2016

# Power Analysis in Applied Linear Regression for Cell Type-Specific Differential Expression Detection

Edmund Glass
*Virginia Commonwealth University*, glasser@vcu.edu

# Power Analysis in Applied Linear Regression for Cell Type-Specific Differential Expression Detection

By

Edmund R. Glass

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Biostatistics
Virginia Commonwealth University

Doctoral Committee:

| | |
|---|---|
| Mikhail G. Dozmorov | Department of Biostatistics (Committee Chair) |
| Nak-Kyeong Kim | Department of Biostatistics |
| Roy T. Sabo | Department of Biostatistics |
| Vladimir I. Vladimirov | Department of Psychiatry |
| Timothy P. York, | Department of Human and Molecular Genetics |

2016

# ACKNOWLEDGEMENTS

Hunter S. Thompson once wrote, “Anything worth doing, is worth doing right.”  The fact of a learning curve dictates that I will miss the target until I get it right.  If I have learned anything it is to not quit until I get it right.  Then really don’t quit!

I am the luckiest person I know.  Each time I believed my situation was untenable, friends and family were there to provide support of their time, effort, and resources.  I have learned that morale absolutely matters!

The Knitting Klatch and the Caregivers belong to a long list of lifesavers that I consider family.  Especially Beth P and Carol C.  I would not have been able to finish this without their efforts and assistance.  Antoine W showed up and worked grueling hours during the worst of times when a family’s needs were the greatest.

To the entire medical and nursing staff at Hampton Sentara and Warwick Forest rehab center: Y’all are all unsung heroes.

To Janet and Jessica B, Kim F and Kathy O, Irma C, Hal and Terrie, Ralph and Teddy-Jo R, and Mr. Joe B: Your gifts are priceless beyond words.

The best friends a guy could hope for are Scott B, Brian M, Pete S, Reed A, and Paul D. You know who you are and what you did!  I have learned that I cannot make it without help.

Powerful sources of inspiration include but are not limited to people with superpowers, who are: Bob J, Theresa H, Katrina L, and Daphne G. Y'all make me want to do everything better. As I age, I find that heroes remain important.

My academic family includes but is not limited to Kyle F, Amanda G, Sid K, Rebecca L, Paul M, and Uma S, Mary-Beth M, and Penny R. Team learning surpasses trying to go it alone! I could not have gotten through this program without this crew. The faculty and staff of the VCU Biostatistics Department will never be forgotten for their leadership, guidance, and patience. Especially Gayle S and Yvonne H. Shout out to Sonny R for having my back!

I am especially grateful to my adviser Mikhail D for the long hours painstakingly showing me the way that science gets done and for putting up with me for as long as he has.

To Candace F for demonstrating that it can be done, for showing the sort of passion and commitment that it takes to do the job right, and showing me that scientists are real folks.

Family. Without the encouragement, material, and morale support, and long list of inspirational examples, in no particular order, my gratitude and love go to, Julie G, Tim B, Katie K, Chris L, Patricia L, David G, Andrew G, Phil G, and Louise G. You guys are the wind under my wings. I know that I can't list all of you here. That would require another entire paper!

To Stony, I hope to live up to your example of leaving a positive and lasting legacy.

To my best friend, unwavering supporter, and an inspiring example of strength, beauty, wisdom, and kindness, I love you Mom. I could not have done this without you.

# TABLE OF CONTENTS

CHAPTER

# LIST OF FIGURES

# LIST OF TABLES

# ABSTRACT

Power Analysis in Applied Linear Regression for

Cell Type-Specific Differential Expression Detection

By

Edmund R. Glass

Chair: Mikhail G. Dozmorov

The goal of many human disease-oriented studies is to detect molecular mechanisms different between healthy controls and patients. Yet, commonly used gene expression measurements from any tissues suffer from variability of cell composition. This variability hinders the detection of differentially expressed genes and is often ignored. However, this variability may actually be advantageous, as heterogeneous gene expression measurements coupled with cell counts may provide deeper insights into the gene expression differences on the cell type-specific level. Published computational methods use linear regression to estimate cell type-specific differential expression. Yet, they do not consider many artifacts hidden in high-dimensional gene expression data that may negatively affect the performance of linear regression. In this dissertation we specifically address the parameter space involved in the most rigorous use of linear regression to

estimate cell type-specific differential expression and report under which conditions significant detection is probable.

We define parameters affecting the sensitivity of cell type-specific differential expression estimation as follows: sample size, cell type-specific proportion variability, mean squared error (spread of observations around linear regression line), conditioning of the cell proportions predictor matrix, and the size of actual cell type-specific differential expression. Each parameter, with the exception of cell type-specific differential expression (effect size), affects the variability of cell type-specific differential expression estimates. We have developed a power-analysis approach to cell type by cell type and genomic site by site differential expression detection which relies upon Welch's two-sample t-test and factors in differences in cell type-specific expression estimate variability and reduces false discovery. To this end we have published an R package, LRCDE, available in GitHub (http://www.github.com/ERGlass/lrcde.dev) which outputs observed statistics of cell type-specific differential expression, including two-sample t-statistic, t-statistic p-value, and power calculated from two-sample t-statistic on a genomic site-by-site basis.

# 1. Introduction

## 1.1. Background and Motivation for Use of Linear Regression in Deconvolution

In many studies the end goal is to determine by what biological mechanism healthy controls differ from patients positive with some pathology. Differential expression (DE) analysis is applied in an attempt to determine whether differences exist between patients and controls at the genetic level. These studies are complicated by the fact that DE analysis is frequently performed on tissue that is heterogeneous for multiple cell types. A more informative approach to DE would be to separate genetic signal into cell type-specific components (deconvolution) (Hoffmann et al. 2006; Lähdesmäki et al. 2005; Venet et al. 2001). The results presented in this dissertation extend and refine previous work in detecting cell type-specific signal from heterogeneous measures.

There are several motivations for a computational approach to deconvolution. One motivation comes from the fact that taking cell specific measures may be physically impossible due to tissue characteristics (Lähdesmäki et al. 2005; Shen-Orr et al. 2010). Another motivation for a computational approach to deconvolution arises from the fact that taking cell specific measures, even when physically possible, can be prohibitively expensive in terms of time and resources (Gosink, Petrie, and Tsinoremas 2007). Finally, differential expression detection analysis applied only to heterogeneous tissue may miss the signal entirely (Shen-Orr et al. 2010). Thus, deconvolution may uncover cell type-specific signal not seen at the heterogeneous level (appendix A.4). This interest in cell type-specific information is true whether genomic measures are derived from microarray (Stuart et al. 2004), or newer “high throughput” technologies (Anders et al. 2010; Gong and Szustakowski 2013; Tarazona 2011).

Linear regression has been applied to the problem of deconvolution. Linear regression is a statistical method of modeling a linear relationship between a set of outcomes and a set of predictors in which the relationship between predictors and outcomes is believed to be additive.

The linear regression approach to cell type-specific expression estimation appears to have been first mentioned in 2001 by Venet et.al. (Venet et al. 2001). In the early 2000's the Human Genome Project (Lander ES 2001) had just been completed and the cost of sequencing the first human genome was said to be around 3 billion U.S. dollars. By 2012 the cost of sequencing a single human genome had dropped to around $5,000 (Newton and Moore 2014). With the promise of ever decreasing costs of sequencing, the prospect of sequencing individual patients for the purposes of diagnosis and treatment appears to be growing more feasible. Thus, the motivation to understand patient samples at the cellular level has the added impetus of becoming a realistic option.

Developing clinically actionable information with the minimum amount of false positives is the ultimate goal of extracting cell type-specific expression for the purpose of differential expression detection. Linear regression-based cell type-specific differential expression detection and analysis (LRCDE) is the motivating purpose behind these algorithms. The ultimate gold standard for the effectiveness of these techniques will be the degree to which actual patient outcomes are improved by any genetic revelations which result from their application.

In applying linear regression for the purpose of extracting cell specific information from heterogeneous measures, multiplicative regression coefficients (weights) are associated with each predictor variable in the model in such a way as to explain as much variability across outcomes as possible (Kutner et al. 2005). Bioinformatics applies linear regression deconvolution in at least two ways.

In one well explored direction, cell proportions can be estimated (deconvolved using linear regression) (Abbas et al. 2009; Gong and Szustakowski 2013; Houseman et al. 2012; Jaffe and Irizarry 2014; Montaño et al. 2013; Newman et al. 2015) from heterogeneous measures given genetic profiles of purified cell lines (cell signatures) that are homologous to cells in the samples. These purified cell signatures are not from the same samples as the heterogeneous measures. This *in silico* approach to cell proportion estimation is sometimes referred to as "computational microdissection" (Liebner, Huang, and Parvin 2014). Other authors apply single value decomposition (SVD) techniques (Chikina, Zaslavsky, and Sealfon 2015) in order to estimate cell proportions given cell signatures.

In the opposite direction, linear regression may be applied to the problem of estimating cell type-specific expression levels obtained from heterogeneous tissue. Here, relative cell proportions measured on the same samples are used as predictors of heterogeneous gene expression (Erkkilä et al. 2010; Lähdesmäki et al. 2005; Shen-Orr et al. 2010; Stuart et al. 2004). These cell proportions (predictors) may be considered as "weights" of the corresponding cell type-specific gene expression. The cell type-specific gene expression estimates, weighted by the corresponding cell proportions, are the regression coefficients within the linear regression framework. This linear regression approach to disentangle cell type-specific measures from heterogeneous samples is referred to as linear regression deconvolution. This model has been demonstrated to successfully identify features that are differentiated at the cell level in prostate cancer (Stuart et al. 2004). This approach then to estimation of cell type-specific expression for the purpose of differential expression detection is important, but not as well explored. For these reasons, this dissertation focuses on the problem of using linear regression for the estimation of cell type-specific differential expression, and specifically on factors which affect the strength and reliability of those estimates.

## 1.2. Existing Deconvolution Algorithms Using Linear Regression

Several authors have used linear regression similarly in the direction of estimating cell proportions given either cell type-specific genetic profiles from the samples which provide the heterogeneous measures or from so-called cell signatures profiles in which outstanding genetic marks for specific cell types are known to be uniquely expressed (Houseman et al. 2012; Montaño et al. 2013).

Other authors have published algorithms (Erkkilä et al. 2010; Newman et al. 2015; Shen-Orr et al. 2010) starting to address the need for an efficient method of computationally estimating cell type-specific differential expression through linear regression. Each of these approaches is a two-stage process: 1) linear regression based estimation of cell type-specific signal given signal measured on tissue heterogeneous for multiple cell types (the deconvolution step), and 2) estimation of group-wise differential expression based on estimates from the deconvolution step. When estimating cell type-specific differential expression, two pieces of information are required: 1) relative cell type proportions per sample used as regression predictors, and 2) heterogeneous genomic measures per sample used as outcomes. Differential expression is then estimated based on obtained group-wise cell type-specific expression estimates.

DSection method (Erkkilä et al. 2010) assumes that cell proportion measures are imprecise and that this imprecision must be accounted for. DSection uses a Bayesian approach to "de-noise" cell proportion measures prior to application of linear regression deconvolution. The authors of DSection contrast their method to a "gold standard" of using linear regression when cell proportions are precisely known. The DSection authors correctly point out that in real settings no true knowledge of cell proportions is known and that measurements are presumed to be estimates. They also acknowledge that the choice of prior information to use with their Bayesian approach has a subjective component. DSection has been implemented in Matlab commercial platform, making it less accessible to general audience. Although free implementations of Matlab functionality are available, e.g., Octave, they have limitations. These limitations prevent the user to test DSection approach, and are the reason this method was not considered in our approach.

The approach used by Shen-Orr et.al. (Shen-Orr et al. 2010), named csSAM, uses the heterogeneous observations as outcomes in a linear regression model and the measured cell proportions as predictors in the linear model. Two regressions, one per study group, are performed and the difference between coefficient estimates between regressions represents the cell type-specific differential expression estimates. Group (sample) label permutations are performed and false discovery rates (FDR) are estimated per genomic site per cell type. The csSAM authors acknowledge that increasing sample variability will improve cell type-specific expression accuracy, and we have attempted to quantify the lower bound for such variability.

Linear regression has therefore shown to be a promising approach to the problem of estimating cell type-specific expression and subsequently cell type-specific differential expression between study groups. In this dissertation the focus is entirely upon the direction of estimating cell type-specific expression from heterogeneous measures given the assumption of precisely measured cell proportions. We examine the variability around cell type-specific expression estimates and how this variability impacts false discovery rates, area under receiver operator characteristic curves, and observed power.

### 1.3. Drawbacks to the Linear Regression Approach to Deconvolution

Although linear relationships have been demonstrated to model cell type-specific proportion-expression levels (Erkkilä et al. 2010; Lähdesmäki et al. 2005; Shen-Orr et al. 2010; Stuart et al. 2004), variability around cell type-specific expression estimates is not well addressed. There are several known sources of variation in linear regression models (Kutner et al. 2005) and we further explore them here as applied to deconvolution and demonstrate their theoretical effect upon the level of sensitivity and specificity of the derived predictions.

In the application of linear regression deconvolution, at least one regression is performed per genomic site. Thus, no two regressions will have identical residual distributions. Overall sizes of the residuals will vary by genomic site. One genomic site may have residuals tightly clustered around the regression line for a heterogeneous measure that is well modeled by the cell proportions across samples. Another site may have residuals that are far flung from the regression line for a model in which the cell proportions have little bearing on the heterogeneous expression. Widely spread residuals as quantified by the mean squared error (MSE) will result in low sensitivity (poor detection of true positives) and low specificity (higher false positive rates). When calibrated using simulated data with known differential expression between specific "sites" in specific "cells", constructed receiver operator characteristic curves (ROC) will fall nearly on the diagonal with poorly fitting regressions. Other sources of variation come from variability of cell proportion standard deviations across samples and overall poor conditioning of the cell proportion predictor matrix.

### 1.4. Outline

In chapter 2 we describe the linear regression model as it relates to cell type-specific expression estimation and subsequent differential expression detection. We describe and conduct a simulation study with three aims being to illustrate the effects upon differential expression sensitivity of varying: 1) group sample sizes, 2) per regression residual variability (quantified by mean squared error – MSE), and 3) the variability of predictor cell type proportions across samples.  These three parameters have a well explored (Kutner et al. 2005) impact upon the standard error estimates (variability) of linear regression coefficient estimates, which are taken as surrogates for cell type-specific expression in the linear regression model of deconvolution.  The effect of varying these parameters is not explored in the method specified by (Shen-Orr et al. 2010).  In concordance with this published model, part of our simulation study relies upon a permutation method in order to obtain p-values per cell type per gene. We detail the simulation study and explain how this approach is used to create the various ROC curves and areas under ROC curves (AUROCs) which illustrate the effects of varying the indicated three parameters.

Beginning in chapter 3, we conduct simulations in which we compare cell type-specific differential expression detection sensitivity of a permutation based global false discovery rate (FDR) method computed across all genes per cell type to a traditional two-sample t-test statistic computed on a cell type by cell type gene-by-gene basis (LRCDE).

In chapter 4, we perform simulations to compare log base 2-transformed vs. raw (“de-logged”) heterogeneous observations when performing linear regression deconvolution.  This is a question which has been subject to debate (Zhong and Liu 2012).  Since it is the distribution of the regression outcomes (heterogeneous observations) that are being transformed, we explore the effect this has upon the shape of residuals.

Chapter 5 explores conditioning of the cell proportion predictor design matrix (inner product of cell proportions predictors) and the effect that the condition number has on cell type-specific expression estimates and subsequent differential expression detection.  The increased volatility of cell expression estimates resulting from extremely high condition numbers is illustrated. We explore dropping one or more cell proportion predictors as a method of effectively handling a high condition number scenario.

In chapter 6, the significance of this work, its drawbacks, and future directions are discussed.

All analyses were performed using RStudio (RStudio 2015) on the R statistical platform version 3.3.1 running on Ubuntu version 15.10 on an x86_64 laptop with i7-6700k processor and 64Gb RAM.

# 2. **Linear Regression Theory of Deconvolution and Parameters Affecting Differential Expression Detection Sensitivity**

## 2.1. Overview of linear regression in terms of cell type-specific differential expression and parameters affecting detection sensitivity

Chapter 1 provided the background context and motivation for the application of linear regression deconvolution. In chapter 2, we first lay out the details of linear regression in terms of cell type-specific differential expression detection (deconvolution).

The primary question we address in this chapter is how three parameters affect the sensitivity and specificity of cell type-specific differential expression detection. We hypothesize that there are combinations of group sample sizes, mean squared error, and predictor variability under which cell type-specific differential expression predictive power will be such that sensitivity will not exceed the compliment of specificity (1 – specificity) to a practical degree. Such a situation is illustrated as points on the diagonal of a receiver operator characteristic (ROC) curve. When sensitivity equals the compliment of specificity, differential expression detection is little better than flipping a coin.

We describe the simulation process which we use for three aims: illustrating the effect upon cell type-specific expression estimates (regression coefficient estimates) variability of varying: 1) group sample sizes, 2) overall size of residuals (per gene), and 2) variability of a single cell proportion predictor across samples. We do not expect to define hard upper or lower boundaries for any of these three parameters, but rather to drive home the point that these factors interact in such a way as to make analysis on a cell-by-cell and gene-by-gene basis necessary.

### 2.2. Linear regression described in terms of deconvolution

Gene measures taken on tissue which is heterogeneous for multiple cell types is referred to as heterogeneous gene expression measures, or simply heterogeneous measures. We model heterogeneous measures across samples as cumulative contributions of cell type-specific gene expression measures weighted by the corresponding cell proportions of *P* cell types. A theoretical biologically meaningful constraint of this model is that cell proportions for any given sample should sum up to 1, or 100% (Shen-Orr et al. 2010). As proposed, heterogeneous gene expression measures ( $y_{mn}$ , where *n* is gene index, *m* is sample index) are modeled using a linear regression approach:

$$y_{mn} = \sum_{k=1}^{p} \beta_{kn} x_{km} + \varepsilon_{mn} \qquad (1),$$

where $\beta_{kn}$ is the average cell type-specific gene expression for the $k^{th}$ of *p* total cell types, $x_{km}$ is the cell proportion (predictor), and $\varepsilon_{mn}$ is a normally distributed random error. Estimates of the random error are defined as the difference between the observed values $y_{mn}$ and values predicted by the linear regression $\hat{y}_{mn}$ , $\left(y_{mn} - \hat{y}_{mn}\right)$, referred to as residuals.

Linear regression results in linear regression coefficient estimates $\hat{\beta}_{kn}$ , interpreted as cell type-specific gene expression estimates. Intuitively, equation (1) describes a linear relationship between heterogeneous gene expression level $y_{mn}$ and contribution of cell type-specific gene expressions $\beta_{kn}$ weighted by the corresponding cell proportions $x_{km}$ . The model in equation (1) contains no intercept term since we assume zero heterogeneous expression ( $y_{mn} = 0$ ) in the absence of individual cell contributions (Shen-Orr et al. 2010; Stuart et al. 2004). Thus, for each gene we have a total of *P* cell type-specific gene expression estimates (regression coefficients, one per cell type) in the model. Model in eq. 1 is more compactly represented in matrix form, for a single gene *j*:

$$\mathbf{y}_j = \mathbf{X}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon} \qquad (2).$$

The matrix form eq. 2 suggests the form in which $\boldsymbol{\beta}_j$ is estimated:

$$\hat{\boldsymbol{\beta}}_j = (\mathbf{X'X})^{-1}\mathbf{X'y}_j \qquad (3).$$

Fitted regression estimates are then given by:

$$\hat{\mathbf{y}}_j = \mathbf{X}\hat{\boldsymbol{\beta}}_j \qquad (4),$$

which are required in order to calculate residual values needed to estimate the variance of $\hat{\boldsymbol{\beta}}_j$.

Obtaining cell type-specific gene expression estimates carries a quantifiable level of uncertainty.

Statistical inferences based on any supplied statistic are typically qualified with an accompanying measure of uncertainty, such as power levels when a statistic is found to be significant. Such quantification of the uncertainty was missing from supplied FDR measures in the case of csSAM. The name “false discovery rate” implies that FDR itself is a measure of uncertainty. However, since FDR itself is an estimate then it too has some boundary expressing our level or certainty in its measure.

In the case of linear regression, cell specific expression estimates (linear regression coefficient estimates) have a well-defined level of uncertainty attached (Kutner et al. 2005) . This uncertainty can be expressed as a function of sample size, number of cell types, the size of the residuals, and the variability of cell type proportions. The formula for the theoretical variance of the linear regression coefficient for simple linear regression (single predictor vector ***X***) provides an intuitive illustration of how various parameters affect the variance:

$$\operatorname{var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{m} (\mathrm{x}_i - \bar{x})^2} \qquad (5).$$

In practice, the estimated variance of $\hat{\beta}_1$ in eq. 5 uses the mean squared error (MSE) as an estimate of $\sigma^2$, represented as $s^2$:

$$s^2 = MSE = \frac{\sum_{i=1}^{m} (\mathrm{y}_i - \hat{\mathrm{y}}_i)^2}{(\mathrm{M} - \mathrm{P})} \qquad (6).$$

In this simple linear regression context of eq. 5 and eq. 6, *M* is the sample size and *P* is typically equal to 2, since there are two parameters being estimated: an intercept term $\hat{\beta}_0$ and the coefficient of the predictor variable: $\hat{\beta}_1$. Thus, the estimated variance of $\hat{\beta}_1$ in simple linear regression is represented as:

$$\operatorname{var}\hat{\beta}_1 = \frac{\sum_{i=1}^{m} (\mathrm{y}_i - \hat{\mathrm{y}}_i)^2 / (\mathrm{M} - \mathrm{P})}{\sum_{i=1}^{m} (\mathrm{x}_i - \bar{x})^2} \qquad (7)$$

where $y_i - \hat{y}_i$ is the residual for sample *i*, and $x_i - \bar{x}$ is the difference between the predictor for sample *i* and the mean of *x* across all *M* samples. In this way, predictor variability is captured in the denominator of eq. 7, as is sample size M. Residual variability is captured in the numerator of eq. 7. Each component of eq. 7 affects the estimated variance of $\hat{\beta}_1$.

In multivariate linear regression, matrix notation simplifies representation of the variances of all $P$ regression coefficients. The theoretical variance-covariance matrix $\Sigma$ of linear regression coefficients is represented as:

$$\Sigma = \sigma^2 \left( \mathbf{X'X}^{-1} \right) \qquad (8),$$

where the variances of the $k^{th}$ individual $\hat{\beta}_k$ regression coefficients are found on the diagonal of $\Sigma$. In eq. 8 it is less intuitive to see the way in which individual parameters affect the variance estimates of the individual $\hat{\beta}_k$ in matrix form, yet the principles are the same as in eq. 5. Predictor variability is captured in the inverse of the design matrix: $\left( \mathbf{X'X} \right)^{-1}$, analogous to the denominator of eq. 5. As with simple single variable regression, $\sigma^2$ is estimated by MSE, represented by $s^2$ providing the estimated covariance matrix:

$$\hat{\Sigma} = s^2 \left( \mathbf{X'X}^{-1} \right) \qquad (9),$$

where $s^2$ is:

$$s^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{M - P} = \frac{\mathbf{y'y} - \hat{\boldsymbol{\beta}}'\mathbf{X'y}}{M - P} = \frac{\mathbf{y'y} - \mathbf{y}'\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'y}}{M - P} \qquad (10).$$

The primary focus of this dissertation is to evaluate the effects of sample size, residual variability, and MSE on the estimated variances $\hat{\Sigma}$ of the cell type-specific expression estimates $\hat{\beta}_k$ and the effect this has upon differential expression detection sensitivity. (Matrix notation for equations 2,3,4,8, 9 and 10 is attributed to Graybill (Graybill 1969)).

Differential expression analysis implies comparison of two or more groups for detectable gene expression differences. For simplicity, we consider a two-group design, such as a case-control study. To obtain group specific cell type-specific gene expression estimates ( $\hat{\beta}_{kn}$ ), we apply linear regression separately to each group of heterogeneous gene expression measures (two regressions). A difference between these cell type-specific estimates represents the level of gene expression change between the two groups in a given cell type:

$$\hat{\delta}_{kn} = \left(\hat{\beta}\right)_{kn}^{cases} - \left(\hat{\beta}\right)_{kn}^{controls} \tag{11}$$

where $\hat{\delta}_{kn}$ is estimated effect size, *k* is the specific cell type and *n* is the genomic site.

Measuring the cell type-specific gene expression differences between groups using linear regression (LR) requires accurate cell type-specific gene expression estimates. Any factors affecting the variability of cell type-specific gene expression estimates per group will affect the power to detect cell type-specific fold changes between groups.

### 2.3. Simulation study to assess parameter effects with known cell type-specific differential expressions were created for LRCDE testing

Quantification of sensitivity and specific of any discriminative measure, such as differential expression detection, requires that the truth be known *a priori*. The "truth" for the purposes of cell type-specific differential expression detection are which genes in a data set are differentiated in which cell types and to what degree. For these reasons, we simulated data with engineered and known cell type-specific differential expression on pre-identified genes. To establish a "gold-standard" of known cell type-specific differential expressions to benchmark LRCDE estimates of cell type-specific differential expressions, synthetic data with controlled changes (Dozmorov et al. 2010) was constructed in three steps.

First, we created synthetic *P* cell proportions across *M* samples per group with known standard deviation for the target cell type *p*. For the sake of comparable per-group regressions, we simulate the condition where both groups have identical cell proportions (Box 2.3.1).

**Box 2.3.1:** Cell proportions are simulated by creating per group an *M* by *P* matrix of random uniform values, which must sum to 1 across a row of any given sample. Thus, we have two identical *M* by *P* matrices, one per study group so that both control group and case group linear regressions are performed with identical cell proportions as predictors.

We chose a target cell standard deviation (SD) to simulate. The target cell is the cell type in the simulation which has a cell type-specific differential expression applied to it in the synthetic cases cell type-specific expression matrix. We then create a single vector of cell proportions for sample size *m*. This begins with a small proportion for sample 1 and creates evenly spaced proportions until a chosen "high" proportion is reached for sample *m*. The SD of this vector is taken and compared to our target SD. If our target SD is too low, we increase the "high" target proportion and recreate the matrix. This iterative brute force process is repeated until we arrive to within a desired tolerance of our target SD.

Second, we created synthetic matrixes of cell type-specific gene expression estimates for both control and case groups. We applied identical defined differences to half of the "genes" in the target cell type *p* ("true changes"). In this way we have 1000 total "genes" per cell type, with 500 "genes" differentially expressed in the target cell type (Table 2.1). We now have two matrices of cell type-specific expression of identical dimensions with changed applied to half of the genes in a single cell type (Box 2.3.2).

**Box 2.3.2: Simulating cell type-specific expression and differential expression**

1. Choose a base level of expression and create a vector of length N of for cell 1
2. Create p-1 remaining cell type expression vectors.
3. Combine cell expression vectors in ‘Cases’ group matrix
4. Copy ‘Cases’ group matrix to create ‘Controls’ group matrix.
5. Alter some percentage (we use 50%) of the “genes” in cell 1 of ‘Cases’ group by applying a specified level of differential expression.
6. Make a note of actual known ‘gold-standard’ difference between expression levels of altered genes in cell 1 between ‘Cases’ and ‘Controls’. This is ‘diff’.

Finally, the cross-product of both synthetic cell type expression and synthetic cell proportion matrices was taken for each group to produce simulated matrices of heterogeneous “fitted values” analogous to the predicted values obtained from linear regression. Normally distributed “noise” was added to the “fitted values” to simulate residual values obtained from a linear regression. In this way, each of the genes in the heterogeneous matrix have approximately the same mean squared error (MSE) after linear regression. Simulation of data is now complete.

**Box 2.3.3: Simulation of heterogeneous measures:**

1. Multiply cell proportions matrix by cell type-specific expression matrix: cross-product is heterogeneous “fitted” values.
2. Simulate residuals via random normal process.
3. Add simulated residuals to the “fitted” values obtained from step 1.

Having *a priori* known cell type-specific expression differences provided us with a benchmark against which to compare the results of LRCDE analysis which produces estimated cell type-specific expression differences.

**Table 2.1: Simple illustrative example of synthetic cell type-specific expression matrices:**

Synthetic cell type-specific expression matrix with 8 genes and 3 cell types. Here cell 1 is the "target cell type" which has half (4) genes with a differential expression between controls and cases of 2 each. For running simulations we used 1000 genes with 500 differentially expressed between controls and cases. Permuted p-values were measured for detected differential expression for the target cell and these p-values along with a "truth" vector of 0's and 1's was used in order to calculate sensitivity and specificity across the range of observed p-values. This allowed construction of ROC curves and computation of AUROC.

| | Gene 1 | Gene 2 | Gene 3 | Gene 4 | Gene 5 | Gene 6 | Gene 7 | Gene 8 |
|---|---|---|---|---|---|---|---|---|
| Control – cell 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Control – cell 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Control – cell 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Case – cell 1 | 4 | 4 | 4 | 4 | 2 | 2 | 2 | 2 |
| Case – cell 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Case – cell 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

Synthetic data is assembled by joining the two ("cases" and "controls") heterogeneous gene expression matrices into one 2$M$ by $J$ heterogeneous gene expression matrix with a vector of group labels. The two cell proportion matrices, identical for groups of "cases" and "controls" were joined to obtain one 2$M$ by $P$ cell proportion matrix.

Assessing performance of LRCDE via area under receiver operator characteristic curve (AUROC) analysis using synthetic data with known cell type-specific differential expression required an additional "true changes" vector of 0's and '1s, with 1's corresponding to genes which were differentially expressed between our two artificial study groups. After performing LRCDE on simulated data, the "true changes" (truth) vector provided a convenient flag of “known” differentially expressed genes (1’s) versus those not differentially expressed (0’s), enabling construction of sensitivity vs. specificity relationship (ROC curves).

### 2.4. Method of AUROC analysis of LRCDE using simulated data

Simulation proceeds by first performing deconvolution in order to obtain cell type-specific differential expression estimates (eq. 11). These difference estimates are the “effect sizes”. Heterogeneous observations are regressed on cell proportions as predictors in a linear regression model. Linear regression is performed once per study group. The linear regression coefficient estimates are taken as surrogates for estimated cell type-specific group average gene expressions. The subtractive difference between group-wise cell type-specific expression estimates is taken as the differential expression estimates where *k* is the specific cell type and *n* is the genomic site. Once difference estimates are obtained and stored, simulation proceeds with permutation of sample labels of both heterogeneous “observations” and cell proportions. Group membership labels are randomly sampled a specified finite number of times with repeats allowed in order to obtain a distribution agnostic estimate of the null distribution of no difference between controls and cases. For equal group sizes of 10 each, there are a total of 184,756 ways of permuting group labels. It is computationally infeasible to perform all possible permutations, so a random subset of these total group label combinations is taken.

For the purposes of simulation, we focus on a single “target” cell type which has known engineered differential expression between controls and cases on half of the genes. This vector of observed differential expression for target cell type is stored for comparison against subsequent permutation derived differences.

Group membership labels are randomly permuted and the linear regression differential expression estimation step is performed 1000 times. The differential expression estimate from each permutation is stored in a 3 dimensional array for a 1000 by $P$ by $J$ array of difference estimates.

Permutation based p-values are now calculated by dividing the number of times permutation derived differential expression is greater or equal the observed differential expression by the number of permutations. A pseudo count of 1 is added to both numerator and denominator at this step in order to avoid p-values which are identically equal to zero (Smyth and Phipson 2010). In the case of a two-sided test, the absolute value of observed and permuted differential expression is taken prior to summing and dividing. For a one-sided test of up-regulation, the absolute value is not taken. For a one-sided test of down-regulation, the sum of times in which permuted differences are less or equal to observed differences is taken.

Having these permuted p-values and the truth indicator vector which flags the known differentially expressed genes in the target cell allows calculation of sensitivity (true positive rates: TPR) and 1 minus specificity (false positive rates: FPR) (Fawcett 2006). Having TPR and FPR allows construction of ROC curves and computation of AUROC via the pROC package (Robin et al. 2011). ROC curves and AUROC are used to compare performance of LRCDE between sets of parameter values, e.g., group sample sizes, MSE, cell proportion standard deviation, and effect sizes.

We consider an AUROC of greater than 0.8 to be an acceptable indicator of sufficient power (although AUROC is not to be confused with actual statistical power of detected differences which is computed from the t-statistic of Welch's two-sample t-test).

Another issue with AUROC is that AUROC provides no information about the underlying power of differences detected at any given feature site. One vector of p-values with a preponderance of those below a significance threshold (say of 0.05) may provide an identical AUROC to another array of p-values with all p-values being less than, say, 0.2. If the same number of the lowest p-values are associated with the true differences in the "truth" vector in both vectors of p-values, then the ROC curves may appear identical and have identical AUROC. Another drawback to looking at AUROC over the entire range of specificity is that relevant interest may only be in those regions under the ROC curves which lie below some false positive rate (FPR: 1-specificity) threshold. If one would prefer take such an FPR threshold into account during analysis, then one should focus on a partial AUROC (McClish 2015) rather than AUROC for the entire ROC curve.

These aspects of AUROC analysis are one of our motivations for taking a closer look at realized power of individual features per cell type rather than focusing on globally obtained FDR or sensitivity and specificity alone. One motivation for using group label permutation based AUROCs is that group label permutations is precisely the method used by the csSAM algorithm prior to computing FDRs. For this reason, we felt it relevant to explore the effects of parameter variations using similar permutations and permutation based p-values. Chapter 3 outlines the actual t-statistic based power calculation per cell per genomic site and compares this to the performance of FDR at detection of actual known (simulated) differences at the cell type-specific level.

### 2.5. AUROC increases with sample size, predictor variability, and reduced MSE

Using the simulation design described in 2.3 and 2.4, we focused on three aims: illustrate the effect upon deconvolution sensitivity of 1) sample sizes, 2) residual sizes, and 3) variability of cell type-specific proportions across samples.

Distinct values of each parameter for group sample sizes, effect size, residual sizes (quantified by MSE), and target cell proportions standard deviation were used to create synthetic data as in the simulation described in 2.3 and 2.4. Each ROC curve plotline in **Figure 2.1**represents a distinct simulation data set. Each ROC curve was constructed using permutation based p-values and a truth vector as described in section 2.4.

### 2.5.1. Reductions in MSE increase sensitivity and specificity of cell type-specific differential expression detection as quantified by AUROC

It is important to draw the distinction here between the actual variance $\sigma^2$, which cannot be known, vs. the estimated variance $s^2$. The variance is estimated using the mean squared error (MSE). In simulation, we apply known variability to our synthetic residuals in the process of simulating heterogeneous expression levels. We thus manipulate the size of the variance estimate (MSE) by manipulating the known variability of our simulated residuals.

Reductions in MSE consistently resulted in increased AUROC (panel A **Figure 2.1**). Since $s^2$ is the numerator of the linear regression estimate of coefficient variability (eq. 9), it is intuitive to see that there is a directly proportional relationship between $s^2$ (estimated using MSE) and cell type-specific expression estimate variance (eq. 9). Any reduction in cell type-specific expression estimate variance will shrink the confidence interval around the estimate resulting in greater power of differential expression detection for any given fixed size of difference. In actual biological data, no two genes will likely have identical MSE. MSE is therefore a cell type and genomic site specific measure, unique for every single regression conducted during the analysis.

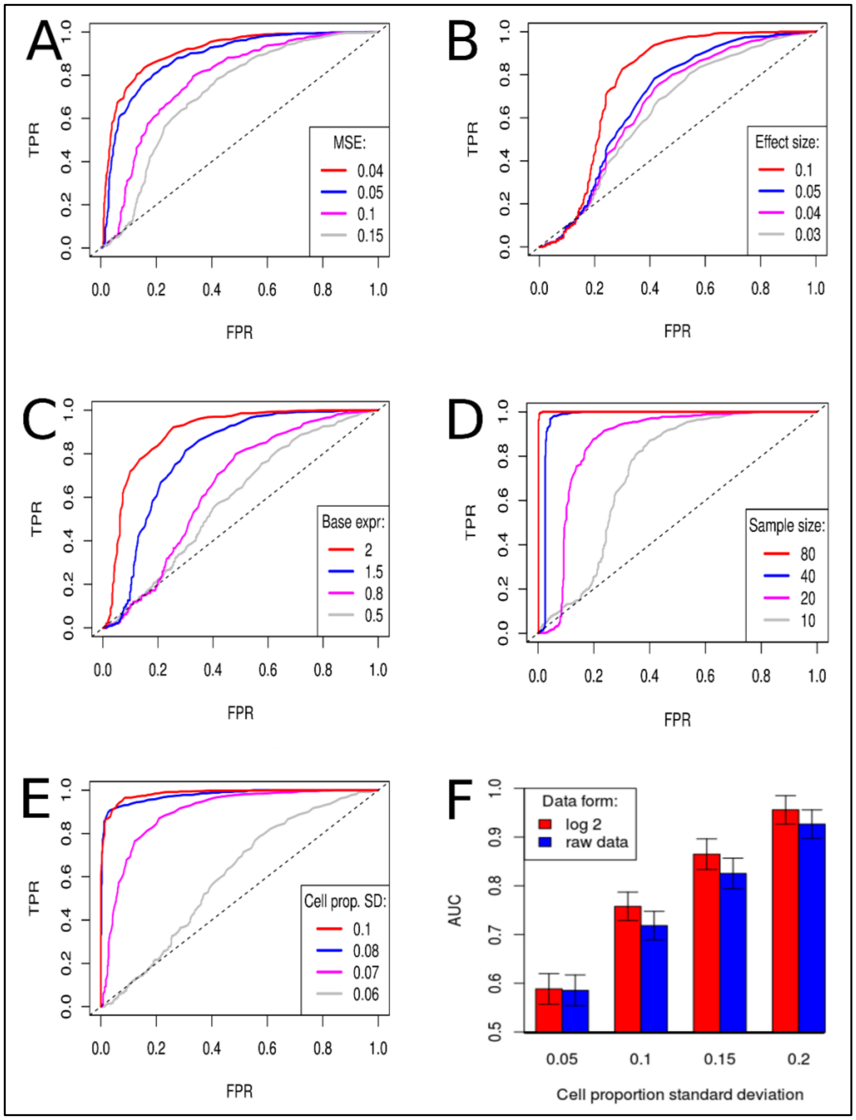


**Figure 2.1: Results of AUROC analysis for various parameters.**

Figure 2.1 panel A) shows increases in AUROC with decreases in mean squared error (MSE). Panel B) show increases in AUROC with increases in effect size. Panel C) shows increase in AUROC with increased base expression at fixed fold change (increasing effect size as base increases). Panel D) shows increase in AUROC with increase in sample size. Panel E) shows increase in AUROC with increase in target cell proportion standard deviation across samples. Panel F) shows higher AUROC for log2 versus “raw” analysis for various levels of cell proportion standard deviation.

### 2.5.2. Increases in sample size per group increase sensitivity and specificity of cell type-specific differential expression detection as quantified by AUROC

Increasing group-wise sample sizes results in increased AUROC as illustrated in the resulting ROC curves (Figure 2.1). For an intuitive understanding of why this is so, a look at the simple linear regression estimation equation (eq. 7) serves to illustrate. Since sample size M is in the denominator of eq. 7 this result not surprising. The effect of increasing sample size is reduced variability around cell type-specific expression estimates, which narrows the confidence intervals around these estimates resulting in greater observed power to detect differential expression. In this way, sample size for a given study group will impact the variability of cell type-specific expression estimates for all cells and all genes for that study group.

### 2.5.3. Increases in predictor cell proportion variability increase sensitivity and specificity of cell type-specific differential expression detection as quantified by AUROC

Larger variability of the target cells proportions across samples as quantified by the standard deviation results in increased AUROC (Figure 2.1). Again, looking to the simple linear regression equation (eq. 7) for a more intuitive feel, cell proportion variability across samples is captured in the denominator. In our multivariate regression this cell proportion variability is captured in the inverse design matrix $\mathbf{X}^{-1}\mathbf{X}$ (eq. 9), thus any increase in cell proportion variability will result in a proportional reduction of the variance of cell type-specific expression estimates for each gene in that particular target cell type. A different cell type with a different standard deviation across samples will have a different AUROC and a different quantification of power given the same sample sizes, MSE, and effect sizes for the same set of genes.

### 2.5.4. Increases in target cell differential expression increase sensitivity and specificity of cell type-specific differential expression detection as quantified by AUROC

Unsurprisingly, increasing effect size (size of cell type-specific differential expression) while holding other parameters fixed results in no change to AUROC. Changes in effect size do not impact the variability of cell type-specific expression estimates variances (results not shown). Unchanged variances equate to unchanged 95% confidence intervals around point estimates. Therefore, simply moving the two group-wise expression points along the x-axis for equal distances in the same direction results in identical observed AUROC.

**Error! Reference source not found.** shows the parameter values used in simulation in order to generate **Figure 2.1**.

**Table 2.2: Parameters used to generate Figure 2.1.**

| | Group sample size | MSE | Target cell SD | Effect size | Base expression |
|---|---|---|---|---|---|
| Panel A | 10 | - | 0.1 | 0.05 | 1.0 |
| Panel B | 10 | 0.01 | 0.06 | - | 1.0 |
| Panel C | 10 | 0.01 | 0.06 | 0.05 | - |
| Panel D | - | 0.45 | 0.1 | 0.2 | 1.0 |
| Panel E | 10 | 0.01 | - | 0.01 | 1.0 |
| Panel F | 10 | 0.01 | - | 0.05 | 1.0 |

## 2.6. Conclusions

The results of the permutation based simulation study (2.3 and 2.4) illustrate the way in which changes in sample sizes, MSE, and cell proportion standard deviation affect the sensitivity of cell type-specific differential expression detection.

These three parameters each affect the variability estimates of cell type-specific expression estimates (eq. 9) and thus affect the sensitivity and specificity (AUROC) of cell type-specific differential expression detection (LRCDE). Given fixed effect size, MSE, and cell proportion variability, any increase in sample size will result in increased AUROC. When other parameters are held fixed, decreased MSE will result in increased AUROC. Holding sample size, MSE, and effect size fixed, increased target cell proportion variability across samples (cell SD) will result in increased AUROC. It is important in any analysis to realize that given fixed sample sizes, observed effect size, and cell SD, detection sensitivity and specificity will vary with MSE per gene site for that target cell. Thus, two identical effect sizes observed within the same cell type may have different observed power: one site may indicate a significant difference while the other does not. It is precisely this reason that we subsequently examine power for each cell type on a gene-by-gene basis.

# 3. Comparing Cell Type-Specific Differential Expression Detection Sensitivity between Two-sample T-test to FDR

## 3.1. Two approaches to analysis of cell type-specific differential expression detection

In chapter 2 we demonstrated the effect that changes in sample size, MSE, and cell SD have upon sensitivity and specificity (AUROC) of cell type-specific differential expression detection (effect sizes). In this dissertation, we propose that a two-sample t-test as a more sensitive and precise way of detecting significant effect sizes than the globally derived FDR approach by applying a gene-by-gene analysis to the detection of cell type-specific differential expression between two groups. We hypothesize that the per-gene approach will be more sensitive because it will take into account gene-specific parameters affecting linear regression.

In order to test this premise and investigate the differences between the two-sample t-test approach vs. the permutation based false discovery rate (FDR) approach, we conduct a simulation study in which sensitivity of detection is contrasted between the two methods across changes in samples size, MSE, and cell SD. We then apply both methods to a study involving stable vs. rejection groups in kidney transplant patients (Shen-Orr et al. 2010).

Worth noting here is that csSAM includes options for median centering and standardizing cell type-specific differential expression estimates. Median centering involves subtracting the median differential expression for an entire cell type from all difference estimates in that cell. Standardization involves dividing by the adjusted standard deviation so that each difference estimate for the entire cell is on the same scale.

Comparison of the two-sample t-test to the FDR approach is conducted through simulation study. Heterogeneous measures are synthesized using synthetic and known cell type-specific proportion measures across samples and cell type-specific differential expressions. Synthetic heterogeneous measures and synthetic cell proportions are then used in linear regression in order to obtain cell type-specific expression estimates. Differential expression estimates are then analyzed using both the two-sample t-test approach and the FDR approach. True positive rates (TPR – sensitivity) of both results are then compared.

Our aims are thus to demonstrate: 1) that two-sample t-test is more sensitive than FDR across a range of effect sizes, and 2) that two-sample t-test remains sensitive to detection of effect sizes missed by FDR when group sample sizes are small.

### 3.2. Two-sample T-test approach to LRCDE

Given the estimated difference between average measures taken on two groups, a two-sample t-statistic (Welch-Satterthwaite test (Welch 1947)) may be calculated and tested against a t-critical value calculated using two-sample degrees of freedom. Any detected differential expression will have variability attached to it as a result of variability around the group-wise cell expression (linear regression coefficient) estimates. In order to test whether an observed difference is statistically significant we apply the two-sample t-test, in which we compute the two-sample t-statistic

$$t_{diff} = \frac{(\hat{\beta}_2 - \hat{\beta}_1) - \delta_0}{(\mathrm{se}_{\mathrm{welch}})} \tag{12}$$

for the observed difference $\hat{\beta}_2 - \hat{\beta}_1$ and compare it against a t-critical value, where

$$se_{welch} = \sqrt{\frac{se_1^2}{n_1} + \frac{se_2^2}{n_2}} \tag{13}$$

and $\delta_0 = 0$ in order to test the null hypothesis of zero difference. In eq. 13, $n_2$ and $n_1$ are case and control group sample sizes respectively, and $se_2^2$ and $se_1^2$ are case and control group variance estimates for $\hat{\beta}_2, \hat{\beta}_1$ respectively. Degrees of freedom used to determine the t-critical value are calculated using Satterthwaite's equation:

$$d.f. = \frac{\left(se_1^2 / n_1\right) + \left(se_2^2 / n_2\right)}{\left(\frac{\left(se_1^2 / n_1\right)^2}{(\mathrm{n}_1 - 1)} + \frac{\left(se_2^2 / n_2\right)^2}{(\mathrm{n}_2 - 1)}\right)} \qquad (14).$$

This results in a potentially non-integral degrees of freedom value. The t-critical value may be interpolated from any t-statistic chart or computed using the 'qt' function available in the R statistical platform. Whenever there is a balanced design and both groups have identical standard errors, then the Satterthwaite degrees of freedom will agree with the pooled degrees of freedom result:

$$d.f._{pooled} = (n_1 + n_2 - 2) \qquad (15).$$

Any difference between group standard errors will result in reduced degrees of freedom in the Satterthwaite equation. In the case that the t-statistic (eq. 12) exceeds the t-critical value, then we reject the null hypothesis of no difference between $\hat{\beta}_2, \hat{\beta}_1$ and conclude that there is significant evidence indicating that a substantive difference exists.

The above approach describes a scenario in which a single test of significance is being conducted. In the case of multiple tests (as with multiple genes), a Bonferroni family-wise error rate (FWER) is calculated for *n* tests as $FWER = \frac{\alpha}{n}$ . The t-statistic p-value is then compared to the FWER and the null hypothesis of no difference between group mean expression is rejected when p-value < *FWER*. This provides a conservative approach to differential expression detection and is less likely to create false positive observations than applying unmodified $\alpha$.

### 3.2.1. Relationship between power and t-critical significance threshold

Power calculation proceeds in the following manner. Having calculated the two-sample t-statistic as in eq. 12 and 13, and a t-critical value based on eq. 14 degrees of freedom, power then is the probability of rejecting the null hypothesis of no difference between group mean expressions. Assuming an up-regulated significantly detected difference (t-critical < t-statistic), then t-critical minus t-statistic represents a negative distance on the same scale as t-critical and the t-statistic: another t-statistic. The upper tail probability of this negative t-statistic is the power, interpreted as the probability of rejecting the null hypothesis of zero difference.

## 3.3. FDR per feature is dependent upon all features in sample

FDR reporting is based upon group label permutations and repeated linear regression differential expression estimates. Therefore, FDR is dependent upon adequate sample sizes per study group in order to properly profile the non-parametric null distribution of no difference between groups. As implemented, it further relies upon having a distribution of differences of various sizes across the range of features (genes). FDR is thus limited by sample size as well as actual observable cell type-specific expression differences.

In exploring the behavior of FDR, we were able to closely replicate FDR plots (Figure 3.1) published in the supplemental material of (Shen-Orr et al. 2010), using their csSAM algorithm. Importantly: we found no indication of the random seed setting used in the R environment used to obtain the published plots. Allowing the random seed to float over repeated analyses of the same data will produce differing lists of low FDR (FDR < 0.3) probe sites. This is due to the randomly sampled group labels used in the permutation process prior to FDR calculation. Thus, our findings will likely not produce the exact same list of significantly detected (low FDR) sites as the authors of csSAM.

In order to replicate published FDR plots, the csSAM parameters for median centering and standardization were set to true. Standardizing each gene difference in a single cell type by dividing by the pooled adjusted standard deviation of expression estimates places all detected differences for that cell on the same scale. Median centering creates symmetry in the distribution of the observed differences around zero. Half the estimates are now negative, the other half positive, and the mid-point is now zero (for an odd number of observations).

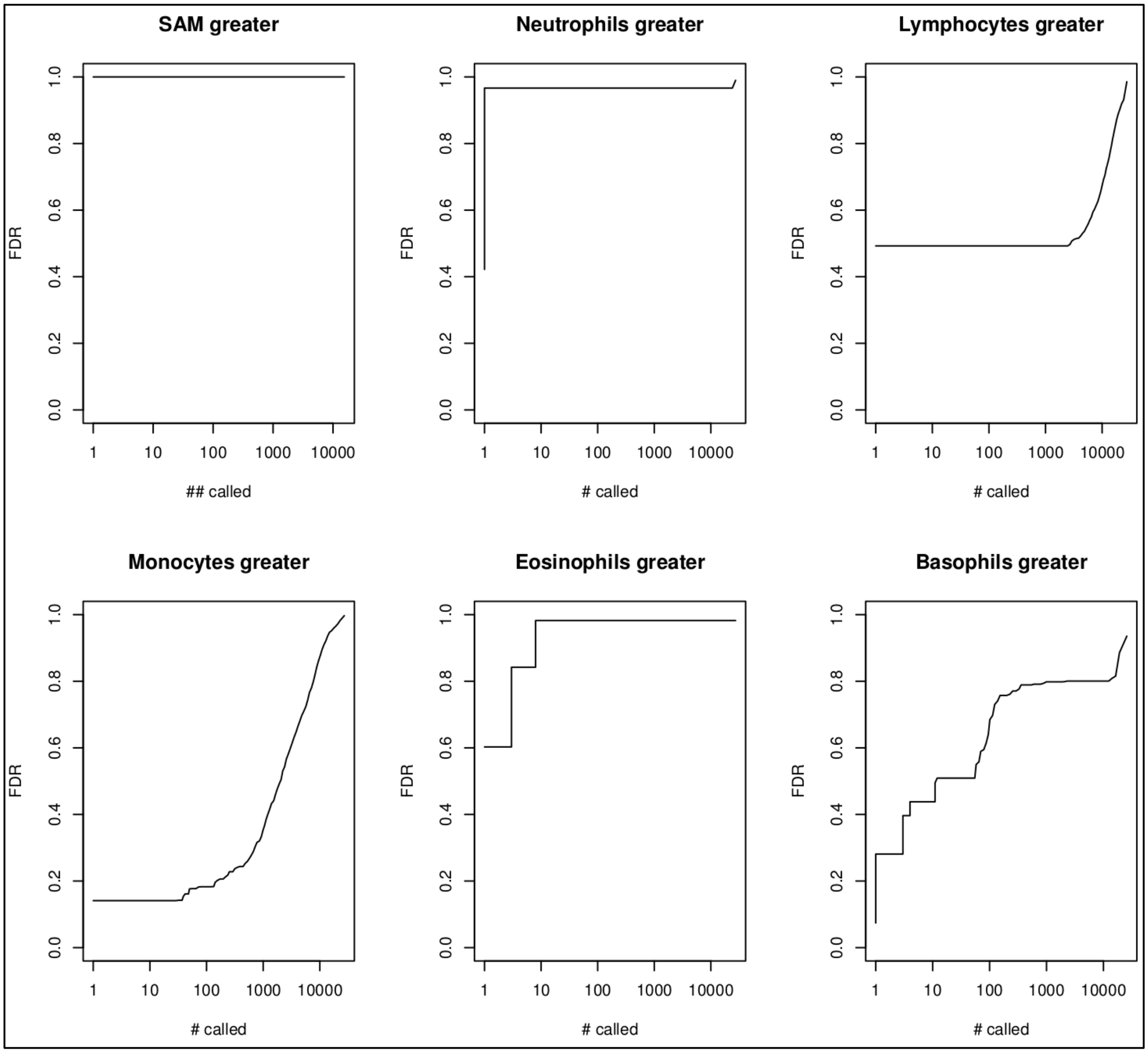
The data set in the Shen-Orr paper are blood samples from kidney transplant patients with 9 stable and 15 experiencing rejection. Included in the kidney data are two matrices: a heterogeneous observations matrix over 54,675 features and a cell proportions matrix for each of the 24 samples over 5 cell types: neutrophils, lymphocytes, monocytes, eosinophils, and basophils.

In the analysis of the kidney transplant data, the authors focus in on low FDR rates noticed in monocytes after testing for up-regulated features alone across all 54,675 features. We noted 1,203 probe sites (882 genes) with FDR below 0.3 in the process of replicating their findings.

Replication of published FDR plots was conducted in order to demonstrate use of the csSAM algorithm consistent with the author’s intentions (Figure 3.1).

We must note that the authors report filtering out all but the top 5000 “most variable” sites in order to produce at least one set of figures in their supplemental material. They report the improvement in FDR rates in some cell types do to this filtering step. This change in FDR based on elimination of some features is an artifact of the way in which FDR is calculated based on the range of observed cell specific differential expression sizes detected. For this reason, pre-filtering (removal) of genomic sites is not recommended. In replicating the reported sites with FDR below 0.3, it is unclear in which way several of their FDR plots are produced.

**Figure 3.1: FDR plots replicated using kidney transplant data from Shen-Orr et al 2010.**



The FDR plots generated using the native csSAM plot mechanism after analyzing all probe sites in original kidney data with median centering, standardization, and non-negative parameters set to TRUE.

### 3.4. Simulation designed to compare two-sample t-statistic to FDR

Data simulation proceeded as described in 2.3 with the exception of applied known cell type-specific differential expression. For comparison of t-statistic sensitivity to FDR, we simulate 40 genes, 20 are differentially expressed across groups along the range of 0.11, 0.12, … , 0.30. The remaining 20 are not differentially expressed across groups. Sample size is fixed at 15 per group, MSE at 1.5, cell SD at 0.1, and cell predictor design matrix condition number at 100. Cell 1 is target cell with differential expression applied.

After obtaining LRCDE estimates of differential expression, t-statistics, t-statistic p-values, and FDRs for cell 1, we compare the true positive detection rates between t-statistic p-values and FDR at prescribed thresholds: 0.3 and 0.05/(20) for FDR and p-values respectively. The alpha level of 0.05 is divided by 40 in order to obtain the FWER for tests over 20 genes. In this way, we apply a conservative threshold to the t-statistic based p-value and a relatively liberal 0.3 threshold for the FDR. Any FWER corrected t-statistic p-value observed below the FWER corrected threshold is "called" as significant, and any FDR below 0.3 is "called" as significant.

### 3.5. Two-sample t-statistic exhibits greater sensitivity than FDR

#### 3.5.1. Simulation study comparison of two-sample t-statistic to FDR

Using the described simulation methodology in 3.4, we contrasted the sensitivity of a two-sample t-statistic vs. FDR in the detection of cell type-specific differential expression (LRCDE). The same random normal heterogeneous matrix was analyzed using both the LRCDE package and csSAM. With alpha critical threshold for t-statistic p-values set to 0.05, then the FWER was 0.05/40, (40 genes being tested). LRCDE using t-statistic p-value compared against FWER indicated 10 out of 20 of the known differentially expressed genes as significantly so. These corresponded to the largest known simulated effect sizes of 0.21 through 0.30. Lowest indicated

FDR from csSAM was non-significant being below the global significance threshold 0.3. FDR was equal to 0.627 uniformly across all 20 known differentially expressed genes with results of testing both LRCDE and csSAM shown in table Table 3.1 (with both median centering and standardization set to FALSE). Testing the same simulated data with median centering and standardization of differential expression estimates set to TRUE resulted in all 40 genes being “called” as significant with both the 20 differentially expressed and the 20 not differentially expressed “detected” at FDR below 0.176 (Table 3.3).

**Table 3.1: Comparison of LRCDE to csSAM on simulated data with 15 samples per group**

Based on FWER of 0.00125, sites 11 through 20 are significantly differentiated as indicated by “p-value”. None of the 20 differentially expressed sites indicated FDR < 0.3. Sites 21 through 40 were not differentially expressed and are not shown. “Site” – probe site name; “Base” – control group cell type specific expression estimate; “Case” – case group expression estimate; “Diff.est” – estimated differential expression estimate; “t-critical” – calculated t-critical value based on Satterthwaite’s degrees of freedom; “t-statistic” – observed t-statistic for differential expression estimated based on group sample sizes, and standard error estimates of base and case expressions; “p-value” – based on observed t-statistic; “FDR” – as calculated by the csSAM algorithm on the same simulated data set. Group sample sizes were 15, MSE target of 1.5, cell SD of 0.1, and cell proportion design matrix condition number target of 100.

| Site | Base | Case | Diff.est | t-critical | t-statistic | p-value | FDR |
|---|---|---|---|---|---|---|---|
| 1 | 1.90 | 2.01 | 0.11 | 1.70 | 1.79 | 4.21e-02 | 0.63 |
| 2 | 1.90 | 2.02 | 0.12 | 1.70 | 1.95 | 3.05e-02 | 0.63 |
| 3 | 1.90 | 2.03 | 0.13 | 1.70 | 2.11 | 2.17e-02 | 0.63 |
| 4 | 1.90 | 2.04 | 0.14 | 1.70 | 2.28 | 1.53e-02 | 0.63 |
| 5 | 1.90 | 2.05 | 0.15 | 1.70 | 2.44 | 1.06e-02 | 0.63 |
| 6 | 1.90 | 2.06 | 0.16 | 1.70 | 2.60 | 7.31e-03 | 0.63 |
| 7 | 1.90 | 2.07 | 0.17 | 1.70 | 2.77 | 4.97e-03 | 0.63 |
| 8 | 1.90 | 2.08 | 0.18 | 1.70 | 2.93 | 3.35e-03 | 0.63 |
| 9 | 1.90 | 2.09 | 0.19 | 1.70 | 3.09 | 2.24e-03 | 0.63 |
| 10 | 1.90 | 2.10 | 0.20 | 1.70 | 3.25 | 1.49e-03 | 0.63 |
| 11 | 1.90 | 2.11 | 0.21 | 1.70 | 3.42 | 9.79e-04 | 0.63 |
| 12 | 1.90 | 2.12 | 0.22 | 1.70 | 3.58 | 6.41e-04 | 0.63 |
| 13 | 1.90 | 2.13 | 0.23 | 1.70 | 3.74 | 4.18e-04 | 0.63 |
| 14 | 1.90 | 2.14 | 0.24 | 1.70 | 3.90 | 2.71e-04 | 0.63 |
| 15 | 1.90 | 2.15 | 0.25 | 1.70 | 4.07 | 1.75e-04 | 0.63 |
| 16 | 1.90 | 2.16 | 0.26 | 1.70 | 4.23 | 1.13e-04 | 0.63 |
| 17 | 1.90 | 2.17 | 0.27 | 1.70 | 4.39 | 7.28e-05 | 0.63 |
| 18 | 1.90 | 2.18 | 0.28 | 1.70 | 4.56 | 4.67e-05 | 0.63 |
| 19 | 1.90 | 2.19 | 0.29 | 1.70 | 4.72 | 3.00e-05 | 0.63 |
| 20 | 1.90 | 2.20 | 0.30 | 1.70 | 4.88 | 1.92e-05 | 0.63 |

Doubling the sample size from 15 per study group to 30 per study group while holding other parameters fixed resulted in both t-statistic based p-values and FDR detecting all 20 differentially expressed simulated genes at FWER of 0.00125 and 0.30 respectively Table 3.2.

**Table 3.2: Comparison of LRCDE to csSAM on simulated data with 30 samples per group**

Based on FWER of 0.00125, sites 1 through 20 are significantly differentiated as indicated by “p-value”. Also; all of the 20 differentially expressed sites indicated FDR < 0.3. Greatest p-value for non-differentially expressed sites was 0.5. FDR for non-differentially expressed sites 21 through 40 had indicated FDR of 0.98. Non-differentially expressed sites not shown. “Site” – probe site name; “Base” – control group cell type specific expression estimate; “Case” – case group expression estimate; “Diff.est” – estimated differential expression estimate; “t-critical” – calculated t-critical value based on Satterthwaite’s degrees of freedom; “t-statistic” – observed t-statistic for differential expression estimated based on group sample sizes, and standard error estimates of base and case expressions; “p-value” – based on observed t-statistic; “FDR” – as calculated by the csSAM algorithm on the same simulated data set. Group sample sizes were 30, MSE target of 1.5, cell SD of 0.1, and cell proportion design matrix condition number target of 100.

| Site | Base | Case | Diff.est | t-critical | t-statistic | p-value | FDR |
|---|---|---|---|---|---|---|---|
| 1 | 1.87 | 1.98 | 0.11 | 1.67 | 9.46 | 1.17E−13 | 0.180 |
| 2 | 1.87 | 1.99 | 0.12 | 1.67 | 10.32 | 4.79E−15 | 0.159 |
| 3 | 1.87 | 2.00 | 0.13 | 1.67 | 11.18 | 2.13E−16 | 0.135 |
| 4 | 1.87 | 2.01 | 0.14 | 1.67 | 12.04 | 1.03E−17 | 0.104 |
| 5 | 1.87 | 2.02 | 0.15 | 1.67 | 12.90 | 5.51E−19 | 0.091 |
| 6 | 1.87 | 2.03 | 0.16 | 1.67 | 13.76 | 3.23E−20 | 0.069 |
| 7 | 1.87 | 2.04 | 0.17 | 1.67 | 14.62 | 2.08E−21 | 0.048 |
| 8 | 1.87 | 2.05 | 0.18 | 1.67 | 15.48 | 1.47E−22 | 0.041 |
| 9 | 1.87 | 2.06 | 0.19 | 1.67 | 16.34 | 1.14E−23 | 0.032 |
| 10 | 1.87 | 2.07 | 0.2 | 1.67 | 17.20 | 9.58E−25 | 0.028 |
| 11 | 1.87 | 2.08 | 0.21 | 1.67 | 18.06 | 8.76E−26 | 0.023 |
| 12 | 1.87 | 2.09 | 0.22 | 1.67 | 18.92 | 8.67E−27 | 0.021 |
| 13 | 1.87 | 2.10 | 0.23 | 1.67 | 19.78 | 9.25E−28 | 0.018 |
| 14 | 1.87 | 2.11 | 0.24 | 1.67 | 20.64 | 1.06E−28 | 0.013 |
| 15 | 1.87 | 2.12 | 0.25 | 1.67 | 21.50 | 1.30E−29 | 0.013 |
| 16 | 1.87 | 2.13 | 0.26 | 1.67 | 22.36 | 1.70E−30 | 0.013 |
| 17 | 1.87 | 2.14 | 0.27 | 1.67 | 23.22 | 2.37E−31 | 0.013 |
| 18 | 1.87 | 2.15 | 0.28 | 1.67 | 24.08 | 3.49E−32 | 0.013 |
| 19 | 1.87 | 2.16 | 0.29 | 1.67 | 24.94 | 5.44E−33 | 0.013 |
| 20 | 1.87 | 2.17 | 0.3 | 1.67 | 25.80 | 8.93E−34 | 0.013 |

In Table 3.1 we see little correlation between FDR and differential expression size estimate (the FDR column in fact has zero variability). However, in Table 3.2 the correlation between FDR and differential expression size estimate has become a strong negative correlation of 0.87. This change in FDR correlation from zero to -0.87 when going from 15 samples per group to 30 samples per group indicates a problem inherent in permutation based methods given small samples sizes. Under these simulated conditions, the sensitivity of FDR has become equal to the sensitivity of the two-sample t-statistic approach to detection by doubling the sample size.

Following the same simulation methodology as described above, we looked at the 500 known differentially expressed genes with 1000 total genes (500 non-differentiated). True positive rates (TPR - sensitivity) were calculated for FDR-based and for t-statistic p-values over three distinct values for each of samples sizes, MSE, and cell type-specific standard deviation of proportions across samples. Each of these results is plotted in Figure 3.2 and values of each manipulated parameter are indicated in figure legends. In Figure 3.2.A the effect of sample size on FDR is evident: at a sample size of 10 per group, there is zero level of true positive detection given any FDR threshold. At a sample size of 18, however, FDR now indicates a TPR of ~0.6 at the lowest threshold level.

Table 3.3 is included in order to illustrate the effect of using the csSAM parameters for median centering and standardizing the cell type-specific differential expression estimates obtained from linear regression. Notice that “sites” 21 through 40 are not differentially expressed, yet FDR is reported as less than 0.2 for all sites in the simulated data set.

**Table 3.3: Results of setting median centering and standardization to TRUE on FDR**

| Site | Base | Case | Diff.est | t-critical | t-statistic | p-value | FDR |
|---|---|---|---|---|---|---|---|
| 1 | 1.90 | 2.01 | 0.11 | 1.70 | 1.79 | 4.22e-02 | 0.18 |
| 2 | 1.90 | 2.02 | 0.12 | 1.70 | 1.95 | 3.05e-02 | 0.15 |
| 3 | 1.90 | 2.03 | 0.13 | 1.70 | 2.11 | 2.17e-02 | 0.12 |
| 4 | 1.90 | 2.04 | 0.14 | 1.70 | 2.28 | 1.53e-02 | 0.11 |
| 5 | 1.90 | 2.05 | 0.15 | 1.70 | 2.44 | 1.06e-02 | 0.09 |
| 6 | 1.90 | 2.06 | 0.16 | 1.70 | 2.60 | 7.31e-03 | 0.08 |
| 7 | 1.90 | 2.07 | 0.17 | 1.70 | 2.77 | 4.97e-03 | 0.07 |
| 8 | 1.90 | 2.08 | 0.18 | 1.70 | 2.93 | 3.35e-03 | 0.06 |
| 9 | 1.90 | 2.09 | 0.19 | 1.70 | 3.09 | 2.24e-03 | 0.05 |
| 10 | 1.90 | 2.10 | 0.2 | 1.70 | 3.25 | 1.49e-03 | 0.04 |
| 11 | 1.90 | 2.11 | 0.21 | 1.70 | 3.42 | 9.79e-04 | 0.03 |
| 12 | 1.90 | 2.12 | 0.22 | 1.70 | 3.58 | 6.41e-04 | 0.03 |
| 13 | 1.90 | 2.13 | 0.23 | 1.70 | 3.74 | 4.18e-04 | 0.02 |
| 14 | 1.90 | 2.14 | 0.24 | 1.70 | 3.90 | 2.71e-04 | 0.02 |
| 15 | 1.90 | 2.15 | 0.25 | 1.70 | 4.07 | 1.75e-04 | 0.01 |
| 16 | 1.90 | 2.16 | 0.26 | 1.70 | 4.23 | 1.13e-04 | 0.01 |
| 17 | 1.90 | 2.17 | 0.27 | 1.70 | 4.39 | 7.28e-05 | 0.01 |
| 18 | 1.90 | 2.18 | 0.28 | 1.70 | 4.56 | 4.67e-05 | 0.01 |
| 19 | 1.90 | 2.19 | 0.29 | 1.70 | 4.72 | 3.00e-05 | 0.01 |
| 20 | 1.90 | 2.20 | 0.30 | 1.70 | 4.88 | 1.92e-05 | 0.01 |
| 21 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 22 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 23 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 24 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 25 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 26 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 27 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 28 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 29 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 30 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 31 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 32 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 33 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 34 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 35 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 36 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 37 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 38 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 39 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |
| 40 | 1.90 | 1.90 | 0 | 1.70 | 0 | 5.00e-01 | 0.18 |

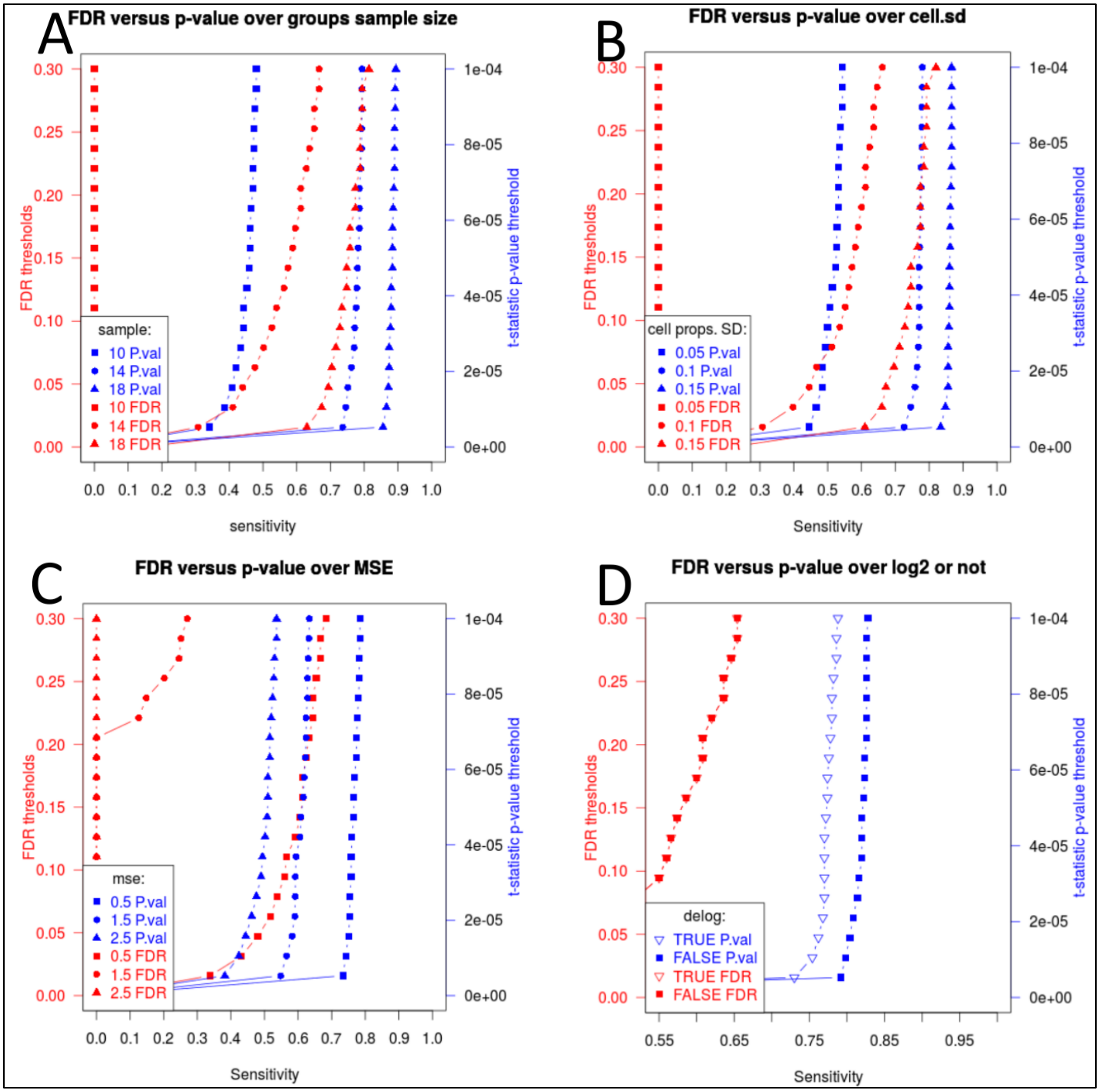


**Figure 3.2: Contrasting FDR vs. two-sample t-statistic**

Panel A) FDR contrasted with two-sample t-statistic over sample sizes.

Panel B) FDR contrasted with two-sample t-statistic over cell proportion variability (cell SD).

Panel C) FDR contrasted with two-sample t-statistic over levels of MSE.

Panel D) FDR contrasted with two-sample t-statistic comparing log2 (FALSE = normal) heterogeneous measures to raw (TRUE = “de-logged”) measures.

Unless specified otherwise in Figure 3.2 legends, log2-transformed data was used, and the following parameters were held constant: per-group sample size - 14, condition number - 100, cell proportion SD - 0.1 (0.6 for C), MSE - 1.5. All two-sample t-statistic p-values were compared against the Bonferroni corrected $\alpha$ with $\alpha$ ranging from 0 to 0.1 (right hand y-axis of all panels in Figure 3.2).

In tables Table 3.1, Table 3.2, and Table 3.3 and in Figure 3.2, the csSAM package was used to obtain FDR and the LRCDE package was used to calculate two-sample t-statistic p-values. The csSAM package relies upon group label permutations in order to calculate FDR. The total possible number of group label permutations is:

$$\text{total possible permutations } = \frac{n!}{r!(n-r)!} \qquad (16),$$

where $n$ would be total sample size (sample size of controls plus sample size of cases). For unequal samples sizes, r could be size of either control or cases group. For unequal group sizes of 9 and 15, such as in the kidney data, there are a total of 1,307,504 ways of permuting group labels. This is computationally infeasible, so a random subset of these total combinations is taken. We ran csSAM using 1000 permutations of group labels in recreating Figure 3.1 and for simulated data used in tables Table 3.1, Table 3.2, and Table 3.3 and in Figure 3.2.

### 3.5.2. Comparing two-sample t-statistic to FDR applied to biological data

After testing both the two-sample t-test and FDR on simulated data, we applied both methods to the kidney transplant dataset (Shen-Orr et al. 2010) analyzed by the authors of csSAM which contains samples from 9 healthy controls and 15 transplant patients on whole blood. Proportions of the 5 cell types: neutrophils, lymphocytes, monocytes, basophils, and eosinophils, are supplied and presumed to comprise a total 100% across all samples. We focused on testing for up-regulation in monocytes since this is where the authors report finding significantly detected (according to FDR below 0.3) differentially expressed probe sites between study groups.

**Table 3.4: The number of cell type-specific differentially expressed probes (genes) identified in kidney transplant gene expression data**.

| Cell | Neutrophils | Lymphocytes | Monocytes | Basophils | Eosinophils |
|---|---|---|---|---|---|
| Mean | 0.592 | 0.281 | 0.098 | 0.025 | 0.004 |
| SD | 0.193 | 0.151 | 0.063 | 0.024 | 0.003 |
| FDR < 0.3 | 0 | 0 | 1203 (882) | 1 | 0 |
| Bonferroni p-value < 0.05 | 3122 | 4975 | 9066 (6018) | 648 | 1263 |
| Overlapping | 0 | 0 | 1187 (877) | 0 | 0 |

Looking at monocytes alone, we found using csSAM 1203 probes (882 genes) with FDR below 0.3 (section 3.3). Applying the Bonferroni corrected t-statistic p-value < 0.05 using LRCDE for up-regulation we found 9,066 features (6018 genes) with significant differences. Table 3.4 summarizes these findings. There were 1187 sites (877 genes) in common between those detected by csSAM with FDR below 0.3 and those detected by LRCDE.

LRCDE detects 877 of the 882 genes detected at FDR < 0.30 as well as 5141 additional genes. This mirrors observations from analysis of simulated data in that LRCDE detects sites missed by FDR. From Table 3.3 we know that csSAM will “detect” sites with FDR < 0.30 which have no true differential expression when median centering and standardization are applied to cell type-specific differential expression estimates. Table 3.3 illustrates the problem with FDR when applying median centering and standardization which we refer to as the “symmetry effect”. Because detected differential expression estimates are median centered about zero, the highest FDR rates will be assigned to the midpoint (median) of detected differences. Lowest FDR rates will be assigned to both the smallest differences and largest differences detected. This effectively disconnects FDR from any meaningful correlation with the sizes of detected differential expression.

### 3.6. Conclusions

Contrasting the sensitivity of two-sample t-statistic p-value against FDR indicates that the t-statistic approach to quantifying the significance of cell type-specific differential expression (effect sizes) is consistently more sensitive than FDR at detection of smaller known effect sizes.

When sample sizes are insufficient, group label permutations fail to provide enough information about a null distribution to allow the algorithm to distinguish small differential expression from larger differential expression resulting in identical and non-significant ( > 0.3) FDRs across a range of effect sizes.

The csSAM algorithm includes a facility for median centering and standardizing the cell type-specific differential expression estimates. The plots in Figure 3.2 were re-created using csSAM native plotting facility with csSAM parameters for median centering and standardization are set to TRUE during analysis. When median centering and standardization are set to TRUE, the csSAM computed FDR sensitivity appears consistently improved across known differentially expressed probe sites (sites 1 through 20 in Table 3.3). However, setting median centering and standardization are set to TRUE also results in increased false discovery based on the FDR < 0.3 threshold (sites 21 through 40 in Table 3.3). Although FDR indicates significant (< 0.3) detection on simulated data with no true differential expression between groups at the cell specific level when setting median centering and standardization of differential expression estimates to TRUE, it does not appear to do so when these two parameters are set to FALSE. Based on the results of these simulations, we do not recommend setting either of these two parameters to TRUE.

Median centering and standardizing differential expression estimates prior to calculating two-sample t-statistics will result in flawed inference since the t-statistic relies upon estimated differences of group means on the original scale. Median centering and standardizing differential expression estimates prior to calculating the two-sample t-statistic will result in meaningless values.

The large overlap in Table 3.4 between detected genes with FDR < 0.3 and the greater number detected by LRCDE provide further evidence of the higher sensitivity of two-sample t-test observed in simulation studies Figure 3.2. Given that csSAM indicates FDR < 0.3 for simulated genes with no differential expression when setting median centering and standardization of differential expression estimates to TRUE may be the reason why FDR appears to detect some probe sites missed by the t-test.

Given larger MSE, low cell proportion variability, small sample sizes, or a combination of these three conditions, FDR fails to detect differential expression below a 0.3 threshold. Given smaller MSE, higher variability across target cell proportions, increased sample sizes, or a combination of these conditions, FDR begins to detect known differential expressions at a 0.3 threshold.

For any combination of parameters, two-sample t-statistic p-values tested against a Bonferroni corrected significance threshold are consistently more sensitive to detection of known differential expression across a range of effect sizes than FDR.

# 4. Log2 Transformation of Heterogeneous Observations

## 4.1. Is greater sensitivity derived by deconvolution of log2 or raw data

We now turn to the question of which provides greater sensitivity and specificity in cell type-specific differential expression detection: deconvolution of 1) log2 transformed heterogeneous measures, or 2) raw ("de-logged") heterogeneous measures. In order to address this question, we conduct simulations using the permutation method in order to obtain AUROCs as well as a simple comparison of true positive rates (sensitivity).

## 4.2. Linear regression expects normally distributed residuals

Genomic feature measurements, particularly microarray, are typically log2 transformed prior to performing any sort of analysis on them for several reasons. Log2 transformation reduces the magnitude of the range of the heterogeneous measures, making them more normally distributed, and it provides easy interpretation of differential expression between groups making ratios symmetrical around 0 (a 1-unit change is equivalent to a two-fold change or doubling of expression) (Ballman 2008).

A primary assumption of linear regression is that the residuals around a regression line are normally distributed. If this assumption is violated, then the coefficient estimates associated with predictor variables may not follow a symmetric t-distribution and inferences based on coefficient estimates, particularly those focused on variability of coefficient estimates, may be misleading. However, even if outcome measures (heterogeneous measures) depart from normality, we may still operate under the assumption that coefficient estimates follow a t-distribution given large enough sample sizes (Kutner et al. 2005) by appeal to the central limit theorem.

### 4.3. Log2 transformation normalizes distribution of heterogeneous measures

The argument for log2 transformation of heterogeneous observations is that genomic measures are typically right skewed and non-normal and that differential expression analysis, at least at the heterogeneous level, should be performed on normal data. Log2 transformed measures take on an approximately normally shaped distribution across samples for any given genomic site. Exceptions are those genes that which are nearly unexpressed across all samples.

### 4.4. Deconvolution of log2 transformed data results in underestimates of cell expressions

Performing linear regression on log2 transformed data has been criticized for breaking the linearity relationship between outcome measures and predictors (Zhong and Liu 2012). As (Zhong and Liu 2012) demonstrate, the relationship between log2 transformed heterogeneous outcomes and coefficients estimates derived from a regression on those outcomes is such that the cell type-specific expression estimates ( $\hat{\beta}_{kn}$ ) are no longer on the same scale as the original "raw" data prior to normalization as shown in equation (3):

$$\log_2\left(y_{mn}\right)=\sum\nolimits_{k=1}^{p}\hat{\beta}_{kn}x_{kn}+\varepsilon_{mn} \qquad (17).$$

In order to obtain true cell type-specific expression estimates from the coefficient estimates ( $\hat{\beta}_{kn}$ ) in equation (4), a back transformation would be required. Zhong and Liu demonstrate that back transforming the coefficients alone will result in underestimates of cell type-specific expression, since, in matrix notation, we have that:

$$\log_2\left(\mathbf{Y}\right)=\log_2\left(\mathbf{X}\hat{\boldsymbol{\beta}}\right)>\mathbf{X}\bullet\log_2(\hat{\boldsymbol{\beta}}) \qquad (18),$$

where $\mathbf{X}$ is the $\boldsymbol{M}$ by $\boldsymbol{P}$ cell proportions matrix of predictors, $\hat{\boldsymbol{\beta}}$ is the $\boldsymbol{P}$ by $\boldsymbol{J}$ matrix of cell-type specific expression estimates, and $\mathbf{Y}$ is the $\boldsymbol{M}$ by $\boldsymbol{J}$ matrix of heterogeneous measures.

Zhong and Liu have shown in (5) that it is strictly mathematically true that log2 transformation of heterogeneous measures prior to linear regression deconvolution will always result in an underestimate of the precise cell type-specific group-wise expressions.

### 4.5. Simulation results indicate LRCDE analysis of log2 measures is more sensitive than analysis of raw measures

We performed AUROC analysis and sensitivity comparison between deconvolution of log2 transformed vs. raw ("de-logged") heterogeneous observations. The result of AUROC analysis are shown in **Figure 2.1** Panel F. This plot is based on calculated AUROC based on permutation based p-values. All parameters were held fixed while target cell proportion standard deviation was iterated over a range of values. The figure shows that analysis of Log2 (normal) data results in consistently greater sensitivity and specificity to a significant degree (see confidence "whiskers" on bar plot).

Comparing analysis of log2 versus "de-logged" heterogeneous measures using LRCDE two-sample t-test resulted in Figure 3.2 Panel D. Analysis of Log2 measures produced higher sensitivity across the range of significance thresholds.

### 4.6. Discussion and Conclusions

The results of AUROC analysis vs. true positive rate detection indicate a modest improvement in sensitivity versus specificity when comparing analysis of log2 (normal) heterogeneous measures against those that have been "de-logged" (exponentiated). However; despite log2 transformation, real biological data will never prove precisely normal (Box 1976). It is the residuals that linear regression presumes follow a normal distribution. Even so; appealing to the central limit theorem will allow for departures from strict normality of heterogeneous measures given large enough samples. The Gauss-Markov theorem provides that outcome measures (heterogeneous observations) need not be normal in order for linear regression to target the true coefficient values in an unbiased way, others have demonstrated (Zhong and Liu 2012) the way in which log2 transforming dependent (heterogeneous) measures prior to regression decouple the coefficient estimates from a linear relationship with the “raw” untransformed measures by placing them on a different scale than the original raw heterogeneous measures.

Our findings indicate that for the purposes of differential expression detection, this may be irrelevant. We have observed through simulation studies that cell type-specific differential expression detection is consistently more precise and more specific (higher AUROC) when performed on normally distributed data than on the same raw data that has been “de-logged” (Figure 2.1.F).

The fact is that cell type-specific differential expression detection performed on log2 data will result in a downward bias on both group-wise estimates. If we then assume that the variances around the group-wise expression estimates are not also downward biased, then any analysis of differential expression between the two groups will tend to be conservative since the differential expression estimates will also be downward biased. This scenario can only result in decreased false discovery, but may fail to detect small but real differences at the cell type-specific level.

One caveat to our argument is that this simulation models an extreme case. Simulating log2 transformed data with a random normal process results in simulated residuals which are nearly perfectly normal. De-logging these simulated normal data (2^data) produces an almost perfectly exponential distribution. Furthermore, we are simulating data using sample sizes which are not large enough to overcome such extreme departures from normality despite the central limit theorem. Real biological data across samples may not follow this exponential distribution and, although not normally distributed, may be less extremely departed from normality than a strict exponential distribution.  In such a situation, the difference in sensitivity between analysis of log2 measures versus “raw” measure may be negligible.

# 5. Addressing collinearity of Cell Proportion Predictor Design Matrix

## 5.1. Biological constraints force cell proportions to sum to 1 across all samples creates linear dependence

In chapter 2 we explored parameters known to affect the estimated variability of linear regression coefficient estimates (cell type-specific expression estimates). In this chapter we closely examine a source of variability not captured in eq. 9 with two specific aims being 1) to demonstrate the instability of differential expression detection resulting from ill-conditioning as quantified by the condition number of the cell proportions design matrix, and 2) to explore and quantify the effects of methods of reducing the effects of ill-conditioning. One way of quantifying the "near singularity" (ill-conditioning) of the design matrix, is to calculate the condition number.

By model specification, cell proportions are assumed to sum to 1 across each sample taken in order to model the biologically relevant situation in which a whole sample is composed of fractions of constituent cell types. This creates perfect linear dependence between any single cell type and the remaining cell types as indicated by infinite variance inflation factors (VIFs).

VIFs are considered one quantification of the degree to which any predictors in a model are related to each other (Kutner et al. 2005). VIFs are calculated by regressing any given predictor on the remaining predictors and taking the coefficient of determination from the regression ( $R^2$ ):

$$VIF_k = 1/(1-R_k^2) \qquad (19),$$

for the *k*-th cell type. Turning again to the simple linear regression formula for the estimated variance of coefficient estimate in order to illustrate the way in which VIF is applied, when the VIF is combined with eq. 7 we have:

$$\operatorname{var}\hat{\beta}_1 = \frac{\sum_{i=1}^{m}(\mathrm{y}_i - \hat{\mathrm{y}}_i)^2 / (\mathrm{M}-\mathrm{P})}{\sum_{i=1}^{m}(\mathrm{x}_i - \overline{x})^2} \bullet VIF \qquad (20).$$

In eq. 20 we see the idea that the estimated variance of coefficient estimates are directly proportional to the VIF. The greater the VIF, the greater the estimated variance. However; given the constraints of the linear regression deconvolution model, in which cell proportions are deliberately scaled in order to sum perfectly to 1, each VIF will be "infinite" since coefficients of determination (eq. 19) from regressions will each equal 1. For this reason, VIFs are meaningless as predictors of multicollinearity for the purposes of the linear regression deconvolution model when using cell proportions as predictors.

Despite these non-informative VIFs, we may have a situation in which the design matrix remains invertible. However, although the design matrix (dot product of the cell proportions matrix) may remain invertible, it may also be close to being less than full rank, or "near-singular".

### 5.2. Ill-conditioning is a measure of multicollinearity

The condition number is a single number quantification of the degree of overall multicollinearity built into a matrix (use "kappa" function in R to obtain condition number). A matrix with very high condition number is said to be "ill-conditioned", as opposed to a matrix with low condition number closer to 1, which is said to be "well-conditioned". When performing linear regression for the purpose of estimating cell proportions from genetic profiles of purified cell lines, there is a high correlation between goodness of fit of the regression line and the conditioning of the predictor matrix (Abbas et al. 2009). In the case of estimating cell proportions from linear regression, researchers have to ability to pick and choose from the tens of thousands of genomic features in order to engineer a cell signature matrix which is well-conditioned (Abbas et al. 2009). In the case of using linear regression for the purpose of estimating cell type-specific expression per group when using measured cell proportions, the purpose is to retain all of the cell's proportions (predictors) in order to determine in which cell and for which features differential expression exists between study groups. Simply discarding entire cells as predictors in order to engineer a well-conditioned predictor matrix is not as straightforward as picking and choosing genomic features from cell signatures.

The condition number (CN) for a matrix is calculated as:

$$CN = \left| \frac{\max(eigenvalue(\mathbf{X'X}))}{\min(eigenvalue(\mathbf{X'X}))} \right| \qquad (7),$$

where x is the *M* (samples) by *P* (number of cell types) matrix of predictor values, and $\mathbf{X'X}$ is the cross-product of the predictor matrix with itself. In the R statistical platform, the 'kappa' function returns the CN when given the argument $\mathbf{X'X}$.

In the case of an invertible cross-product which has a least but non-zero eigenvalue that is small enough to result in an extremely high CN, then we have a "near singular" matrix. The "rule of thumb" frequently mentioned is that CNs greater than 10 constitute very high multicollinearity and are thus near-singular.

The result of having a predictor matrix with a very high CN is that small changes (perturbations) to the values of predictors will result in unpredictable fluctuations in coefficient (cell type-specific expression) estimates. According to theory, given a target cell type in one matrix with one set of numbers producing a standard deviation across samples and a different matrix with a different set of numbers for the same theoretical cell type producing the same standard deviation across samples may result in drastically different set of coefficient estimates. This is a form of coefficient estimate variability which is not captured in eq. 9, or eq. 20 for that matter.

### 5.3. Simulation study to highlight the effect of ill-conditioned predictor matrix

We calculated power for each of 100 iterations using simulation described in 2.3 and 2.4 with 500 changed "genes" and 500 unchanged "genes" in target cell 1 of the cases groups. On each iteration the random number seed in R was allowed to float between iterations. Group sample sizes were fixed at 10. Target cell proportion standard deviation was fixed at 0.2. Effect size was fixed at 0.1. The cell proportions matrix was re-created on each iteration with a consistent cell standard deviation on the target cell and approximately the same condition number for each iteration. The MSE was targeted to be the same for each "gene" in the heterogeneous data without having precisely identical residuals for each "gene". In this way, we were able to observe calculated power for 100 iterations over "perturbed" cell proportions.

This process of creating 100 iterations was repeated for cell proportion condition numbers (kappas) of 100, 200, 500, 1000, 2000, 3000, 4000, 5000, 10000, 50000, and 75000.

Figure 5.1 shows boxplots of calculated power for each of 100 iterations over each of these listed condition numbers. The range of power when condition number is 100 is from 0.98 to 1.0. The loss of stability of calculated power becomes clear from a condition number of 2000 to 3000. Beyond a condition number of 5000, the erratic behavior of observed power has grown to the point that the range is from 0.38 to 1.0.

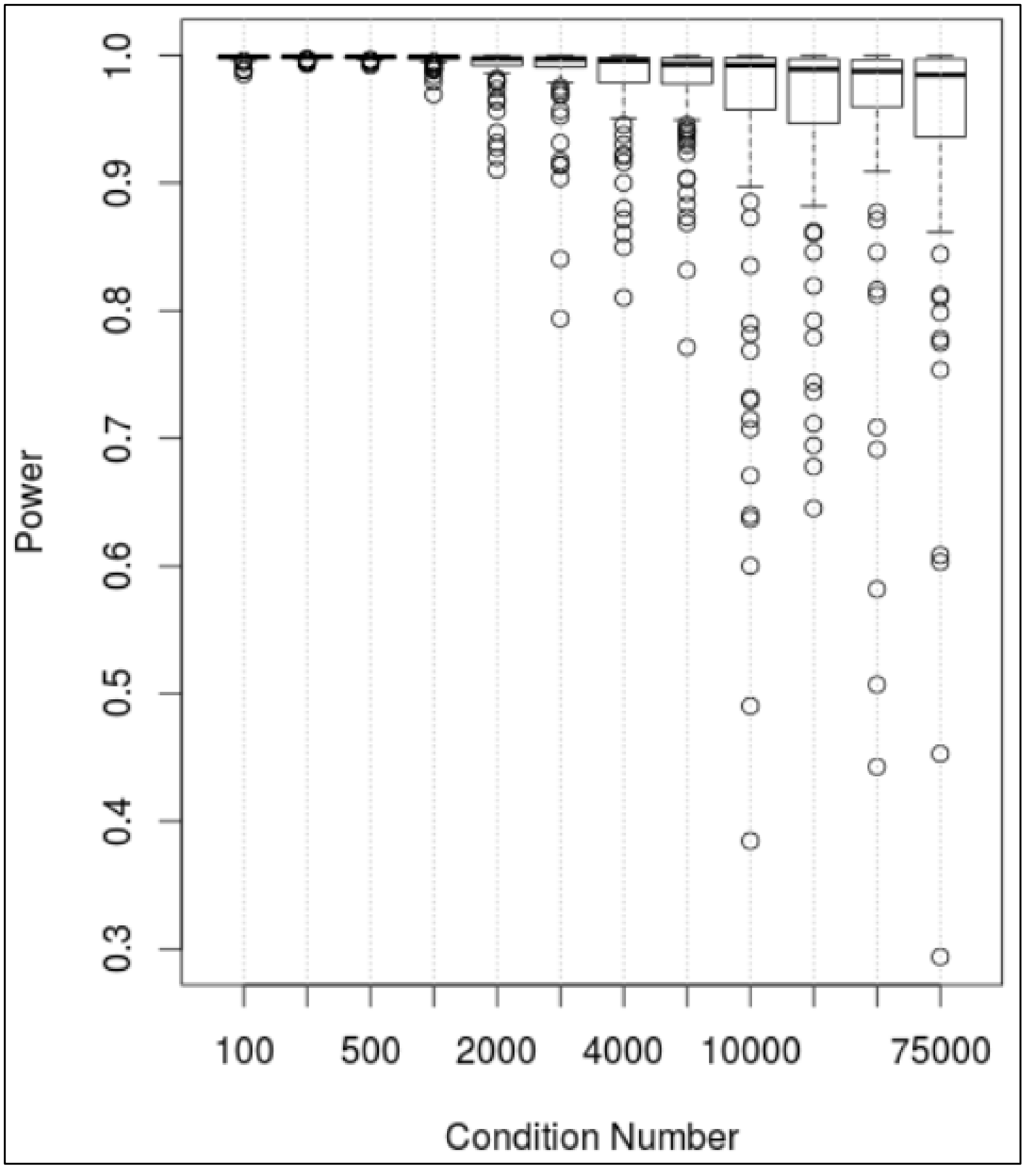


**Figure 5.1: Power over condition numbers of cell predictor design matrix**

Condition numbers observed in the kidney transplant blood data (Gaujoux and Seoighe 2013; Shen-Orr et al. 2010) on the cell proportions matrix cross-product were greater than 57,000 for stable controls and a condition number of greater than 85,000 for transplant cases.

### 5.4. Dropping cell proportion predictors in order to address ill-conditioning

Following the same simulation technique outlined in 5.2, we tested dropping cell proportion predictors from the predictor matrix as a means of addressing ill-conditioning of the design matrix.  Simulating for 5 cell types, we conducted several tests in order to compare which produced a design matrix resulting in the least volatility while retaining high sensitivity to differential expression detection across 100 iterations of identical parameters.

We tested dropping a single cell type and calculating the determinant of the design matrix given the 4 remaining cell types.  On each of the 100 iterations, we dropped the cell type which resulted in a design matrix with the least determinant.

We then tested dropping a single cell type which resulted in the least condition number of the remaining 4 cell types.

We also tested dropping the cell type with the lowest standard deviation across samples.  Finally, we tested dropping the cell type with the least mean proportion across samples.

The results of each 100 iteration tests of dropping a single cell type's proportions can be observed in Figure 5.2.  Dropping the cell predictor with the least mean proportion across samples resulted in the best observed stabilization of power calculations across 100 iterations of simulated cell proportions with identical parameters.  The one observed outlier with the least power out of 100 occurred when dropping the cell predictor with the least mean proportions resulted in a modest increase in the condition number from the original matrix containing all 5 cell predictors.

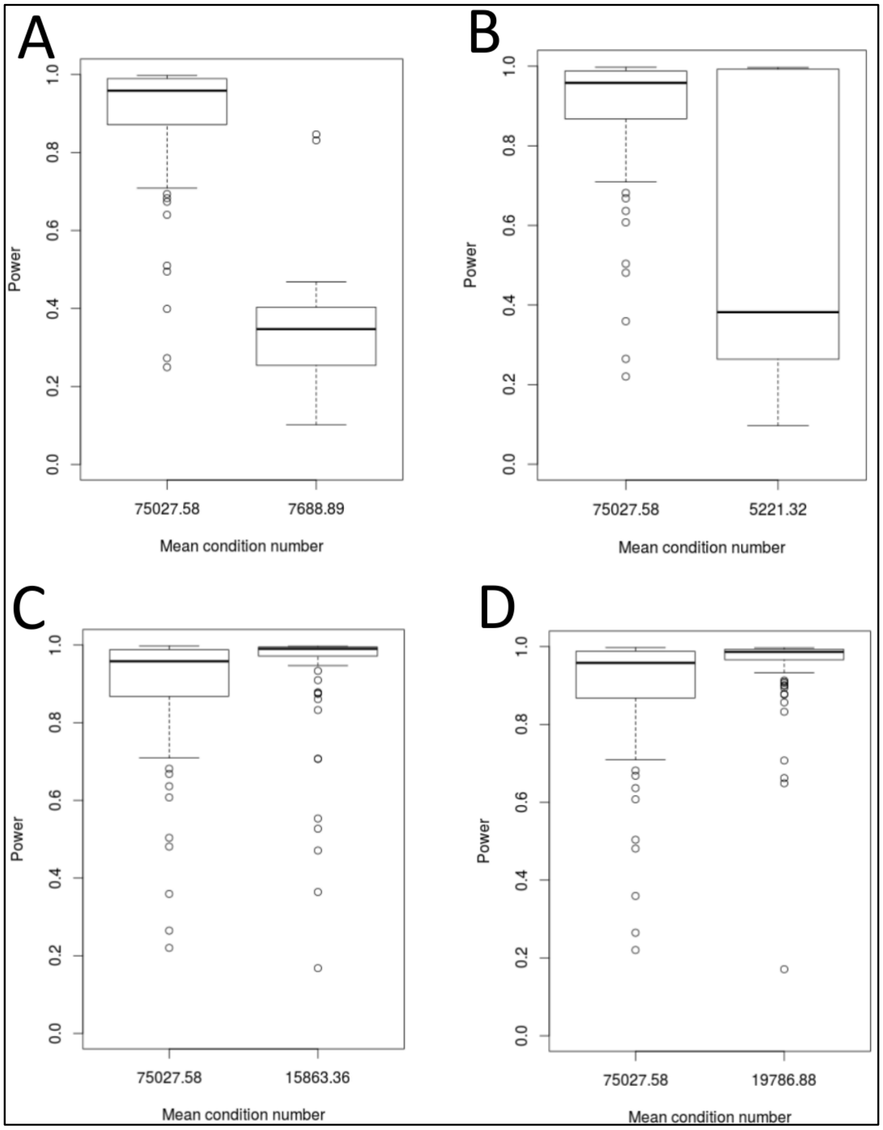


**Figure 5.2: Effect of dropping cell proportion predictors on stability of calculated power**

Panel A) Dropped cell predictor resulting in lowest determinant across iterations.

Panel B) Dropped cell predictor resulting in least condition number across iterations.

Panel C) Dropped cell predictor with least standard deviation across samples.

Panel D) Dropped cell predictor with least mean proportion across samples.

### 5.5. Single regression deconvolution in an attempt to address ill-conditioning

Adding a group membership predictor and group by cell type proportions interaction term variables allows deconvolution and cell type-specific differential expression detection to take place in a single regression (single-step regression) for both groups rather than two separate regression ("dual regression - one per group). In this way, the standard error of the coefficient estimate for target cell proportions and group membership is the surrogate value for differential expression estimates for the target cell type. This is meant to simplify computations. In fact, the "lm" function in the R statistical platform calculates t-statistics for these interaction coefficients as well as associated p-values without additional "by hand" calculations.

Comparing calculated two-sample t-statistic p-values from dual regression to the t-statistic p-values computed by R for single-step regression resulted in p-values plotted in Figure 5.3.

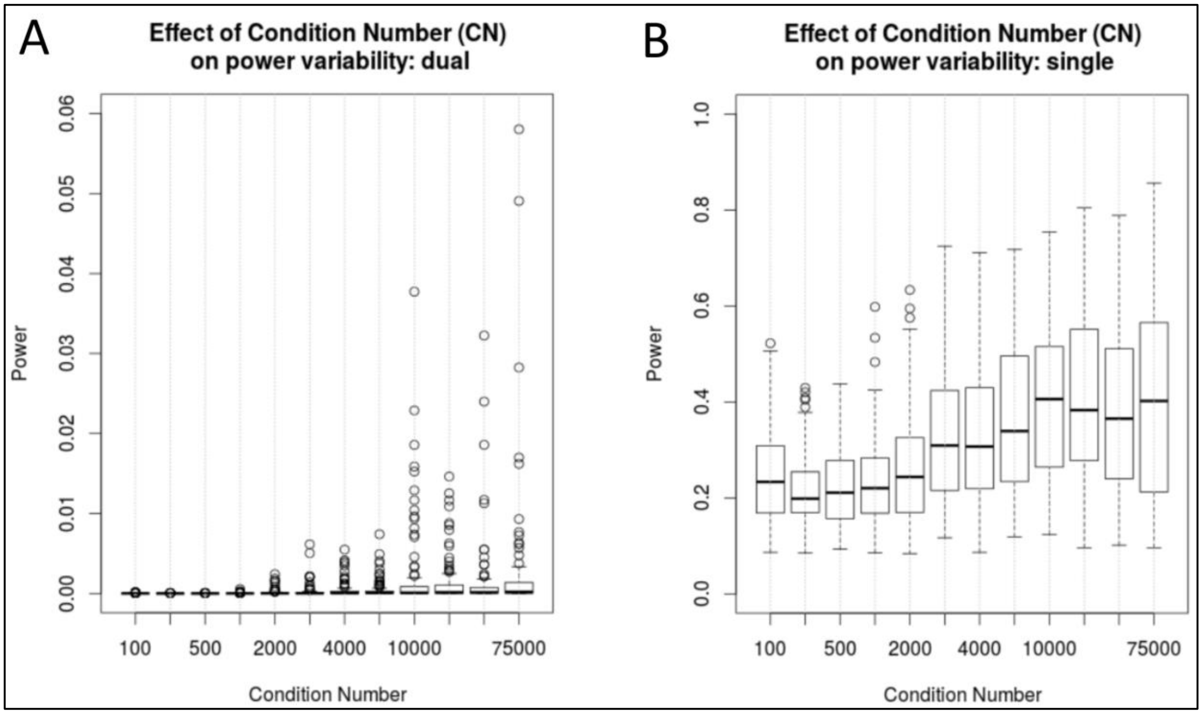


**Figure 5.3: T-statistic p-values for dual vs. single regression deconvolution**

Panel A) One regression per study group. Panel B) One regression step for both study groups.

### 5.6. Conclusions

(Abbas et al. 2009) pointed out that the condition number is "*a high-fidelity marker for the ability of a basis matrix to accurately deconvolve a mixture*" when estimating cell proportions when given heterogeneous observations and a cell signature matrix. In the Abbas paper they refer to the condition number of the cell signatures predictor matrix. When predicting cell proportions, (Abbas et al. 2009) specifically selected genomic features which both uniquely characterized specific types of cells and also resulted in a well-conditioned (low condition number) predictor matrix.

What our findings here illustrate is that the condition number is just as important when going in the opposite direction and predicting cell type-specific expressions from a predictor matrix of measured cell proportions.

Ill-conditioning of the cell proportion predictor design matrix, as quantified by high condition number, creates instability and thus unreliability of linear regression coefficient estimates which translates into unreliable two-sample t-statistics associated with effect size estimates. Linear regression coefficient estimates are taken as cell type-specific expression estimates. Given the demonstrated instability of power measurements tied to the conditioning of the cell proportions design matrix, it is clear that ill-conditioning is reason for caution when interpreting any measure of significant differential expression detection.

Based on preliminary results shown in 5.3, it appears that dropping a single cell proportion predictor with the least mean proportions across samples so as to retain the bulk of the original sample will result in an acceptable level of stability of power calculations. The caveat to this approach is that one must remain vigilant to test the condition number after dropping the cell with the least mean proportions. It may be necessary to select the 2nd least mean cell proportions predictor to drop for the case in which the condition number is not appreciably reduced by dropping the least mean proportion predictor.

Combining cell type-specific estimation and differential expression detection into a single regression step, vs. one regression per group, fails to decrease volatility and does not improve sensitivity as witnessed by the reduced significance level of t-statistic p-values for simulated known differentially expressed genes.  This may be due to the reduced parsimony of the model since combining these steps into a single regression more than doubles the number of parameters being estimated.

# 6. Discussion

## 6.1. Novel contributions of this research

### 6.1.1. Power of LRCDE will depend upon combinations of parameters

Cell type-specific differential expression power will require either lower MSE, higher target cell proportion variability across samples, or both given smaller sample sizes. There is no strict lower bound on these three parameters as changes in any one will affect values of the others depending upon the effect size one wishes to call as significant. For practical purposes, there may be a lower bound on either MSE, or cell proportion variability beyond which larger samples sizes may provide little meaningful gain.

### 6.1.2. Greater sensitivity of two-sample t-test vs. FDR

We have demonstrated that although the dual regression approach reporting false discovery rates (FDR) seems to target known differential expression, it appears uniformly limited in sensitivity when compared against p-values derived from Welch's two-sample t-test. Furthermore; small sample sizes harm FDR computation by lacking enough data points to adequately quantify the null-distribution against which the observed differential expression is compared. Although known differential expression may be detected and signaled by some FDR less than 1.0, a higher FDR in such instance may give the impression that the difference is not significant. Furthermore; FDR fails to distinguish between varying effect sizes given smaller samples.

Operating on simulated data in which the cell type-specific differences are engineered and known *a priori* and reported with FDR > 0.3, two-sample t-statistic p-values are more sensitive to these same differences and show as significant under several test conditions: low sample sizes, higher MSE, and lower cell proportion standard deviation across samples.

### 6.1.3. Ill-conditioning of cell proportion based design matrix results in instability of cell type specific expression estimates from linear regression

The reliability of any measure of significance of detected cell type-specific differential expression will be affected by extremely high conditioning (condition numbers greater than ~2,000) of the group-wise cell proportion predictor matrix. The variability introduced by ill-conditioning of the design matrix introduces a level of variability around cell type-specific expression estimates which is not captured in the coefficient variance calculation (eq. 9) and must be taken into consideration when making biological inferences. Our findings indicate that this is a concern which cannot be safely ignored, particularly when estimating cell type-specific expression levels from measured cell proportions.

When using linear regression in order to estimate cell proportions, it is straightforward to pick and choose feature sites which result in a well-conditioned design matrix (Houseman et al. 2012).

### 6.1.4. Addressing ill-conditioning to reduce instability of LRCDE

When estimating cell type-specific expression levels from cell proportions, our preliminary findings indicate that dropping a single cell proportion predictor with least mean proportions across samples appears to have the greatest effect upon stabilizing estimates of significance across multiple perturbations of the cell proportions matrix. The caveat to this observation is that the choice of dropping a particular cell proportion predictor requires attention to the resulting design matrix condition number. Dropping the least mean cell proportions predictor does not guarantee an improvement in conditioning of the design matrix. Attention must be paid to both: 1) retaining the bulk of sample proportions across all samples, and 2) minimizing the resulting condition number of the design matrix. The linear regression deconvolution model faithfully targets known cell type-specific differential expressions. Thus: further investigation into methods of addressing ill-conditioning resulting from high multicollinearity in this model are warranted.

A method of addressing the ill-conditioning of cell proportions predictor matrices without harming LRCDE power is needed in order to reduce the instability of any resulting measure of detected differential expression significance.

### 6.2. Drawbacks of power analysis dual regression approach to deconvolution

The issue of back-transformation remains when performing LRCDE analysis of log2 transformed heterogeneous measures. We have demonstrated that higher AUROC is attained when performing LRCDE analysis on normally distributed heterogeneous measures versus measures that have been “de-logged”. Other authors have pointed out that performing linear regression deconvolution on log2 transformed measures will result in consistent downward bias of cell type-specific expression estimates. We argue that this will additionally downward bias the observed effect size (size of cell type-specific differential expression) and therefore observed power will be conservatively under estimated. Even though LRCDE analysis of log2 transformed measures results in higher AUROC for true cell type-specific differences, this downward bias must be taken into account if one designs to estimate actual fold change or overall true effect size. However; this observation was made under simulated conditions in which the raw (“de-logged”) measures take on a nearly exponential distribution which may not reflect biological data and perhaps exaggerates the difference between real world measures and the log2 transformation of those same measures.

### 6.3. Future Directions

A generalized linear model approach is the next logical step in the development of the regression approach to performing deconvolution.  Although we have shown that increased sensitivity is achieved by analyzing log2 heterogeneous measures versus raw data, it would represent an improvement to apply a statistical model which more accurately models the means of non-transformed heterogeneous measures as regression outcomes.

By finding an appropriate link function which adequately models the distribution of raw measures across samples, then non-log2 transformed “raw” observations could be analyzed potentially providing unbiased cell type-specific expression estimates which in turn could provide unbiased differential expression estimates.

# A. APPENDICES

## A.1 Analysis of Heterogeneous observations versus cell type-specific

There are two theoretical scenarios which exist when considering differential expression detection at the heterogeneous level versus using linear regression as a means to detecting cell type-specific differential expression.

In the a situation where there exists differential expression of a genomic site in a single cell type, that difference is theoretically detectable at the heterogeneous level when relying upon a simple t-test.

The situation in which cell type-specific expression differences between study groups is undetectable at the heterogeneous level may occur when a same genomic site is up-regulated in one cell type while equally down-regulated in another cell type Table A.1.

Table A.1: Theoretical example of indistinguishable differences at heterogeneous level

Given a 2 fold up-regulation of cell 2, then there must be some amount of down-regulation of the same gene in at least one or more other cell types to mask the change at the heterogeneous level.

|  | Heterogeneous Expression gene J |  | Cell 1 Prop |  | cell expression |  | Cell Prop |  | cell expression |  | Cell Prop |  | cell expression |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| controls | 3 | = | 0.2 | x | 2.5 | + | 0.3 | x | 3 | + | 0.4 | x | 4 |
|  | 3 | = | 0.2 | x | 2.5 | + | 0.3 | x | 3 | + | 0.4 | x | 4 |
|  | 3 | = | 0.2 | x | 2.5 | + | 0.3 | x | 3 | + | 0.4 | x | 4 |
|  | 3 | = | 0.2 | x | 2.5 | + | 0.3 | x | 3 | + | 0.4 | x | 4 |
|  | 3 | = | 0.2 | x | 2.5 | + | 0.3 | x | 3 | + | 0.4 | x | 4 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| cases | 3 | = | 0.2 | x | 0.1 | + | 0.3 | x | 6 | + | 0.4 | x | 2.95 |
|  | 3 | = | 0.2 | x | 0.1 | + | 0.3 | x | 6 | + | 0.4 | x | 2.95 |
|  | 3 | = | 0.2 | x | 0.1 | + | 0.3 | x | 6 | + | 0.4 | x | 2.95 |
|  | 3 | = | 0.2 | x | 0.1 | + | 0.3 | x | 6 | + | 0.4 | x | 2.95 |
|  | 3 | = | 0.2 | x | 0.1 | + | 0.3 | x | 6 | + | 0.4 | x | 2.95 |
|  |  |  | Fold: |  | 0.04 |  | Fold: |  | 2 |  | Fold: |  | 0.7375 |
|  |  |  | diff: |  | -2.4 |  | diff: |  | 3 |  | diff: |  | -1.05 |

# B. APPENDICES

### B.1 How to use LRCDE package

```
# Code to get started using LRCDE
library(lrcde) # Load the lrcde package
# setwd("/home/your.user.name/output.directory") # change this to your own setup
# Custom parameters to model:
    n.samps = 15 # Sample size per group
    # Mean Squared Error to model (actual average MSE per gene will be a small fraction of this)
     mse2model.vec  = c( 0.05)
# Cell proportion parameterss to model:
    # Target cell standard deviation (across samples) to model:
    cell.sd.2.model = c( 0.08 )
    # Condition number (kappa) to target (resulting kappa will be approximate):
    kappa.2.model  = c( 71500 )
    # Number of cells to simulate:
    n.cells  = c( 3 )
    # The "target" cell (the one with the fold change) for simulations
    cell.p = 1
# Cell expression params to model:
    # Base level cell expressions to model:

    base.expr.vec    = c( 2  )

    # Cell level differential expression to model:

    diff.2.model.vec = c( 0.01, 0.02, 0.03, 0.04, 0.05, 0.06,
0.07, 0.08, 0.09, .1)






# Sim cell level expression (gold-standard).

cell.expr = custom.sim.cell.expr(   n.cells           # Number
of cell types being simulated

                                  , base.expr.vec     # The
"base" expression level to model

                                  , diff.2.model.vec  #
Differential expression to model

                                  , cell.p            # Target
cell to modify in cases

                                  , length( mse2model.vec ) )


# Sim residuals:

set.seed(seed2set)

resids = custom.resids.synthetic(    mse2model.vec    # Actual
MSE will be small fraction of this
```

```
                                   , groups           # groups
vector
                                   , diff.2.model.vec # Included
to get matrix size correct
                                   , base.expr.vec    # Included
to get matrix size correct
                                   , adjuster=1
                                   , n.cells )     # Scaling
factor for MSE target


# Sim het obs:
het.obs = het.from.synthetic(   cell.props     # The entire cell
proportions matrix
                              , cell.expr      # Cell type-
specific expressions matrix
                              , resids         # Simulated
residuals matrix
                              , groups )       # groups
membership vector


colnames( het.obs ) = 1:dim( het.obs )[2]         # LRCDE
expects to see feature names
colnames( cell.props ) = 1:dim( cell.props )[2]   # LRCDE
expects to see cell type names
```

```
################################################################
    # Use these for LRCDE since power calculation is meaningless
if
    #      differences are transformed but standard errors are
not:
    stdz=FALSE; medCntr=FALSE; nonNeg=TRUE
################################################################
    method2use="dual" # Which type of deconvolution to run (dual
is only thing implemented)
    lrcde.output.file   = paste0( "lrcde_sim_example.csv"  )
    alternative='two.sided' # One of "two.sided", "greater", or
"less"
    # Run LRCDE:
    return.list = lrcde(  het.obs, cell.props, groups
                          , output.file = lrcde.output.file
                          , medCntr     = medCntr
                          , stdz        = stdz
                          , nonNeg      = nonNeg
                          , method      = method2use
                          , direction   = alternative
    )
    result.frame = return.list[[1]]
    result.frame
```

### B.2 Script used in Figure 3.2:

```
options(stringsAsFactors = FALSE)

library( lrcde )

library( csSAM )

rm(list=ls()) # Start with clean environment

n.samps.vec =c(  10, 14, 18  )

n.genes = 1000                # Number of "genes" to simulate.
Half of these will be "folded".

n.perms = 1000                # For csSAM permutations

kappa.2.model   = c( 100  )   # Condition number (kappa) to
model in cell proportions

mse2model   = c( 1.5 )   # Target MSE to model in residuals
(will actually be a fraction of this)

cell.sd.2.model = c( .1 )   # Target cell type proportion
standard deviation

n.cells  = c( 5 )             # Number of cells to simulate:

cell.p = 1                    # The "target" cell (the one with
the fold change) for simulations

base.expr    = c( 2  )        # Base level cell expressions to
model:

diff.2.model = seq( 0.001, 1.0, length.out= 500 ) # Range of
differential expressions to model

counter=0
```

```
for( n.samps in n.samps.vec ) {
  counter = counter + 1 # For indexing tpr.p.vec and tpr.f.vec
  # Simulate data:
  seed2set = (11221963)
  # Group indicator vector (group membership indicator):
  groups = c(rep(1,n.samps), rep(2,n.samps) )
  # Customized cell expression:
  sim.cell.expr.perms = function(  n.cells, base.expr,
diff.2.model, cell.p, n.genes ) {
    cells.cntl = matrix( rep(base.expr, n.cells*n.genes),
nrow=n.cells )
    cells.case = cells.cntl
    haf.genes = n.genes / 2

cells.case[cell.p,1:haf.genes]=cells.case[cell.p,1:haf.genes] +
diff.2.model
    cells.expr = rbind(cells.cntl, cells.case)
    return( cells.expr )
  }

cell.expr=sim.cell.expr.perms(n.cells,base.expr,diff.2.model,cel
l.p,n.genes )
  haf.genes = n.genes / 2
  truth = c(  rep(1, haf.genes), rep(0, haf.genes)    )
  # Only looping for permutations:
  resids = matrix( , ncol=n.genes, nrow= 2*n.samps )
  for( p in 1:n.genes) {
    resids[, p ] = custom.resids.synthetic( mse2model
                                             , groups,
diff.2.model=2, base.expr
                                             # 'adjuster' is a
scaling factor for target MSE
                                             , adjuster=1,
n.cells )
  }
  auroc.frame = data.frame();   power.frame = data.frame()
  # For replicability: set seed before synthesizing cell
proportions
  set.seed( seed2set )
  cell.props.1 = cell.props.target( n.cells
                                     , n.samps
                                     , cell.sd.2.model
                                     , kappa.2.model )



  # Stack control and cases (identical) cell proportions:
  cell.props          = rbind( cell.props.1, cell.props.1 )
  colnames( cell.props ) = 1:dim( cell.props )[2]   # LRCDE
expects to see cell type names

################################################################
###########
  eigen.values = eigen(t(cell.props.1)%*%cell.props.1)$values
  min(eigen.values); max(eigen.values)              # min and
max eigen values
  abs(max(eigen.values)/min(eigen.values))          # Calculated
condition number.
  kappa(t(cell.props.1)%*%cell.props.1, exact=TRUE) # The R
version of condition number.
  apply(cell.props.1, 1, sum)                       # Sum of
proportions per sample
  det(t(cell.props.1)%*%cell.props.1)               # The
determinant

################################################################
###########
  # Sim het obs:
  het.obs = het.from.synthetic(    cell.props
                                   , cell.expr
                                   , resids
                                   , groups )
  colnames( het.obs ) = 1:dim( het.obs )[2]         # LRCDE
expects to see feature names
  colnames( cell.props ) = 1:dim( cell.props )[2]   # LRCDE
expects to see cell type names
  method2use="dual"
  lrcde.output.file   = paste0( "ROC_for_power.csv"  )
  alternative='two.sided' # One of "two.sided", "greater", or
"less"
  # Use these for LRCDE since power calculation is meaningless
if
  #      differences are transformed but standard errors are
not:
  # DO NOT standardize (stdz) or median center (medCntr)
difference estimates !!!
  stdz=FALSE; medCntr=FALSE; nonNeg=TRUE
  return.list = lrcde(  het.obs, cell.props, groups
                        , output.file = lrcde.output.file
                        , medCntr     = medCntr
                        , stdz        = stdz
                        , nonNeg      = nonNeg
                        , method      = method2use
                        , direction   = alternative
  )
  auc.frame = return.list[[1]]
  p.vec = auc.frame$p.val.t[1:n.genes]
  # Run canned csSAM:
  # csSAM appears to produce better (more significant) FDRs when
median centering
  # and standardizing cell type-specific difference estimates.
  stdz=TRUE; medCntr=TRUE; nonNeg=TRUE
  G = het.obs                # Use exact same data as was fed
to LRCDE
  cc = cell.props
  y <- factor(groups)
  numset = nlevels(y)
  n <- summary(y, maxsum=Inf) # number of samples in each class
  numgene = ncol(G)
  numcell = ncol(cc)
  geneID = colnames(G)
  cellID = colnames(cc)
  deconv <- list()
  # run analysis
  set.seed( seed2set )
  for (curset in levels(y))
    deconv[[curset]]= csfit(cc[y==curset,], G[y==curset,])

  rhat <- array(dim = c(numcell,numgene))
  rhat[, ] <- csSAM(deconv[[1]]$ghat, deconv[[1]]$se,
                    n[1], deconv[[2]]$ghat, deconv[[2]]$se,
n[2],
                    standardize=stdz, medianCenter=medCntr,
nonNeg=nonNeg)
  falseDiscovR <- fdrCsSAM( G,cc,y,n,numcell,numgene, rhat,
                            nperms =
n.perms,standardize=stdz,alternative=alternative,
                            medianCenter=medCntr, nonNeg=nonNeg)
  sigGene <- findSigGene( G, cc, y, rhat, falseDiscovR )
  fdr.vec.500 = sigGene[ cell.p, 1:haf.genes ]
  # End csSAM analysis
  text2parse = paste0( "p.vec.", n.samps , " =
p.vec[1:haf.genes]" )
  eval( parse(text=text2parse) );
  fdr.vec.500[is.na(fdr.vec.500)] = 0
  text2parse = paste0( "f.vec.", n.samps, " = fdr.vec.500" )
  eval( parse(text=text2parse) );
} # End loop over parameter
################################################################
##############
# Setup series of TPRs for both based on a series of thresholds:
FDR and t-test
len=20
p.thresh.max=0.1/1000
f.thresh.max=0.3
```

```
p.thresholds = seq( 0, p.thresh.max, length.out = len )

f.thresholds = seq( 0, f.thresh.max, length.out = len

# Loop over thresholds:

for( n.samps in n.samps.vec ) {

  text2parse = paste0( "tpr.p.vec.", n.samps," = vector()" )

  eval( parse(text=text2parse) );

  text2parse = paste0( "tpr.f.vec.", n.samps," = vector()" )

  eval( parse(text=text2parse) );

  for( t  in 1:length( p.thresholds ) ) {

    text2parse = paste0( "p.pos = ifelse( p.vec.", n.samps

                        ,"   <= p.thresholds[ t ] , 1 , 0 )" )

    eval( parse(text=text2parse) );

    text2parse = paste0( "f.pos = ifelse( f.vec.", n.samps

                        ,"   <= f.thresholds[ t ] , 1 , 0 )" )

    eval( parse(text=text2parse) );

    text2parse = paste0( "tpr.p.vec.", n.samps

                        , "[ t ] = sum(   p.pos ) / haf.genes"
)

    eval( parse(text=text2parse) );

    text2parse = paste0( "tpr.f.vec.", n.samps

                        , "[ t ] = sum(   f.pos ) / haf.genes"
)

    eval( parse(text=text2parse) );
  } }
# Double axis plots in a loop # Two axis plots:
par(mar=c(5, 4, 4, 5.5) )
tot.col = length( n.samps.vec )
symbols = c(15,16,17)
lo.x=0
main.title = paste0( "FDR versus p-value over groups sample
size" )
space = 3
tot = length( n.samps.vec )
col.p = c( "blue" )
col.f = c( "red"  )
count = 0
for( n.samps in n.samps.vec ) {
  count=count+1
  text2parse = paste0( "  plot( tpr.f.vec.", n.samps
                        ,", f.thresholds , axes=FALSE,
pch=symbols[ count ], xlim=c(lo.x,1), ylim=c(0,f.thresh.max),
xlab='', ylab='', type='b', col=col.f, main=main.title) " )
  eval( parse(text=text2parse) );
  if(count==1){
    fdr.ax = 2
    axis( fdr.ax , ylim=c( 0, f.thresh.max ), col=col.f ,
col.axis=col.f, las=1)
```

```
    mtext("FDR thresholds"               , col=col.f  ,
side=fdr.ax , line=space-.5 )
  }
    par(new=TRUE)
    text2parse = paste0( "  plot( tpr.p.vec.", n.samps
                       ,", p.thresholds ,axes=FALSE,
pch=symbols[ count ], xlim=c(lo.x,1), ylim=c(0,p.thresh.max),
xlab='', ylab='', type='b',col=col.p) " )
  eval( parse(text=text2parse) );
  if(count==1){
    p.ax=4
    axis( p.ax   , ylim=c( 0, p.thresh.max ), col=col.p, las=1)
## las=1 makes horizontal labels
    mtext("t-statistic p-value threshold", col=col.p , side=p.ax
, line=space+.25 )
  }
    if( count != tot ) par(new=TRUE)
}
TPR=seq( 0, 1, length.out = 20 )
axis( 1, pretty(range( c(lo.x,1) ), 10 ))
mtext("sensitivity", side=1, col="black", line=space)
box()
## Add Legend
legend( "bottomleft", title="sample:", title.col="black"
        , legend=c( paste( n.samps.vec , "P.val"), paste(
n.samps.vec ,"FDR") )

        , text.col = c( rep(col.p, tot.col ), rep(col.f, tot.col
) )

        , col      = c( rep(col.p, tot.col ), rep(col.f, tot.col
) )

        , pch=c( rep( symbols , 2 )  )
```

# BIBLIOGRAPHY

Abbas, Alexander R., Kristen Wolslegel, Dhaya Seshasayee, Zora Modrusan, and Hilary F. Clark. 2009. “Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus” edited by P. Tan. *PLoS ONE* 4(7):e6098.

Anders, S., W. Huber, Anders S, and Huber W. 2010. “Differential Expression Analysis for Sequence Count Data.” *Genome Biol* 11(10):R106. Retrieved (papers2://publication/uuid/7525A374-3511-418E-B510-04C5305E9033).

Ballman, Karla V. 2008. “Genetics and Genomics: Gene Expression Microarrays.” *Circulation* 118(15):1593–97.

Box, George E. P. 1976. “Statistics.” *Journal of the American Statistical Association* 71(356):791–99.

Chikina, M., E. Zaslavsky, and S. C. Sealfon. 2015. “CellCODE: A Robust Latent Variable Approach to Differential Expression Analysis for Heterogeneous Cell Populations.” *Bioinformatics* 31(January):1584–91. Retrieved (http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btv015).

Dozmorov, Mikhail G., Joel M. Guthridge, Robert E. Hurst, and Igor M. Dozmorov. 2010. "A Comprehensive and Universal Method for Assessing the Performance of Differential Gene Expression Analyses." *PloS one* 5(9).

Erkkilä, Timo et al. 2010. "Probabilistic Analysis of Gene Expression Measurements from Heterogeneous Tissues." *Bioinformatics* 26(20):2571–77.

Fawcett, Tom. 2006. "An Introduction to ROC Analysis." *Pattern Recognition Letters* 27(8):861–74.

Gaujoux, Renaud and Cathal Seoighe. 2013. "CellMix: A Comprehensive Toolbox for Gene Expression Deconvolution." *Bioinformatics (Oxford, England)* 29(17):2211–12.

Gong, Ting and Joseph D. Szustakowski. 2013. "DeconRNASeq: A Statistical Framework for Deconvolution of Heterogeneous Tissue Samples Based on mRNA-Seq Data." *Bioinformatics* 29(8):1083–85.

Gosink, Mark M., Howard T. Petrie, and Nicholas F. Tsinoremas. 2007. "Electronically Subtracting Expression Patterns from a Mixed Cell Population." *Bioinformatics* 23(24):3328–34.

Graybill, Franklin. 1969. *Matrices with Applications in Statistics*. 2nd Editio. edited by P. Bickel, W. Cleveland, and R. Dudley. Belmont: Wadsworth.

Hoffmann, Martin et al. 2006. "Robust Computational Reconstitution - a New Method for the Comparative Analysis of Gene Expression in Tissues and Isolated Cell Fractions." *BMC bioinformatics* 7:369.

Houseman, Eugene Andres et al. 2012. "DNA Methylation Arrays as Surrogate Measures of Cell

Mixture Distribution.” *BMC bioinformatics* 13(1):86. Retrieved (http://www.biomedcentral.com/1471-2105/13/86).

Jaffe, Andrew E. and Rafael a Irizarry. 2014. “Accounting for Cellular Heterogeneity Is Critical in Epigenome-Wide Association Studies.” *Genome biology* 15(2):R31. Retrieved March 24, 2014 (http://www.ncbi.nlm.nih.gov/pubmed/24495553).

Kutner, Nachtsheim, Neter, and Li. 2005. *Applied Linear Statistical Models*. Fifth Edit. edited by B. Gordon and L. Stone. New York: McGraw-Hill/Irwin.

Lähdesmäki, Harri, Llya Shmulevich, Valerie Dunmire, Olli Yli-Harja, and Wei Zhang. 2005. “In Silico Microdissection of Microarray Data from Heterogeneous Cell Populations.” *BMC bioinformatics* 6:54. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/15766384).

Lander ES, Linton LM. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Articulo* 411(6838). Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/11237011).

Liebner, David A., Kun Huang, and Jeffrey D. Parvin. 2014. “MMAD: Microarray Microdissection with Analysis of Differences Is a Computational Tool for Deconvoluting Cell Type-Specific Contributions from Tissue Samples.” *Bioinformatics* 30(5):682–89.

McClish, D. K. 2015. “Analyzing a Portion of the ROC Curve.” *Medical decision making : an international journal of the Society for Medical Decision Making* 9(3):190–95.

Montaño, Carolina M. et al. 2013. “Measuring Cell-Type Specific Differential Methylation in Human Brain Tissue.” *Genome biology* 14(8):R94. Retrieved November 25, 2014 (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4054676&tool=pmcentrez&rendertype=abstract).

Newman, Aaron M. et al. 2015. “Robust Enumeration of Cell Subsets from Tissue Expression Profiles.” *Nature Methods* 12(MAY 2014):1–10.

Newton, Isaac and Gordon Moore. 2014. “The $1,000 Genome.” *Nature* 507:294–95.

Robin, Xavier et al. 2011. “pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves.” *BMC bioinformatics* 12(1):77. Retrieved (http://www.biomedcentral.com/1471-2105/12/77).

RStudio. 2015. “RStudio: Integrated Development for R. RStudio, Inc., Boston, MA.” http://www.rstudio.com. Retrieved (http://www.rstudio.com).

Shen-Orr, Shai S. et al. 2010. “Cell Type-Specific Gene Expression Differences in Complex Tissues.” *Nature methods* 7(4):287–89.

Smyth, Gordon K. and Belinda Phipson. 2010. “Permutation P-Values Should Never Be Zero: Calculating Exact P-Values When Permutations Are Randomly Drawn.” *Statistical applications in genetics and molecular biology* 9(1):Article39.

Stuart, R. O. et al. 2004. “In Silico Dissection of Cell-Type-Associated Patterns of Gene Expression in Prostate Cancer.” *Proceedings of the National Academy of Sciences* 101(2):615–20. Retrieved (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=327196&tool=pmcentrez&rendertype=abstract).

Tarazona, Sonia. 2011. “Differential Expression in RNA-Seq (NOISeq Users Guide).” *Gene Expression* 21(March):2213–23. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/21903743).

Venet, D., F. Pecasse, C. Maenhaut, and H. Bersini. 2001. “Separation of Samples into Their Constituents Using Gene Expression Data.” *Bioinformatics (Oxford, England)* 17 Suppl 1:S279–87.

Welch, B. L. 1947. “The Generalization of "Student’s' Problem When Several Different Population Variances Are Involved.” 34(1/2):28–35.

Zhong, Yi and Zhandong Liu. 2012. “Gene Expression Deconvolution in Linear Space.” *Nature methods* 9(1):8–9; author reply 9.

# VITA

Edmund Glass is a U.S. Citizen who grew up in Hampton, Virginia working in his parents Civil Engineering & Land Surveying business. Ed graduated from Phoebus High School where he was able to interact with a main-frame computer though a command-line interface on tractor-feed paper via a handset-cradle modem. Attending Thomas Nelson Community College, Ed graduated with an Associate's Degree in Fine Art. Ed earned his Bachelor's in Computer Science from Virginia Commonwealth University in 2001. Returning to VCU in 2011, he joined the VCU Biostatistics program.

Ed maintains interests in physical fitness, martial arts, computer security, and teaching.