Gene Set Creation Algorithm for Microarray Studies with Low Sample Size

Erik Langenborg, Kevin Sun, Lingfeng Cao, Christopher Overall, and Abigail Flower
University of Virginia, el4gf, kws8de, lc2ub, co4p, aaf4q@virginia.edu

***Abstract* - This study aims to validate a novel approach for discovering cell-specific gene sets based on gene expression profiles. More than a billion people suffer from neurological disorders, and understanding how cells interact in the context of different pathologies could lead to viable medical interventions. To this end, the field requires reliable biological markers to discriminate distinct cell types present in a heterogeneous population. Distilling these markers is complex, hampered by problems such as the curse of dimensionality when relating large counts of genes to few cell type observations. This research presents a method for overcoming these challenges in a dataset of microarray gene expression values. A bootstrapped LDA algorithm was used to construct new gene sets and validated using GSVA. Gene sets from a non-bootstrapped LDA process provided a baseline to compare gene set performance in the face of genetic noise. Results showed a marked improvement in half of the validation examples, while the other sets were not significantly enriched.**

Index Terms - Bootstrap, Curse of Dimensionality, LDA, PCA

INTRODUCTION

A cell has both a type and a state. The more permanent aspect of a cell's identity is referred to as its type while the more transient aspect of a cell is referred to as its state [1]. There exist at least 55 known cell types in the human body [2]. Each cell type may assume different states, which are determined by its environmental stimuli, position in taxonomy during development, cell cycle stage, and spatial context (Figure I) [1].

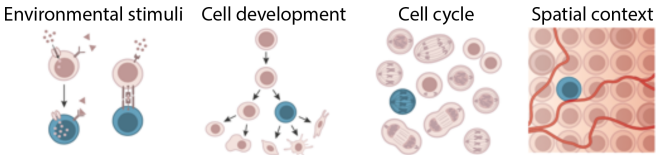

FIGURE I [1]
Factors that influences a cell's identity

Vacillating processes, such as cell-cycle or circadian rhythms can cause cells to change states in an oscillatory manner, in which the cell may change back to its original state. Time-dependent processes such as cell differentiation can cause cells to change states in a unidirectional manner in which cells may develop into finer, more permanent subtypes. Cells may change states randomly due to stochastic, or environmentally controlled, molecular events. These factors make identifying markers that shape a cell's identity challenging [1].

Genomic experiments that measure a cell's molecular profile along with computational methods can infer facets of a cell's identity. A cell can be illustrated as a set of basis vectors that span a space of cell identities (FIGURE II) [1]. Though computational methods are able to form such basis vectors directly, this idealized mathematical representation of a cell does not fully capture the true nature of this space; particularly, basis vectors are defined as independent, but facets of cell identity are most likely dependent on each other. For example, although cell cycle phases are mostly invariant across systems, the ability of a cell to enter the cell cycle and the duration of the phase is most likely dependent on its cell type and other temporal processes such as differentiation [1].

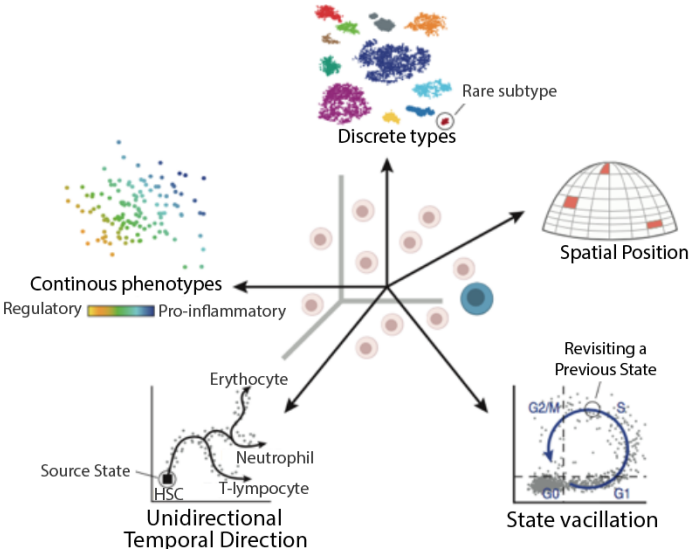

FIGURE II [1]
Basis vector representations of cell identity facets

Aside from the problem of basis vector dependencies, trying to classify a mixture of cells is computationally challenging. Human cells can express over 20,000 genes and trying to classify a mixture of cells to its proper type by representing each cell in a high dimensional space introduces a problem known as the curse of dimensionality. When the number of predictors outweighs the number of

observations, using distance metrics to cluster similar observations becomes unstable as distances between points become exponentially similar as the dimension of space increases. Thus, researchers often use dimensionality reduction approaches, such as principal component analysis (PCA), prior to cell classification [1]. By using PCA to reduce the dimensions of genetic features, researchers have not only successfully classified cell types, but have also discovered new genetic markers for cell identity and new functionally distinct cell subtypes [3]-[5].

PCA works by identifying orthogonal principal components that are directionally aligned to the maximal variations in the data. Principal components that do not represent much of the variance can be removed to reduce the dimensionality of the problem, with little loss of information. A shortcoming of PCA is that it does not preserve class discrimination information. A technique similar to PCA that preserves class discrimination information is linear discriminant analysis (LDA). Unlike PCA, which looks for the dimension with the greatest variance, LDA components lie along dimensions such that the separation of means of projected classes are maximized and the variance within projected classes are minimized. LDA assumes data is Gaussian and may not classify well if discriminatory information is not in the mean, but in the variance, of the data. Regardless, researchers have had success using LDA and LDA variants in supervised classification of microarray data and identifying informative genes [6]-[8]. Many of these researchers applied a pre-selection method that reduced the number of genes by two orders of magnitude prior to LDA classification. Such techniques include filter methods, such as the t-statistic or wrapper methods, such as a forward-backward search algorithms [7]-[8].

Feature reduction and classification methods, such as LDA, can identify a set of genes that discriminate between cell types, which may provide insight into their underlying mechanisms. Gene set enrichment analysis (GSEA) was created to identify sets of genes that may be associated with a particular biological pathway, disease phenotype, or cell type by using statistical methods to determine significantly differentially enriched gene sets [9]. GSEA can help researchers understand the functional profile and underlying biological processes of a gene set [10]. A more powerful unsupervised variation of GSEA is gene set variation analysis (GSVA). GSVA estimates the gene expression level distribution across all samples using a nonparametric kernel and then computes the Kolmogorov-Smirnov statistic in order to yield differential expressions of gene sets [11]. Researchers have had success using GSVA to identify differentially activated gene sets in cancerous cells [12].

DATA, METHODOLOGY, AND RESULTS

Training data was acquired from Immgen's online database, a collaborative effort from immunologists and computational biologists to study gene regulatory networks in immune cells. The mouse dataset of 137 microarray experiments measure the transcriptomes of eight general cell types and 44 more specific subtypes from both the nervous and immune systems. The eight general cell types were microglia, macrophage, T cells, Nk cells, neutrophils, monocytes, dendritic cells, and B cells. The dataset was unbalanced, with microglia at the lower extreme (2% of the data) and with dendritic cells at the upper extreme (26% of the data). Each sample contained 20,270 gene expression values. Testing data was from the Amit dataset, acquired from NCBI's Gene Expression Omnibus (GEO), a public repository for gene expression data [13]. The test dataset contained 18 samples of four general cell types (macrophages, microglia, monocytes, and neutrophils). LDA was chosen over PCA for feature extraction to preserve class discrimination between cell types. However, preliminary results showed that LDA classifications can overfit the data. A bootstrapped, leave-one-out cross validation (LOOCV), down-sampled variant of LDA was developed to overcome this obstacle.

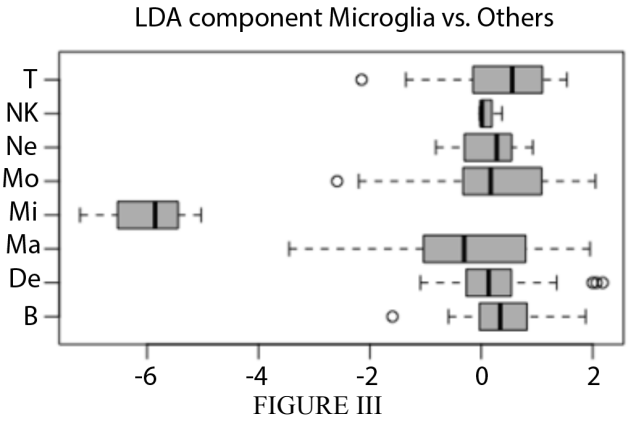

FIGURE III

A preliminary result shows the LDA component created when testing microglia against non-microglial cells separates microglia well. Similar preliminary results were obtained for the other cell types.

The fundamental operation for the pipeline (FIGURE IV) was a bootstrap pass, which produced sets of related gene pairs and their associated performance score. After many bootstrap iterations, gene sets were built from the scored pairs by grouping up the best related genes, and finally assessed with GSVA enrichment analysis. The process was repeated for each cell type and the enrichment values of the final gene sets were validated against the Amit data. Since high false positive rates of significant genes are common in high dimensional data [14], each bootstrap started by downsampling the available genes to a percentage of the available total. Every observation was classified as either the target cell type or other and LDA was performed on the results. The most significant gene loadings were built into 200 gene pairs, since these genes appeared relevant to this bootstrap. The F-measure performance was ascertained by LDA prediction with LOOCV, associating a performance metric with the gene pairs.

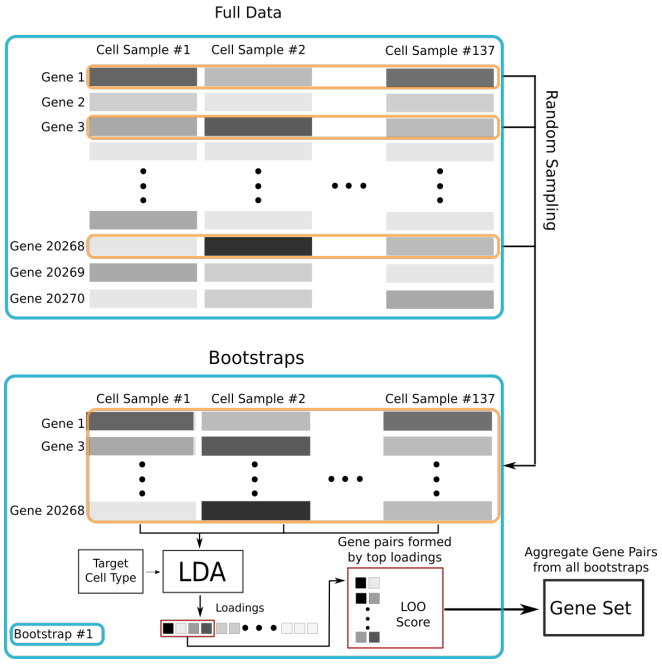

FIGURE IV

Bootstrapped LDA LOOCY algorithm for creating gene sets

The process merged every bootstrap's gene pairs with a final score, computed as an average weighted by frequency of appearance. Gene sets were built from these pairs by considering the most significant combinations and linking pairs which shared a common gene. By incrementally increasing the required scores to consider a pair, the process constructed unique gene sets that conformed to hyperparameter requirements for desired gene counts. Finally, the results sets underwent GSVA and the most differentially enriched candidate was selected for the cell type.

The gene sets created by the bootstrapped LDA methods were evaluated against the base non-bootstrapped method by differential enrichment scores (FIGURE V). In the training data, almost all of the three bootstrap LDA variants yielded higher scores than the base model. Only the microglial gene set had a lower score (-1.52%) when 66% of the genes were randomly selected over 100 bootstraps. In the Amit data, the bootstrapped LDA methods outperformed the base model when creating gene sets for monocytes and neutrophils, but did not outperform the base model when creating gene sets for macrophage. For the microglial gene sets, two out of the three bootstrapped LDA methods outperformed the base model; when 66% of the genes were randomly selected over 100 bootstraps, the microglia gene set created had a lower score (-6.67%) than the base microglial gene set. Overall, eight out of the 12 test cases in the validation set show signs that bootstrapping genes prior to feature extraction may result in more meaningful gene sets.

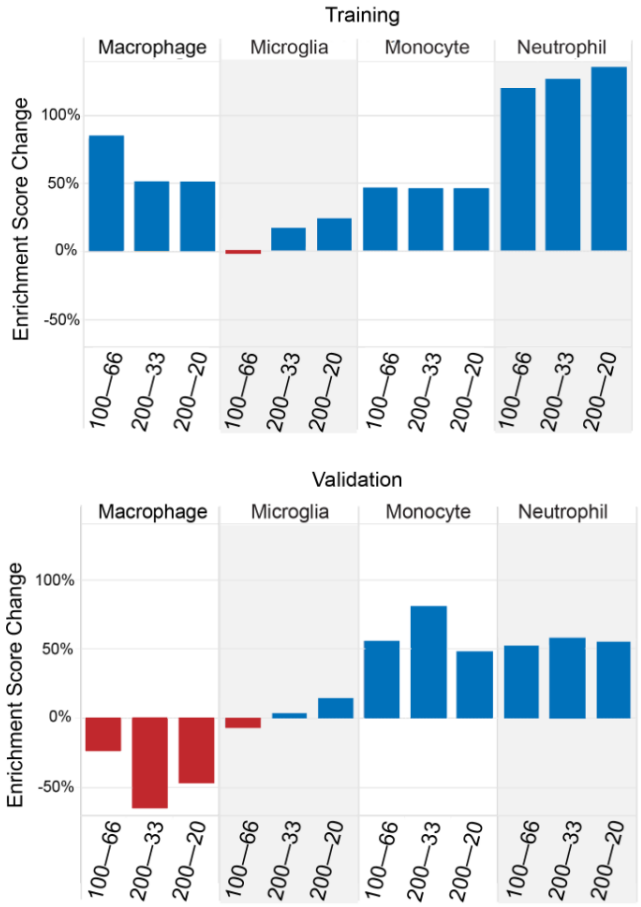

FIGURE V

The lattice plots show the percent change in enrichment scores of bootstrapped methods compared to the baseline non-bootstrapped method. Green represents when bootstrap outperformed non-bootstrap while red represents when it failed to outperform. The first number refers the number of bootstraps and the second number refers to the percentage of genes used within each bootstrap.

To determine how well the newly created gene sets generalize, average differential enrichment scores were calculated against all cell types for all gene sets (FIGURE VI). In the training set, the gene sets created for discriminating microglia, monocyte, and neutrophil does separate them well. However, the gene set created for discriminating macrophage actually discriminates neutrophil better (the differential enrichment score was 160% greater for neutrophil than macrophage). Additionally, the macrophage gene set differentiated every other cell type better. In the Amit validation set, the gene sets created for differentiating macrophage, microglia, and neutrophils do discriminate them well, though for the macrophage gene set the differential enrichment score for macrophage was only 27% higher than the score for microglia - the second highest score. In both the training and validation data set, three-fourths of the gene sets generalized well.

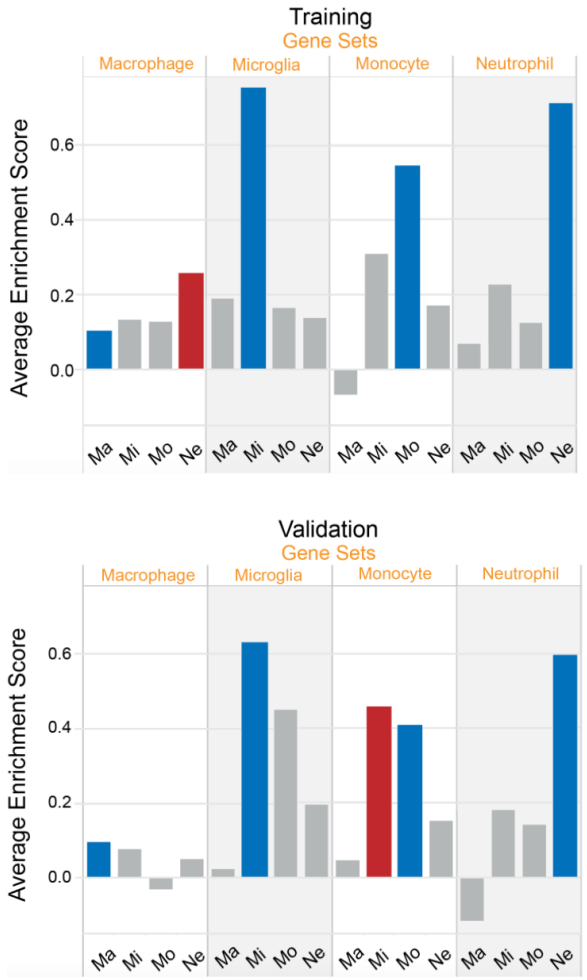

FIGURE VI

These lattice plots show average differential enrichment scores of cell type samples from gene sets created with the bootstrapped LDA method. Lower abbreviation refers to gene sets generated for a given cell type, while upper labels refers to observations in the given data set. Blue indicates the target cell while red denotes another type with greater enrichment. Macrophage (Ma), Macroglia (Mi), Monocyte (Mo), Neutrophil (Ne).

DISCUSSION

This research shows promise in a novel gene set creation method that may be used to further genomics research. Bootstrapping provides a reduction of experimental and technical noise, resulting in constructed gene sets better able to discriminate between cell types. Gene sets corresponding to each of the four general cell types lacked any intersection with each other, which is unusual since most well-known gene sets contain some overlapping genes. Furthermore, the gene sets this research yielded were much smaller (between 15 and 58 genes) than the gene sets found in the Molecular Signature Database (typically hundreds or thousands of genes). More exploration is needed to explain the characteristics of the produced gene sets. Since the bootstrap LDA methods performed worse than the base LDA method for macrophage and only three-fourths of the gene sets discriminated their appropriate cell types well, the proposed system may be improved to yield better results.

One possible enhancement could be replacing microarray data with single cell sequencing data. Microarray data is a bulk average of signals from individual cells taken from a heterogenous population of cells and thus only partial information is gained. Bulk averages suffer problems such as the Simpson's paradox problem and misleading time series studies of gene expression (FIGURE VII) [15]. Single cell sequencing provides a higher resolution of cellular differences by yielding gene expression data in thousands of individual cells in a single experiment and are not susceptible to the shortcomings of bulk-averaged gene expression measurements.

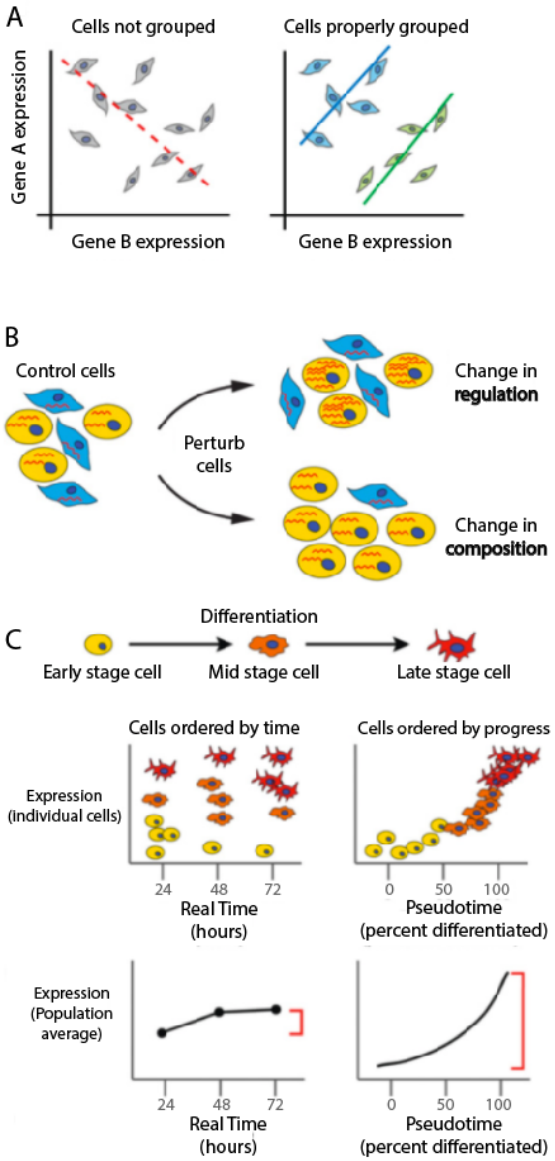

FIGURE VII [15]

A) Failing to properly subgroup the data by cell type first may lead to incorrect correlation analysis, known as the Simpson's Paradox problem. B) Bulk averaged measurements cannot distinguish whether the cause of changes in gene expression was due to a changes in cell compositions or changes in regulation. C) Tracking the time-average gene expression levels to develop qualitative analysis may be misleading since a population of differentiating cells do not change states in a synchronized way.

Another possible improvement is to include a different feature extraction algorithm. LDA assumes Gaussian data and performs best if the mean contains discriminatory information rather than the variance. Due to the complexity of genetic data, these assumptions may be invalid. If discriminatory information is contained in the variance of the data, then PCA may be a more appropriate feature extraction approach. Nonlinear feature extraction methods such as QDA [16] may yield better results if the true underlying features are nonlinear. However, nonlinear methods may be more computationally expensive and less interpretable.

The introduction of other methods aside from GSVA to validate the gene sets may also improve the results. This research assumes that the differential enrichment scores calculated by GSVA are the gold standard of discriminating cell types, but inherent biases to the method may distort the results. An ensemble of machine-learning classification techniques could mitigate this bias with additional performance metrics for gene pair or gene set evaluation in the pipeline. By predicting cell types with several methods, individual biases could be tempered and result in more discriminatory gene sets.

Overall, the study hopes that future research can build on these results and utilize them to further medical research.

REFERENCES

[1] A Wagner, A Regev, N Yosef. November 2016. "Revealing the Vectors of Cellular Identity with Single-Cell Genomics." *Nature Biotechnology* 34(11), pp. 1145-1160

[2] E Bianconi, A Piovesan, F Facchin, A Beraudi, R Casadei, F Frabetti, L Vitale, MC Pelleri, S Tassani, F Piva, S Perez-Amodio, P Strippoli, S Canaider. December 2013. "An Estimation of the Number of Cells in the Human Body." *Annals of Human Biology* 40(6), pp 463 – 471

[3] AA Pollen, TJ Nowakowski, J Chen, H Retallack, C Sandoval-Espinosa, CR Nicholas, J Shuga, SJ Liu, MC Oldham, A Diaz, DA Lim, AA Leyrat, JA West, AR Kriegstein. September 2015. "Molecular Identity of Human Outer Radial Glia during Cortical Development." Cell 163(1), pp. 55-67

[4] D Usoskin, A Furlan, S Islam, H Abdo, P Lonnerberg, D Lou, J Hjerling-Leffler, J Haeggstrom, O Kharchenko, PV Kharchenko, S LInnarsson, P Ernfors. January 2015. "Unbiased Classification of Sensory Neuron Types by Large-Scale Single-Cell RNA Sequencing." *Nature Neuroscience* 18(1), pp. 145-153

[5] L Chu, N Leng, J Zhang, Z Hou, D Mamott, D Vereide, J Choi, C Kendziorski, R Stewart, JA Thomson. August 2016. "Single-cell RNA-seq Reveals Novel Regulators of Human Embryonic Stem Cell Differentiation to Definitive Endoderm." *Genome Biology* 17(173). DOI: 10.1186/s13059-016-1033-x

[6] D Huang, Y Quan, M He, B Zhou. December 2009. "Comparison of Linear Discriminant Analysis Methods for the Classification of Cancer Based on Gene Expression Data." *Journal of Experimental & Clinical Cancer Research* 28(149). DOI: 10.1186/1756-9966-28-149

[7] FF González-Navarro, LA Belanche-Muñoz, KA Silva-Colón. December 2013. "Effective Classification and Gene Expression Profiling for the Facioscapulohumeral Muscular Dystrophy." *Plos ONE* 8(12). DOI: 10.1371/journal.pone.0082071

[8] EB Huerta, B Duval, JK Hao. May 2013. "A Hybrid LDA and Genetic Algorithm for Gene Selection and Classification of Microarray Data." *Neurocomputing* 73(13-15), pp. 2375-2383

[9] A Subramanian, P Tamayo, VK Mootha, S Mukherjee, BL Ebert, MA Gillette, A Paulovich, SL Pomeroy, TR Golub, ES Lander, JP Mesirov. October 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *PNSA* 102(43), pp. 15545-15550

[10] VK Mootha, CM Lindgren, KF Eriksson, A Subramanian, S Sihag, J Lehar, P Puigserver, E Carlsson, M Ridderstrale, E Laurila, N Houstis, MJ Daly, N Patterson JP Mesirov, TR Golub, P Tamayo, B Spiegelman, ES Lander, JN Hirschhorn, D Altshuler, LC Groop. July 2003. "PGC-1alpha-Responsive Genes Involved in Phosphorylation are Coordinately Downregulated in human Diabetes." *Nature Genetics* 34(3), pp. 267-273

[11] S Hänzelmann, R Castelo, J Guinney. January 2013. "GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data." *BMC Bioinformatics* 14(7). DOI: 10.1186/1471-2105-14-7

[12] P Kickingereder, F Sahm, A Radbruch, W Wick, S Heiland, A von Deimling, M Bendszus, B Wiestler. November 2015. "IDH Mutation Status is Associated with a Distinct Hypoxia/Angiogenesis Transcriptome Signature which is Non-Invasively Predictable with rCBV Imaging in Human Glioma." *Scientific Reports* 5(16238). DOI: 10.1038/srep16238

[13] Y Lavin, D Winter, R Blecher-Gonen, E David, H Keren-Shaul, M Merad, S Jung, I Amit. December 2014. "Tissue-Resident Macrophage Enhancer Landscapes Are Shaped by the Local Microenvironment." *Cell* 159(6), pp 1312-1326

[14] Y Wang, DJ Miller, R Clarke. March 2008. "Approaches to Working in High-Dimensional Data Spaces: Gene Expression Microarrays." *British Journal of Cancer* 12(6), pp.1023-1028

[15] JM Arevalillo & H Navarro. November 2011. "A New Method for Identifying Bivariate Differential Expression in High Dimensional Microarray Data using Quadratic Discriminant Analysis." *BMC Bioinformatics* 12(12). DOI: 10.1186/1471-2105-12-S12-S6

[16] C Trapnell. October 2015. "Defining Cell Types and States with Single-Cell Genomics." *Genome Research* 25(10), pp. 1491-1498