2024

# Monte-Carlo method for identifying aberrantly expressed genes in cancer

Matthew M. Beltran
*Virginia Commonwealth University*

## Copyright Page

# Monte-Carlo method for identifying aberrantly expressed genes in cancer

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

By
Matthew Beltran
B.S. in Physics, Virginia Commonwealth University, 2019
Virginia Commonwealth University

Advisor
Richard Inho Joh, PhD
Assistant Professor, Department of Physics
School of Humanities and Sciences

Committee Members
LaMont Cannon, PhD
Assistant Professor, Center for Biological Data Science

Jason Reed, PhD,
Professor, Department of Physics

Joseph Reiner, PhD
Professor and Department Chair, Department of Physics

Virginia Commonwealth University
Richmond, Virginia
May 2024

# Acknowledgments

This work would not have been possible without the constant support and motivation of my friends and family. I owe a lot of gratitude to my advisor, Dr. Richard Joh, who offered inspiration, guidance, and understanding in all aspects of my work. I cannot describe how often when facing obstacles in the different stages of my work our discussions gave me renewed perspective to be able to move forward.

I want to acknowledge Dr. Joseph Reiner, Dr. Jason Reed, and Dr. LaMont Cannon for serving on my committee.

I also want to thank my fellow graduate student Puranjan Ghimire. I am so glad to have gone through this entire experience with a friend. Research has the tendency to be very isolating and I am truly not sure how I would have managed to push through the uncertainty and self-doubt without being able to discuss these things openly with someone who understands the same experience.

I have to thank my family who gave me the curiosity from a young age to pursue an interest in science. I am very grateful to my parents and grandparents for their continued moral and financial support, without which it would not have been possible to pursue higher education.

Lastly, I want to thank my fiancé, Nadia, who probably has had the closest perspective on the ups and downs that came during my PhD journey. Our dreams gave me a reason to give my best effort day after day.

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| AEG | Aberrantly Expressed Gene |
| BP | Biological Process |
| cDNA | complementary DeoxyriboNucleic Acid |
| CCLE | Cancer Cell-Line Encyclopedia |
| CDK | Cyclin-Dependent Kinase |
| CE | Contour Effectiveness |
| CER | Characteristic Expression Region |
| ChIP | Chromatin ImmunoPrecipitation |
| CPDF | Cumulative Probability Density Function |
| OS | Outlier Score |
| DAVID | Database for Annotation, Visualization, and Integrated Discovery |
| DBSCAN | Density BaSed Clustering ANalysis |
| DE | Differentially Expressed |
| DEG | Differentially Expressed Gene |
| DNA | DeoxyriboNucleic Acid |
| EMT | Epithelial to Mesenchymal Transition |
| exAEGs | exclusively Aberrantly Expressed Genes |
| exDEGs | exclusively Differentially Expressed Genes |
| FDR | False Discovery Rate |
| FACS | Fluorescence Activated Cell Sorting |
| FC | Fold-Change |
| FE | Fold Enrichment |
| GEO | Gene Expression Omnibus |
| GMM | Gaussian Mixture Model |

| | |
|---|---|
| GO | Gene Ontology |
| GRN | Gene Regulatory Network |
| GTEx | Genotype-Tissue Expression project |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| limma | linear models and differential expression for microarray data |
| MAGE | Monte-carlo method for Aberrant Gene Expression |
| MC | Monte-Carlo |
| miRNA | micro RiboNucleic Acid |
| ML | Machine Learning |
| MRG | Master Regulator Gene |
| mRNA | messenger RiboNucleic Acid |
| mTOR | mammalian Target Of Rapamycin |
| NB | Negative-Binomial |
| NCBI | National Center for Biological Information |
| NGS | Next-Generation Sequencing |
| OS | Outlier Score |
| PDF | Probability Density Function |
| PCA | Principal Components Analysis |
| RNA | RiboNucleic Acid |
| SC | Single-Cell |
| SD | Standard Deviation |
| SVM | Support Vector Machine |
| TAD | Topologically Associated Domains |
| TCGA | The Cancer Genome Atlas |
| TF | Transcription Factor |
| TPM | Transcripts Per Million |

# Abstract

Gene expression provides insight into the functional variations on the cellular level that shape biological phenomena. Several recent sequencing technologies have produced an abundance of expression profiles spurring entirely new disciplines of biological study. With this myriad of data, the new task is deciding how to assess and extract meaningful insights. Identifying genes with expression changes in disease conditions is often the first step in finding potential biomarkers for diagnosis, and targets for pharmaceutical treatments. Parametric statistical tests at the individual gene level have been the conventional approach for finding differentially expressed genes. These tests exhibit high statistical power but rely on distributional assumptions that are difficult to validate. Which has led to a vast number of selected genes, with very few being effective in clinical applications. Alternatively, machine learning algorithms have been developed to identify patterns in high-dimensional data that can be easily applied to gene expression analysis.

Here we present a novel algorithm for identifying aberrantly expressed genes in cancer. By comparing the expression pattern of individual genes to the cumulative pattern of the whole profile, we have developed a robust classification tool. We provide evidence that aberrant expression is effective in reporting biologically relevant gene signatures that may be overlooked by traditional methods. Due to the general assumptions used in our approach, we demonstrate its ability to assess gene expression from multiple technologies (microarray, RNA-Seq, scRNA-Seq) and for multiple insights (disease associations, treatment associations, cell/tissue variability). Lastly, we apply our method to single-cell RNA profiles, where robust identification of AEGs is possible with fewer samples than the conventional approaches. We hope these results inspire further research into developing a generalized framework for assessing gene expression patterns that can lead to the improvement of clinical outcomes and the development of personalized medicine.

# 1. Introduction

Gene expression provides a view into the functional role of genes and their importance in regulating cellular processes. Recent advances in high-throughput sequencing technology have provided an unprecedented ability to study physiological responses to stimuli on the cellular level. Next-generation sequencing (NGS) technologies (e.g. Illumina, PacBio, nanopore) have brought the cost of whole-genome and transcriptome sequencing from over $100,000 less than 15 years ago, to only a few hundred dollars today [1,2]. The increased accessibility of performing RNA-seq has led to massive amounts of data analyzing expression responses to many effects most notably in disease progression. Many of these data are made publicly available in databases such as the NCBI Gene Expression Omnibus (GEO), the Cancer Genome Atlas (TCGA), and the Genotype Tissue Expression project (GTEx), where they can be used to offshoot additional studies [3]. This abundance of biological data has raised new questions about how to analyze and extract meaningful conclusions to better our understanding of biological processes and the molecular mechanisms that drive them.

## 1.1. Gene regulatory networks

It is important to consider that the expression of a gene is not an independent quantity. Rather, each gene is interconnected by regulatory networks. Genes that are associated with a common biological process (BP) will often display shared expression patterns (coexpression). These patterns reflect the positive and negative feedback loops that regulate the particular BP [4]. One mechanism for maintaining these feedback loops is the ability of genes to regulate expression by coding transcriptional factors (TF). A TF is a protein that binds to a target gene's promoter region to either enhance expression (up-regulate) often by recruiting the RNA polymerase machinery that begins transcription, or repress expression (down-regulate) by

inhibiting the binding of RNA polymerase [5]. TFs can be thought of as expression switches, turning the target gene ‘on’ or ‘off’. Typically their binding DNA fragments, known as motifs, are small, and they can be deployed to affect many downstream targets. TFs are capable of regulating target genes locally within the same DNA molecule (*cis*-effects), and distally (trans-effects) [6]. A gene regulatory network (GRN) contains the set of interacting genes (represented by nodes) and the interactions from regulator to target (represented by edges).

GRNs vary in size and complexity from a few, up to thousands of genes. Additionally, many genes operate within multiple GRNs. The functional similarities and genetic overlap between GRNs give rise to the hierarchical organization [7,8] with broadly defined networks made up of thousands of nodes that encompass smaller networks describing unique processes. As an example, a broad GRN such as metabolic process contains the subnetwork nitrogen compound metabolic process, which contains a specific network ammonia oxidation [9]. There has been significant progress in assembling comprehensive GRNs to describe the interactions of genes and their function within BPs. This has given rise to several publicly accessible databases for pathways such as GO (Gene Ontology), KEGG (Kyoto Encyclopedia of Genes and Genomes), Reactome, Molecular signatures database (MsigDB), etc. [10–13]. These databases are often used as references to determine the statistical enrichment of a gene set (i.e. DEGs of a sample) and what processes they are involved in. There is considerable agreement among databases, however, it has still been shown that the choice of database for reference can yield conflicting results when used in enrichment analysis and predictive models [14].

## 1.2. Transcriptional coherence

To adapt to changing environmental conditions, cells often need to alter the expression of hundreds and thousands of genes rapidly and robustly. Clustering analysis of expression

profiles has shown the association of genes within GRNs usually corresponds with a correlation of expression levels (co-expression) across samples [15–17]. However, co-expression can still occur between genes that are not in common GRNs. Several causes have been found to contribute to co-expression, aside from common regulatory elements. This leads to difficulty in determining co-regulation when observing co-expression alone.

It has been shown that gene location has a significant effect on transcription and neighboring genes are frequently co-expressed [18]. This is in part due to the chromatin structure of DNA, which can either be ‘tightly packed’ heterochromatin, or ‘loosely packed’ euchromatin. The spatially constricted packing of heterochromatin makes it unlikely for RNA polymerase to bind, reducing transcription in the region. The opposite is true for euchromatin, so these regions tend to be highly active transcription sites. Since chromatin structure often spreads along the genome 1-dimensionally, this results in neighboring genes frequently occupying the same chromatin domain and a subsequent contribution to co-expression by gene proximity [19]. This has been expanded on with the use of Hi-C chromosome conformation capture techniques, in which the 3D organization of the genome can be detected by crosslinking spatially localized DNA segments [20]. Analyzing the Hi-C interaction/contact data has shown chromatin regions (domains) with high levels of interaction considered topologically associated domains (TADs) [21,22]. Recent efforts have been made to assess the functional relevance of TADs and their potential role in organizing GRNs [23].

Two other powerful technologies for understanding the structure and organization of GRNs are chromatin immunoprecipitation sequencing (ChIP-seq) and bisulfite-sequencing (bis-seq). Specific DNA-binding proteins or DNA structural proteins (histones) can be targeted with antibodies by ChIP and their connected DNA fragments can be sequenced by ChIP-seq [24]. This allows a detailed mapping of epigenetic features throughout the genome. ChIP-seq is often supplemented with bisulfite-induced DNA modification which converts unmethylated cytosine into uracil [25]. Since DNA methylation mostly occurs at CpG sites (cytosine followed by

guanine), any methylated-cytosines will not be converted to uracil and can be identified through sequencing [26]. These procedures applied together can provide a high-resolution mapping of DNA modifications, which can be further analyzed for their role in epigenetic transcriptional regulation.

Understanding the underlying mechanisms that coordinate together to construct GRNs remains an active area of research. Many studies have aimed at integrating multi-omics datasets to infer undiscovered GRNs and how the hierarchy of co-regulated GRNs [27,28].

## 1.3. Differential expression (statistical approaches)

To identify key GRNs that regulate a particular function or process, we often turn to RNA-seq, which is the read count of individual genes. RNA-seq data typically contain samples from two different conditions. Differential expression (DE) is one of the most common analyses used to understand the transcriptional mechanisms that change by conditions of interest. Common examples of sample conditions include diseased/healthy, treatment/control, or time series among others. When the expression level of a gene compared across two conditions is significantly increased (up-regulated) or decreased (down-regulated), that gene is considered differentially expressed. Deciding whether the difference is significant or not is a matter of statistical analysis. Therefore, the traditional approach to identifying differentially expressed genes (DEGs) is through the use of parametric statistical tests. For each of these tests, the null hypothesis is that there is no difference in the expression of an individual gene between the two conditions.

### 1.3.1. DE normalization

Before DE is assessed, it is a good practice to normalize the raw expression values to account for technical variations that may negatively affect downstream analysis. One example

of such technical variations includes sequencing depth in an RNA-Seq experiment (referred to as library size). Since RNA-Seq relies on PCR amplification before sequencing, each sample measured will have a different total number of reads due to the varying amounts of amplification. To account for differences in sequencing depth, the most commonly used expression measures are adjusted per million reads. Another technical variation arises from the varying length of each gene. Most sequencing methods do not sequence full-length RNA transcripts. Rather, RNA transcripts or their cDNA counterparts are fragmented via heat, acoustic shearing, enzymatically, or chemically [29]. These fragments are then sequenced and their motifs are realigned with the reference genome/transcriptome so that each fragment is mapped to a specific gene. Transcripts of larger length will naturally have more fragments and subsequently, more reads mapped to the respective gene. For this reason, expression levels are often normalized by dividing each gene's expression by its transcript length. Lastly, when making comparisons between separate samples, more genes can be actively expressed in one sample than the other. After normalizing for the total number of reads and gene length, the genes that are consistently expressed in both samples may falsely appear to be DE since there are fewer active genes that the read counts are normalized over. Therefore, it is necessary to divide by the total number of genes expressed within each sample library [30]. Transcripts-per-million (TPM), given by equation 1.1 with the raw read counts as *x* and gene length *l* for gene *i* in a profile of *n* genes, is a commonly used normalization method, which takes most of these considerations into account [31].

$$TPM_i = \frac{x_i/l_i}{\sum_j^n (x_j/l_j)} \times 10^6 \quad \textbf{(1.1)}$$

| Raw Reads | | | | | |
|---|---|---|---|---|---|
| ***RNAseq reads*** | **Gene size (kb)** | **Lung 1** | **Lung 2** | **Colon 1** | **Colon 2** |
| **Gene A** | **15** | **30** | **60** | **54** | **85** |
| **Gene B** | **11** | **11** | **24** | **2** | **3** |
| **Gene C** | **8** | **15** | **30** | **0** | **0** |
| **Gene D** | **5** | **4** | **8** | **12** | **15** |
| **Gene E** | **10** | **21** | **41** | **22** | **31** |

**Table 1.1. Demonstration of normalization considerations in gene expression profiles.**

| TPM | | | | | |
|---|---|---|---|---|---|
| **RNAseq reads** | **Gene size (kb)** | **Lung 1** | **Lung 2** | **Colon 1** | **Colon 2** |
| **Gene A** | **15** | **25.72** | **25.59** | **42.95** | **47.07** |
| **Gene B** | **11** | **12.86** | **13.96** | **2.17** | **2.27** |
| **Gene C** | **8** | **24.12** | **23.99** | **0.00** | **0.00** |
| **Gene D** | **5** | **10.29** | **10.24** | **28.63** | **24.92** |
| **Gene E** | **10** | **27.01** | **26.23** | **26.25** | **25.75** |

**Table 1.2. Demonstration of transcripts-per-million (TPM) normalization of the raw reads from Table 1.**

### 1.3.2. Poisson and negative-binomial distributions

As discussed, data preparation procedures can have a significant effect on the results of all downstream analyses. After the expression data has been prepared, DEGs are often identified using parametric statistical tests. First, the probability distribution of the expression is assumed. Since gene expression is often measured in discrete read counts, discrete probability distributions (e.g. Poisson, negative binomial, etc.). Then a test statistic is determined for every gene based on the estimated parameters fitting the gene's actual expression to the assumed distribution shape. The Poisson distribution is commonly used to model biological count data due to its simplicity since it has the constraint that the mean and variance are equal [32]. Therefore, the distribution's shape is governed by a single parameter, lambda. Unfortunately, the assumption of mean/variance equivalence is often invalid in the majority of gene expression profiles [33]. Regardless, the Poisson distribution has been used heavily and adapted (mixture models) for DE analysis [34,35]. For the more probable case of overdispersion, in which the expression variance is greater than the mean, the negative-binomial (NB) distribution has been employed extensively. The NB distribution models the number of Bernoulli trials (coin-flips) that will occur before *r* number of successes [36]. The NB is governed by the parameters *r* and *p*, the probability for success in each trial. Both the mean and variance of the NB distribution can be found using equations 1.2 and 1.3 respectively. Since *p* must be less than 1, the variance will always be greater than the mean (overdispersed).

$$\mu = \frac{r(1-p)}{p} \quad \textbf{(1.2)}$$

$$\sigma = \frac{r(1-p)}{p^2} \quad \textbf{(1.3)}$$

### 1.3.3. DESeq and edgeR

DESeq2 and edgeR are among the most commonly used DE analysis toolkits for RNA-seq studies. Both methods utilize the NB distribution for DE identification, have built-in low-expression gene filtering procedures, provide DE data visualization (volcano, MA plots), and are available as open-source software packages in R. The methods differ in their normalization and dispersion estimate schemes. Pre-normalized data, commonly TMM (trimmed mean of M-values), is required by edgeR [37]. DESeq2 accepts a count matrix of reads per gene, which is then normalized internally. For estimating the dispersion with a small number of replicate samples edgeR uses a quantile adjustment to account for differences in library size [38]. Whereas, DESeq handles library sizes by scaling the mean and subsequent variance parameters [39]. The update of DESeq2 contributed a conservative shrinkage of FC values based on mean expression [40]. Depending on the particular data, both DESeq and edgeR share significant overlap in DEG identification [37]. They have also been shown to have high statistical power, however, in samples with small numbers of replicates they may suffer from high false-discovery-rates (FDR) [41].

### 1.3.4. limma

Linear models and differential expression for microarray data (limma) is another commonly used DE analysis package that was developed originally for use in examining microarray expression data. Microarrays can capture relative differences in expression based on fluorescence intensity compared between arrays [42]. The intensity measurements from microarrays are continuous values, which distinguishes them from the discrete count data from RNA-seq. limma has been updated to be capable of performing similar DE analysis on RNA-seq data using the established linear models [43,44]. Limma operates by fitting a linear model to the

expression of individual genes followed by empirical Bayes statistical tests used to borrow information across genes [45]. This provides a simple framework for analyzing gene expression from many types of study designs. DEGs predicted by limma have been shown to overlap significantly with other DE tools such as DESeq and edgeR [37,46].

### 1.3.5. SAMSeq

DESeq, edgeR, and limma are all examples of parametric statistical tests for DE. These have been shown to work exceptionally well (high statistical power and low FDR) when the distributional assumptions are valid [47,48]. However, in cases where the model assumptions are not met these methods tend to maintain high power at the expense of high FDR [49]. Nonparametric statistical tests provide a more reliable alternative that generates consistent results in cases where parametric assumptions are unverified. SAMSeq uses the Wilcoxon/Mann-Whitney statistic which is based on the ranked expression [50]. The Wilcoxon statistic assumes an equal sample size (sequencing depth) in order to make comparisons between conditions. To solve this issue SAMSeq performs resampling of read counts from each gene to ensure equal sample size. SAMSeq has been shown to have similar performance to parametric tests in many cases and better performance in atypical RNA-seq studies (lncRNAs) [41,51,52].

## 1.4. Differential expression (machine learning approaches)

Conventionally, statistical approaches have been employed to capture DE in both microarray and RNA-seq studies. Recently, significant advances in machine learning (ML) and artificial intelligence algorithms have spurred the development of ML approaches for analyzing gene expression data. In addition to DE analysis, ML algorithms have been developed for

sample classification, dimensionality reduction, feature selection, and missing data imputation [53]. ML methods for DE analysis include support vector machines (SVM), decision trees, random forests, clustering, and various forms of neural networks [54]. The overarching theme of ML methods is finding patterns within data by making generalized observations (e.g. convolutions, dimensional reduction, etc.). While parametric statistical tests assume specific distributional characteristics accurately describe the data, ML assumes more general features are present such as distinct clusters or heteroscedasticity across features/genes. This may be beneficial in cases where it is difficult to make predictions on the distributional characteristics required to validate statistical assumptions.

### 1.4.1. Principal component analysis

Principal component analysis (PCA) is frequently used for dimensionality reduction. This is particularly applicable in gene expression analysis, where expression profiles often include thousands of genes, measured across sometimes hundreds or thousands of samples. Removing redundant or insignificant genes/samples can improve the performance and meaningfulness of downstream analysis. PCA constitutes the reconfiguration of data axes in order to maximize variance and subsequently the information explained by each axis [55]. The new axes are ordered by their amount of variance explained so that once a sufficient amount of variation is explained all further components can be disregarded. Principal components are determined by finding the maximum eigenvalue of the covariance matrix with the constraint that all components are orthogonal [56].

### 1.4.2. Clustering (k-means, hierarchical, density-based)

The expression of genes within a transcript profile do not often act as independent randomly distributed values, but rather as groups/clusters of distinct patterns. This is likely due

to the co-expression of genes within GRNs and of genes with similar functions. Clustering methods attempt to categorize genes based on expression similarity. The majority of clustering algorithms are considered “unsupervised” since the input data does not need to include predetermined cluster labels. Some methods such as k-means search for a predefined number of clusters, while others may return any number of clusters based on the chosen parameters (DBSCAN) [57]. Hierarchical clustering algorithms produce a more complete view by returning a branched sequence of possible clusters, displaying k-means clustering for all values of k [58]. All methods employ a particular proximity measure (e.g. euclidean distance, Pearson correlation, spearman correlation) [59]. Cluster analysis of gene expression data was popularized by Eisen et al. in which they were able to find distinct functional gene clusters in an *S. Cerevisiae* expression profile as well as a time series during the mitotic cycle [15].

### 1.4.3. Support vector machines

Support vector machines (SVMs) are a popular tool for data classification. The goal of SVM is to identify an appropriate cutoff to separate two classes of data based on prelabeled (supervised) training data. The single-dimension cutoff is easily expandable by identifying a hyperplane separating classes of high-dimensional data. The hyperplane is determined by an optimization algorithm that maximizes the spatial margin between the classes of training data [60]. To allow for non-linear hyperplanes SVMs map the input data to a feature space using a variety of kernel functions, such as the polynomial kernel [61]. SVMs have been applied to identify gene signatures in cancer [62].

## 1.5. Pathway Enrichment Analysis

After gene signatures have been identified using any of the previous methods, the next question is often whether the gene signatures are biologically relevant and if so, what

conclusions can be drawn from them. Genetic pathways represent the series of interactions that facilitate a particular biological process within cells [63]. These processes may result in the production of cellular products (e.g. proteins, metabolites), or regulate cellular homeostasis. Understanding what pathways the signature genes we have identified play a role in is a critical first step in describing the biological significance of the gene set. There are numerous publicly accessible databases (KEGG, Reactome, Gene Ontology) listing our current understanding of all biological processes/pathways and the connectivity of the genes that facilitate them [10–12].

To determine whether the signature gene set is related to a particular pathway, a Fisher's exact test is performed on each pathway [64]. The ratio of signature genes that are part of the pathway of interest is compared to the ratio of all reference genes that are part of the same pathway. The reference gene set can be the set of all genes analyzed by the study or simply the entire reference genome for the species studied. For example, if we identify 145 signature genes and 22 of them are related to signal transduction, then we compare that to the human reference genome consisting of 19,256 total coding genes with 1251 related to signal transduction [65]. Therefore, about 15% of the signature genes are related to the signal transduction pathway, whereas if these genes were selected at random we would expect only 6.5% to be associated with signal transduction. Using Fisher's exact test, we can determine a p-value (probability of null hypothesis) as $1.3 \times 10^{-4}$, which is the probability that randomly selecting genes would result in the same ratio related to this particular pathway. Since this is far below the typical threshold of 0.05 we can say that our signature gene set is likely related to signal transduction.

P-values are a good metric for statistical significance for a single test. However, when performing pathway enrichment, we are calculating thousands of p-values. There are nearly 30,000 biological processes within the GO database [66]. If we use the standard p-value threshold of 0.05, then even if we select genes at random we would still likely find about 1,500 GO terms with significant enrichment. This is the problem of multiple testing which has been

addressed in several ways by correcting/adjusting the p-value based on the number of tests. The simplest and most conservative adjusted p-value is the Bonferroni correction, which simply multiplies the original p-value by the number of tests [67]. Using the previous example, the unadjusted p-value would need to be ~$1.7 \times 10^{-6}$ in order for the adjusted p-value to be below 0.05. For the signal transduction pathway, the Bonferroni adjusted p-value is 0.15, so we could no longer say that there is a likely association with signal transduction among our signature gene set. The Bonferroni correction significantly controls the family-wise error rate or the likelihood that at least one reported false-positive. However, it has been considered too conservative, because of the simultaneous loss of true-positive discoveries [68]. For this reason, the Benjamini and Hochberg method is more often implemented. In this procedure, p-values are ranked and the adjusted p-value is determined by multiplying the unadjusted p-value by the number of tests divided by its rank [69]. Using the Benjamini and Hochberg correction on the signal transduction example, the adjusted p-value would be 0.03. Therefore, after controlling for the false discovery rate we can say there is a significant association, albeit much less significant than without the correction.

## 1.6. Uncovering heterogeneity with scRNA-Seq

While bulk-RNA-Seq profiles are useful for identifying changes in gene expression throughout an entire tissue, many transcription signals may be missed by averaging over the numerous cell types that make up a particular bulk tissue sample. Single-cell sequencing provides the potential to study the intra-sample heterogeneity by sequencing distinct cell types separately. In scRNA-Seq cell types are often sorted by size, granularity, and fluorescent tags in a process called fluorescence-activated cell sorting (FACS) using flow cytometry [70]. Subsequently, cells can be separated and sequenced individually or combined with cells of the same type and sequenced as a pseudobulk sample. ScRNA-Seq has seen promising use in

situations where the cell type of interest is relatively sparse compared to other cell types within the local tissue. For example, in early embryonic development, each cell type is by definition sparse. Single-cell sequencing has been used to understand the genetic lineage and developmental programs that regulate early pluripotent cells into later stages of development and differentiation [71]. Similar studies have been performed on early-development tumors or metastases. Circulating metastatic tumors are another case where the cell type of interest is relatively sparse compared to the surrounding cell types. Efficiently isolating, sequencing, and analyzing individual cells has allowed insight into the transcriptional response mechanisms required throughout the various stages of the metastatic process [72,73].

Alongside the many benefits of the high-resolution perspective on the transcriptome, there have been many difficulties in the reliable interpretation of scRNA-Seq data. Within bulk samples, there is the common issue of analyzing genes with low read counts. There is a tendency for lowly expressed genes to have high fold-changes which can easily be interpreted as significant by downstream statistical analyses. In some cases, these genes are simply removed from the profile to prevent skewing downstream results. This has also been addressed by shrinking the FCs of low-expression genes in DESeq2 [40]. Unfortunately, this issue is exacerbated in single-cell studies, since transcript levels are significantly less than for bulk samples. This can lead to high FDRs when assessing DE in scRNA-Seq profiles [74]. Because read counts are low, technical variation between single-cell samples is also significantly greater than in bulk, which makes it harder to make accurate predictions of true DE [75]. This issue also presents itself as the “zero inflation” problem, where many genes have zero transcripts in a single cell. There is disagreement on how to handle unexpressed genes which may be the majority of the profile [76].

## 1.7. Sources of variation in gene expression data

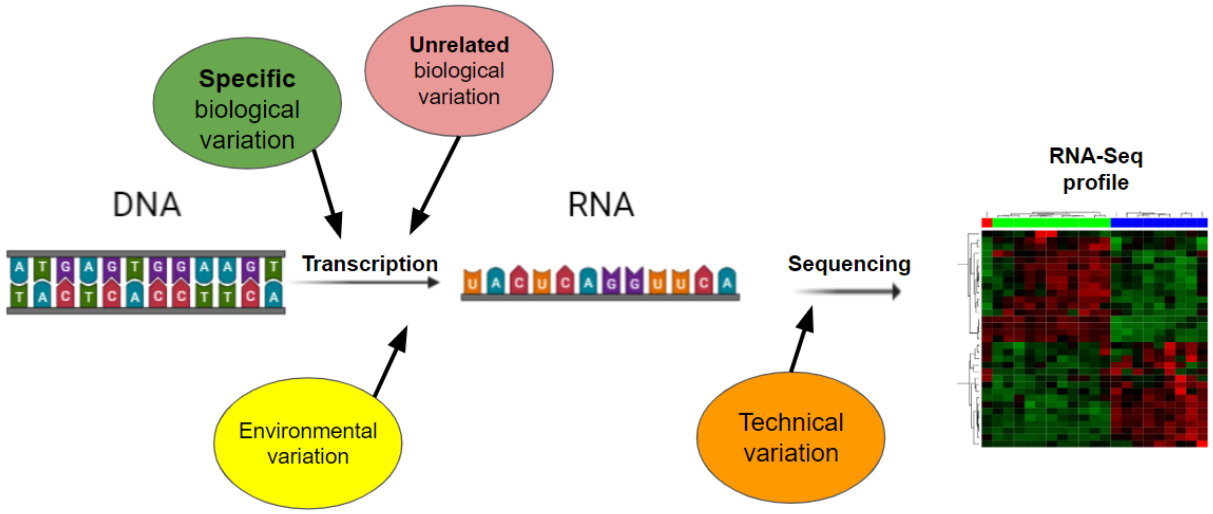


**Figure 1.1. Types of variations within RNA-Seq data.**

The goal of DE is to identify the genes responsible for a significant portion of the variation between samples. DE can be used as a tool to screen for potential biomarkers, or drug targets based on the assumption that the variation from DEGs is due to the physiological differences (biological variation) between samples. Variations in gene expression can be induced from many sources depending on how samples were obtained and what sequencing technologies were used. This is especially true for data mining studies, where comparisons between profiles from different laboratories and different protocols are quite common. For example, both cell lines and patient-derived xenografts are commonly used models for studying cancer, and both have extremely dissimilar environments (*in vitro* vs *in vivo*) that will likely induce unique cellular responses. Naturally, both models have shown different predictive efficacy in preclinical trials [77]. Technical variations are those introduced by the measurement procedure. A historical example of RNA-Seq would be the GC effect, where regions of the genome with a higher fraction of G and C nucleotides would receive a higher coverage of sequencing compared to regions of sparse GCs [78]. It has been shown that these biases arise largely during PCR amplification and their effect differs based on PCR protocols [79]. Technical variation has been shown in some cases to be larger than biological variation [75,80]. Work has

been done to correct for these technical variations, however, there is still no consensus on the best practices and it is often up to the authors to decide the appropriate workflows. Although studies attempt to control for non-biological variations between samples, they can never be completely eliminated.

Even if a study perfectly controls for all environmental sources of variation as well as technical variation between samples, there is still the possibility that the accurately identified biological variations are irrelevant to the mechanism of interest. An example of undesired biological variation when studying cancer is tumor purity or the fraction of tumor cells to healthy cells within the microenvironment [81]. Tumor purity can vary with a standard deviation of up to 20% depending on the histology which can lead to significant differences in expression signals across samples [82]. A further complication arises from the correlation versus causation dilemma or in this case the conflation of disease-causing pathways and disease-induced pathways [83]. Although recognizing the pathways induced by a disease can help to understand the stages of progression. This is often not the primary interest, which is instead to understand what pathways contribute to the onset and progression.

### 1.7.1. Large-scale shifts in gene expression

Genes of similar functions are linked by GRNs and tend to be coexpressed due to regulation by common mechanisms. TFs represent one such mechanism, by coding DNA binding proteins that target downstream gene promoters and enhancers to either increase (activator) or decrease (repressor) transcription in the target [84]. TFs that regulate multiple pathways, usually by regulating downstream TFs, are considered master regulator genes (MRGs) [85]. MRGs have been associated with various diseases most notably in cancer. The mammalian target of rapamycin (mTOR) is an MRG that has seen recent significant interest for its role in cancer metabolism and facilitating stem-like features that promote metastasis [86].

Another prominent example is the family of cyclin-dependent kinases (CDKs) which regulate cell-cycle transitions [87]. CDKs have been shown to regulate significant parts (~10%) of the genome in yeast and many associated processes are conserved in mammals [88,89]. Large-scale down-regulation has been seen during cancer progression [90]. In addition to MRG effects, copy number loss has been shown to result in large-scale down-regulation [91].

## 1.8. Transcriptional reprogramming in cancer

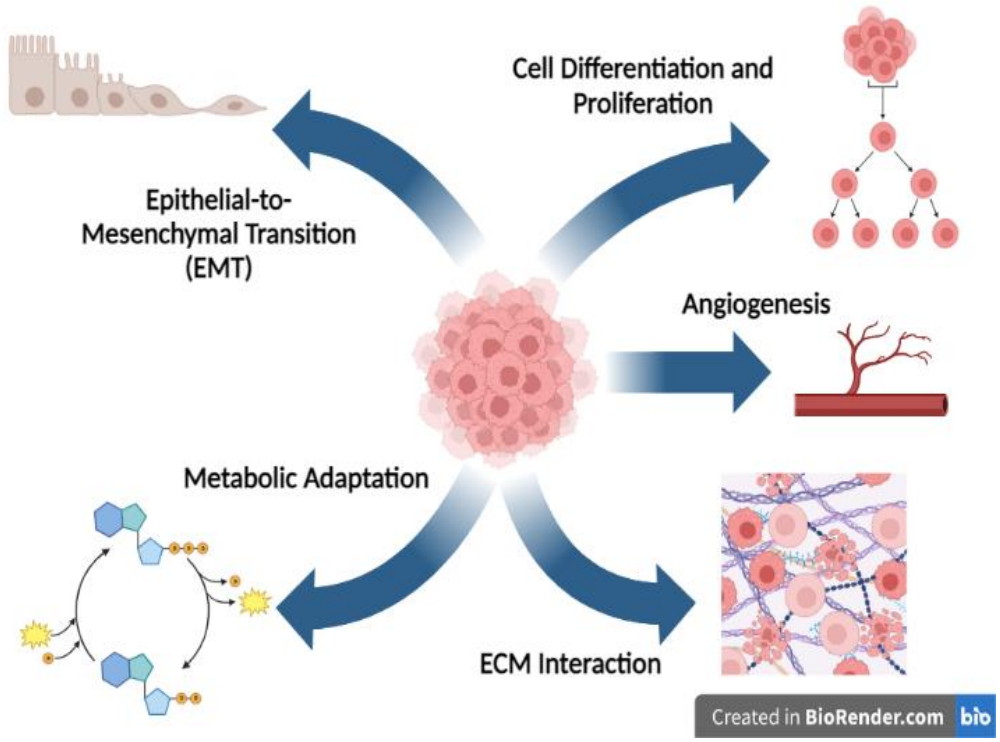


**Figure 1.2. Hallmarks of cancer.**

Cancer is a complex disease that originates from the accumulation of mutations that either provide a loss-of-function in tumor-suppressing genes or a gain-of-function in tumor-promoting oncogenes [92]. Although their progression mechanisms are diverse depending on the tissue of origin and histology, there are several common capabilities acquired during tumorigenesis. These hallmarks of cancer include resisting cell death, proliferative signaling, angiogenesis, invasion, and metastasis [93]. Understanding the transcriptional programs that

drive these interactions and the genes that control them is a key part of developing therapeutic strategies for cancer treatment.

### 1.8.1. Cancer metabolism and angiogenesis

Malignancy depends on the continued growth and proliferation of tumor cells, which then require higher amounts of energy and nutrients to be sustained. One of the hallmarks of cancer is the upregulation of multiple metabolic pathways including glycolysis, glutaminolysis, lipid metabolism, and mitochondrial biogenesis [94]. In addition to increased metabolic rate certain cancer types have been shown to adapt their metabolic processes throughout tumor progression. In the early stages of development, tumors tend to rely on faster, albeit less efficient, glycolysis. While fully formed tumors rely on the slower and more efficient oxidative phosphorylation, which is seen to provide drug resistance and be better suited for cellular differentiation and subsequent invasion [95]. In conjunction with altered metabolic pathways, tumors have been shown to promote angiogenesis (formation of new blood vessels) by simultaneously upregulating angiogenic activators such as the vascular endothelial growth factor and downregulating inhibitors [96]. The newly formed blood vessels then serve to provide oxygen and other nutrients to promote further tumor development, as well as increase the metastatic potential.

### 1.8.2. EMT, pluripotency, and the metastatic cascade

Metastasis is responsible for the majority (roughly 90% by some estimates) of cancer-related deaths [97]. The metastatic cascade is a complicated process in which cancer cells undergo detachment from the primary tumor, intravasation into the bloodstream, extravasation to secondary sites, and proliferation in order to form secondary/metastatic tumors [98].

Throughout this process, metastatic cells experience a wide range of forces and mechanical stresses that are uncommon in their traditional microenvironment [99]. Survival in these diverse conditions requires an adaptive cellular response resulting in changes in morphology, signaling, and metabolism [100,101]. The epithelial-to-mesenchymal transition (EMT) is a well-studied regulatory program that enables cellular detachment and intravasation. Epithelial cells are characterized by their immobility, cellular adherence, and cell-to-cell interactions [102]. Whereas, mesenchymal cells have stem-like properties such as pluripotency and motility [103]. The coordinated alteration of gene expression is necessary to induce, propagate, and sustain cells undergoing EMT and metastasis. For instance, integrin proteins, which play a role in cellular adhesion, have been shown to have altered expression in cancer which promotes cellular detachment and subsequent migration [104]. Similarly, the down-regulation of tumor suppressor E-cadherin, which facilitates epithelial adhesion, has been observed across multiple metastatic tumor types [105]. Several families of transcription factors (TFs) have been implicated in regulating EMT such as zinc-fingers, and basic helix-loop-helix [106].

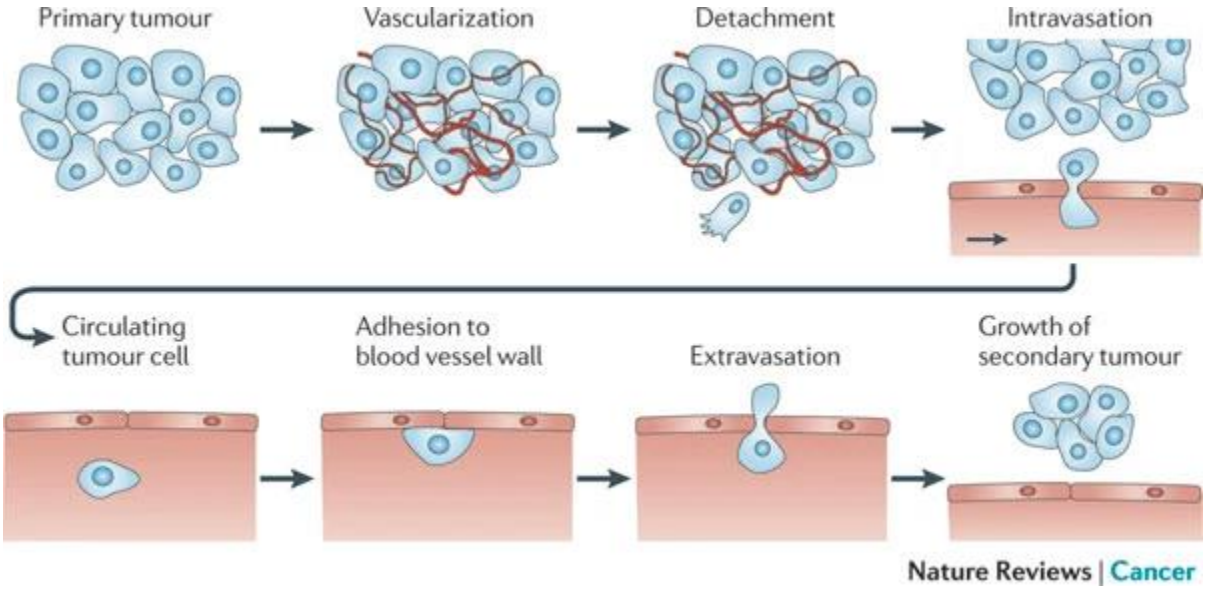


**Figure 1.3. Overview of the metastatic process.** *Courtesy of Wirtz et. Al. Reproduced with permission from Springer Nature* [99].

EMT is only one component of metastatic potential, which goes to show that metastasis is a highly complex biological process with many unanswered questions in understanding its potential, development, survival, and effectiveness [97,107]. Evidence suggests that the vast majority of metastatic cells do not survive long enough or are ineffective at forming secondary tumors [107]. One of the major factors in cell survival may be the ability of cells to sense and adapt to the various environmental forces experienced outside of their native tissues [99,108]. Understanding the mechanisms that determine these adaptive features may uncover potential therapeutic targets and preventative measures to reduce metastatic potential.

During the primary tumor growth cells experience increasing compressive forces due to the tumor growth and stiffening of the neighboring healthy tissue [108]. The level of tumor confinement combined with microenvironment stiffness has shown an increase in collective (multi-cell) intravasation which poses a risk due to their greater potential to form secondary tumors [108,109]. After intravasation metastatic cells are exposed to shear stresses from flow differentials within the circulatory system [110]. High shear stresses have been shown to kill tumor cells at the level obtained during intense exercise [111]. Before settling in a secondary location, metastatic cells either adhere to the endothelial cells that make up the blood vessel barrier or become restricted within smaller capillaries [99,112]. Once metastatic cells extravasate out of the circulatory system and into their secondary location, they can either remain dormant or reactivate the growth of new tumors [108].

Recent interest has been in understanding the mechanisms that allow cells to sense and adapt to external forces (mechanosensing). Many questions remain as to what genes play a role in mechanosensing and how their regulatory programs are altered in cancer.

# 2. MAGE: Monte Carlo method for Aberrant Gene Expression*

**Submitted for review at PLOS ONE.*

## 2.1. Abstract

Identifying genes which are aberrantly expressed is an important first step in the diagnosis and treatment of many diseases. Conventionally, differential expression (DE) analysis is used to screen gene expression profiles to identify functionally associated genes. DE often relies on the variance and fold change in expression from individual genes, which does not take into account the gene expression profiles of all other genes. When the overall gene expression is skewed, DE does not capture outliers in gene expression. To address this, we have developed a non-parametric DE method based on the probability density for an entire expression profile to select genes that deviate from the global distribution between two gene expression profiles with multiple replicates. Rather than assuming a particular distribution of expression per gene, our method assumes that aberrantly expressed genes (AEGs) will exhibit expression patterns distinguishable from non-AGEs which make up the majority of the profile.

Here we introduce our nonparametric method (MAGE: Monte Carlo method for aberrant gene expression) and demonstrate that MAGE can identify AEGs different from conventional DE analyses. The main feature of MAGE is (1) identifying outliers based on the expression profile of all genes rather than performing DE analyses on a per-gene basis and (2) consideration of the variance in expression between two different conditions. We also compared our results with traditional DE analysis as well as density-based clustering methods. MAGE

produces consistent results in a variety of conditions and performs conservatively with the addition of noise. Furthermore, we have studied the biological significance of the identified signature genes and assessed the potential to gain insight into AEG-associated biological processes and pathways.

## 2.2. Introduction

One of the first steps for tackling human diseases is to understand the molecular players behind the emergence of diseases, and disease-associated key genes can be potential therapeutic targets. A common practice is to utilize expression profile data (often RNA-seqs) to identify differentially expressed genes (DEG) by comparing differences in expression levels between healthy and diseased tissues or control and treatment samples [113,114]. When a gene's expression is significantly different from that of the control, it is considered to be differentially expressed, and DEGs can become candidate biomarkers for diagnosis and prognosis of human diseases. Additionally, the transcriptional, translational, and post-translational regulation of DEGs can be further targeted for potential use in drug development.

Most methods for DEG identification fall in the category of parametric statistical significance tests, where gene expression levels are assumed to be from a known probability distribution with a few parameters [115,116]. Since expression levels are typically measured by read counts from RNA-seqs, discrete probability distributions (e.g. Poisson [35,117] and negative-binomial [38,40] for DESeq and EdgeR) are commonly employed to characterize expression profiles. The Poisson distribution, as governed by a single parameter, is preferred for its simplicity, however, it is limited by the constraint that the variance and mean are equal. This is often not applicable as the biological variation among distinct replicates is larger than the expected technical variation of sample preparation [118]. Because of this constraint, the Poisson distribution often does not fully account for the deviation in many expression profiles

and may result in higher false-positive discovery rates than other probability distributions [39,119]. While the negative-binomial distribution (ND) includes separate parameters for the mean and variance, the number of samples is sometimes too small to effectively evaluate both parameters. With each DE method based on a different set of assumptions, this has led to a large number of context-dependent health- and disease-associated DEGs [120], and challenges exist in devising methods for narrowing the large number of DEGs robustly. Prior studies suggest different genes may not conform to the same distribution, and an approach that can handle many types of probability distributions and extract biologically relevant gene signatures is needed [121].

Our approach takes inspiration from several classes of nonparametric machine learning-based methods previously used to identify DEGs. Namely, gaussian mixture models (GMM) which have been used for outlier detection in RNA-seq datasets [122,123], operate on a similar premise of considering the underlying probability density function (PDF) from each gene. Other methods of detection of outliers involve distance-based clustering methods such as k-nearest neighbor [124], and density-based approaches such as DBSCAN [125,126]. Information entropy is also used to discover cluster genes in a noise-resistant manner [127,128] . Previous algorithms have been developed for identifying outlier genes among RNA-seq profiles. Aberrant gene expression in rare disorders has been assessed using OUTRIDER [129], which utilizes an auto-encoder to account for unknown covariation between genes, followed by a statistical p-value determined using ND. Similarly, ABEILLE [130] utilizes a variational auto-encoder and introduces an anomaly score that is determined using an isolation forest approach. This removes the distributional assumptions used in OUTRIDER and other parametric DEG methods, however, it introduces the need to use multiple predetermined thresholds for AE classification.  Ordensity is another algorithm to find outlier genes reliably from microarray data [131].

With this in mind, we have developed a method that identifies AEGs by estimating the 2-dimensional (2D) PDF of each gene and comparing it with the combined 2D PDF (cumulative PDF) from all genes. We assume that true AEGs will exhibit deviated expression compared to those of all others. It is also vital to consider the computational costs of conducting complex analysis for the identification of DEGs due to the sheer size of the omics datasets that are frequently encountered. Our method performs robustly while remaining computationally cost-effective so that it can be applied to expression profiles containing thousands of genes across many samples.

## 2.3. Methods

### 2.3.1. Aberrant expression vs differential expression

It is important to note the distinction between traditional DEGs and AEGs, and aberrant expression may or may not take the form of differential expression as we are interested in the outlier genes based on the overall gene expression patterns across two different conditions. In many cases, there is a significant overlap between DEGs and AEGs, and this overlap is especially evident in simple expression profiles examining small effects between conditions. We have observed that as the expression behavior becomes more complex, the agreement between DEGs and AEGs decreases.

### 2.3.2. Setting up MAGE

The input data is an expression profile from two sample conditions (e.g. cell/tissue types, diagnosis, treatment/control pairs, etc.) each of which contains *n* number of biological replicates, and *n* must be greater than 2. Figure 2.1 illustrates the overview of MAGE and its workflow.

MAGE does not have built-in normalization, so the input expression profile should be in pre-normalized units such as TPM, RPKM, etc. to account for differences in transcript length and sequencing depth. Further filtering can be performed by removing genes with near-zero reads in a majority of samples. Filtering of low read count genes is recommended to significantly reduce processing times and avoid noisy expression. Similar to other methods of finding gene signatures, our algorithm is capable of producing results for any profiles with two conditions and more than one replicate sample per condition.

First, we calculate the mean and standard deviation (SD) from the expression of a gene at a given condition (eqs. 1-2). Then we calculate PDF over the 2D expression region (eqs. 3-5) and estimate the cumulative sum of PDFs of individual genes (eqs. 6-7). The resulting cumulative PDF (CPDF) approximates a two-dimensional PDF or density matrix for the whole profile. We choose the range of the grid to cover CPDF beyond the mean expression values, and we account for this by adding four times the average SD of each gene to the respective grid range. The height and width of each grid are determined by dividing the adjusted ranges by the number of bins as an external parameter (eq. 5). The number of grid indices is an adjustable parameter that can be increased for higher precision or decreased for faster computation. Figure S2.1 shows CPDFs from selected individual genes while varying the total number of genes, and Figure S2.2 displays the entire CPDF ($N \cong 15{,}000$). It is possible that given a large enough number of grid points, each grid may have very few contributing genes. To avoid this, the total number of grid points should be significantly less than the number of genes, and the range of the grid should be set high enough to fully capture the data structure.

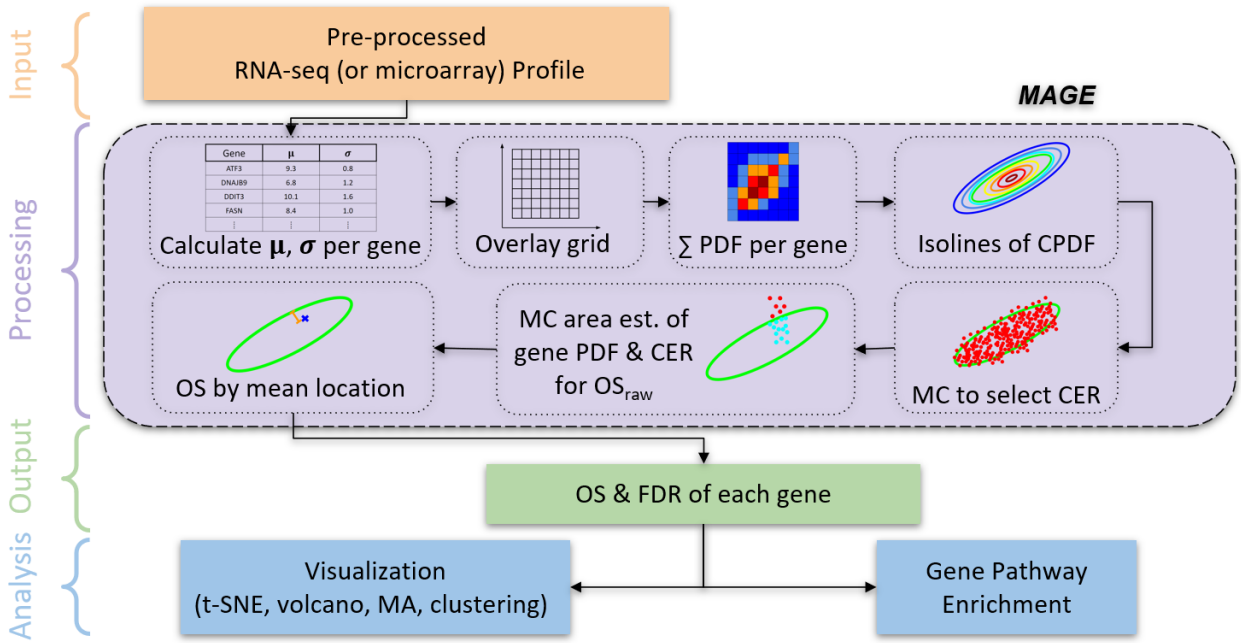


**Figure 2.1. MAGE analysis overview.** MAGE accepts RNA-seq or microarray profiles from two conditions and at least two replicates per condition. MAGE provides output as outlier scores based on the cumulative PDF of all genes and the distance between the CER and the mean coordinate of each gene. MC sampling is performed twice to determine CER from gene expression profiles and to quantify the outlier score and FDR for individual genes. Further analysis includes visualization techniques and gene pathway enrichment to identify the functional relevance of AEGs.

$$\bar{x}_i = \frac{\sum_{j=1}^{n} x_{i,j}}{n} \quad \textbf{(2.1)}$$

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{n} (x_{i,j} - \underline{x}_i)}{n}} \quad \textbf{(2.2)}$$

$$\bar{\sigma} = \frac{\sum_{i=1}^{m} \sigma_i}{m} \quad \textbf{(2.3)}$$

$$\bar{x}_A = \frac{\sum_{i=1}^{m} x_{i,A}}{m} \quad \textbf{(2.4)}$$

$$Grid\ spacing = \frac{\bar{x}max - \bar{x}_{min} + 4\bar{\sigma}}{D} \quad \textbf{(2.5)}$$

$$PDF_{i,A,k} = \frac{e^{-\frac{1}{2}(\frac{g_k - \bar{x}_A}{\sigma_{i,A}})^2}}{\sigma_{i,A}\sqrt{2\pi}} \quad \textbf{(2.6)}$$

$$CPDF_k = \sum_{i=1}^{m}\left(PDF_{i,A,k} \times PDF_{i,B,k}\right) \quad \textbf{(2.7)}$$

$x_{i,j}$ is the read count of the i[th] gene in the j[th] sample. $n$ and $m$ represent the number of replicate samples and the number of genes, respectively. $\sigma_i$ is the standard deviation of i[th] gene across replicates, and $\bar{\sigma}$ is the mean standard deviation of all genes. $D^2$ is the total number of grid indices, and $g_k$ is the location of grid index k. $PDF_{i,A,k}$ is the probability density function of the ith gene in condition A at grid index k, while $CPDF_k$ is the cumulative probability density of all genes evaluated at grid index k.

**Figure 2.2. Selection of the characteristic expression region (CER) and estimation of the $OS_{raw}$.** (A) For a given contour, Monte Carlo (MC) sampling (fixed number $N_{mc1}$) is performed for each gene from its 2D PDF. Then we calculate the fraction of all MC sampled points (total $N_{gene} \times N_{mc1}$ points) inside the contour. Blue and red points represent gene means and MC sampled points, respectively. CER is selected to match the target fraction of MC sampled points lying inside. (B) For individual genes, further points ($N_{mc2}$ per gene) are sampled from the individual gene PDF. The fraction of these points outside CER (denoted red) determines $OS_{raw}$.

### 2.3.3. Determining the characteristic expression region (CER)

From the probability density matrix, we generate a contour plot to find isolines across the CPDF. We define CER as an optimized contour of CPDF, which captures the gene expression profile of the majority of genes. CER from the CPDFs can be determined using contour lines (Figure S2.3), and an iterative optimization algorithm finds the optimal contour level as the CER. First, we select a fixed number of contour lines, which is an adjustable parameter. For each contour line, we quantify contour effectiveness ($CE$) by the fraction of enclosed genes, and the contour line that matches the target containment fraction is CER. CER can consist of multiple unconnected components (Figure S2.4). To calculate the actual fraction of CPDF containment, we perform Monte Carlo (MC) sampling (10 points per gene) from each gene's PDF (Figure 2.2A, S2.3B). Since most profiles contain thousands of genes, sampling only a few points from each gene's PDF still results in a significant (> 10,000) sample size for area estimation. We quantify the total fraction of MC points inside a given contour. Then we compare the actual containment fraction (A) with the target containment fraction (T) (eq. 8). The target containment fraction is a user-set parameter, and we use 95%. $CE$ is quantified as

$$CE = 1 - |T - A| \quad \textbf{(2.8)}$$

where $T$, and $A$ are the target fraction of containment, the actual fraction of containment, and the size (area) of the contour relative to the size of the grid, respectively. Figures S2.5 illustrate the performance of CER and other isolines.

### 2.3.4. Quantification of raw outlier score ($OS_{raw}$)

We consider genes that are likely to be found outside CER to be aberrantly expressed, which we term as AEGs. To quantify this, we implemented a second round of MC sampling. The expression of each gene follows its PDF, and CER can have an arbitrary shape. A larger number of points (1000 in this study as an adjustable parameter) are sampled from each gene's PDF. The ratio of points falling inside and outside of CER is quantified as the raw outlier score

($OS_{raw}$) (Figure 2.2B). For an individual gene, $OS_{raw}$ represents the probability of that gene's expression to be outside of the CER as

$$OS_{raw} = \frac{N_{out}}{N_{in} + N_{out}} \quad \textbf{(2.9)}$$

where $P$ is the number of sampled points inside/outside of CER.

## 2.3.5. Adjusting high variance genes based on mean location

Two genes may have the same outlier probability, but the variance of one gene is greater than that of the other. Simple quantification of $OS_{raw}$ does not account for the degree of certainty affected by the variance of points. Figure S2.6 illustrates two genes with similar $OS_{raw}$ and significantly different variances in expression. A gene with very high variance can lead to significant $OS_{raw}$ regardless of the sample mean values. To adjust for these differences in variance, we chose to adjust $OS_{raw}$ since it is less certain that genes with high variance are meaningfully deviating from the CER. Since $OS_{raw}$ has a maximum value of 1, we devised a ranked adjustment which also ranges from 0 to 1. We implement this by subtracting the adjustment from the $OS_{raw}$ to determine $OS$ (eq. 10). The correction is determined as

$$OS_i = OS_{raw,i} - \frac{{d_i}^{rank}}{m_{inner}} \quad \textbf{(2.10)}$$

where ${d_i}^{rank}$ is the nearest distance from i$^{th}$ gene mean expression to CER, and $m_{inner}$ is the number of genes closer to CER than ith gene. The adjustment is determined by finding the nearest distance from each gene (mean in both conditions) to the CER and rrank-orderingall genes with a mean expression falling within the CER. Therefore, genes that are near the center of the CER will be penalized the most whereas genes near CER undergo sa mall reduction in $OS$. Figure S2.7 illustrates the change in $OS$, and the adjustment largely eliminates the effect of

high variance on $OS$. Figure S2.8 shows that $OS$ is adjusted due to the distance to CER, and genes interior to CER is unlikely to be called as an outlier after this adjustment.

### 2.3.6. Filtering low/high expression $OS$ values

Often there is a significant correlation of expression between 2 samples. Because our method considers the density of the expression probability, genes that are very lowly or highly expressed in both samples will receive a high $OS_{raw}$. Housekeeping genes may fall into this category since their expression tends to be consistent across many conditions [132]. These genes may or may not be biologically relevant in the mechanism of interest. Therefore, we provide the option to ignore these genes by setting the $OS$ for lowly/highly expressed genes to zero. The threshold is determined based on the target containment. For example, if the target containment is 95%, then the lowest 2.5%, and highest 2.5% genes will receive an $OS$ of zero. Filtered genes are shown in Figure S2.9 where genes with high/low expression are automatically eliminated from MAGE analysis.

### 2.3.7. FDR estimation

To estimate the rate of false discoveries in datasets where the true FDR is unknown, we used a standard permutation-based method by running MAGE while using permuted sample classes. Randomly assigning treatment/control labels guarantees a true null hypothesis and therefore any predicted gene signatures are considered falsely discovered. The ratio of false gene signatures to total gene signatures is found to determine the FDR (eq. 11-12). To avoid overestimation of FDR, we used the modification proposed by Xie et. Al. of only using non-signature genes for the estimation of FDR [133].

$$S = \sum_{i=1}^{m}(t_i > C) \quad \textbf{(2.11)}$$

$$\widehat{FDR} = \frac{S'}{S + S'} \quad \textbf{(2.12)}$$

where $S$ is the number of gene signatures. $t_i$ and $C$ are the test statistic (FC, p-value, $OS$, etc.) for gene I and the test statistic classification threshold, respectively. FDR estimation as a function of $OS$ is shown in Figure S2.9.

## 2.3.8. Effects of noise

RNA-seq data is susceptible to many sources of noise which can arise from errors during transcription or splicing [134], PCR amplification biases [79], and barcode swapping during library preparation [135]. Noise can either introduce false positives by elevating the test statistics across a broad set of genes, or true-positive genes may be missed as the statistical power is reduced. We tested the performance of MAGE against the standard 2 sample t-test to analyze the $\gamma$ - T3 breast cancer profile with different amounts of Gaussian noise introduced (Figure S2.9). The noise was introduced by adding a normally distributed random number with SD ($\sigma_n$) for each gene (eq. 13) as

$$x'_{i,j} = x_{i,j} + \sigma_n z \quad \textbf{(2.13)}$$

where $z$ is a normally distributed random variable.

The effect of noise is shown by a widening of the CER (Figure S2.10), which results in an overall reduction in $OS$. Figures S2.11-S2.12 show the performance of MAGE and identified AEGs in two different datasets. DE analysis (as described below) identifies more DEGs with the presence of noise, but AEG identification remains conservative for small to moderate levels of noise introduced. These results suggest that MAGE performs well with noisy data.

## 2.3.9. Density-based clustering vs MAGE

MAGE works similarly in principle to density-based classification methods, and we performed a side-by-side comparison between a widely used density-clustering algorithm, DBSCAN [125]. DBSCAN performs solely on individual data points without consideration for variance, and this constitutes a major downside in the analysis of gene expression as there is no way to incorporate multiple replicates to improve predictions. For the input to DBSCAN, we used the mean expression of each gene within both conditions. Figure 2.3 shows the comparison between MAGE and DBSCAN. DBSCAN can find the outlier without the consideration of sample variance whereas MAGE can identify interior genes within CER as AEGs depending on variance.

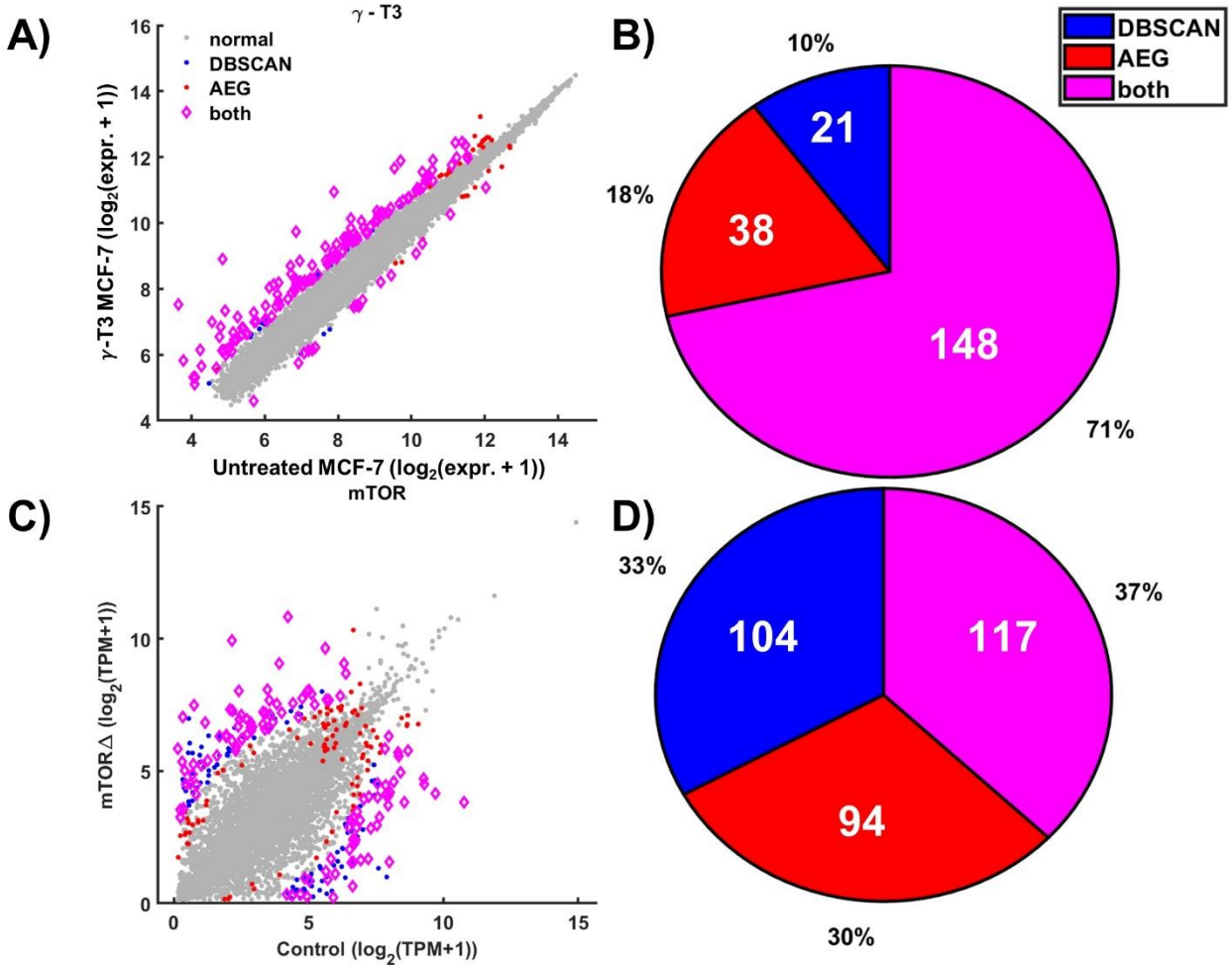


**Figure 2.3. Comparison of DBSCAN and MAGE.** (A) Gene mean values from the $\gamma$-T3 breast cancer profiles (data from GSE21946). Blue, red, and magenta marked points indicate genes

identified by DBSCAN ($\varepsilon = 0.3, minpts = 200$), AEG ($OS > 0.65$), and both, respectively. (B) Overlap between DBSCAN and MAGE for $\gamma$-T3 breast cancer profile genes. (C) Gene mean values from the mTOR KO mouse profile (data are from GSE134316). Signature genes were identifi by DBSCAN ($\varepsilon = 0.7, minpts = 80$), AEG ($OS > 0.1$). (D) Overlap between DBSCAN and MAGE for mTOR KO mouse profile genes.

### 2.3.10. Pathway/GO enrichment

To evaluate the biological significance of identified signature genes, we performed gene set enrichment of the gene ontology (GO) biological process (BP) terms [10] and KEGG pathways [136] using DAVID [65,137]. Significantly enriched pathways were selected and sorted using the Benjamini and Hochberg false discovery rate to account for multiple testing.

### 2.3.11. Identification of DEGs

To compare the performance of MAGE, we also identified DEGs by assessing the logarithmic fold-change (FC) along with a standard 2-sample t-test [138]. The test statistic t was determined for each gene (eq. 14) and used to find the 2-tailed p-value representing the probability of a deviation in mean expression for a single gene across the 2 sample conditions. DEGs were selected by finding genes with a p-value below 0.05 and an FC above a specified threshold.

$$t = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{{\sigma_a}^2}{m_a} + \frac{{\sigma_b}^2}{m_b}}} \quad \textbf{(2.14)}$$

### 2.3.12. Data collection and preparation

We used NCBI GEO (GSE21946), which consists of human breast cancer samples (MCF-7 cells) subjected to treatment with gamma-tocotrienol ($\gamma$ - T3) [139] and GSE134316 mouse mTor knockout cells. Each profile was log-transformed, and we filtered out low read count genes. GSE21946 is a microarray dataset, and data contains expression levels from 22,277 genomic loci. We averaged the expression of multiple probes from the same gene, reducing the profile to 14,054 genes. Genes were filtered to ensure they contained non-zero expression in 6 out of 8 samples, leaving 13,639 genes for DE and AE analysis. $OS$ and FDR values were determined using MAGE, and Figure 2.4 illustrates the CER and distribution of $OS$. Figure S2.9A shows the FDR as a function of $OS$ threshold for AEG classification. Mouse RNA-seq profiles (GSE134316) have 3 healthy control samples and 3 mTOR knockout samples taken from mouse bone marrow [140]. The initial set of 49,431 genes was filtered to remove genes containing zero expression in all samples, and the remaining 34,358 genes were log-shifted.

### 2.3.13. Code availability

All data preparation, processing, and figures were performed in *MATLAB version R2021a* [141]. All code used in this study is available on GitHub github.com/beltranmm/MAGE

## 2.4. Results and Discussion

### 2.4.1. Analysis of human breast cancer microarray data

To compare the performance of MAGE and the conventional DEG identification, we used both approaches to analyze a simple two-condition microarray expression profile. This dataset, obtained from the NCBI GEO (GSE21946), consists of human breast cancer samples

(MCF-7 cells) subjected to treatment with gamma-tocotrienol ($\gamma$ - T3) [139], an antioxidant and form of vitamin E known for its demonstrated inhibition of tumor growth in various types of cancers [142]. Since this data consists of only a single drug treatment that targets a single pathway, there are few differentially expressed genes. Figure 2.4 shows the identification of AEGs, which are distributed mostly outside of the CER, and the distribution of $OS$..

Figure 2.5 compares AEGs and DEGs based on FC, p-value, mean expression and $OS$. We see significant AEG/DEG agreement since there is a correlation between FC and $OS$ (Figure 2.5C). MAGE assigns higher $OS$ to genes with higher mean expression (Figure 2.5B) while the conventional method favors genes with lower expression. The preference for selecting higher expressed genes is explained by the disproportionate number of lowly and highly expressed genes. Since our method considers CPDF of all genes, the higher expression region will naturally have a lower density, resulting in a narrowing of the CER (Figure 2.4C), and is, therefore, more likely to have a higher $OS$.

Next, we compared the biological significance of both DEGs and AEGs (Table S2.1) by selecting genes with low (<0.05) AE FDR and high (> 0.05) DE FDR. As well as genes with low DE FDR and high AE FDR. We performed GO enrichment analysis using DAVID on each gene set (Table S2.2). Figure 2.5D shows the top 10 enriched terms based on the set of AEGs and DEGs respectively. The GO analysis showed a prevalence of terms related to cellular stress response, protein transport, gene expression regulation, and apoptotic processes in common. Notably, several of the top AEG enriched terms are related to the ER stress response to unfolded protein accumulation which has been recognized as a mechanism of tumor progression in multiple cancers, but most prominently in breast cancer [143]. Gamma-tocotrienol has been suggested to possess inhibitory effects on cyclin-dependent kinases (CDKs), which play a major role in regulating the cell cycle. CDKs control cell-cycle progression by initiating phosphorylation of the RB protein which in turn regulates E2F transcription factors

and induces transcription in the set of genes responsible for the G1/S cell-cycle transition [144]. This possibly explains the prevalence of AEGs related to cell cycle phase transition. Additionally, $\gamma$ - T3 has been shown to have a synergistic effect when combined with chemotherapy to increase apoptosis among breast cancer cell lines including MCF-7 cells [145]. Apoptotic regulation appears enriched for both AEGs and DEGs. As expected, the enrichment results from DEGs were very similar to the AEG signatures (Figure S2.14). This supports that for experiments with small changes in the expression profile, MAGE performs similarly to the conventional t-test and both methods can identify disease-relevant signature genes.

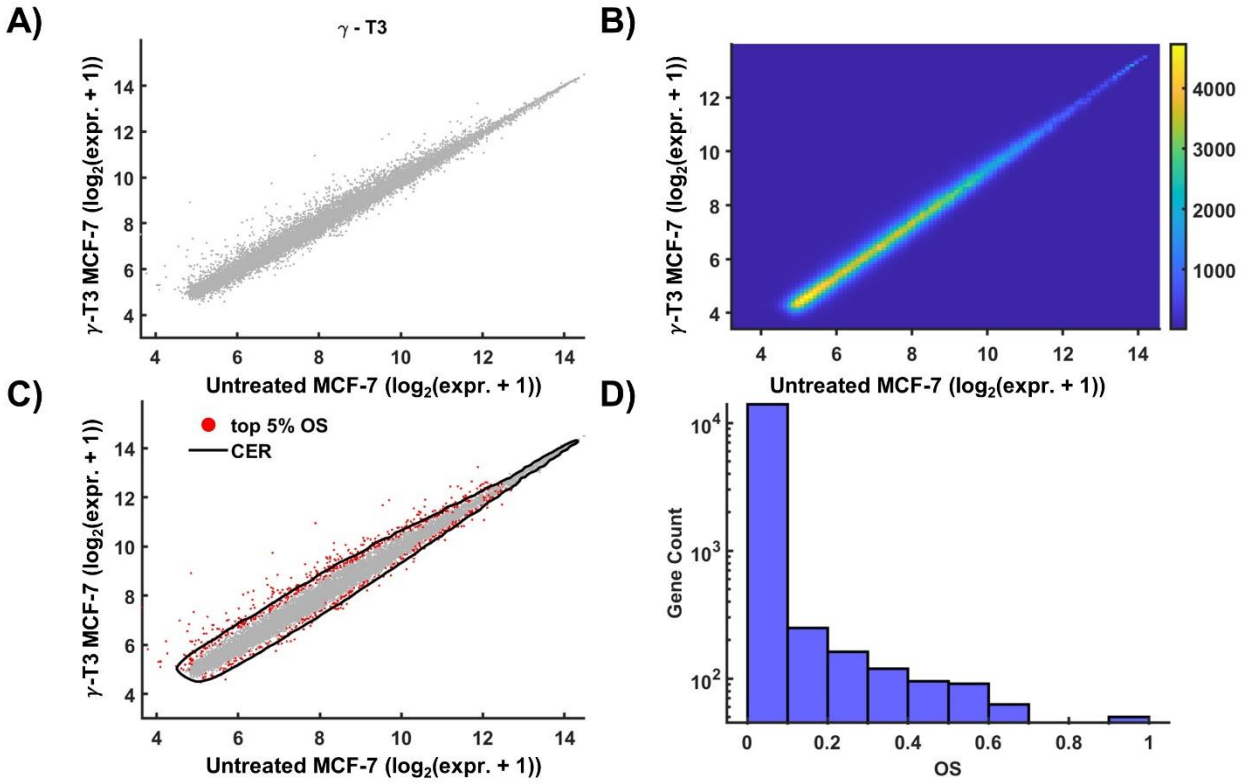


**Figure 2.4. MAGE applied to the breast cancer $\gamma$ – T3 treatment profile.** (A) Mean RNA-seq TPM levels of each gene per sample type after filtering. (B) CPDF from all genes. (C) CER based on the CPDF. (D) The distribution of $OS$ from 1,000 MC sampling per gene.

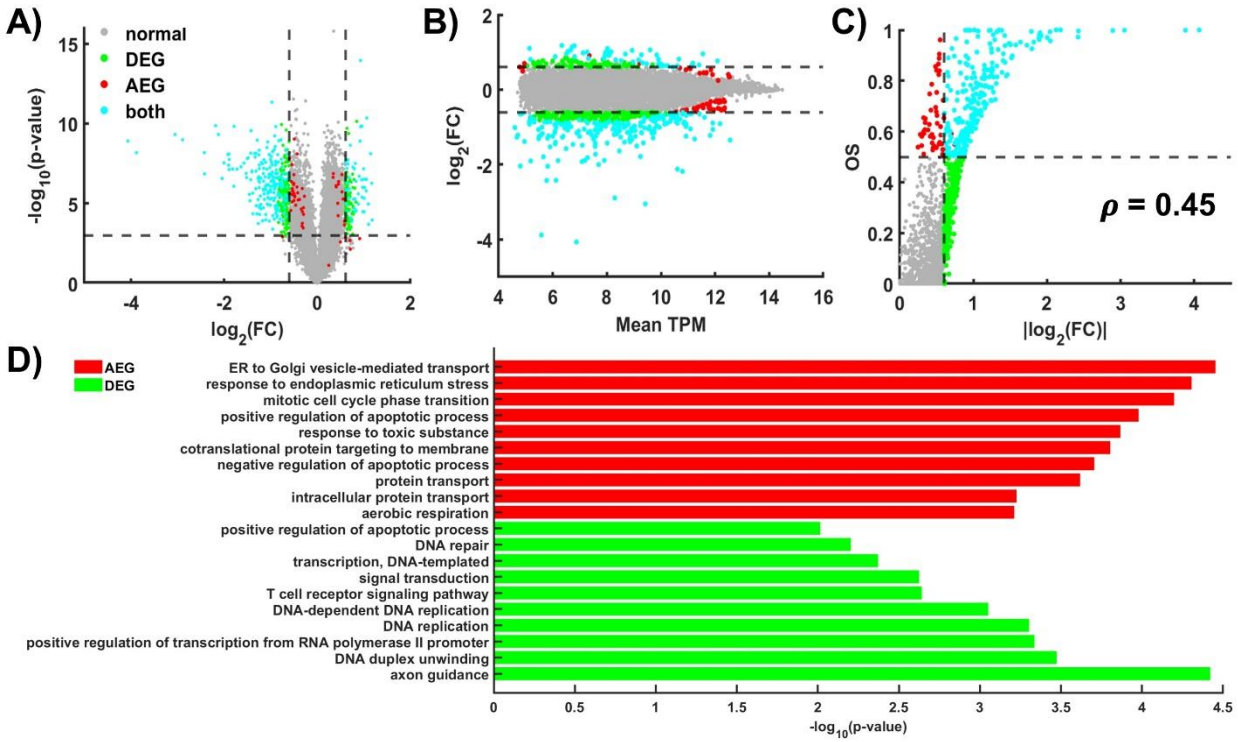


**Figure 2.5. Comparison of MAGE and DEG on the breast cancer $\gamma$ – T3 treatment profile.** (A) Volcano plot based on the raw/uncorrected p-value of DEG and FC. (B) TPM vs FC scatter plot. (C) Relationship between FC and $OS$. Colors indicate if a gene is identified as DEG and/or AEG. (D) Top 10 GO terms enriched by AEGs and DEGs.

### 2.4.2. Analysis of *mus musculus* mTor knockout RNA-seq data

To test the performance of MAGE with highly different gene expression profiles, we used mouse RNA-seq profiles (GSE134316) from healthy and mTOR knockout samples taken from mouse bone marrow [140]. The mechanistic target of rapamycin (mTOR) is one of the master regulators for growth and nutrient signaling, which regulates the global transcriptome [146]. Because of the categorization of mTOR as a master regulator, we expected to see a significant change in the expression of many genes between the two sample conditions. Figure 2.6 shows

the expression profile with and without mTor, and as expected, there was a large variation in gene expression between the two conditions (Figure 2.6A) as well as an increase in the reported FDRs from MAGE (Figure S2.9B). Since there are a larger number of genes with high FC and moderate expression levels, the CER is stretched to encompass many of the genes that are often identified as DE (Figure 2.6B-C).

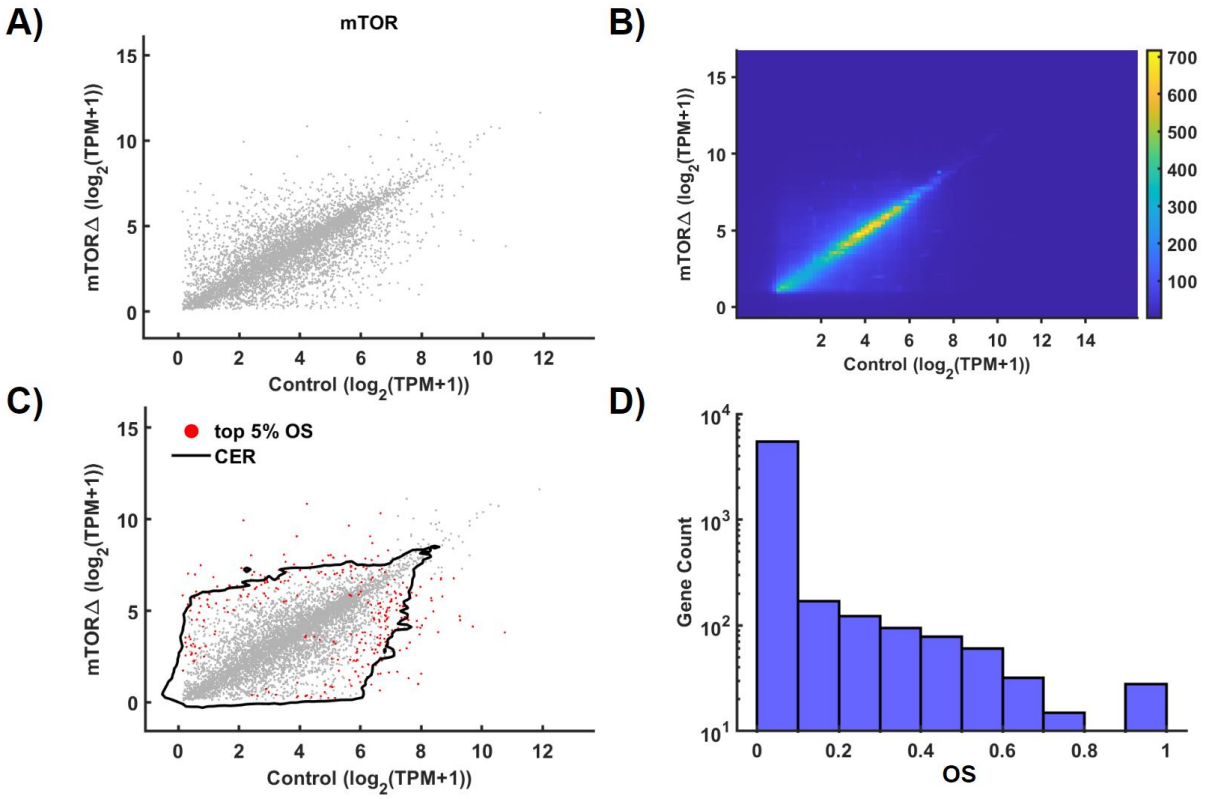


**Figure 2.6. MAGE applied to the mTOR KO mouse profile.** (A) Mean RNA-seq TPM levels of each gene per sample type after filtering. (B) CPDF from all genes. (C) CER based on the CPDF. (D) The distribution of $OS$ from 1,000 MC sampling per gene.

Figure 2.7 compares AEGs and DEGs (Table S2.3), which shows significant agreement between the methods. There is still a positive ~0.47 correlation between FC and $OS$ (Figure 2.7C), which is similar to the ~0.45 correlation in the $\gamma$ – T3 profile. To assess biological significance, we performed GO analysis on AEGs and DEGs. We found the set of AEGs to be

enriched for genes related to cytoplasmic translation, cellular respiration, and oxidative phosphorylation (Table S2.4). This association was significantly less for the set of DEGs (Figure 2.7D). Since mTOR has been associated with regulating both general and preferential mRNA translation [147] as well as the mitochondrial energetic adaptation [148], we consider this to be a biologically relevant gene set that is identified using MAGE. To support this, we also found an enrichment of AEGs and DEGs associated with cell division/cell cycle, and protein transport and folding. All of which have been associated either directly or indirectly with mTOR regulation [149].

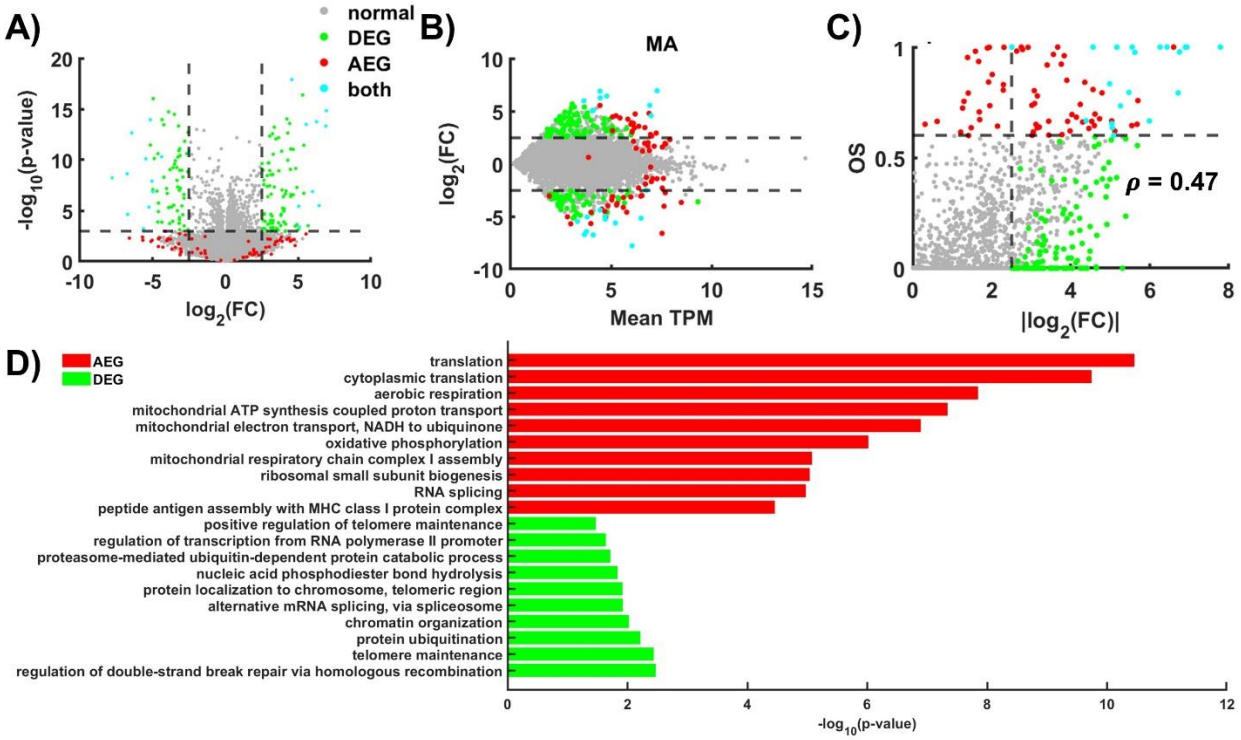


**Figure 2.7. Comparison of MAGE and DEG on the mouse mTor KO profile.** (A) Volcano plot based on the raw/uncorrected p-value of DEG and FC. (B) TPM vs FC scatter plot. (C) Relationship between FC and $OS$. Colors indicate if a gene is identified as DEG and/or AEG. (D) Top 10 GO terms enriched by AEGs and DEGs.

## 2.5. Discussion

### 2.5.1. Interpretation of MAGE expression

It is important to note that differentially expressed genes and aberrantly expressed genes are different. Differential expression implies significant up/down-regulation of a gene which is usually measured by the logarithmic fold-change. Here we define an aberrant expression as a deviation from the group. Therefore, our main assumption is that the majority of genes should not be considered gene signatures and that the genes of biological interest will be found further from the majority. The sets of DEGs and AEGs in many cases will contain significant overlap. However, depending on the types of samples being studied these sets may be significantly different. Both differential expression and aberrant expression may provide potential insight into the features responsible for the biological variation of samples. Therefore, MAGE is not intended to be a replacement or improvement to differential expression analysis, but to be used as an alternative analysis, particularly in cases where differential expression may not be the most informative.

We have also assumed a log-normal distribution for the mean expression of a gene across samples. Although this may not capture the long-tailed nature of some genes, we use this assumption to construct a probabilistic landscape of the cumulative probability distributions from every gene in the experiment.

### 2.5.2. Potential application in single-cell sequencing

Alterations in the transcriptional programs that take part in the onset and progression of diseases can go unnoticed when bulk tissue samples contain highly heterogeneous mixtures of cell types. Recent studies have utilized flow cytometry and subsequent RNA-seq to examine transcriptional evolution throughout disease progression and identify biomarkers that only occur

in individual cell types [150]. However, scRNA-seq introduces many difficulties in generating informative and consistent results. This is because read counts taken from single-cell profiles are notoriously low which can often produce a high expression bias in the DE classification [74]. Furthermore, in the commonly used parametric methods for DE analysis, the assumptions necessary for reliable results are often not met in single-cell studies. Stochastic switching between gene network ‘on’ and ’off’ states becomes more apparent at the low levels of RNA species present in most single-cell profiles. This leads to bi-modal behavior exhibited in scRNA-seq that is often unobserved in bulk [151]. MAGE is designed to work well in the presence of bi-modal expression distributions and can easily identify genes found between states. The ability of MAGE to identify aberrant genes without the need to verify prior distribution assumptions would seem to make this a promising method for scRNA-seq analysis.

## 2.6. Conclusion

We have presented a novel methodology aimed at addressing the limitations of conventional DEG analysis. By analyzing the expressional probability overlap between the genes of 2 samples, our method offers a unique perspective, focusing on the identification of genes exhibiting aberrant expression patterns rather than solely examining differential expression. Through extensive validation using diverse datasets, we demonstrated the robustness and applicability of this approach across various experimental conditions related to cancer.

This methodological shift towards evaluating gene expression based on the deviations of genes relative to the entire profile offers promising insights into understanding disease mechanisms. The ability to identify functionally relevant gene signatures, those exhibiting aberrant expression patterns, presents opportunities for novel biomarker discovery. Furthermore, our approach's adaptability for large-scale omics datasets while remaining

computationally feasible enhances its potential for widespread application in diverse biological studies. However, while our method showcases significant promise, there remain avenues for further refinement and validation. Future research should focus on refining the algorithm, particularly its adaptability to single-cell sequencing data, and exploring its utility across various disease contexts beyond cancer. Additionally, continued efforts in validating identified gene signatures and their biological relevance will be essential for translating these findings into clinical applications.

In essence, MAGE represents a paradigm shift in identifying aberrant gene expression. Rather than viewing our method as an improvement to existing DE methods, we believe it should be considered as a complementary technique capable of returning consistent results in ambiguous scenarios where assumptions are uncertain. Its potential to unveil biologically significant gene signatures holds promise for advancing our understanding of disease mechanisms and fostering the development of precision medicine strategies.

# 3. Single-cell applications of MAGE

## 3.1. Motivations

Single-cell RNA sequencing (scRNA-Seq) is a genomic approach to quantify the cell-to-cell heterogeneity of gene expression that is often overlooked in bulk samples. As discussed earlier, the analysis of scRNA-Seq data proves to be challenging due to the large portion of untranscribed and lowly transcribed genes as well as technical variation among samples. This causes the sample variability to be significantly higher than what is normally seen in bulk profiles [152]. Conventional DE approaches rely on distributional assumptions that are difficult to validate in single-cell profiles and are often not well-suited for reliable prediction [153]. Because of our previous results demonstrating conservative AEG identification in the presence of noise using MAGE, we believed scRNA-Seq may be a perfect case where MAGE is more robust than DEG methods in identifying biologically relevant genes. One of the major challenges with single-cell RNA profiles is that sample-to-sample variability makes it difficult to get a reliable number of reads for lowly transcribed transcripts as many genes exhibit zero reads. Therefore, sample size plays a critical role in determining how informative subsequent analysis can be [154]. Here we test if the MAGE pipeline works robustly for scRNA-Seq samples, and we assess the number of samples required to produce consistent results.

## 3.2. Data collection and preprocessing

To test the MAGE analysis, we used scRNA-Seq profiles of single cells taken from mouse blood vessels in the brain and lung. Cell types were sorted using FACS into two main groups of endothelial cells and mural cells for both the brain and lung. Single-cells were isolated and sequenced using Illumina HiSeq technology [155]. We collected the data from the NCBI GEO ascension code GSE98816 for the brain profile and GSE99235 for the lung profile. Both

brain and lung profiles consist of roughly 20,000 genes with reads measured in 3,186 brain and 1,504 lung cells. Both profiles were filtered to remove genes that did not contain more than 1 read in at least 100 samples. Note, that the threshold for the number of reads did not make a significant difference as expected from the zero inflation problem common to single-cell data [76]. The brain profile contained mural, endothelial, and other extra cell types, but we only analyzed the larger samples of mural and endothelial cells, which make up the majority of brain cells. After filtering, the brain profile consisted of 10,461 genes and 2,929 samples, and the lung profile contained 9,973 genes and still 1,504 samples. Genes were then cross-referenced and kept only if found in both profiles leaving 9,237 genes for subsequent analysis. In total across both brain and lung, there were 2,219 mural cells and 2,214 endothelial cells.

## 3.3. DEG and AEG identification

To compare DEG and AEG identification, AEGs were identified using MAGE as described in Chapter 2. MAGE parameters were set to default values with the target containment of 0.95, grid density of 100, and 5 contours per iteration. The upper and lower 2.5% extremes of the expression were disregarded for OS quantification. The mean expression within brain and lung cells is shown in Figure 3.1, along with the determined CER and genes within the highest 5% of OS values. We noticed that the variance within these profiles is significantly larger compared to the previously examined bulk samples. This forces the CER to be quite large and encompasses the vast majority of gene mean values, lowering the overall OS distribution. We had tested several other single-cell profiles previously and saw similar results. DEG analysis was performed as described in Chapter 2 as well. The correlation between OS and FC is still present but slightly reduced compared to the previously analyzed bulk samples (Fig. S3.1 B). Interestingly, the genes with the highest OS tend to be found near the outer periphery of the volcano plot (Fig. S3.1 A). This means that at a given p-value, OS tends to increase with FC.

The higher OS in higher expressed genes, first observed in the previous microarray and RNA-Seq profiles, is also apparent in the single-cell profile. Additionally, we notice a skew of higher expression in the brain compared to the lung (Fig. S3.1 A, C). We performed the sample permutation FDR test as described in Chapter 2 to decide an appropriate OS classification threshold for AEGs (Fig. S3.2).

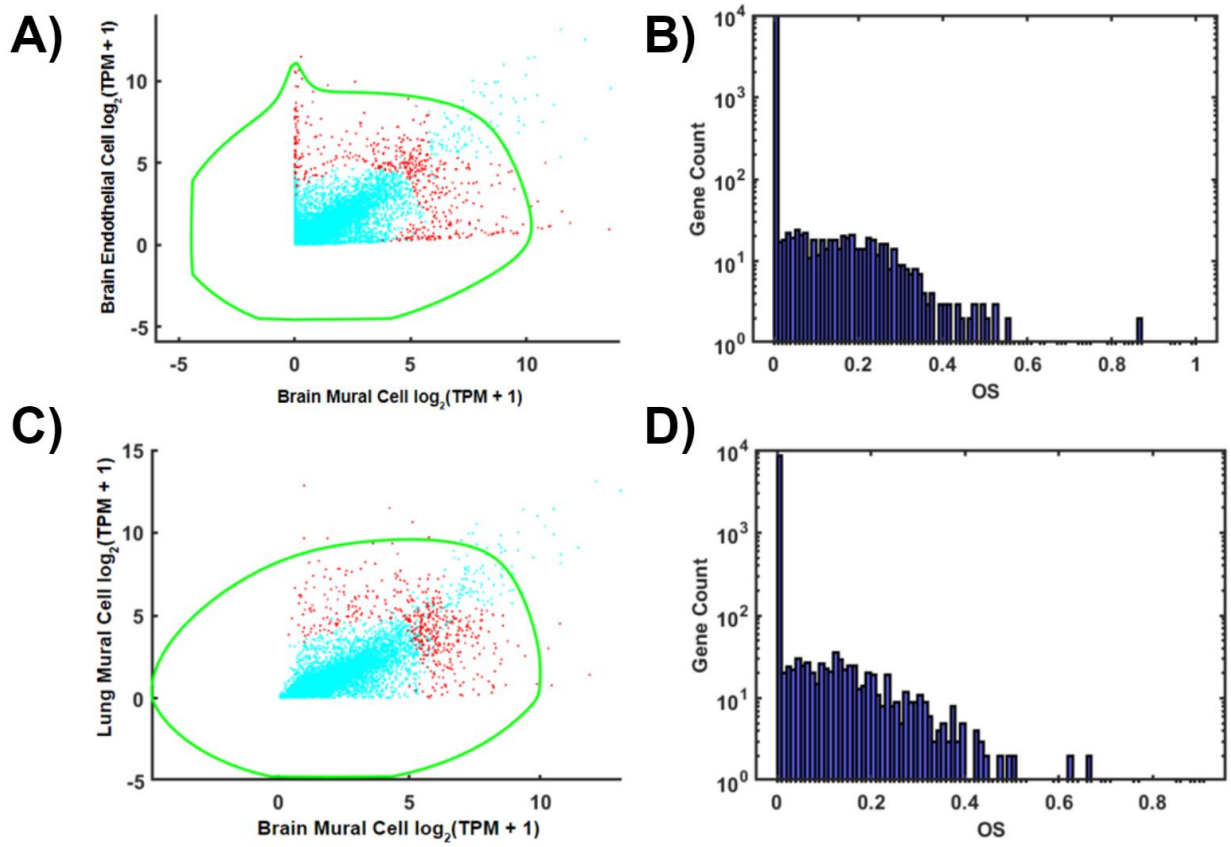


**Figure 3.1. Comparing brain mural, brain endothelial, and lung mural cells using scRNA-seq.** (A-B) MAGE analysis of brain mural and endothelial cells. (A) Mean expression of each gene and CER used to assess AE. (B) Distribution of OS. (C-D) MAGE analysis of brain mural and lung mural cells. (C) Mean expression of each gene and CER used to assess AE. (D) Distribution of OS.

## 3.4. AEGs conserved at lower sample size

Next, we asked how the sample sizes affect the robust AEG identification and compared its performance with that of conventional DE analysis. An important question in DE and AE analysis is how many samples are required for reliable DEG/AEG identification. A greater number of samples is always better but with a diminishing return on preparation costs. Previous studies analyzing bulk profiles have noted that the optimal sample size is highly dependent on the variability of the data and in cases of low variability relatively few samples (<10) are necessary for reliable results [47]. Since SC profiles tend to have significantly higher variability, we expected to see the need for a higher number of samples to provide consistent predictions. We were also interested in investigating any differences in the number of required samples between AE and DE. To test this we performed a random sample permutation in both of the SC profiles and assessed the change in the OS distribution as well as the CER used to quantify OS. To quantify CER similarity, we used MC area estimation to find the area overlap between the CER found using all samples and the CER found using a subsampling. CER similarity was determined using equation 3.1. To quantify any shifts in OS distribution we simply calculated the percent difference between the mean OS found using all samples compared to using subsampling. To test for the number of samples required for stable AE/DE prediction, we identified the top 5% AEGs and DEGs using each number of random samples and found the percentage of agreement with the top 5% found using all samples. To consider the variability of these metrics we performed 4 trials for each with separate sampling in each trial.

$$CER\ similarity = \frac{pts_i \cap pts_{ref}}{pts_i \cup pts_{ref}} \quad \textbf{(3.1)}$$

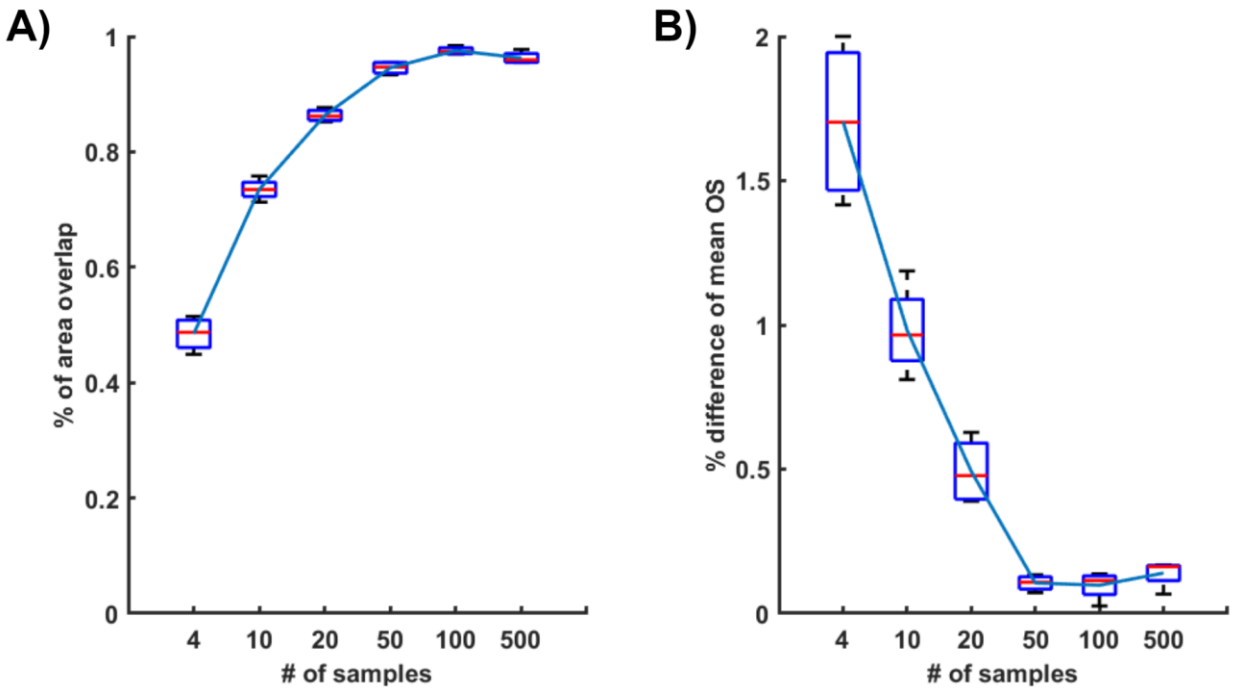


**Figure 3.2. MAGE performance consistency by subsampling.** (A) CER similarity in terms of the area of overlap between the reference CER (all samples) and the CER determined with the indicated number of randomly subsampled genes. The box plot was from 4 subsamplings. (B) Percent difference comparing the mean OS from the reference and with the indicated number of randomly subsampled genes. Data is brain and lung mural cells scRNA-Seq.

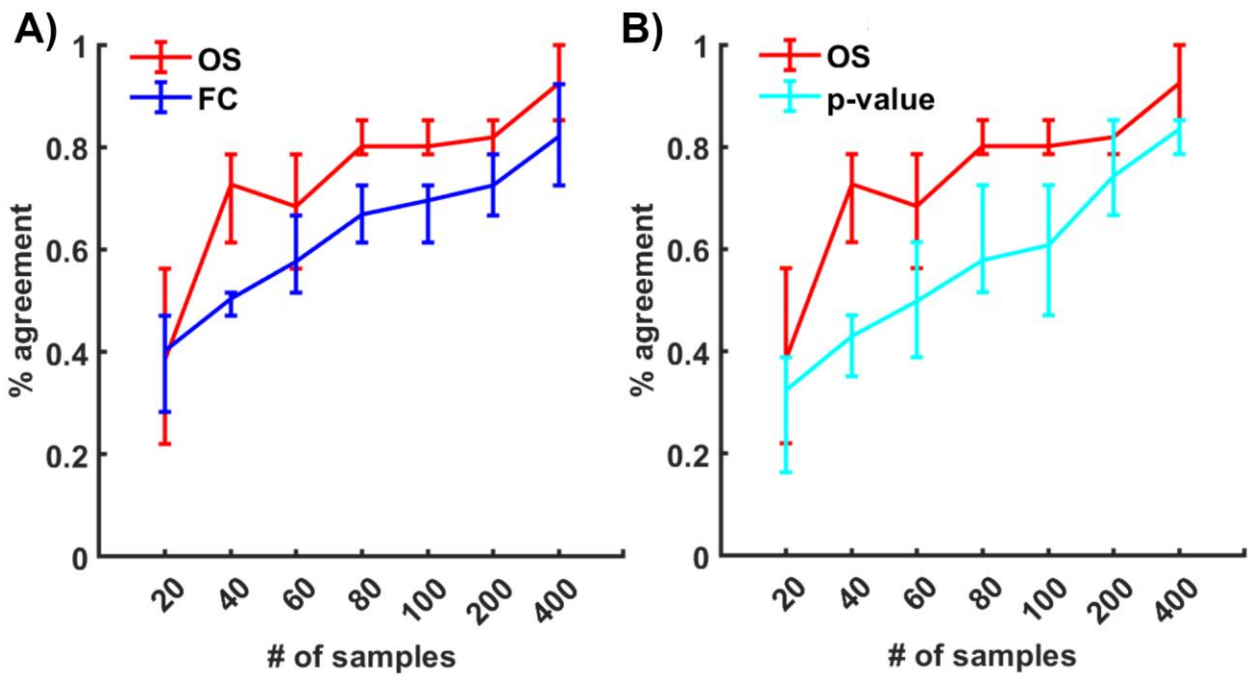


**Figure 3.3. Robustness of AEG and DEG identification by the number of samples.** (A) Robustness of AEG and DEG identification between brain and lung mural cells by the fraction of overlap of the top 5% genes for AEG (red) and DEG (blue), respectively. Lines represent mean values over the 4 trials, and error bars represent the maximum and minimum out of the 4 trials. (B) Consistency of genes determined by the lowest 5% p-value (cyan) and highest OS (red).

Figure 3.2 shows that the increase in both the CER similarity and the mean OS stagnated before 100 samples. This demonstrates that for this particular data, MAGE requires only 100 samples and any further sampling will not lead to significant improvement in performance. Comparing the 5% classification consistency graphs (Fig. S3.3), we notice that genes classified by OS (AEGs) tend to be much more stable with 50 samples (73% +/- 6%) compared to both FC (52% +/- 9%) and p-value (40% +/- 5%) (DEGs). To examine this further we ran the analysis again focusing on the range of 20-500 samples (Fig. 3.3, S3.4) and noticed that OS classification reaches 75% +/- 5% consistency with only 40 samples, while genes classified by FC or p-value do not reach similar consistency until 200 samples are used.

## 3.5. Pathway enrichment

To test whether identified AEGs are biologically relevant, we performed the pathway enrichment of AEGs and compared it to that of DEGs. After quantifying the OS, FC, and p-values we performed pathway enrichment to investigate what gene functional associations are selected by AE and DE. For both the mural cells and endothelial cells, we identified exclusive (not DE) AEGS (exAEGs) by sorting for genes with an OS greater than 0.1 and an FC less than 2. Here the exAEGs represent the set of genes identified by AE but missed in conventional DE approaches. We also sorted for exclusive DEGs (exDEGs) by selecting genes with an OS below 0.1 and an FC above 2.5. These thresholds selected roughly 200 genes for each list. We performed pathway enrichment on each set separately using DAVID [65]. Table 3.1 shows the pathway enrichment results for exAEGs. The exDEGs did not provide any significant (FDR < 0.05) enrichment results. We interpret this as promising evidence that AE is able to identify functionally relevant genes that are missed by DE analysis, while also retaining the most significant functional DEGs. We investigated some of the individual exAEGs and one of the clear examples is CXCL12, which contained an OS of 0.36 and an FC of only 0.70, making it one of the highest AEGs will low FC. CXCL12 plays a significant role in brain development, particularly through angiogenesis [156].

| Term | Count | % | PValue | Fold Enrichment | FDR |
|---|---|---|---|---|---|
| GO:0001525~angiogenesis | 22 | 11.3 | 6.7E-10 | 5.3 | 1.3E-06 |
| GO:0007155~cell adhesion | 21 | 10.8 | 6.8E-08 | 4.3 | 6.8E-05 |
| GO:0019221~ cytokine-mediated signaling pathway | 9 | 4.6 | 2.2E-05 | 7.4 | 1.5E-02 |
| GO:0003197~endocardial cushion development | 5 | 2.6 | 3.4E-05 | 23.4 | 1.7E-02 |
| GO:0071711~basement membrane organization | 6 | 3.1 | 4.6E-05 | 14.1 | 1.8E-02 |
| GO:0030335~positive regulation of cell migration | 13 | 6.7 | 1.7E-04 | 3.7 | 5.8E-02 |

**Table 3.1. Pathway enrichment results from single-cell mouse endothelial brain/lung cell exAEGs identified by MAGE (OS > 0.1, FC < 2).**

## 3.6. Conclusion

Through our investigation of scRNA-Seq profiles from mouse brain and lung blood vessels, we have demonstrated that MAGE performs similarly to traditional methods for DE analysis in capturing meaningful gene expression patterns. We also highlighted the identification of relevant AEGs that are not considered DE. This evidence supports the use of MAGE as a robust tool with potential for exploratory analysis of gene expression on the single-cell level.

Notably, our analysis revealed that AE, as quantified by MAGE, is more stable for reliable predictions using fewer number of samples, highlighting its efficiency and applicability in studies with limited sample sizes. The quantification of AE holds promise for advancing our knowledge of complex biological systems and ultimately driving discoveries in health and disease.

# 4. Future work and conclusion

We have introduced and tested a novel methodology for identifying genes relevant to cancer progression. By combining concepts from previous statistical and machine learning approaches we have developed a robust quantification of aberrant gene expression patterns and verified that these patterns are relevant to the physiological differences within the data.

## 4.1. Summary of MAGE contributions

We hope that the ideas brought about during the development of MAGE will continue to be explored by future studies. These ideas include:

- Exploration of AE as a metric for screening potential biomarkers, and therapeutic targets, and for gaining mechanistic insights into transcriptional programs. Rather than simply looking for changes in individual genes, AEGs are selected based on their deviation from all genes within the profile. This shift in perspective may prove to show interest in genes previously overlooked and understudied.
- Consolidation of the many available approaches for analyzing gene expression data. One can argue that whenever there exist many solutions to the same problem, likely none of them truly work. It is possible that by consolidating ideas from the methods that work best in specific cases, we can derive a general framework for analysis that applies to all expression profiles. In MAGE we have tried to incorporate some of the fundamentals of machine learning (particularly density-based clustering) and include some of the considerations of variability that is a key part of statistical analysis.

## 4.2. Limitations and possible improvements to AE analysis

AE analysis is still a recent concept for understanding gene expression. There still exist several limitations to the validity and significance of AEGs. In our development of MAGE, we sought to overcome or at least mitigate some of these limitations. However, future improvements are certainly possible. The limitations we identified include:

- Assumption of a Gaussian distribution for each gene's mean expression. We considered the idea of using other distributions (e.g. truncated Gaussian, Log-normal, and Gamma) to model the PDF for individual genes. However, to keep the analysis simple and consistent across data sets, all of our analysis was implemented with a Gaussian PDF. Other studies have looked at the use of multiple distributions to assess expression outliers. Evidence from these studies suggests that expression distributions can vary between genes even within the same profile and less than half of genes can truly be considered normally distributed [121]. Including a prior validation of optimal distributional assumptions before assessing AE would likely improve the functional relevance of AEG predictions. This may be especially true in single-cell datasets with high variability, but also large sample sizes.
- Disregarding the upper and lower extremes of the expression for AE. Since MAGE purely looks at the deviation of each individual gene's PDF from the CPDF of the whole profile, the genes that are unexpressed in both conditions or highly expressed in both conditions would always appear to be AE. To focus more on the genes with high FC we chose to disregard these sets from subsequent analysis. Perhaps considering FC as a direct parameter in the AE quantification algorithm would eliminate the need for an *ad hoc* adjustment.
- MAGE is more computationally demanding than conventional DE assessment. MAGE relies on multiple rounds of Monte-Carlo estimations for every gene. The

built-in MATLAB function "*inpolygon"* which is used to check whether or not each sampled point is within the CER is a fairly time-consuming algorithm [141]. For profiles containing more than a few thousand genes, the total processing time can become prohibitive compared to a simple t-test. Computational efficiency is always a consideration for big data pipelines and any improvement in compute times could lead to the further adoption of AE assessment in gene expression analysis.

- OS is highly dependent on user-selected parameters which makes cross-study comparisons difficult. Because MAGE, and ML algorithms in general, rely on user-specified parameters that may be subjectively chosen, it is nearly impossible to compare results from multiple studies performed using differing parameters. Part of the convenience and widespread use of statistical metrics such as the p-value, is the ease of interpretation. Of course, these metrics are also susceptible to unethical practices such as p-hacking and the problem of multiple testing as discussed in Chapter 1. We attempted to improve comparability by reporting FDR values determined by sample permutation. However, these methods are not as commonly used or interpreted.

# Appendix

## Supplementary figures for Chapter 2

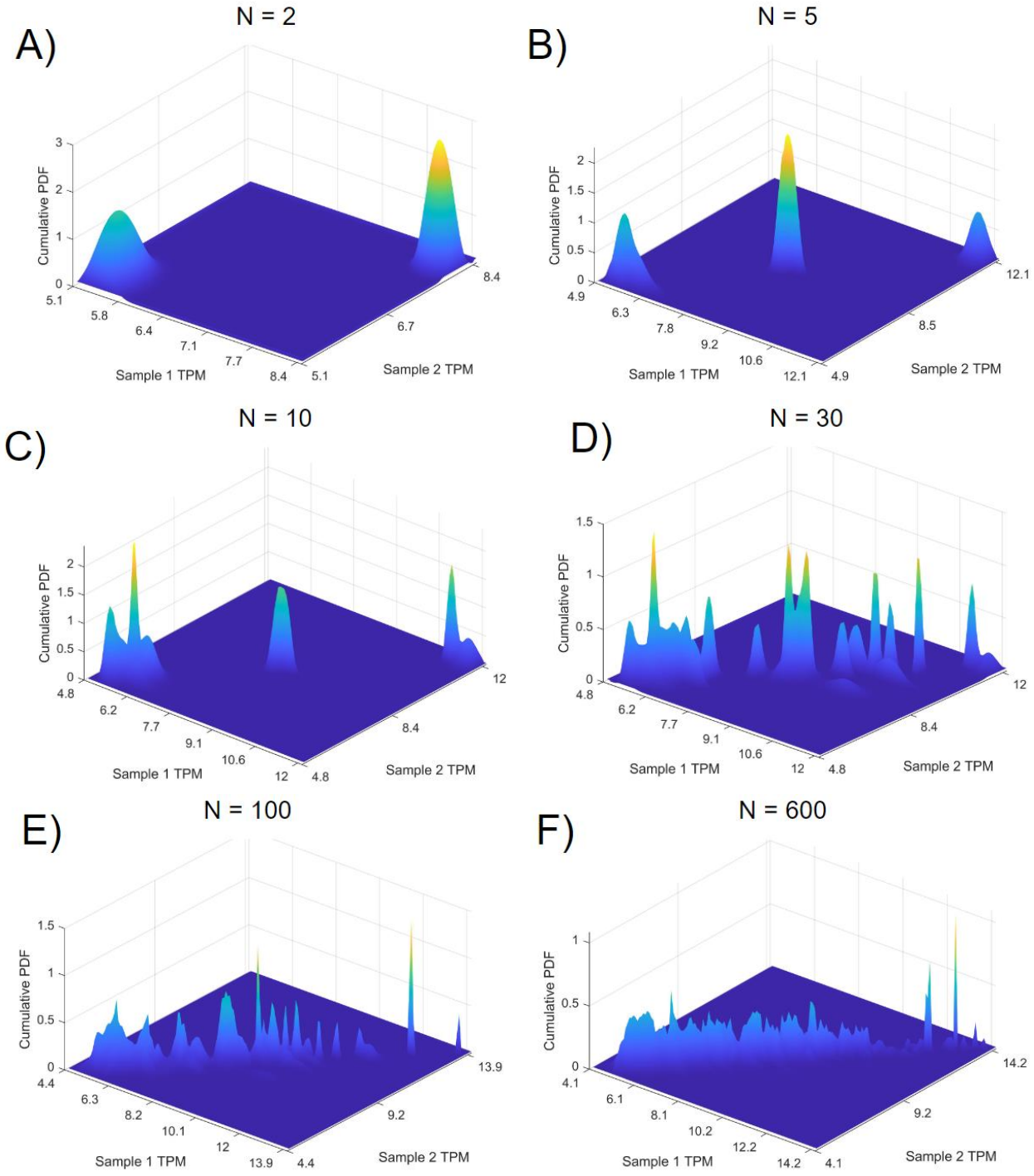


**Figure S2.1. Cumulative PDF as a function of the number of genes.** Surface plots of the probability density matrices formed by running the topological analysis of the breast cancer $\gamma$-T3 treatment profile (GSE21946) with the indicated number of selected genes.

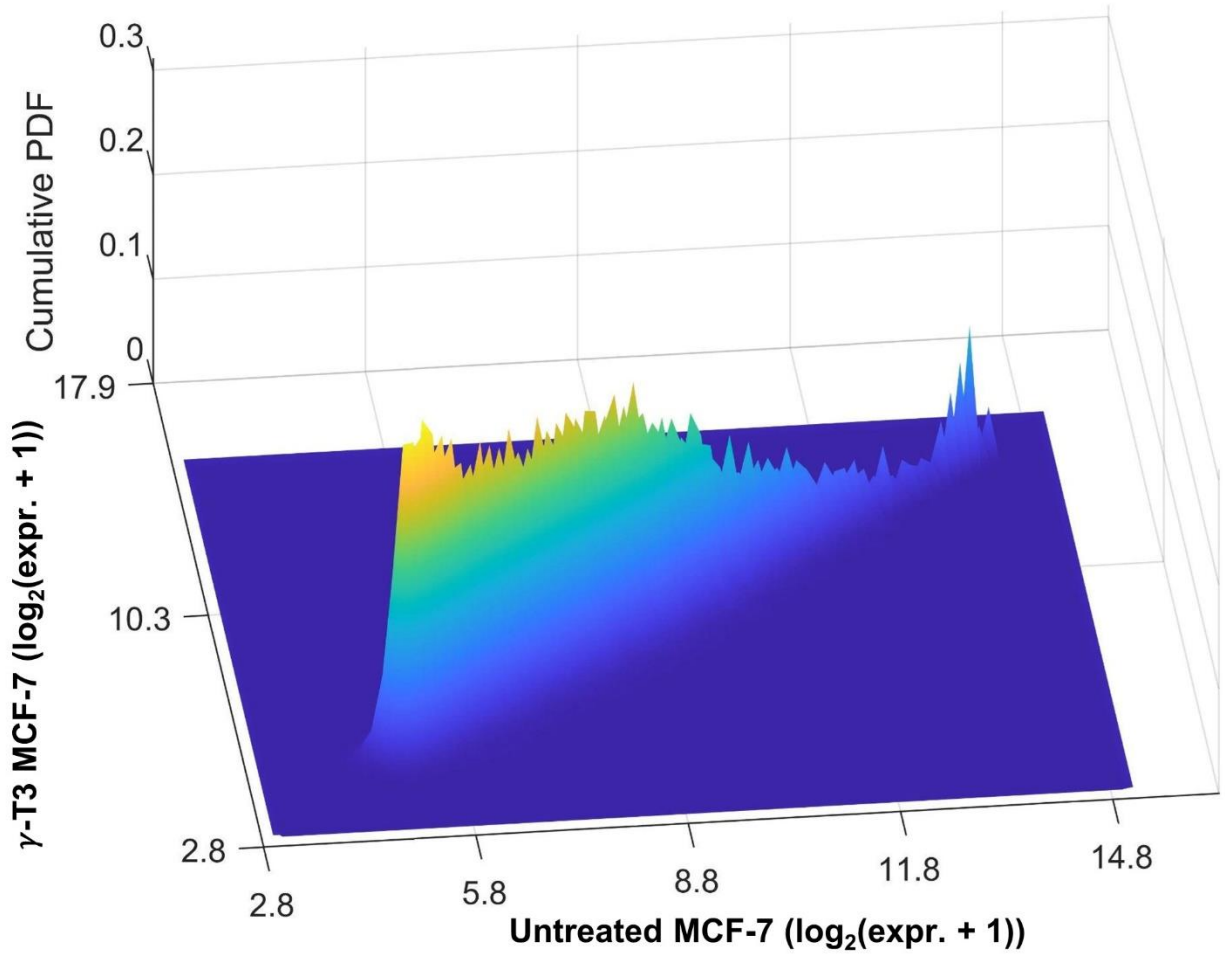


**Figure S2.2. Cumulative PDF from the breast cancer $\gamma$-T3 treatment profile against control (data from GSE21946).**

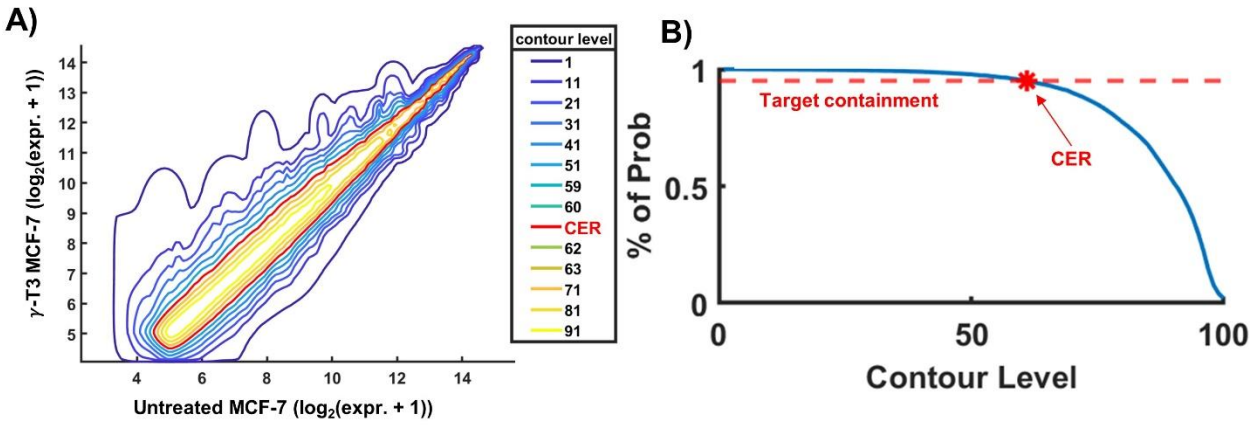


**Figure S2.3. Probability containment of CPDF contours and selection of CER.** (A) Representative contours as indicated by color were selected at various heights/levels of CPDF using the breast cancer $\gamma$-T3 treatment profile. The optimal contour selected as the CER is shown in red. (B) The fraction of contained genes at each contour level.

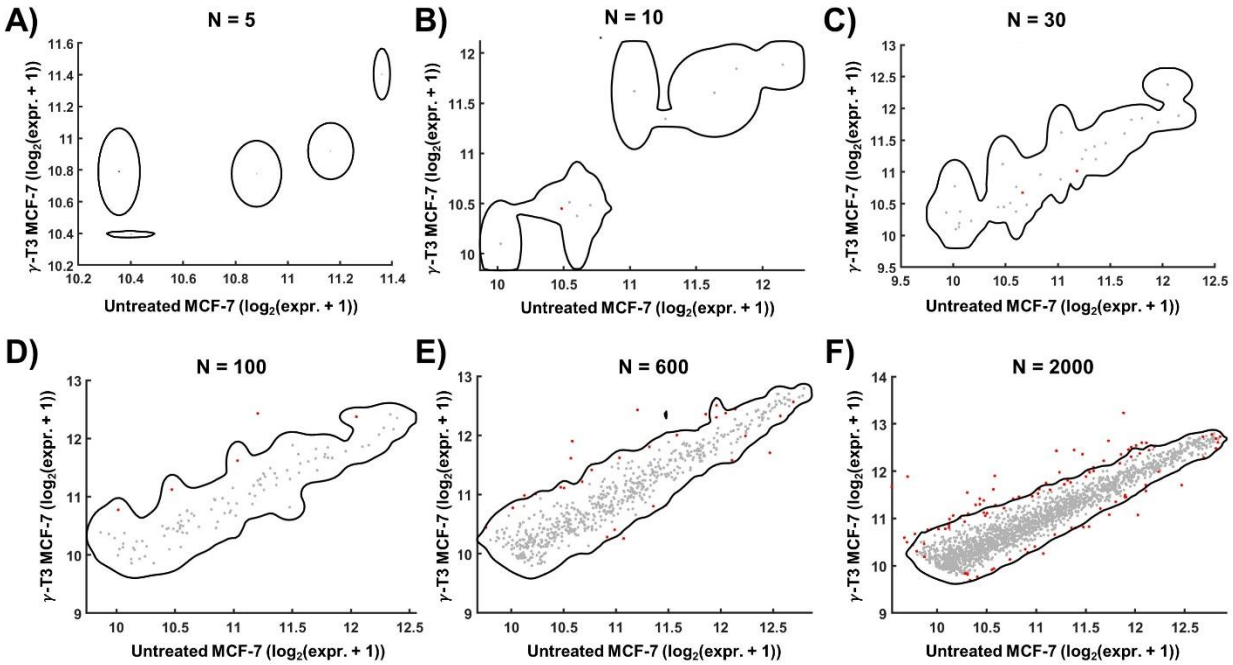


**Figure S2.4. CER with different numbers of genes using breast cancer $\gamma$-T3 treatment profile.** The indicated number of genes was randomly selected. Data points represent mean expression values for individual genes and genes with the highest 5% of outlier scores displayed in red. The black curve represents the CER boundary.

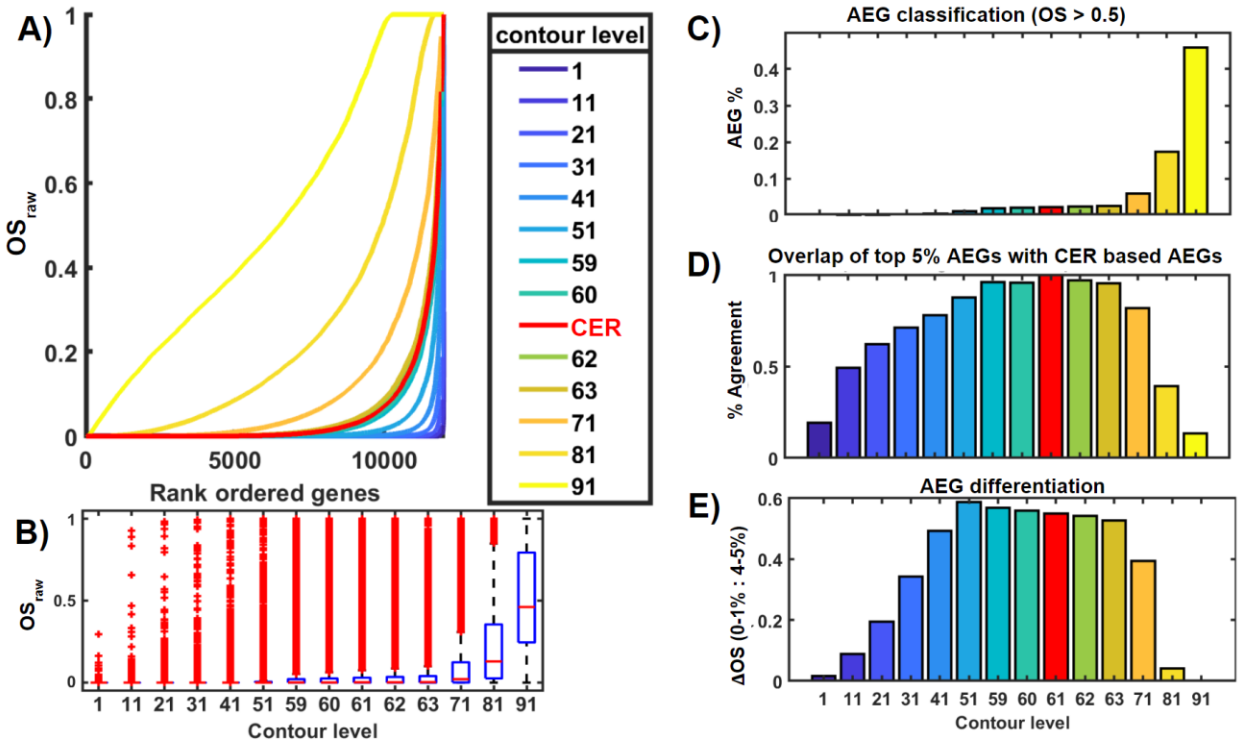


**Figure S2.5. Comparison of AEGs from CER and other CPDF contours.** (A) Distributions of $OS$ by indicated contour as the CER. Each curve represents the cumulative distribution of $OS$. (B) Distribution of $OS_{raw}$ based on each contour level. (C) AEG identification by MAGE using $OS$ as a cutoff ($OS$ < 0.5). (D) Overlap of top 5% $OS_{raw}$ genes between the optimal contour and indicated contours. (E) The difference of $OS_{raw}$ of genes in the highest 1% and 5% as a selectivity for AEGs (data from GSE21946).

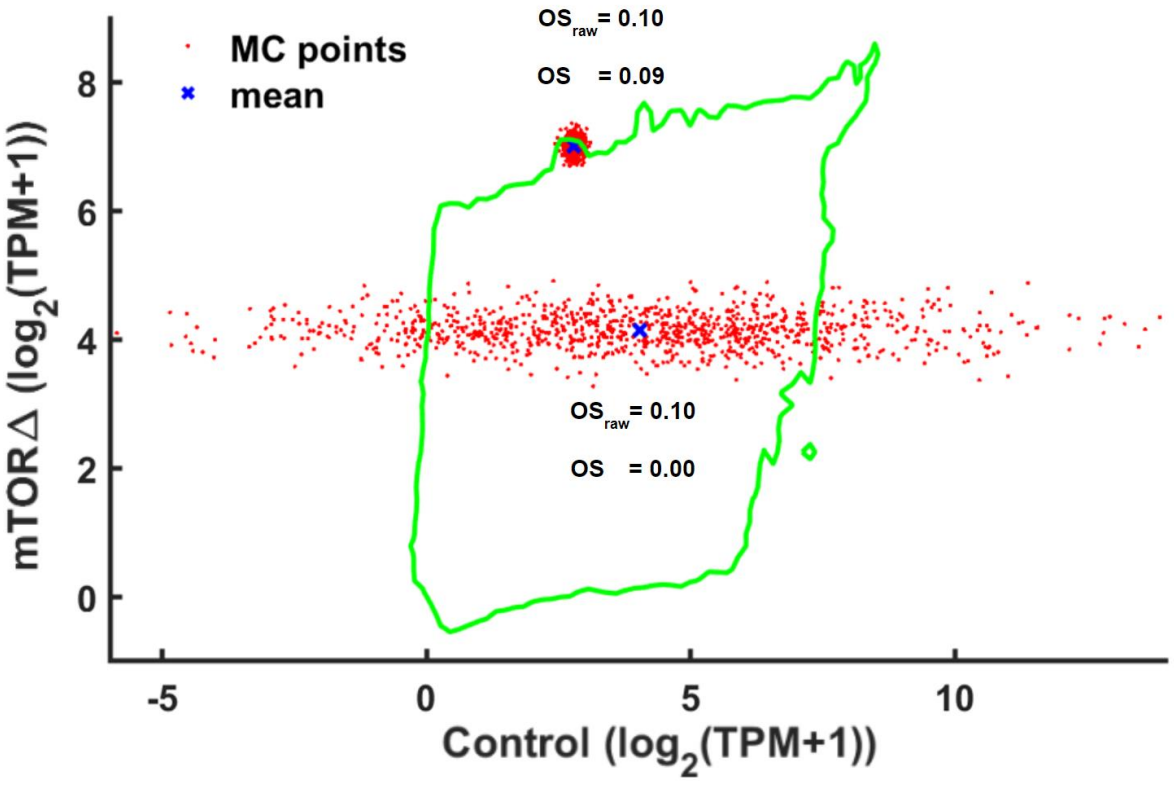


**Figure S2.6. The effect of mean expression and variance on $OS$.** Scatter plot of MC sampled points for two genes with similar $OS_{raw}$~0.1. The mean expression of each gene is shown in blue and 1,000 randomly selected points from each gene's PDF are shown in red. The estimation of $OS$ then adjusted by the distance from CER. Data are from GSE134316.

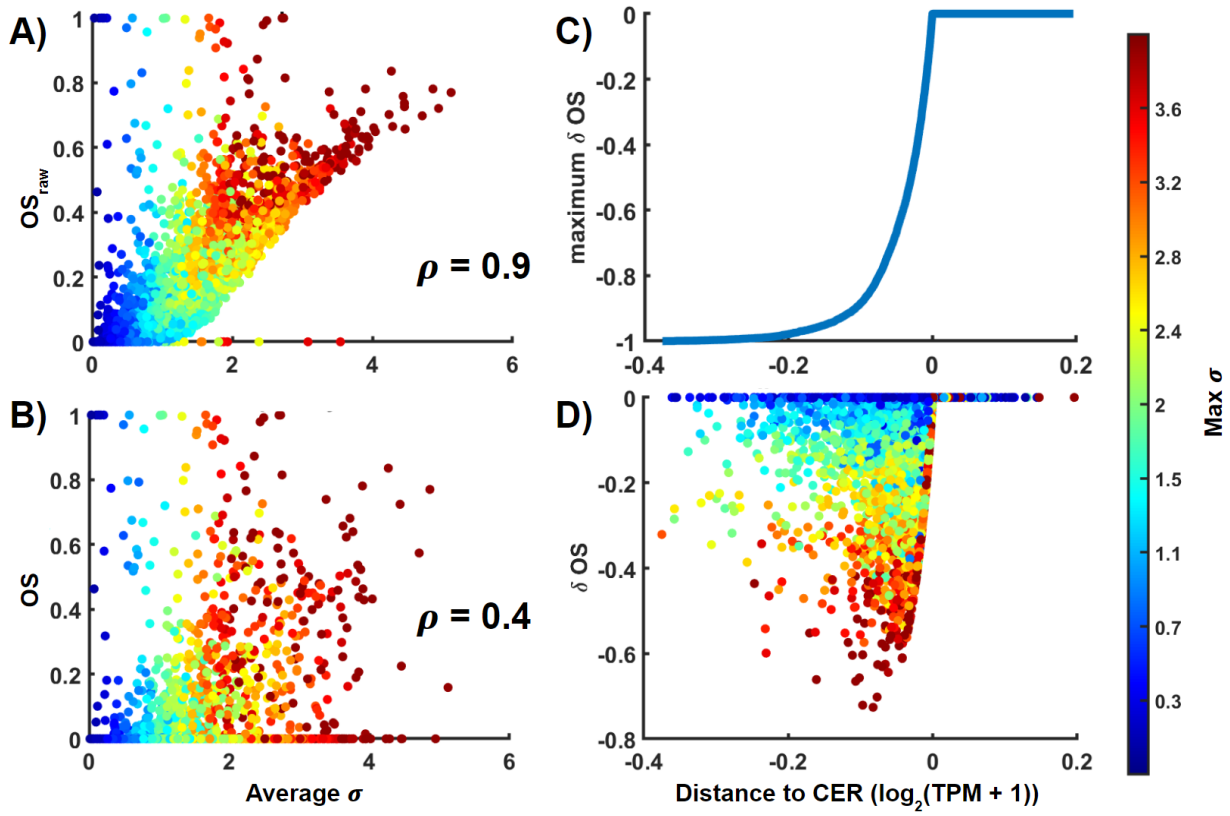


**Figure S2.7. Correction of $OS_{raw}$ by the distance to CER.** (A) Scatter plot of $OS_{raw}$ and SD (denoted as $\sigma = \sqrt{\sigma_x^2 + \sigma_y^2}$) of all genes in the mTOR KO profile. (B) Scatter plot of $OS$ and SD. ρ indicates the Pearson correlation. (C) The correction term $\delta OS = OS_{raw} - OS$ based on eq.10 as a function of the distance to CER. (D) Scatter plot of $\delta OS$ and mean SD. Color corresponds to the maximum SD of each gene calculated in both conditions.

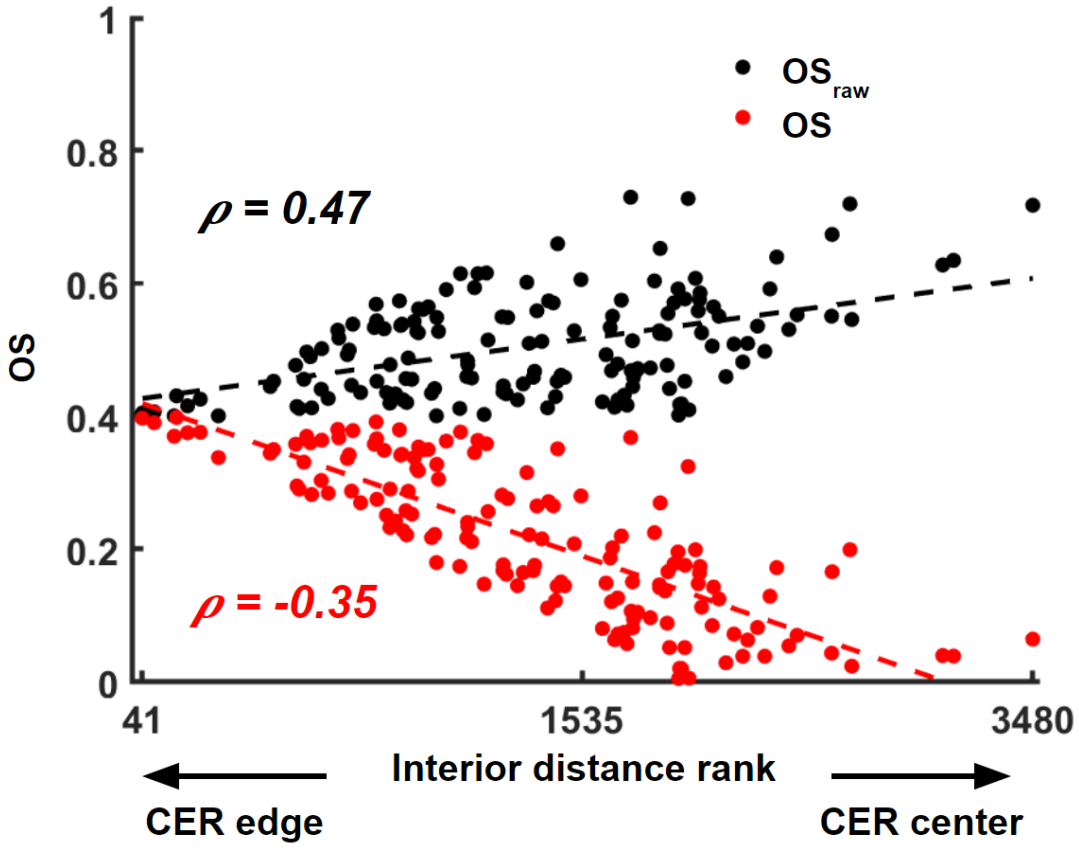


**Figure S2.8. $OS_{raw}$ and $OS$ as a function of interior distance.** $OS$ of Genes with $OS_{raw} > 0.4$ and $OS > 0$ were plotted with the distance to CER (ranking by distance to CER). A smaller distance indicates genes are near the edge of the CER, while genes with large distances are found near the center.

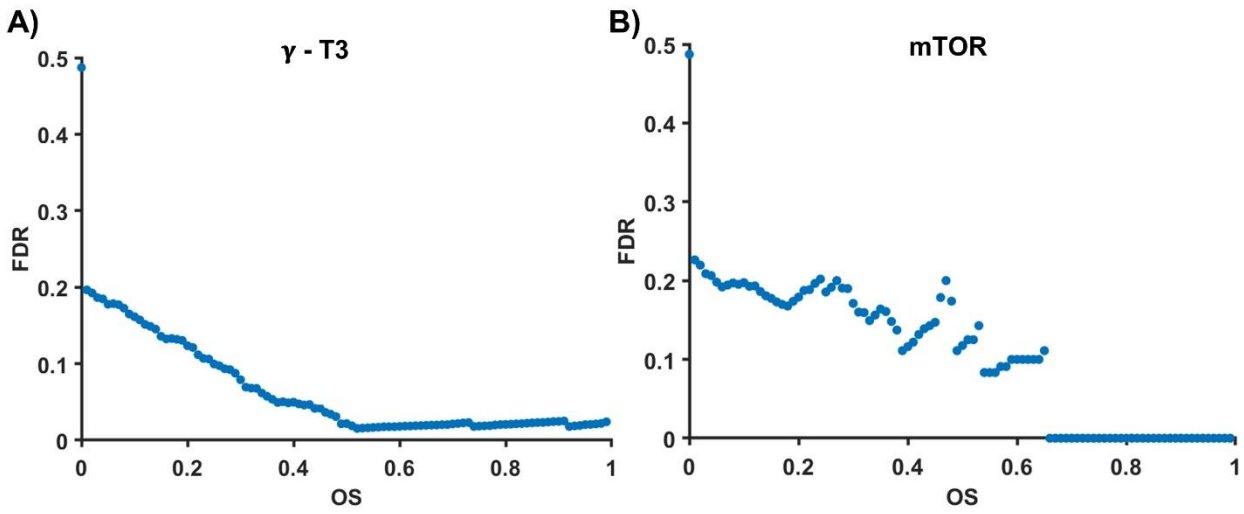


**Figure S2.9. FDR determined by sample permutation.** FDR was estimated (using eqs. 11 and 12) in both (A) breast cancer $\gamma$-T3 treatment and (B) mTOR KO mouse profiles.

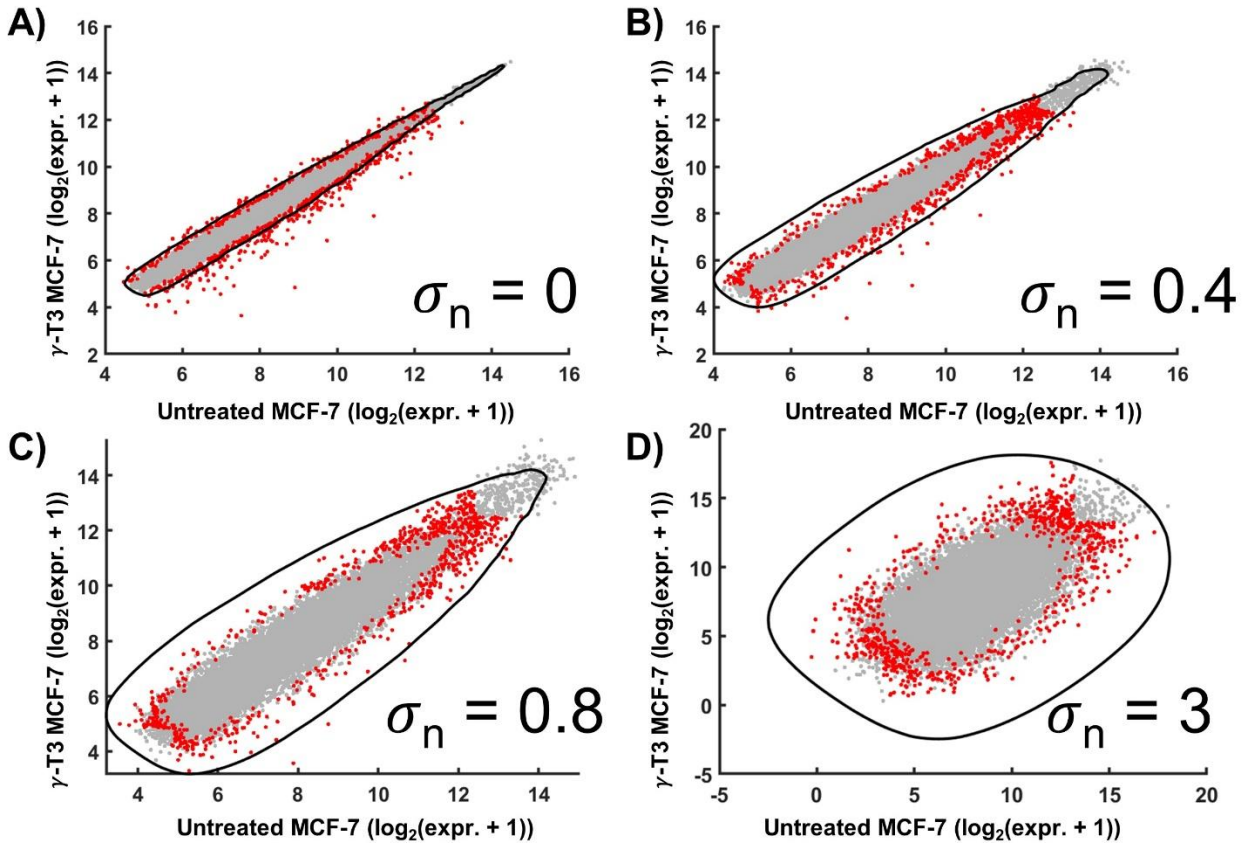


**Figure S2.10. Effect of Gaussian random noise on CER.** For each gene, Gaussian random noise with indicated SD was added to both x- and y-data. After adding normally distributed noise with varying SD $\sigma_n$. The mean expression of each gene is shown plotted in both conditions. Genes with the highest 5% of $OS$ values are shown in red. The CER from each profile is shown in green (data from GSE21946).

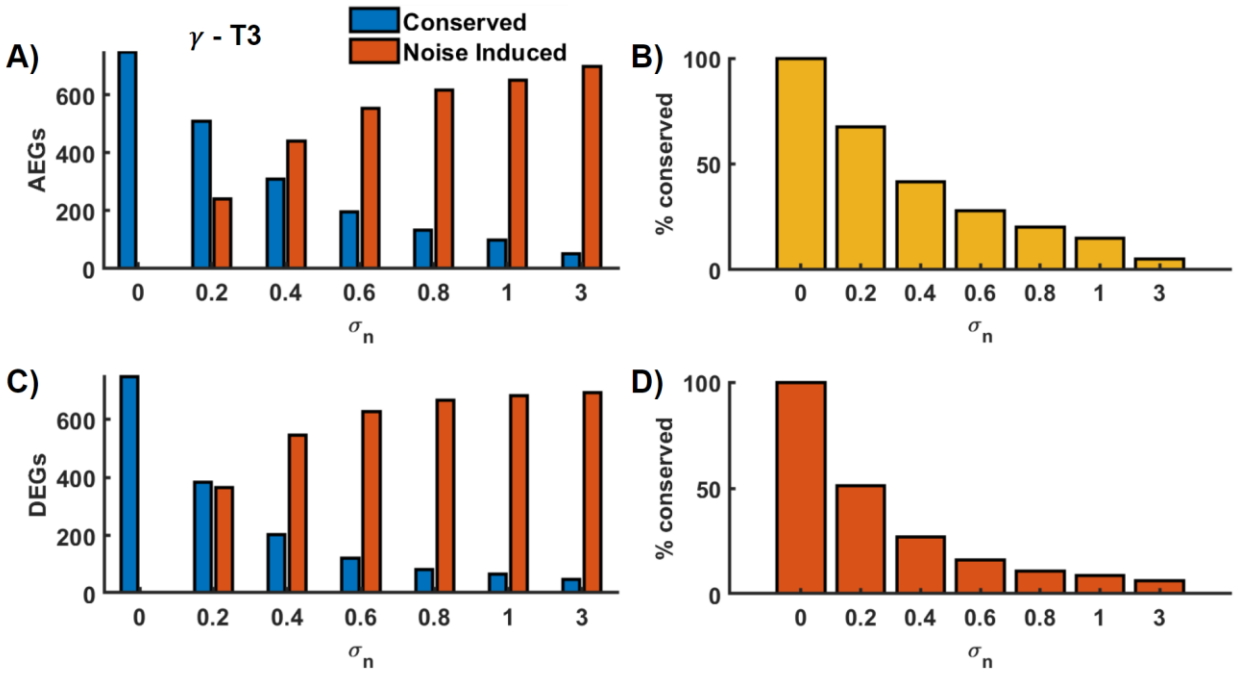


**Figure S2.11. Performance MAGE with varying levels of noise introduced in breast cancer $\gamma$-T3 treatment profile.** (A) The number of AEGs with indicated Gaussian noise. (B) Fraction of overlaps between no-noise-AEGs and AEGs with indicated noise. (C) The number of DEGs with indicated Gaussian noise. (D) Fraction of overlaps between no-noise-DEGs and DEGs with indicated noise (data from GSE21946).

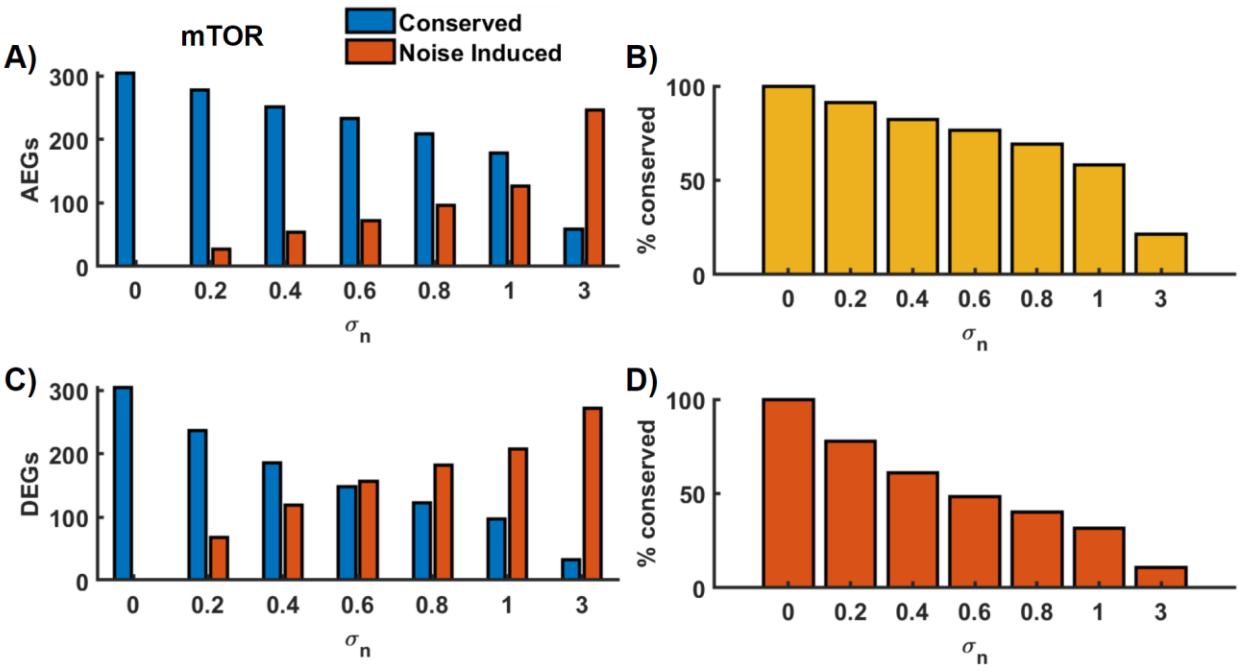


**Figure S2.12. Performance MAGE with varying levels of noise introduced in mTOR KO mouse profile.** (A) The number of AEGs with indicated Gaussian noise. (B) Fraction of overlaps between no-noise-AEGs and AEGs with indicated noise. (C) The number of DEGs with indicated Gaussian noise. (D) Fraction of overlaps between no-noise-DEGs and DEGs with indicated noise (data from GSE134316).

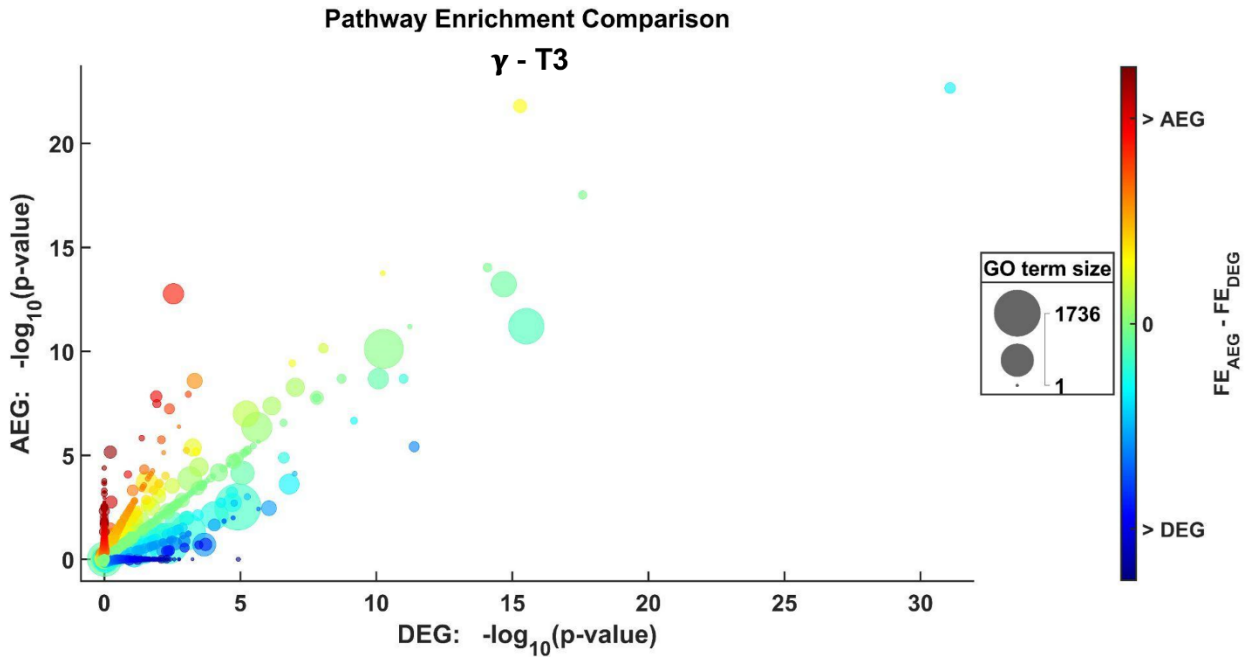


**Figure S2.13. Comparison of breast cancer $\gamma$-T3 treatment profile pathway enrichment of AEGs found using MAGE and DEGs found using t-test.** Bubbles represent individual GO terms. The size of the bubble represents the total number of genes associated with the individual GO term from the DAVID *Homo sapiens* reference gene set. Color represents the difference in the fold-enrichment from the set of AEGs and DEGs.

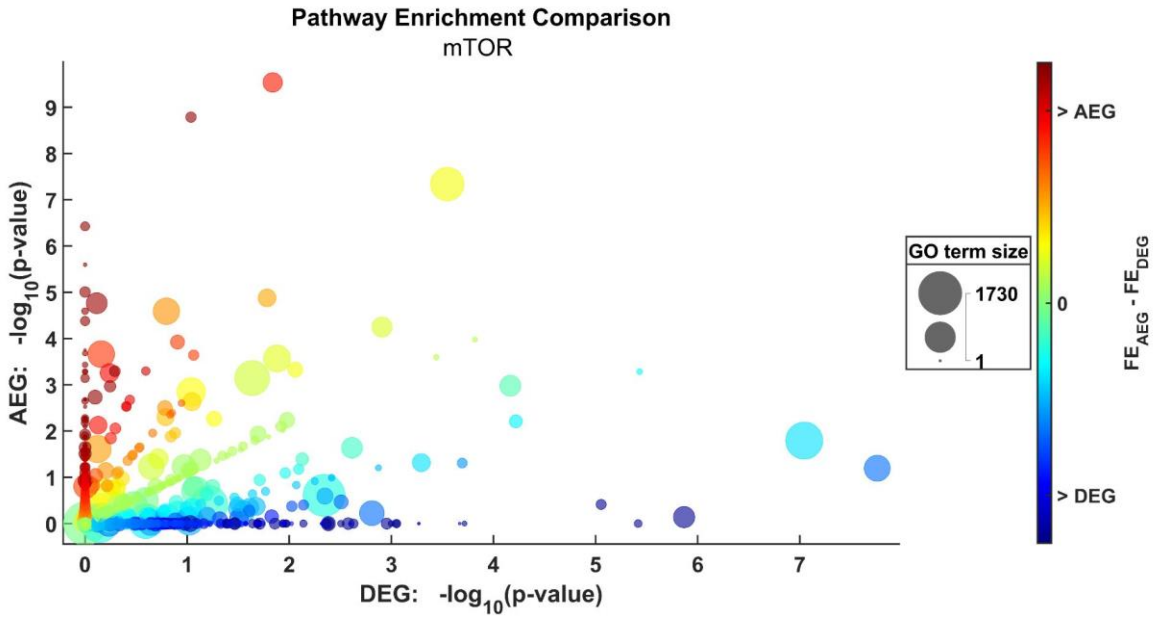


**Figure S2.14. Comparison of mTOR KO mouse profile pathway enrichment of AEGs found using MAGE and DEGs found using t-test.** Bubbles represent individual GO terms. The size of the bubble represents the total number of genes associated with the individual GO term from the DAVID *Mus musculus* reference gene set. Color represents the difference in the fold-enrichment from the set of AEGs and DEGs.

# Supplementary tables for Chapter 2

| Gene symbol | OS | AE FDR | FC | DE FDR |
|---|---|---|---|---|
| *AA393940* | 1 | 0 | 0 | 0.165541 |
| *ADAR* | 1 | 0 | 0 | 0.165541 |
| *ADIPOR2* | 1 | 0 | 0 | 0.165541 |
| *ARPC1B* | 1 | 0 | 0 | 0.165541 |
| *ASS1* | 1 | 0 | 0 | 0.165541 |
| *ATF4* | 1 | 0 | 0 | 0.165541 |
| *C11orf58* | 0.891 | 0 | -0.02694 | 0.165541 |
| *C4orf46* | 1 | 0 | 0 | 0.165541 |
| *CBX1* | 1 | 0 | 0 | 0.165541 |
| *CCNG1* | 1 | 0 | 0 | 0.165541 |
| *CCNI* | 0.798 | 0 | -0.10553 | 0.153901 |
| *CCT7* | 1 | 0 | 0 | 0.165541 |
| *CDK2AP1* | 1 | 0 | 0 | 0.165541 |
| *CKS1B* | 1 | 0 | 0 | 0.165541 |
| *CKS2* | 1 | 0 | 0 | 0.165541 |
| *COX7A2L* | 1 | 0 | 0 | 0.165541 |
| *DSP* | 1 | 0 | 0 | 0.165541 |
| *EIF3D* | 1 | 0 | 0 | 0.165541 |
| *EIF3G* | 1 | 0 | 0 | 0.165541 |
| *EIF3I* | 1 | 0 | 0 | 0.165541 |
| *FUT8* | 1 | 0 | 0 | 0.165541 |
| *HDGF* | 1 | 0 | 0 | 0.165541 |
| *HMGN3* | 1 | 0 | 0 | 0.165541 |
| *IFITM3* | 1 | 0 | 0 | 0.165541 |
| *KCMF1* | 1 | 0 | 0 | 0.165541 |
| *LAMTOR5* | 0.953 | 0 | -0.03048 | 0.165541 |
| *LGALS1* | 1 | 0 | 0 | 0.165541 |
| *LOC101927180* | 1 | 0 | 0 | 0.165541 |
| *LOC101928747* | 1 | 0 | 0 | 0.165541 |
| *LOC101930400* | 0.896 | 0 | 0.075425 | 0.16041 |
| *MIR3620* | 0.864 | 0 | -0.11204 | 0.152738 |
| *MREG* | 1 | 0 | 0 | 0.165541 |
| *NARS* | 1 | 0 | 0 | 0.165541 |
| *NDUFB5* | 1 | 0 | 0 | 0.165541 |
| *NDUFC1* | 1 | 0 | 0 | 0.165541 |
| *NDUFS3* | 1 | 0 | 0 | 0.165541 |
| *NDUFS6* | 1 | 0 | 0 | 0.165541 |
| *NMD3* | 1 | 0 | 0 | 0.165541 |
| *NUTF2P4* | 1 | 0 | 0 | 0.165541 |
| *OAT* | 1 | 0 | 0 | 0.165541 |
| *OLA1* | 1 | 0 | 0 | 0.165541 |
| *PFN2* | 1 | 0 | 0 | 0.165541 |
| *PLEKHF2* | 1 | 0 | 0 | 0.165541 |
| *POLR1D* | 1 | 0 | 0 | 0.165541 |
| *PPP2CA* | 1 | 0 | 0 | 0.165541 |
| *PPT1* | 1 | 0 | 0 | 0.165541 |
| *PRMT1* | 1 | 0 | 0 | 0.165541 |
| *PSMD8* | 1 | 0 | 0 | 0.165541 |
| *PSME1* | 1 | 0 | 0 | 0.165541 |
| *RAP1B* | 1 | 0 | 0 | 0.165541 |
| *RPA3* | 1 | 0 | 0 | 0.165541 |
| *RPL36* | 1 | 0 | 0 | 0.165541 |
| *RPS6* | 0.968 | 0 | -0.03068 | 0.165541 |
| *S100A11* | 1 | 0 | 0 | 0.165541 |
| *SARAF* | 1 | 0 | 0 | 0.165541 |
| *SARS* | 1 | 0 | 0 | 0.165541 |
| *SEC13* | 1 | 0 | 0 | 0.165541 |
| *SEPHS2* | 1 | 0 | 0 | 0.165541 |
| *SLC25A24* | 1 | 0 | 0 | 0.165541 |
| *SLC35B1* | 1 | 0 | 0 | 0.165541 |
| *SLC9A3R1* | 1 | 0 | 0 | 0.165541 |
| *SNHG4* | 0.917 | 0 | 0.073908 | 0.16041 |
| *SNORD73A* | 0.985 | 0 | -0.01237 | 0.165541 |
| *SRP19* | 1 | 0 | 0 | 0.165541 |
| *SSR2* | 1 | 0 | 0 | 0.165541 |
| *STRAP* | 1 | 0 | 0 | 0.165541 |
| *TCEAL4* | 1 | 0 | 0 | 0.165541 |
| *TMEM147* | 1 | 0 | 0 | 0.165541 |
| *TMEM14B* | 1 | 0 | 0 | 0.165541 |
| *TMEM59* | 1 | 0 | 0 | 0.165541 |
| *TUBA1A* | 1 | 0 | 0 | 0.165541 |
| *UFC1* | 1 | 0 | 0 | 0.165541 |
| *VAMP8* | 1 | 0 | 0 | 0.165541 |
| *VBP1* | 1 | 0 | 0 | 0.165541 |
| *VTI1B* | 1 | 0 | 0 | 0.165541 |
| *YTHDF1* | 1 | 0 | 0 | 0.165541 |
| *BNIP3* | 0.543049 | 0.016461 | -0.11769 | 0.151095 |
| *RPS6KB1* | 0.583403 | 0.017778 | -0.07984 | 0.159589 |
| *RAC1* | 0.641 | 0.019324 | 0.022463 | 0.165541 |
| *PRKAR1A* | 0.447672 | 0.029703 | -0.04233 | 0.165088 |
| *IDI1* | 0.429818 | 0.031847 | 0.07455 | 0.16041 |

**Table S2.1. MAGE and t-test results for breast cancer gT3 treatment profile.** Filtered for most significant exAEGs (AE FDR < 0.05, DE FDR > 0.15).

| GO Terms | Ref. count | AEG count | DEG count | AEG FE | DEG FE | AEG p-value | DEG p-value | AEG FDR | DEG FDR |
|---|---|---|---|---|---|---|---|---|---|
| *ER to Golgi vesicle-mediated transport* | 135 | 17 | 14 | 7.8 | 6.3 | 5.66E-10 | 3.80E-07 | 1.28E-06 | 4.30E-04 |
| *response to endoplasmic reticulum stress* | 88 | 14 | 12 | 9.9 | 8.2 | 1.54E-09 | 2.15E-07 | 1.74E-06 | 4.30E-04 |
| *cargo loading into COPII-coated vesicle* | 15 | 7 | 6 | 28.9 | 24.2 | 7.33E-08 | 3.11E-06 | 5.52E-05 | 1.76E-03 |
| *cellular response to oxidative stress* | 97 | 12 | 7 | 7.7 | 4.6 | 4.57E-07 | 5.42E-03 | 2.58E-04 | 4.72E-01 |
| *positive regulation of apoptotic process* | 336 | 20 | 19 | 3.7 | 3.4 | 2.42E-06 | 1.32E-05 | 9.18E-04 | 4.99E-03 |
| *negative regulation of apoptotic process* | 540 | 26 | 17 | 3.0 | 1.9 | 2.44E-06 | 1.79E-02 | 9.18E-04 | 7.62E-01 |
| *protein transport* | 444 | 22 | 15 | 3.1 | 2.0 | 1.15E-05 | 1.61E-02 | 3.72E-03 | 7.28E-01 |
| *intracellular protein transport* | 337 | 18 | 13 | 3.3 | 2.3 | 3.57E-05 | 1.06E-02 | 1.01E-02 | 6.08E-01 |
| *mitotic cell cycle phase transition* | 25 | 6 | 4 | 14.9 | 9.7 | 4.26E-05 | 7.82E-03 | 1.01E-02 | 5.53E-01 |
| *negative regulation of transcription from RNA polymerase II promoter* | 1016 | 35 | 34 | 2.1 | 2.0 | 4.47E-05 | 1.66E-04 | 1.01E-02 | 4.16E-02 |

**Table S2.2. Pathway enrichment results for breast cancer $\gamma$-T3 treatment profile.** Top 10 GO terms based on AEG FDR.

| Gene symbol | OS | AE FDR | FC | DE FDR |
|---|---|---|---|---|
| *Csnk1d* | 0.757895 | 0 | 4.693602 | 0.23871 |
| *H2afy* | 0.822 | 0 | 3.307704 | 0.236548 |
| *Rpl8* | 0.787947 | 0 | 6.94058 | 0.142857 |
| *Adcy5* | 0.579987 | 0 | -3.76108 | 0.231813 |
| *Tubb5* | 0.631631 | 0 | 3.010809 | 0.234703 |
| *Snord22* | 0.614577 | 0 | 5.172483 | 0.232704 |
| *Gm25176* | 0.755 | 0 | -7.64855 | 0.185185 |
| *Gm23865* | 0.583048 | 0 | -4.7266 | 0.2407 |
| *Rps15a-ps7* | 0.82 | 0 | 7.771004 | 0.238095 |
| *Gm4332* | 1 | 0 | -7.94763 | 0.266667 |
| *Rpl36a-ps2* | 1 | 0 | -6.60329 | 0.177215 |
| *Gm9794* | 0.677208 | 0 | 3.144276 | 0.23384 |
| *Cct7* | 0.883 | 0 | 6.363228 | 0.173469 |
| *Gm4604* | 0.845 | 0 | 4.538961 | 0.233591 |
| *Rpl32* | 0.994 | 0 | 7.029737 | 0.162791 |
| *Gm8203* | 0.93 | 0 | 3.403778 | 0.234177 |
| *Trim28* | 0.784 | 0 | 6.570147 | 0.1875 |
| *Kmt2b* | 0.562786 | 0 | 4.89639 | 0.226044 |
| *Eif4g2* | 0.829 | 0 | 3.418035 | 0.233546 |
| *Fus* | 0.72 | 0 | -2.97726 | 0.234445 |
| *Rps26-ps1* | 0.662526 | 0 | 6.014689 | 0.224638 |
| *Rplp1* | 0.727315 | 0 | -2.89532 | 0.233069 |
| *Scml2* | 0.953 | 0 | -7.78658 | 0.277778 |
| *Cbx5* | 0.531 | 0.033333 | -4.0276 | 0.229692 |
| *Mir6340* | 0.533673 | 0.033333 | 5.303201 | 0.22695 |
| *Eef1a1* | 0.528 | 0.033333 | 10.21094 | 0 |
| *mt-Te* | 0.531684 | 0.033333 | -5.78094 | 0.210227 |
| *Igha* | 0.538209 | 0.038462 | 5.713956 | 0.217617 |
| *mt-Rnr2* | 0.55421 | 0.04 | 4.789175 | 0.23516 |
| *Fmr1nb* | 0.545334 | 0.04 | -0.11441 | 0.205227 |
| *Gm29266* | 0.465365 | 0.047619 | -5.33991 | 0.226277 |
| *Snora75* | 0.47576 | 0.05 | -0.55656 | 0.209889 |
| *Gm22721* | 0.478789 | 0.05 | -7.62701 | 0.214286 |
| *Rasl10a* | 0.484465 | 0.05 | -3.92246 | 0.230053 |

**Table S2.3. MAGE and t-test results for mTOR KO mouse profile.** Filtered for most significant AEGs (AE FDR < 0.05).

| GO Terms | Ref. count | AEG count | DEG count | AEG FE | DEG FE | AEG p-value | DEG p-value | AEG FDR | DEG FDR |
|---|---|---|---|---|---|---|---|---|---|
| *cytoplasmic translation* | 97 | 14 | 3 | 14.4 | 2.4 | 1.43E-11 | 3.55E-01 | 1.96E-08 | 1 |
| *translation* | 348 | 20 | 8 | 5.7 | 1.8 | 2.17E-09 | 1.59E-01 | 1.09E-06 | 1 |
| *aerobic respiration* | 72 | 11 | 1 | 15.3 | 1.1 | 2.39E-09 | 1.00E+00 | 1.09E-06 | 1 |
| *mitochondrial respiratory chain complex I assembly* | 66 | 8 | 1 | 12.1 | 1.2 | 4.13E-06 | 1.00E+00 | 9.45E-04 | 1 |
| *aging* | 168 | 9 | 4 | 5.4 | 1.8 | 2.86E-04 | 3.66E-01 | 3.28E-02 | 1 |
| *RNA splicing* | 289 | 10 | 7 | 3.5 | 1.9 | 2.50E-03 | 1.68E-01 | 2.06E-01 | 1 |
| *cell cycle* | 659 | 16 | 7 | 2.4 | 0.8 | 2.55E-03 | 8.54E-01 | 2.06E-01 | 1 |
| *apoptotic process* | 670 | 16 | 13 | 2.4 | 1.5 | 2.98E-03 | 1.53E-01 | 2.26E-01 | 1 |
| *positive regulation of G2/M transition of mitotic cell cycle* | 30 | 4 | 1 | 13.3 | 2.6 | 3.23E-03 | 1.00E+00 | 2.26E-01 | 1 |
| *negative regulation of endoplasmic reticulum calcium ion concentration* | 9 | 3 | 1 | 33.3 | 8.6 | 3.39E-03 | 1.00E+00 | 2.26E-01 | 1 |

**Table S2.4. Pathway enrichment results for mTOR KO mouse profile.** Top 10 GO terms based on AEG FDR.

## Supplementary figures for Chapter 3

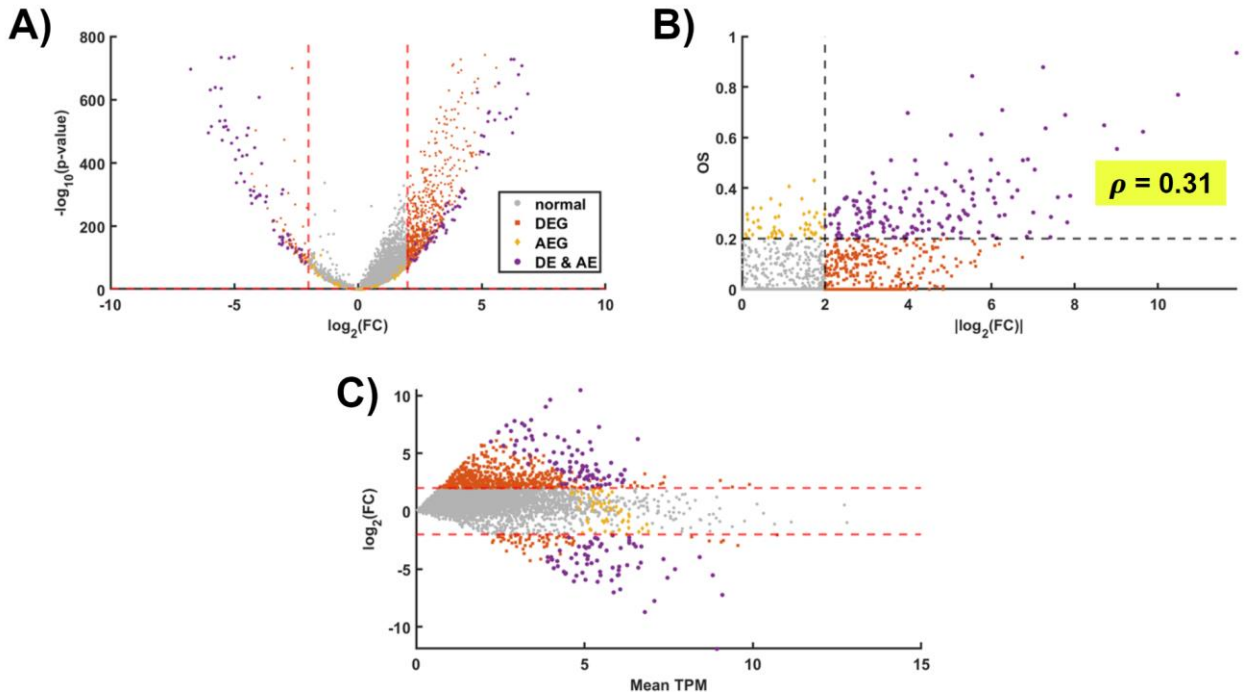


**Figure S3.1. Comparison of brain and lung mural cells.** (A) Volcano plot from the DEG analysis. (B) FC versus OS of all genes. (C) TPM versus FC of all genes. Color indicates normal genes (grey), AEGs (yellow) DEGs (red), and both (purple).

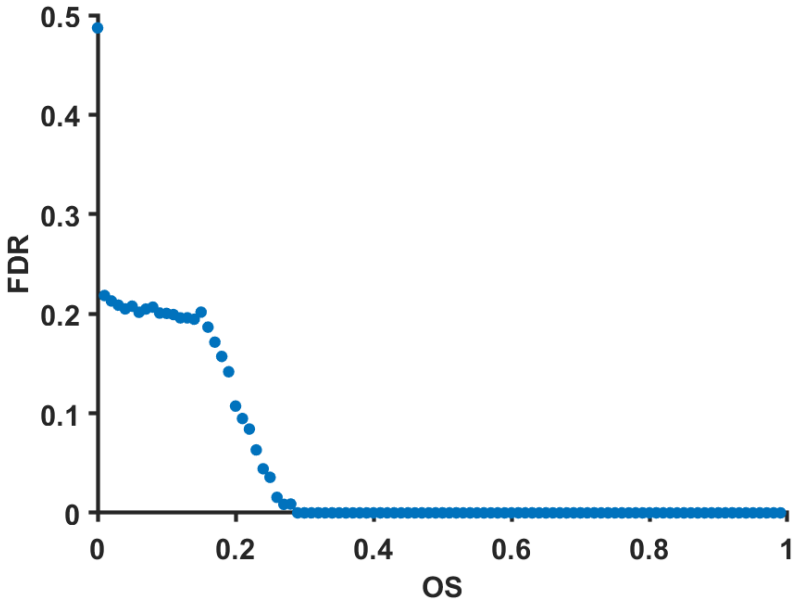


**Figure S3.2. FDR as a function of OS in brain and lung mural cells.**

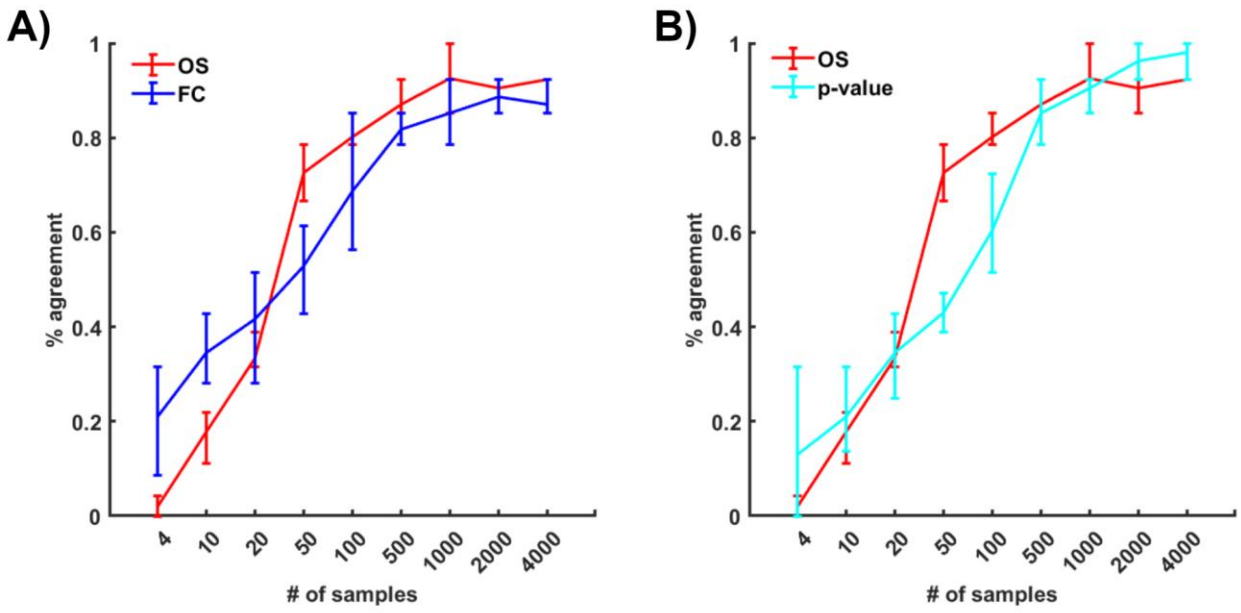


**Figure S3.3. Robustness of AEG and DEG identification by the number of samples.** (A) Robustness of AEG and DEG identification between brain and lung mural cells by the fraction of overlap of top 5% genes by OS (AEGs) (red) and FC (DEGs) (blue), respectively. Lines represent mean values over the 4 trials, and error bars represent standard deviations. (B) Consistency of genes determined by the lowest 5% p-value (DEGs) (cyan) and highest OS (AEGs) (red).

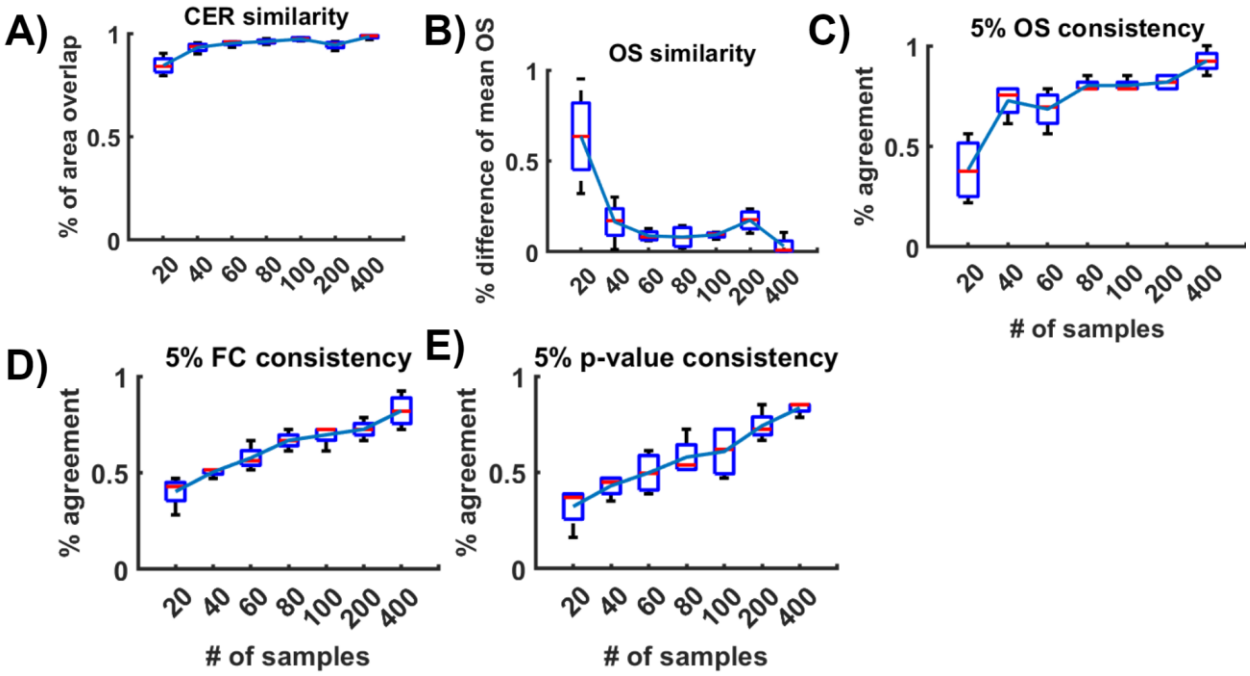


**Figure S3.4. MAGE analysis with different numbers of samples using brain and lung mural cells.** (A) CER similarity in terms of area overlap between the reference CER (all samples) and when the indicated number of samples were used. (B) The difference of OS when using all samples and when the indicated number of samples were used. (C-E) Overlap between top 5% genes based on OS, FC, and p-value when using all samples and when the indicated number of samples were used, respectively.

# List of References


1. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! Genome Biol. 2011;12: 125. doi:10.1186/gb-2011-12-8-125

2. Weymann D, Laskin J, Roscoe R, Schrader KA, Chia S, Yip S, et al. The cost and cost trajectory of whole-genome analysis guiding treatment of patients with advanced cancers. Mol Genet Genomic Med. 2017;5: 251–260. doi:10.1002/mgg3.281

3. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, et al. Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun. 2018;9: 1366. doi:10.1038/s41467-018-03751-6

4. Schlitt T, Brazma A. Current approaches to gene regulatory network modelling. BMC Bioinformatics. 2007;8: S9. doi:10.1186/1471-2105-8-S6-S9

5. Ma J. Transcriptional activators and activation mechanisms. Protein Cell. 2011;2: 879–888. doi:10.1007/s13238-011-1101-7

6. Mattioli K, Oliveros W, Gerhardinger C, Andergassen D, Maass PG, Rinn JL, et al. Cis and trans effects differentially contribute to the evolution of promoters and enhancers. Genome Biol. 2020;21: 210. doi:10.1186/s13059-020-02110-3

7. Erwin DH, Davidson EH. The evolution of hierarchical gene regulatory networks. Nat Rev Genet. 2009;10: 141–148. doi:10.1038/nrg2499

8. Yu H, Gerstein M. Genomic analysis of the hierarchical structure of regulatory networks. Proc Natl Acad Sci. 2006;103: 14724–14731. doi:10.1073/pnas.0508637103

9. Carbon, Seth, Mungall, Chris. Gene Ontology Data Archive. Zenodo; 2018. doi:10.5281/ZENODO.5608599

10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25: 25–29. doi:10.1038/75556

11. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28: 27–30.

12. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, et al. Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 2007;8: R39. doi:10.1186/gb-2007-8-3-r39

13. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27: 1739–1740. doi:10.1093/bioinformatics/btr260

14. Mubeen S, Hoyt CT, Gemünd A, Hofmann-Apitius M, Fröhlich H, Domingo-Fernández D. The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. Front Genet. 2019;10: 1203. doi:10.3389/fgene.2019.01203

15. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci. 1998;95: 14863–14868.

16. D'haeseleer P. How does gene expression clustering work? Nat Biotechnol. 2005;23: 1499–1501. doi:10.1038/nbt1205-1499

17. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14: 1083–1086. doi:10.1038/nmeth.4463

18. Hurst LD, Pál C, Lercher MJ. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet. 2004;5: 299–310. doi:10.1038/nrg1319

19. Michalak P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. 2008;91: 243–248. doi:10.1016/j.ygeno.2007.11.002

20. Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi–C: A comprehensive technique to capture the conformation of genomes. Methods. 2012;58: 268–276. doi:10.1016/j.ymeth.2012.05.001

21. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485: 376–380. doi:10.1038/nature11082

22. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-Dimensional Folding and Functional Organization Principles of the *Drosophila* Genome. Cell. 2012;148: 458–472. doi:10.1016/j.cell.2012.01.010

23. Long HS, Greenaway S, Powell G, Mallon A-M, Lindgren CM, Simon MM. Making sense of the linear genome, gene function and TADs. Epigenetics Chromatin. 2022;15: 4. doi:10.1186/s13072-022-00436-9

24. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10: 669–680. doi:10.1038/nrg2641

25. Li Y, Tollefsbol TO. Combined chromatin immunopreciptation and bisulfite methylation sequencing analysis. Methods Mol Biol Clifton NJ. 2011;791: 239–251. doi:10.1007/978-1-61779-316-5_18

26. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci. 1992;89: 1827–1831. doi:10.1073/pnas.89.5.1827

27. Zarayeneh N, Ko E, Oh JH, Suh S, Liu C, Gao J, et al. Integration of multi-omics data for integrative gene regulatory network inference. Int J Data Min Bioinforma. 2017;18: 223–239. doi:10.1504/IJDMB.2017.10008266

28. Hu X, Hu Y, Wu F, Leung RWT, Qin J. Integration of single-cell multi-omics for gene regulatory network inference. Comput Struct Biotechnol J. 2020;18: 1925–1938. doi:10.1016/j.csbj.2020.06.033

29. Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. Wiley Interdiscip Rev RNA. 2017;8: 10.1002/wrna.1364. doi:10.1002/wrna.1364

30. Khan Y, Hammarström D, Ellefsen S, Ahmad R. Normalization of gene expression data revisited: the three viewpoints of the transcriptome in human skeletal muscle undergoing load-induced hypertrophy and why they matter. BMC Bioinformatics. 2022;23: 241. doi:10.1186/s12859-022-04791-y

31. Zhao Y, Li M-C, Konaté MM, Chen L, Das B, Karlovich C, et al. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. J Transl Med. 2021;19: 269. doi:10.1186/s12967-021-02936-w

32. Gart JJ. The Poisson Distribution: The Theory and Application of Some Conditional Tests. In: Patil GP, Kotz S, Ord JK, editors. A Modern Course on Statistical Distributions in Scientific Work. Dordrecht: Springer Netherlands; 1975. pp. 125–140. doi:10.1007/978-94-010-1845-6_11

33. Di Y, Schafer DW, Cumbie JS, Chang JH. The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. Stat Appl Genet Mol Biol. 2011;10. doi:10.2202/1544-6115.1637

34. Zuyderduyn SD. Statistical analysis and significance testing of serial analysis of gene expression data using a Poisson mixture model. BMC Bioinformatics. 2007;8: 282. doi:10.1186/1471-2105-8-282

35. Sun S, Hood M, Scott L, Peng Q, Mukherjee S, Tung J, et al. Differential expression analysis for RNAseq using Poisson mixed models. Nucleic Acids Res. 2017;45: e106. doi:10.1093/nar/gkx204

36. Weisstein EW. Negative Binomial Distribution. Wolfram Research, Inc.; [cited 12 Apr 2024]. Available: https://mathworld.wolfram.com/

37. Liu S, Wang Z, Zhu R, Wang F, Cheng Y, Liu Y. Three Differential Expression Analysis Methods for RNA Sequencing: limma, EdgeR, DESeq2. J Vis Exp JoVE. 2021. doi:10.3791/62528

38. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26: 139–140. doi:10.1093/bioinformatics/btp616

39. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11: R106. doi:10.1186/gb-2010-11-10-r106

40. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550. doi:10.1186/s13059-014-0550-8

41. Li D, Zand MS, Dye TD, Goniewicz ML, Rahman I, Xie Z. An evaluation of RNA-seq differential analysis methods. PLOS ONE. 2022;17: e0264246. doi:10.1371/journal.pone.0264246

42. Slonim DK, Yanai I. Getting Started in Gene Expression Microarray Analysis. PLoS Comput Biol. 2009;5: e1000543. doi:10.1371/journal.pcbi.1000543

43. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47. doi:10.1093/nar/gkv007

44. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15: R29. doi:10.1186/gb-2014-15-2-r29

45. Smyth GK. limma: Linear Models for Microarray Data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, editors. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York, NY: Springer; 2005. pp. 397–420. doi:10.1007/0-387-29362-0_23

46. Tong Y. The comparison of limma and DESeq2 in gene analysis. Zhu T, Anpo M, Sharifi A, editors. E3S Web Conf. 2021;271: 03058. doi:10.1051/e3sconf/202127103058

47. Ching T, Huang S, Garmire LX. Power analysis and sample size estimation for RNA-Seq differential expression. RNA. 2014;20: 1684–1696. doi:10.1261/rna.046011.114

48. Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am J Bot. 2012;99: 248–256. doi:10.3732/ajb.1100340

49. Li Y, Ge X, Peng F, Li W, Li JJ. Exaggerated false positives by popular differential expression methods when analyzing human population samples. Genome Biol. 2022;23: 79. doi:10.1186/s13059-022-02648-4

50. Li J, Tibshirani R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013;22: 519–536. doi:10.1177/0962280211428386

51. Baccarella A, Williams CR, Parrish JZ, Kim CC. Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance. BMC Bioinformatics. 2018;19: 423. doi:10.1186/s12859-018-2445-2

52. Assefa AT, De Paepe K, Everaert C, Mestdagh P, Thas O, Vandesompele J. Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data. Genome Biol. 2018;19: 96. doi:10.1186/s13059-018-1466-5

53. Bhandari N, Walambe R, Kotecha K, Khare SP. A comprehensive survey on computational learning methods for analysis of gene expression data. Front Mol Biosci. 2022;9. doi:10.3389/fmolb.2022.907150

54. Schaack D, Weigand MA, Uhle F. Comparison of machine-learning methodologies for accurate diagnosis of sepsis using microarray gene expression data. PLOS ONE. 2021;16: e0251800. doi:10.1371/journal.pone.0251800

55. Ringnér M. What is principal component analysis? Nat Biotechnol. 2008;26: 303–304. doi:10.1038/nbt0308-303

56. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Transact A Math Phys Eng Sci. 2016;374: 20150202. doi:10.1098/rsta.2015.0202

57. Density-based clustering - Kriegel - 2011 - WIREs Data Mining and Knowledge Discovery - Wiley Online Library. [cited 14 Apr 2024]. Available: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.30?casa_token=_z6QH07NdBEAAAAA:Ysczgc9BujFmOBvPRxz9E8SzFunX8Bhb0rXd7YPz8_yr9KBxCt8ZHbDDuFFvYPxw4GSPInon1Bdi_Q

58. Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa L da F, et al. Clustering algorithms: A comparative approach. PLoS ONE. 2019;14: e0210236. doi:10.1371/journal.pone.0210236

59. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. IEEE Trans Knowl Data Eng. 2004;16: 1370–1386. doi:10.1109/TKDE.2004.68

60. Dagher I. Quadratic kernel-free non-linear support vector machine. J Glob Optim. 2008;41: 15–30. doi:10.1007/s10898-007-9162-0

61. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Their Appl. 1998;13: 18–28. doi:10.1109/5254.708428

62. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn. 2002;46: 389–422. doi:10.1023/A:1012487302797

63. Biological Pathways Fact Sheet. [cited 4 Apr 2024]. Available: https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet

64. Reimand J, Isser R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019;14: 482–517. doi:10.1038/s41596-018-0103-9

65. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022;50: W216–W221. doi:10.1093/nar/gkac194

66. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47: D330–D338. doi:10.1093/nar/gky1055

67. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt. 2014;34: 502–508. doi:10.1111/opo.12131

68. Nicholson KJ, Sherman M, Divi SN, Bowles DR, Vaccaro AR. The Role of Family-wise Error Rate in Determining Statistical Significance. Clin Spine Surg. 2022;35: 222. doi:10.1097/BSD.0000000000001287

69. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57: 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

70. Hu P, Zhang W, Xin H, Deng G. Single Cell Isolation and Analysis. Front Cell Dev Biol. 2016;4: 116. doi:10.3389/fcell.2016.00116

71. Jiang M, Xu X, Guo G. Understanding embryonic development at single-cell resolution. Cell Regen. 2021;10: 10. doi:10.1186/s13619-020-00074-0

72. Song Q, Ruiz J, Xing F, Lo H-W, Craddock L, Pullikuth AK, et al. Single-cell sequencing reveals the landscape of the human brain metastatic microenvironment. Commun Biol. 2023;6: 1–13. doi:10.1038/s42003-023-05124-2

73. Zhu S, Zhang M, Liu X, Luo Q, Zhou J, Song M, et al. Single-cell transcriptomics provide insight into metastasis-related subsets of breast cancer. Breast Cancer Res BCR. 2023;25: 126. doi:10.1186/s13058-023-01728-y

74. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, et al. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021;12: 5692. doi:10.1038/s41467-021-25960-2

75. Kim B, Lee E, Kim JK. Analysis of Technical and Biological Variability in Single-Cell RNA Sequencing. In: Yuan G-C, editor. Computational Methods for Single-Cell Data Analysis. New York, NY: Springer; 2019. pp. 25–43. doi:10.1007/978-1-4939-9057-3_3

76. Jiang R, Sun T, Song D, Li JJ. Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome Biol. 2022;23: 31. doi:10.1186/s13059-022-02601-5

77. Voskoglou-Nomikos T, Pater JL, Seymour L. Clinical Predictive Value of the in Vitro Cell Line, Human Xenograft, and Mouse Allograft Preclinical Cancer Models1. Clin Cancer Res. 2003;9: 4227–4239.

78. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40: e72. doi:10.1093/nar/gks001

79. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011;12: R18. doi:10.1186/gb-2011-12-2-r18

80. Bryant PA, Smyth GK, Robins-Browne R, Curtis N. Technical Variability Is Greater than Biological Variability in a Microarray Experiment but Both Are Outweighed by Changes Induced by Stimulation. PLOS ONE. 2011;6: e19556. doi:10.1371/journal.pone.0019556

81. Molania R, Foroutan M, Gagnon-Bartsch JA, Gandolfo LC, Jain A, Sinha A, et al. Removing unwanted variation from large-scale RNA sequencing data with PRPS. Nat Biotechnol. 2023;41: 82–95. doi:10.1038/s41587-022-01440-w

82. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015;6: 8971. doi:10.1038/ncomms9971

83. Porcu E, Sadler MC, Lepik K, Auwerx C, Wood AR, Weihs A, et al. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. Nat Commun. 2021;12: 5647. doi:10.1038/s41467-021-25805-y

84. Papavassiliou Athanasios G. Transcription Factors. N Engl J Med. 1995;332: 45–47. doi:10.1056/NEJM199501053320108

85. Cai W, Zhou W, Han Z, Lei J, Zhuang J, Zhu P, et al. Master regulator genes and their impact on major diseases. PeerJ. 2020;8: e9952. doi:10.7717/peerj.9952

86. Xia P, Xu X-Y. PI3K/Akt/mTOR signaling pathway in cancer stem cells: from basic research to clinical application. Am J Cancer Res. 2015;5: 1602–1609.

87. Malumbres M. Cyclin-dependent kinases. Genome Biol. 2014;15: 122. doi:10.1186/gb4184

88. Wittenberg C, Reed SI. Cell cycle-dependent transcription in yeast: promoters, transcription factors, and transcriptomes. Oncogene. 2005;24: 2746–2755. doi:10.1038/sj.onc.1208606

89. Zegerman P. Evolutionary conservation of the CDK targets in eukaryotic DNA replication initiation. Chromosoma. 2015;124: 309–321. doi:10.1007/s00412-014-0500-y

90. Danielsson F, Skogs M, Huss M, Rexhepaj E, O'Hurley G, Klevebring D, et al. Majority of differentially expressed genes are down-regulated during malignant transformation in a four-stage model. Proc Natl Acad Sci U S A. 2013;110: 6853–6858. doi:10.1073/pnas.1216436110

91. Wei R, Zhao M, Zheng C-H, Zhao M, Xia J. Concordance between somatic copy number loss and down-regulated expression: A pan-cancer study of cancer predisposition genes. Sci Rep. 2016;6: 37358. doi:10.1038/srep37358

92. Jassim A, Rahrmann EP, Simons BD, Gilbertson RJ. Cancers make their own luck: theories of cancer origins. Nat Rev Cancer. 2023;23: 710–724. doi:10.1038/s41568-023-00602-5

93. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144: 646–674. doi:10.1016/j.cell.2011.02.013

94. Phan LM, Yeung S-CJ, Lee M-H. Cancer metabolic reprogramming: importance, main features, and potentials for precise targeted anti-cancer therapies. Cancer Biol Med. 2014;11: 1–19. doi:10.7497/j.issn.2095-3941.2014.01.001

95. ZHENG J. Energy metabolism of cancer: Glycolysis versus oxidative phosphorylation (Review). Oncol Lett. 2012;4: 1151–1157. doi:10.3892/ol.2012.928

96. Nishida N, Yano H, Nishida T, Kamura T, Kojiro M. Angiogenesis in Cancer. Vasc Health Risk Manag. 2006;2: 213–219.

97. Chaffer CL, Weinberg RA. A Perspective on Cancer Cell Metastasis. Science. 2011;331: 1559–1564. doi:10.1126/science.1203543

98. Seyfried TN, Huysentruyt LC. On the Origin of Cancer Metastasis. Crit Rev Oncog. 2013;18: 43–73.

99. Wirtz D, Konstantopoulos K, Searson PC. The physics of cancer: the role of physical interactions and mechanical forces in metastasis. Nat Rev Cancer. 2011;11: 512–522. doi:10.1038/nrc3080

100. Yang J, Weinberg RA. Epithelial-Mesenchymal Transition: At the Crossroads of Development and Tumor Metastasis. Dev Cell. 2008;14: 818–829. doi:10.1016/j.devcel.2008.05.009

101. Bergers G, Fendt S-M. The metabolism of cancer cells during metastasis. Nat Rev Cancer. 2021;21: 162–180. doi:10.1038/s41568-020-00320-2

102. Yeung KT, Yang J. Epithelial–mesenchymal transition in tumor metastasis. Mol Oncol. 2017;11: 28–39. doi:10.1002/1878-0261.12017

103. Bhartiya D. Are Mesenchymal Cells Indeed Pluripotent Stem Cells or Just Stromal Cells? OCT-4 and VSELs Biology Has Led to Better Understanding. Stem Cells Int. 2013;2013: 547501. doi:10.1155/2013/547501

104. Hamidi H, Ivaska J. Every step of the way: integrins in cancer progression and metastasis. Nat Rev Cancer. 2018;18: 533–548. doi:10.1038/s41568-018-0038-z

105. Pećina-Šlaus N. Tumor suppressor gene E-cadherin and its role in normal and malignant cells. Cancer Cell Int. 2003;3: 17. doi:10.1186/1475-2867-3-17

106. Peinado H, Olmeda D, Cano A. Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype? Nat Rev Cancer. 2007;7: 415–428. doi:10.1038/nrc2131

107. Luzzi KJ, MacDonald IC, Schmidt EE, Kerkvliet N, Morris VL, Chambers AF, et al. Multistep Nature of Metastatic Inefficiency: Dormancy of Solitary Cells after Successful Extravasation and Limited Survival of Early Micrometastases. Am J Pathol. 1998;153: 865–873. doi:10.1016/S0002-9440(10)65628-3

108. Gensbittel V, Kräter M, Harlepp S, Busnelli I, Guck J, Goetz JG. Mechanical Adaptability of Tumor Cells in Metastasis. Dev Cell. 2021;56: 164–179. doi:10.1016/j.devcel.2020.10.011

109. Haeger A, Krause M, Wolf K, Friedl P. Cell jamming: Collective invasion of mesenchymal tumor cells imposed by tissue confinement. Biochim Biophys Acta BBA - Gen Subj. 2014;1840: 2386–2395. doi:10.1016/j.bbagen.2014.03.020

110. Follain G, Herrmann D, Harlepp S, Hyenne V, Osmani N, Warren SC, et al. Fluids and their mechanics in tumour transit: shaping metastasis. Nat Rev Cancer. 2020;20: 107–124. doi:10.1038/s41568-019-0221-x

111. Regmi S, Fu A, Luo KQ. High Shear Stresses under Exercise Condition Destroy Circulating Tumor Cells in a Microfluidic System. Sci Rep. 2017;7: 39975. doi:10.1038/srep39975

112. Osmani N, Follain G, García León MJ, Lefebvre O, Busnelli I, Larnicol A, et al. Metastatic Tumor Cells Exploit Their Adhesion Repertoire to Counteract Shear Forces during Intravascular Arrest. Cell Rep. 2019;28: 2491-2500.e5. doi:10.1016/j.celrep.2019.07.102

113. Walsh CS, Ogawa S, Karahashi H, Scoles DR, Pavelka JC, Tran H, et al. ERCC5 Is a Novel Biomarker of Ovarian Cancer Prognosis. J Clin Oncol. 2008;26: 2952–2958. doi:10.1200/JCO.2007.13.5806

114. Yap YL, Zhang XW, Smith D, Soong R, Hill J. Molecular gene expression signature patterns for gastric cancer diagnosis. Comput Biol Chem. 2007;31: 275–287. doi:10.1016/j.compbiolchem.2007.06.001

115. Fang Z, Martin J, Wang Z. Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. Cell Biosci. 2012;2: 26. doi:10.1186/2045-3701-2-26

116. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. PLoS ONE. 2017;12: e0190152. doi:10.1371/journal.pone.0190152

117. Auer PL, Doerge RW. A Two-Stage Poisson Model for Testing RNA-Seq Data. Stat Appl Genet Mol Biol. 2011;10. doi:10.2202/1544-6115.1627

118. McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, et al. RNA-seq: technical variability and sampling. BMC Genomics. 2011;12: 293. doi:10.1186/1471-2164-12-293

119. Anjum A, Jaggi S, Varghese E, Lall S, Bhowmik A, Rai A. Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach. J Comput Biol. 2016;23: 239–247. doi:10.1089/cmb.2015.0205

120. Stupnikov A, McInerney CE, Savage KI, McIntosh SA, Emmert-Streib F, Kennedy R, et al. Robustness of differential gene expression analysis of RNA-seq. Comput Struct Biotechnol J. 2021;19: 3470–3481. doi:10.1016/j.csbj.2021.05.040

121. de Torrenté L, Zimmerman S, Suzuki M, Christopeit M, Greally JM, Mar JC. The shape of gene expression distributions matter: how incorporating distribution shape improves the interpretation of cancer transcriptomic data. BMC Bioinformatics. 2020;21: 562. doi:10.1186/s12859-020-03892-w

122. Salkovic E, Sadeghi MA, Baggag A, Salem AGR, Bensmail H. OutSingle: a novel method of detecting and injecting outliers in RNA-Seq count data using the optimal hard threshold for singular values. Bioinformatics. 2023;39: btad142. doi:10.1093/bioinformatics/btad142

123. Liu Z, Song Y, Xie C, Tang Z. A new clustering method of gene expression data based on multivariate Gaussian mixture models. Signal Image Video Process. 2016;10: 359–368. doi:10.1007/s11760-015-0749-5

124. McKinney BA, White BC, Grill DE, Li PW, Kennedy RB, Poland GA, et al. ReliefSeq: A Gene-Wise Adaptive-K Nearest-Neighbor Feature Selection Tool for Finding Gene-Gene Interactions and Main Effects in mRNA-Seq Gene Expression Data. PLOS ONE. 2013;8: e81527. doi:10.1371/journal.pone.0081527

125. Ester M, Kriegel H-P, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.

126. Georgieva O. Iterative Clustering for Differential Gene Expression Analysis. In: Rojas I, Valenzuela O, Rojas F, Herrera LJ, Ortuño F, editors. Bioinformatics and Biomedical Engineering. Cham: Springer International Publishing; 2022. pp. 389–398. doi:10.1007/978-3-031-07802-6_33

127. Maheshwari R, Mishra AC, Mohanty SK. An entropy-based density peak clustering for numerical gene expression datasets. Appl Soft Comput. 2023;142: 110321. doi:10.1016/j.asoc.2023.110321

128. Liu H-M, Yang D, Liu Z-F, Hu S-Z, Yan S-H, He X-W. Density distribution of gene expression profiles and evaluation of using maximal information coefficient to identify differentially expressed genes. PLOS ONE. 2019;14: e0219551. doi:10.1371/journal.pone.0219551

129. Brechtmann F, Mertes C, Matusevičiūtė A, Yépez VA, Avsec Ž, Herzog M, et al. OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. Am J Hum Genet. 2018;103: 907–917. doi:10.1016/j.ajhg.2018.10.025

130. Labory J, Le Bideau G, Pratella D, Yao J-E, Ait-El-Mkadem Saadi S, Bannwarth S, et al. ABEILLE: a novel method for ABerrant Expression Identification empLoying machine LEarning from RNA-sequencing data. Bioinformatics. 2022;38: 4754–4761. doi:10.1093/bioinformatics/btac603

131. Irigoien I, Arenas C. Identification of differentially expressed genes by means of outlier detection. BMC Bioinformatics. 2018;19: 317. doi:10.1186/s12859-018-2318-8

132. Joshi CJ, Ke W, Drangowska-Way A, O'Rourke EJ, Lewis NE. What are housekeeping genes? PLoS Comput Biol. 2022;18: e1010295. doi:10.1371/journal.pcbi.1010295

133. Xie Y, Pan W, Khodursky AB. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. Bioinformatics. 2005;21: 4280–4288. doi:10.1093/bioinformatics/bti685

134. Varabyou A, Salzberg SL, Pertea M. Effects of transcriptional noise on estimates of gene and transcript expression in RNA sequencing experiments. Genome Res. 2021;31: 301–308. doi:10.1101/gr.266213.120

135. Griffiths JA, Richard AC, Bach K, Lun ATL, Marioni JC. Detection and removal of barcode swapping in single-cell RNA-seq data. Nat Commun. 2018;9: 2667. doi:10.1038/s41467-018-05083-x

136. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51: D587–D592. doi:10.1093/nar/gkac963

137. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4: 44–57. doi:10.1038/nprot.2008.211

138. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 2003;4: 210. doi:10.1186/gb-2003-4-4-210

139. Patacsil D, Tran AT, Cho YS, Suy S, Saenz F, Malyukova I, et al. Gamma-Tocotrienol induced Apoptosis is Associated with Unfolded Protein Response in Human Breast Cancer Cells. J Nutr Biochem. 2012;23: 93–100. doi:10.1016/j.jnutbio.2010.11.012

140. Fan C, Zhao C, Zhang F, Kesarwani M, Tu Z, Cai X, et al. Adaptive responses to mTOR gene targeting in hematopoietic stem cells reveal a proliferative mechanism evasive to mTOR inhibition. Proc Natl Acad Sci. 2021;118: e2020102118. doi:10.1073/pnas.2020102118

141. MathWorks - Makers of MATLAB and Simulink. [cited 8 Feb 2024]. Available: https://www.mathworks.com/

142. Komiyama K, Iizuka K, Yamaoka M, Watanabe H, Tsuchiya N, Umezawa I. Studies on the Biological Activity of Tocotrienols. Chem Pharm Bull (Tokyo). 1989;37: 1369–1371. doi:10.1248/cpb.37.1369

143. Sisinni L, Pietrafesa M, Lepore S, Maddalena F, Condelli V, Esposito F, et al. Endoplasmic Reticulum Stress and Unfolded Protein Response in Breast Cancer: The Balance between Apoptosis and Autophagy and Its Role in Drug Resistance. Int J Mol Sci. 2019;20: 857. doi:10.3390/ijms20040857

144. Ding L, Cao J, Lin W, Chen H, Xiong X, Ao H, et al. The Roles of Cyclin-Dependent Kinases in Cell-Cycle Progression and Therapeutic Strategies in Human Breast Cancer. Int J Mol Sci. 2020;21: 1960. doi:10.3390/ijms21061960

145. Idriss M, Younes M, Najem SA, Hodroj MH, Fakhoury R, Rizk S. Gamma-Tocotrienol Synergistically Promotes the Anti-proliferative and Pro-apoptotic Effects of Etoposide on Breast Cancer Cell Lines. Curr Mol Pharmacol. 2022;15: 980–986.

146. Laplante M, Sabatini DM. Regulation of mTORC1 and its impact on gene expression at a glance. J Cell Sci. 2013;126: 1713–1719. doi:10.1242/jcs.125773

147. Yang M, Lu Y, Piao W, Jin H. The Translational Regulation in mTOR Pathway. Biomolecules. 2022;12: 802. doi:10.3390/biom12060802

148. Bennett CF, Latorre-Muro P, Puigserver P. Mechanisms of mitochondrial respiratory adaptation. Nat Rev Mol Cell Biol. 2022;23: 817–835. doi:10.1038/s41580-022-00506-6

149. Saxton RA, Sabatini DM. mTOR Signaling in Growth, Metabolism, and Disease. Cell. 2017;168: 960–976. doi:10.1016/j.cell.2017.02.004

150. Stetson LC, Balasubramanian D, Ribeiro SP, Stefan T, Gupta K, Xu X, et al. Single cell RNA sequencing of AML initiating cells reveals RNA-based evolution during disease progression. Leukemia. 2021;35: 2799–2812. doi:10.1038/s41375-021-01338-7

151. Dobrzyński M, Nguyen LK, Birtwistle MR, von Kriegsheim A, Blanco Fernández A, Cheong A, et al. Nonlinear signalling networks and cell-to-cell variability transform external signals into broadly distributed or bimodal responses. J R Soc Interface. 2014;11: 20140383. doi:10.1098/rsif.2014.0383

152. Sun S, Zhu J, Ma Y, Zhou X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol. 2019;20: 269. doi:10.1186/s13059-019-1898-6

153. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics. 2019;20: 40. doi:10.1186/s12859-019-2599-6

154. Davis A, Gao R, Navin NE. SCOPIT: sample size calculations for single-cell sequencing experiments. BMC Bioinformatics. 2019;20: 566. doi:10.1186/s12859-019-3167-9

155. He L, Vanlandewijck M, Mäe MA, Andrae J, Ando K, Gaudio FD, et al. Single-cell RNA sequencing of mouse brain and lung vascular and vessel-associated cell types. Sci Data. 2018;5: 180160. doi:10.1038/sdata.2018.160

156. Guyon A. CXCL12 chemokine and its receptors as major players in the interactions between immune and nervous systems. Front Cell Neurosci. 2014;8. doi:10.3389/fncel.2014.00065

# Vita

Matthew Beltran received a Bachelor's of Science majoring in Physics with minors in Math and Computer Science from Virginia Commonwealth University in 2019. He has pursued a PhD in Nanoscience and Nanotechnology, within the VCU physics department, working on gene expression analysis methods and related experiments.