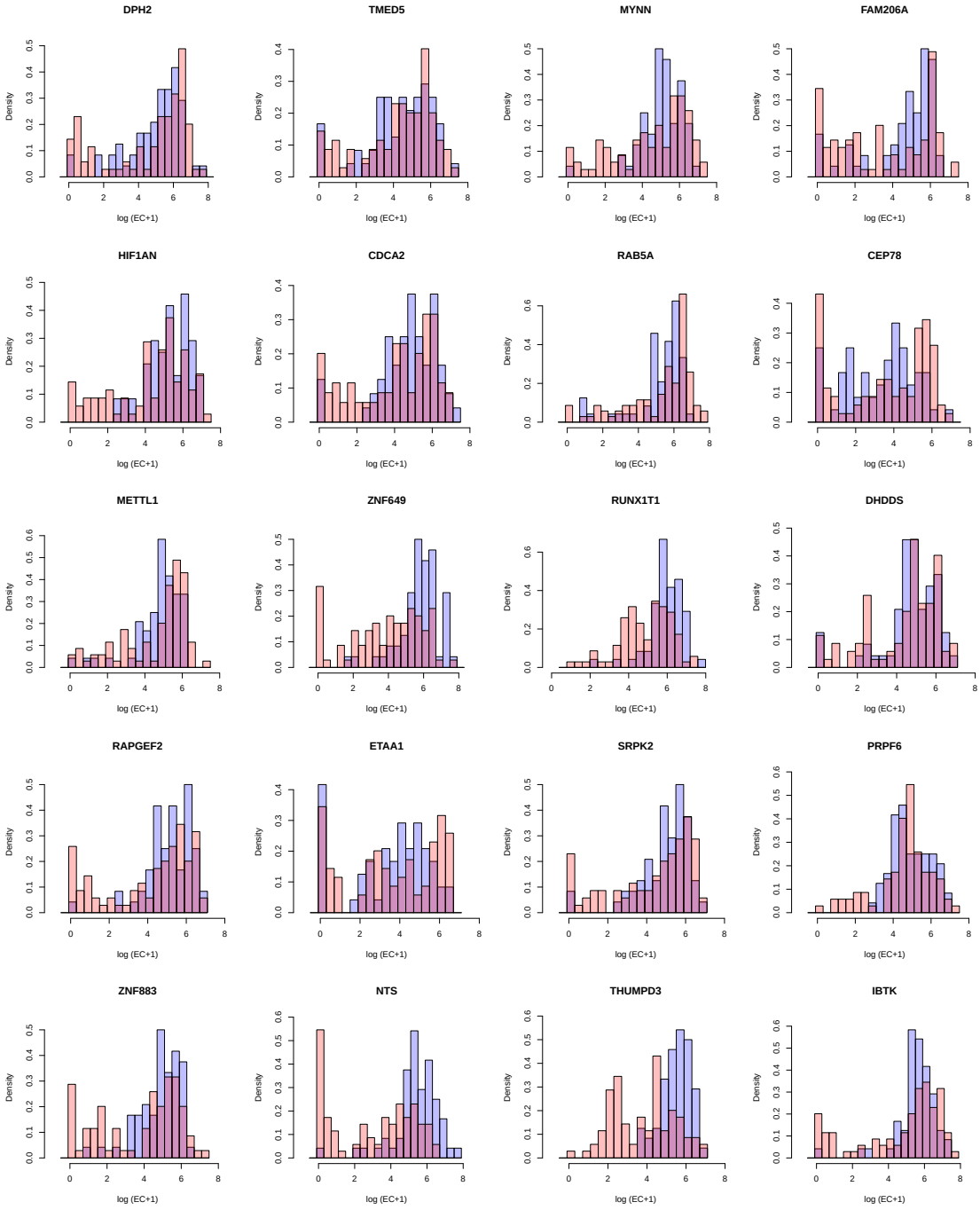


Figure C.15: Top 20 DM genes for H1 vs H9 ranked by Bayes factor score

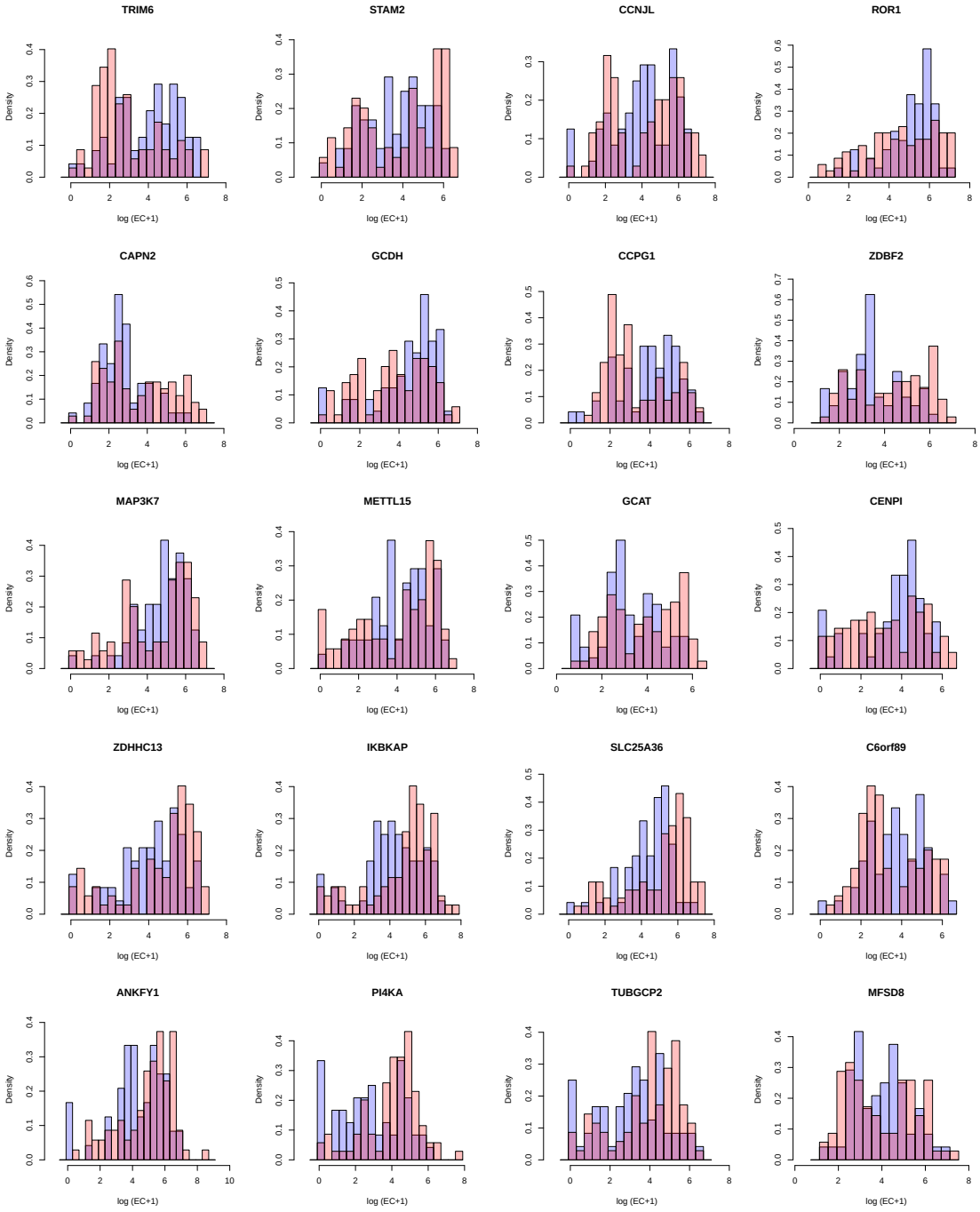


Figure C.16: Top 20 DB genes for H1 vs H9 ranked by Bayes factor score

D DROPOUTS RATES IN A PILOT STUDY USING SPIKE-IN CONTROLS

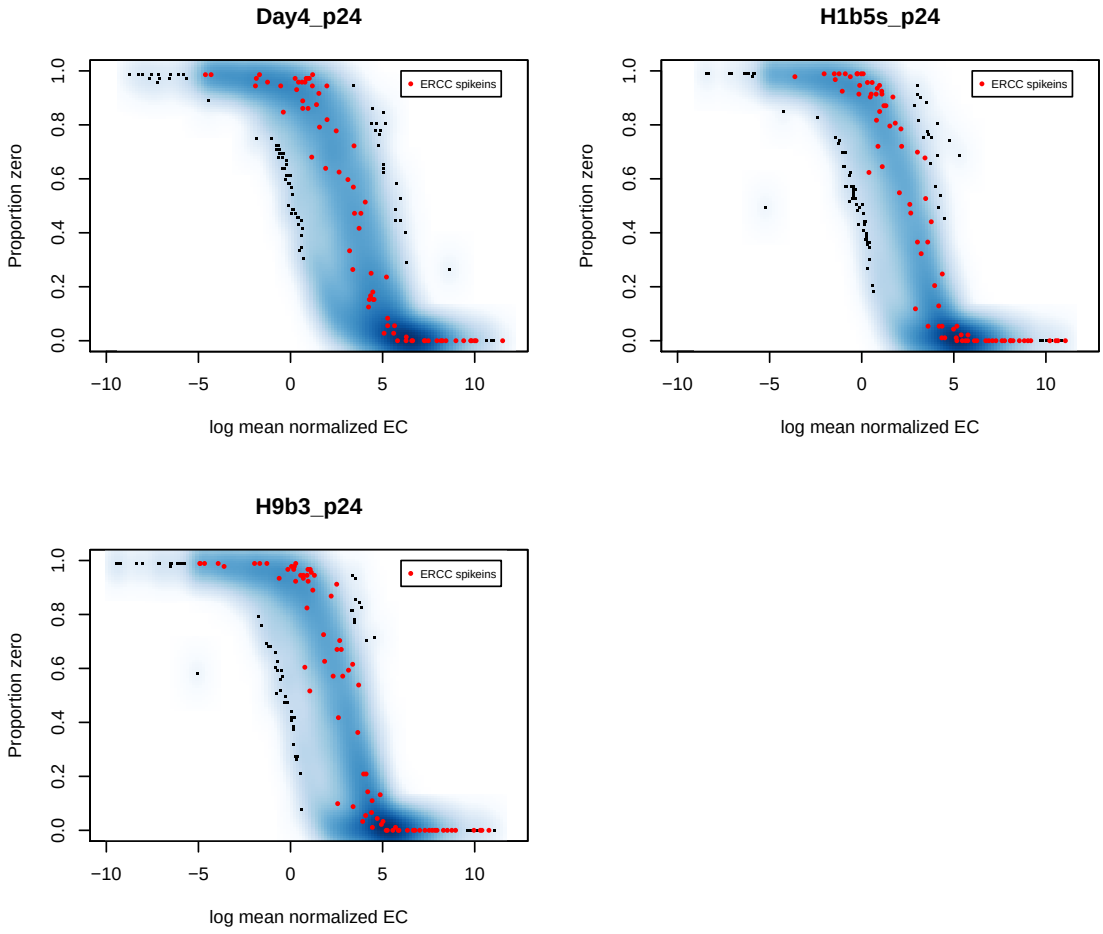


Figure D.1: Density scatterplot of the detection rate (proportion of cells with nonzero measurements in that gene) versus the log average expression level in a pilot experiment to investigate the properties of ERCC spike-ins. Blue points are endogenous genes and red points are the 92 ERCC control RNAs.

REFERENCES

Adzhubei, I. A., et al. 2010. A method and server for predicting damaging missense mutations. *Nature Methods* 7(4):248–249.

Anders, S., and W. Huber. 2010. Differential expression analysis for sequence count data. *Genome Biology* 11(10):R106.

Antoniak, C. E. 1974. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics* 1152–1174.

Barretina, J., et al. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483(7391):603–607.

Bender, R., T. Augustin, and M. Blettner. 2005. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine* 24:1713–1723.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.

Birkbak, N. J., et al. 2013. Tumor mutation burden forecasts outcome in ovarian cancer with brca1 or brca2 mutations. *PloS One* 8(11):e80023.

Birtwistle, M. R., J. Rauch, A. Kiyatkin, E. Aksamitiene, M. Dobrzyński, J. B. Hoek, W. Kolch, B. A. Ogunnaike, and B. N. Kholodenko. 2012. Emergence of bimodal cell population responses from the interplay between analog single-cell signaling and protein expression noise. *BMC Systems Biology* 6(1):109.

Blei, D. M., and J. D. McAuliffe. 2008. Supervised topic models. In *Advances in neural information processing systems 20*, ed. J. C. Platt, D. Koller, Y. Singer, and S. Roweis, 121–128. Cambridge, MA: MIT Press.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bozic, I., et al. 2010. Accumulation of driver and passenger mutations during tumor progression. *Procedings of the National Academy of Sciences* 107(43):18545–18550.

Brennecke, P., S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, et al. 2013. Accounting for technical noise in single-cell rna-seq experiments. *Nature Methods* 10(11):1093–1095.

Buettner, F., K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle. 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*.

Bush, C. A., and S. N. MacEachern. 1996. A semiparametric bayesian model for randomised block designs. *Biometrika* 83(2):275–285.

Chapman, M. A., et al. 2011. Initial genome sequencing and analysis of multiple myeloma. *Nature* 471(7339):476–472.

Chen, C.-L., et al. 2010a. Impact of replication timing on non-cpg and cpg substitution rates in mammalian genomes. *Genome Research* 20(4):447–457.

Chen, X., and L. Wang. 2009. Integrating biological knowledge with gene expression profiles for survival prediction of cancer. *Journal of Computational Biology* 16(2): 265–278.

Chen, X., L. Wang, and H Ishwaran. 2010b. An integrative pathway-based clinical-genomic model for cancer survival prediction. *Statistics & Probability Letters* 80(17-18):1313–1319.

Ciriello, G., et al. 2012. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research* 22:398–406.

Clark-Langone, K., C. Sangli, J. Krishnakumar, and D. Watson. 2010. Translating tumor biology into personalized treatment planning: analytical performance characteristics of the oncotype dx(r) colon cancer assay. *BMC Cancer* 10(1):691.

Dahl, D. B. 2009. Modal clustering in a class of product partition models. *Bayesian Analysis* 4(2):243–264.

Dawson, J., and C. Kendziorski. 2012. Survival-supervised latent dirichlet allocation models for genomic analysis of time-to-event outcomes. *Department of Biostatistics and Medical Informatics Technical Report #225, submitted*.

Dees, N. D., et al. 2012. MuSiC: Identifying mutational significance in cancer genomes. *Genome Research* 22(8):1589–1598.

DeRisi, J.L., V.R. Iyer, and P.O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680–6.

Dettling, M., and P. Buhlmann. 2002. Supervised clustering of genes. *Genome Biology* 3(12).

Ding, L., et al. 2008. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 455(5).

Eddy, J. A., J. Sung, D. German, and N. D. Price. 2010. Relative expression analysis for molecular cancer diagnosis and prognosis. *Technology in Cancer Research and Treatment* 9(2):149–159.

Efron, B., et al. 2001. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 96(456):1151–1160.

Escobar, M. D., and M. West. 1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90(430):577–588.

Forbes, S. A., et al. 2011. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Research* 39(supp 1):D945–D950.

Fraley, C., and A. E. Raftery. 2002. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* 97:611–631.

———. 2006. MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics. (revised 2009).

Fraley, C., A. E. Raftery, T. B. Murphy, and L. Scrucca. 2012. MCLUST version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation. Tech. Rep., no. 597, Department of Statistics, University of Washington.

Frigyesi, A., and M. Höglund. 2008. Non-negative matrix factorization for the analysis of complex gene expression data: Identification of clinically relevant tumor subtypes. *Cancer Informatics* 6:275–292.

Futreal, P. A., et al. 2004. A census of human cancer genes. *Nature Reviews Cancer* 4(3):177–183.

Ghosh, D., and Z. Yuan. 2010. Combining multiple models with survival data: the PHASE algorithm. Technical Report, Penn State University, Department of Statistics.

Gonzalez-Perez, A., and N. Lopez-Bigas. 2012. Functional impact bias reveals cancer drivers. *Nucleic Acids Research* gks743.

Griffiths, T. L., and M. Steyvers. 2004. Finding scientific topics. *PNAS* 101(1): 5228–5235.

Hartigan, J. A. 1990. Partition models. *Communications in Statistics-Theory and Methods* 19(8):2745–2756.

Hasin, Y., et al. 2008. High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. *PLoS Genetetics* 4(11):e1000249.

Hausser, J., and K. Strimmer. 2009. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research* 10:1469–1484.

Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* 89(22):10915–10919.

Hu, X., H. M. Stern, L. Ge, C. O'Brien, L. Haydu, C. D. Honchell, P. M. Haverty, B. A. Peters, T. D. Wu, L. C. Amler, J. Chant, D. Stokoe, M. R. Lackner, and G. Cavet. 2009. Genetic alterations and oncogenic pathways associated with breast cancer subtypes. *Molecular Cancer Research* 7(4):511–522.

Hubert, Lawrence, and Phipps Arabie. 1985. Comparing partitions. *Journal of classification* 2(1):193–218.

Kærn, M., T. C. Elston, W. J. Blake, and J. J. Collins. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics* 6(6):451–464.

Kenfield, S. A., et al. 2008. Comparison of aspects of smoking among the four histological types of lung cancer. *Tobacco Control* 17(3):198–204.

Kharchenko, P. V., L. Silberstein, and D. T. Scadden. 2014. Bayesian approach to single-cell differential expression analysis. *Nature Methods*.

Kim, J. K., and J. C. Marioni. 2013. Inferring the kinetics of stochastic gene expression from single-cell rna-sequencing data. *Genome Biology* 14(1):R7.

Kinzler, K. W., and B. Vogelstein. 1997. Gatekeepers and caretakers. *Nature* 386(6627):761–763.

Koren, A., et al. 2012. Differential relationship of DNA replication timing to different forms of human mutation and variation. *American Journal of Human Genetics* 91: 1033–1040.

Korthauer, K., J. Dawson, and C. Kendziorski. 2013. Predicting cancer subtypes using survival-supervised latent dirichlet allocation models. *Advances in Statistical Bioinformatics: Models and Integrative Inference for High-Throughput Data* 366.

Korthauer, K. D., and C. Kendziorski. 2015. MADGiC: a model-based approach for identifying driver genes in cancer. *Bioinformatics* btu858.

Kumar, P., S. Henikoff, and P. C. Ng. 2009. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature Protocols* 4(7):1073–1081.

Larson, D. R. 2011. What do expression dynamics tell us about the mechanism of transcription? *Current Opinion in Genetics & Development* 21(5):591–599.

Lawrence, M. S., et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499(7457):214–218.

Leng, N., L.-F. Chu, C. Barry, Y. Li, J. Choi, X. Li, P. Jiang, R. M. Stewart, J. A. Thomson, and C. Kendziorski. 2015. Oscope: a statistical pipeline for identifying oscillatory genes in unsynchronized single cell rna-seq. *Submitted*.

Li, B., and C. N. Dewey. 2011. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics* 12(1):323.

Li, H., and J. Gui. 2004. Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics* 20 Suppl. 1:i208–i215.

Li, H., and Y. Luan. 2003. Kernel cox regression models for linking gene expression profiles to censored survival data. In *Pacific symposium on biocomputing*, 65–76.

Li, L., and H Li. 2004. Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics* 20(18):3406–3412.

Liu, X., X. Jian, and E. Boerwinkle. 2011. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation* 32(8):894–899.

Ma, S., X. Song, and J. Huang. 2007. Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics* 8(60).

MacEachern, S. N. 1994. Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics-Simulation and Computation* 23(3): 727–741.

MacEachern, S. N., and P. Müller. 1998. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics* 7(2):223–238.

Mackay, A., B. Weigelt, A. Grigoriadis, B. Kreike, R. Natrajan, R. A'Hern, D. S. P. Tan, M. Dowsett, A. Ashworth, and J. S. Reis-Filho. 2011. Microarray-based class discovery for molecular classification of breast cancer: Analysis of interobserver agreement. *Journal of the National Cancer Institute* 103(8):662–673.

Marinov, G. K., B. A. Williams, K. McCue, G. P. Schroth, J. Gertz, R. M. Myers, and B. J. Wold. 2014. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and rna splicing. *Genome Research* 24(3):496–510.

Mook, S., L. J. Van't Veer, E. Rutgers, M. Piccart-Gebhart, and F. Cardoso. 2007. Individualization of therapy using mammaprint: from development to the mindact trial. *Cancer Genomics Proteomics* 4(3):147–155.

Munsky, B., G. Neuert, and A. van Oudenaarden. 2012. Using gene expression noise to understand gene regulation. *Science* 336(6078):183–187.

National Cancer Institute and National Human Genome Research Institute. 2011. The cancer genome atlas. http://cancergenome.nih.gov/.

Ng, P. C., and S. Henikoff. 2001. Predicting deleterious amino acid substitutions. *Genome Research* 11(5):863–874.

Ohnishi, Y., W. Huber, A. Tsumura, M. Kang, P. Xenopoulos, K. Kurimoto, A. K. Oleś, M. J. Araúzo-Bravo, M. Saitou, A.-K. Hadjantonakis, et al. 2014. Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nature Cell Biology* 16(1):27–37.

Pang, H., D. Datta, and H. Zhao. 2010. Pathway analysis using random forests with bivariate node-split for survival outcomes. *Bioinformatics* 26(2):250–258.

Pleasance, E. D., et al. 2009. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 463(7457):184–190.

R Core Team. 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reva, B., Y. Antipin, and C. Sander. 2011. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research* 39:e18.

Sanchez, A., and I. Golding. 2013. Genetic determinants and cellular constraints in noisy gene expression. *Science* 342(6163):1188–1193.

Shahrezaei, V., and P. S. Swain. 2008. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences* 105(45):17256–17261.

Shalek, A. K., R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, et al. 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*.

Shalek, A. K., R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J. T. Gaublomme, N. Yosef, et al. 2014. Single-cell rna-seq reveals dynamic paracrine control of cellular variation. *Nature*.

Shapiro, E., T. Biezuner, and S. Linnarsson. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics* 14(9):618–630.

Shen, R., Q. Mo, N. Schultz, V. E. Seshan, A. B. Olshen, J. Huse, M. Ladanyi, and C. Sander. 2012. Integrative subtype discovery in glioblastoma using icluster. *PLoS ONE* 7(4):e35236.

Shotwell, M. S. 2013. profdpm: An R package for MAP estimation in a class of conjugate product partition models. *Journal of Statistical Software* 53(8):1–18.

Shotwell, M. S., and E. H. Slate. 2011. Bayesian outlier detection with dirichlet process mixtures. *Bayesian Analysis* 6(4):665–690.

Singer, Z. S., J. Yong, J. Tischler, J. A. Hackett, A. Altinok, M. A. Surani, L. Cai, and M. B. Elowitz. 2014. Dynamic heterogeneity and dna methylation in embryonic stem cells. *Molecular Cell* 55(2):319–331.

Sjoblom, T., et al. 2006. The consensus coding sequences of human breast and colorectal cancers. *Science* 314(5797):268–274.

Sparano, J.A., and S. Paik. 2008. Development of the 21-gene assay and its application in clinical practice and clinical trials. *Journal of Clinical Oncology* 26(5): 721–728.

Stegle, O., S. A. Teichmann, and J. C. Marioni. 2015. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* 16(3):133–145.

Tamborero, D., A. Gonzalez-Perez, and N. Lopez-Bigas. 2013. Oncodriveclust: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* 29(18):2238–2244.

Tothill, R. W., A. V. Tinker, J. George, R. Brown, S. B. Fox, S. Lade, D. S. Johnson, M. K. Trivett, D. Etemadmoghadam, B. Locandro, N. Traficante, S. Fereday, J. A. Hung, Y.-E. Chiew, I. Haviv, Australian Ovarian Cancer Study Group, D. Gertig, A. deFazio, and D. D. L. Bowtell. 2008. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research* 14(16):5198–5208.

Treutlein, B., D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake. 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature* 509(7500):371–375.

Vandin, F., E. Upfal, and B. Raphael. 2012. De novo discovery of mutated driver pathways in cancer. *Genome Research* 22:375–385.

Vaske, C. J., et al. 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics* 26(12): i237–i245.

Vogelstein, B., and K. W. Kinzler. 2004. Cancer genes and the pathways they control. *Nature Medicine* 10(8):789–799.

Vogelstein, B., et al. 2013. Cancer genome landscapes. *Science* 339(6127):1546–1558.

Woo, Y. H., and W.-H. Li. 2012. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nature Communications* 3: 1004.

Wood, L. D., et al. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* 318(5853):1108–1113.

Wu, A. R., N. F. Neff, T. Kalisky, P. Dalerba, B. Treutlein, M. E. Rothenberg, F. M. Mburu, G. L. Mantalas, S. Sim, M. F. Clarke, et al. 2014. Quantitative assessment of single-cell rna-sequencing methods. *Nature Methods* 11(1):41–46.

Yost, S. E., et al. 2013. Mutascope: sensitive detection of somatic mutations from deep amplicon sequencing. *Bioinformatics* 29(15):1908–1909.

Youn, A., and R. Simon. 2011. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* 27(2):175–181.