Improved Modeling and Analysis of Gene Expression

Anna Bair (annabair@mit.edu)
Matthew McDermott (mmd@mit.edu)
Professor Peter Szolovits (psz@mit.edu)

***Abstract*—Analyzing gene expression data, the production of proteins encoded by a DNA sequence, is extremely valuable for understanding disease and the effects of various treatments. However, analysis can be challenging because gene expression datasets are high-dimensional, noisy, and sometimes not normally distributed. This paper demonstrates two main methods to better analyze gene expression data: a method of improved hypothesis testing and a method of correlational analysis. We show that our improved hypothesis testing is able to perform a more meaningful measure of statistical significance on non-normally distributed data. We also demonstrate the power of analyzing correlative relationships between genes and identify significantly differentially expressed genes between a disease and control state.**

I. Introduction

In order to analyze the cellular effects of various drugs or disease-causing mutations, broadly referred to as perturbations, biological researchers often want to examine how much certain genes are being translated into proteins within a cell. This kind of measurement is called gene expression. Each gene consists of a segment of DNA, which is transcribed into mRNA and then translated into specific proteins. Therefore, gene expression can be measured by analyzing the amount of mRNA produced by transcription of each gene. Researchers use the gene expression data to identify which genes have modified expression patterns in response to perturbations so that they can focus further research efforts on significantly impacted genes.

Gene expression data can provide valuable insights into what is happening inside cells: while DNA is static, measurements of mRNA transcript are dynamic, and thus can provide information about the current state of a cell. This information can be useful for understanding disease and for predictive medicine, among other applications.

While gene expression is a useful measure, analysis can be challenging for a few reasons. For one, it suffers from the curse of dimensionality. Humans have around 20,000 genes, which yields an incredibly high-dimensional dataset. Another difficulty when working with genetic data is that the data is noisy. It can be hard to measure the gene expression levels precisely, and measurements can also contain human error [1].

Due to these challenges, accurate ways of measuring gene expression are extremely valuable. Current methods are often technically limited or make significant assumptions about the underlying data distribution [8]. In this paper, we identify new methods for performing differential gene extraction, or the identification of genes whose expression levels differ significantly between two or more states. We introduce a *zeroth order* method for improved hypothesis testing on raw expression data and a *first order* method using correlational analysis.

Our zeroth order method is a refinement of standard hypothesis testing. Like the standard *t*-test, our method allows us to determine whether the expression level of a given gene is significantly different than expected. However, a significant limitation of standard hypothesis testing is the assumption of normality, and gene expression data is often not normally distributed, as can be seen in Figure 1. In order to allow for a wider variety of data distributions, we improved modeling of the underlying data and implemented an improved measure of significance. Rather than assuming the underlying distribution is normal, we fit a mixture model so as to account for truncated, censored, and multi-modal data. As well, we implement a probabilistic *p*-value which is more representative of significance for non-normal distributions than is the standard *p*-value.

The first order method analyzes the relationships between genes. We use the pairwise correlations between genes to determine the most significant differences between two different disease states. By analyzing the relationships between genes in different disease states, we can find structure and information about regulatory relationships that may not be discovered by zeroth order analysis methods [3]. We use data provided by our collaborators at MGH to perform this analysis. By analyzing the differential gene expression between cells from healthy control subjects and cells from patients with Schizophrenia, we are able to identify genes whose

correlative relationships to other genes are significantly altered in the Schizophrenia cells.

We offer three main contributions:

- Since gene expression data is not always normally distributed, we demonstrate a method to model non-normal distributions, including mixtures with truncated or censored components.
- We use a probabilistic p-value to better perform hypothesis testing for non-normal distributions.
- We use correlative analysis to identify genes with significantly different patterns of co-expression in patients with Schizophrenia as opposed to healthy control subjects.

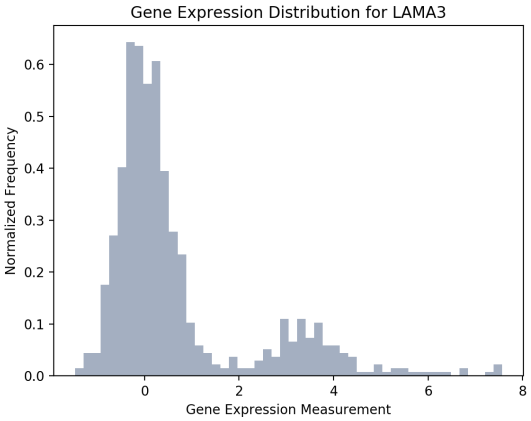

Fig. 1. This bimodal gene expression data is from an MGH Dataset consisting of two patients with Schizophrenia and two healthy control subjects.

II. Related Work

Many of the current methods used for differential gene extraction rely on univariate methods, which are tests that compare the expression level of an individual gene [10]. Univariate tests include Welch's t-test, Significance Analysis of Microarrays (SAM) [18], and limma (Linear Models for Microarray Data) [3], [8], [14], [18], [19]. Of these univariate methods, Welch's t-test, and SAM assume that the distribution of expression levels for each gene is normal [18], [19]. In addition to these parametric tests, which make assumptions about the underlying distribution, are nonparametric tests, which make no assumptions about the underlying distribution. However, the statistical power of non-parametric tests is much weaker.

The limitations of these methods indicate a need for a richer analysis of gene expression datasets. One such example is [3], which allows for multivariate analysis. Another is Gene Set Enrichment Analysis, which focuses on sets of biologically related genes rather than just individual genes [17].

We suggest an improvement upon the current tradeoff between the power of parametric tests and the flexibility of non-parametric tests with our zeroth order method. Rather than assuming normality, we allow for mixture models that can represent an underlying distribution that is non-normal. We build upon the EM Algorithm, which can be used to create Gaussian Mixture Models, but with the assumption that each underlying mixture component is normal [4]. We use [6] and the MixEM package [13] to implement a formulation of the EM algorithm that can incorporate truncated and censored data.

This mixture model itself is not sufficient to fit an unknown distribution, since the number and type of each distribution must be specified before the EM Algorithm optimizes the parameters of each distribution. Therefore, we implemented an automatic distribution selection algorithm based on that of [2]. Our algorithm randomly generates candidate sets of distributions (for example, two truncated distributions) and evaluates the Bayesian Information Criterion (BIC) for each model. The BIC offers a tradeoff between maximizing the log-likelihood but penalizing the number of parameters in the distribution, which helps avoid overfitting.

As well, we perform a richer multivariate analysis through our first order method. Rather than comparing the expression level of a single gene to an expected distribution as in the univariate methods, we use the correlative relationships between pairs of genes to model significant differences between disease states. Papers such as [5] and [15] demonstrate the power of analyzing gene expression correlations and networks rather than on an individual basis. This literature indicates that our first order correlation analysis is similar to other powerful analysis techniques.

III. Data

There were two main sources of data used in this work. We used synthetic data for our zeroth order method of improved hypothesis testing, and we used data from Massachusetts General Hospital (MGH) for our first order correlation analysis.

For our zeroth order method, we generated synthetic data. This allowed us to know precisely what the underlying distribution was when testing our models. In order to test our methods, we wanted to be able to construct a variety of shapes of the data and compare our results to what we would expect.

For our first order method, we used a dataset from MGH consisting of gene expression data from four

patients: two with Schizophrenia (SCZ) and two Healthy Control Subjects (HCS). For each subject, Neural Progenitor Cells (NPC) were derived from a sample of skin cells. Plates were prepared in which each subject's NPCs were exposed to each of 60 compounds. These compounds include DMSO, which we use as a "control" perturbation, and other psychoactive substances such as clozapine and duloxetine. Then, the L1000 technique [16] was used to measure the amount of mRNA produced by the transcription of each gene. This procedure generated 758 HCS and 756 SCZ measurements 978 genes, where each measurement includes a perturbagen and a dosage amount.

IV. Methods

Previous work has both demonstrated a need for improvements upon the standard *t*-test and validated the power of correlational analyses. Here, we describe two methods, a zeroth order method that demonstrates improved hypothesis testing, and a first order method that performs correlational analysis.

A. Zeroth Order

The zeroth order method addresses the problem of non-normal distributions. Since standard hypothesis testing assumes an underlying normal distribution, it may not always be the most appropriate test to use on other distributions. Thus, we implement two main additions to standard hypothesis testing in order to obtain more accurate results on non-normal distributions: improved modeling of the underlying distribution and a probabilistic variant of the *p*-value calculation.

Improved Null Distribution Modeling: In order to better represent our data, we improve the model of the underlying distribution by allowing for truncated, censored, and bi-modal distributions.

Truncated and censored distributions can occur in biological data due to the effects of limited measuring instruments. Truncated distributions are like normal distributions, except with values above and/or below certain thresholds removed. Censored distributions are similar, but rather than removing the data, the mass of the censored data is placed at the extremal values of the remaining distribution. Example plots of truncated and censored distributions are illustrated in figure 2.

As well, gene expression data can sometimes be multi-modal due to patient differences in the underlying population. In order to model multi-modal distributions, we used the Expectation-Maximization (EM) Algorithm [4]. The EM Algorithm is commonly used for Gaussian Mixture Models, which are mixtures of normal distributions. We used an existing implementation of the EM Algorithm and modified it so that it could not only create mixtures of normal distributions, but also truncated and censored distributions [13], [6], [12]. In order to choose from among candidate distributions, we used a model selection algorithm that uses the Bayesian Information Criterion (BIC) to find the most likely distribution without overfitting. Our algorithm was based on a formulation described by [2] that generates many possible model formulations and then calculates the BIC, which balances an maximal log-likelihood of the data belonging to that model with a regularization term. In order to prevent overfitting, the regularization term penalizes the overall number of distribution parameters.

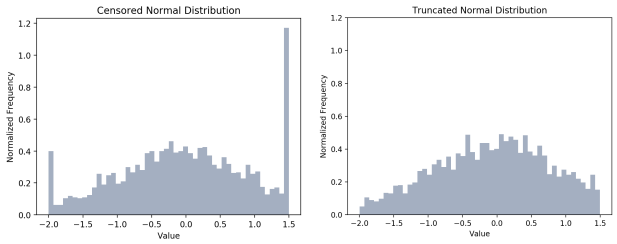

Fig. 2. The left figure shows a censored distribution, in which values outside of the specified region are assigned to the nearest value within the region. The right figure shows a truncated distribution, in which values outside of the specified region are simply removed.

Improved p-value measurement: Once we were able to model a variety of distributions, we needed to update our method of significance testing. The normal *p*-value can be described as follows:

$$p = \mathbb{P}_{\mathsf{n}}(n > X)$$

This standard measure of *p*-value describes the probability of obtaining a value as extreme or more extreme than the observed value, by chance. Since the underlying distribution is assumed to be normal, this measurement can be done by determining the area under the normal curve within some number of standard deviations from the mean (assuming a two-tailed test). However, this standard method of calculating *p*-value does not produce appropriate results for bi-modal distributions. As can be seen in the blue line in figure 3, the assumption of normality results in misrepresentative *p*-values. In order to have a more appropriate measure of significance, we implemented a probability-based measure:

$$p = \mathbb{P}_{\mathsf{n}}(p_{\mathsf{n}}(n) > p_{\mathsf{n}}(X))$$

The result of using this probability-based *p*-value is shown by the orange line in figure 3. It is clearly much more representative of the shape of the underlying distribution than are the standard *p*-values. While the

standard p-value calculation does not capture the decreased likelihood of obtaining results in this region, our method does. One clear region of interest is in between the two modes of the data. While the standard model assigns these values a relatively high p-value, our model correctly detects them as rare and assigns them lower p-values.

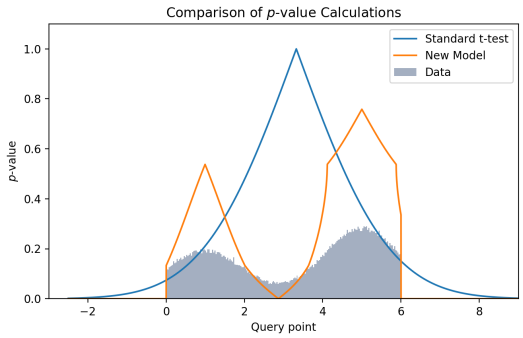

Fig. 3. For a bimodal truncated distribution, the p-value as calculated by a standard t-test misrepresents the distribution in two key ways. It does not assign a probability of zero to points outside the region of truncation. Also, it does not account for the bi-modality of the data: although query points near 3 are actually close to the mean of the distribution, they are still unlikely to occur since they are in a valley between the two modes.

B. First Order

The first order method allows us to look at the correlative relationships between genes. By comparing the correlations between different genes in different disease states, we might be able to gain information not present in the zeroth order analysis. Using the zeroth order analysis, we would compare the expression level of a given gene in a disease state to the distribution of expression levels in the control state and determine how significant the expression level is for the disease state. However, for first order analysis, we compare the correlative relationships between all genes in both the control case and in the disease case and can then extract information about which correlative strengths change the most.

For this work, we analyzed gene expression data of cells from Healthy Control Subjects (HCS) and Schizophrenia patients (SCZ). For each disease state, we used the pairwise correlation between all genes to create a correlation matrix. This matrix can be visualized as a graph, in which nodes are genes and weighted edges are the strength of correlation between two genes. A representative sample of genes is shown in Figure 4 for each state. In order to determine what genes differed the most between the HCS state and the SCZ state, we subtracted the two correlation matrices and summed the edges touching each node. This allows us to determine which correlative relationships between genes change the most between the two states.

In order to determine whether the given gene expression levels were more different than could be expected by chance, we had to establish a baseline. After calculating the amount of difference between the two conditions for each gene, we used a permutation test to determine statistical significance. We performed this permutation test by combining all of the HCS and SCZ data and then randomly splitting the data into two equal-sized matrices. Then, we perform correlation analysis on the differential expression of genes between the two random matrices. We repeated this random permutation 100 times, and each time recorded the maximum difference in expression level of any gene between the two matrices. This was our baseline, since theoretically there shouldn't be a significant difference between two such randomly shuffled matrices.

After establishing this baseline as a null distribution, we determined the p-value for each of the correlation sum values found for each gene. We identified four genes that had significantly different expression pattens in the SCZ state and the HCS state.

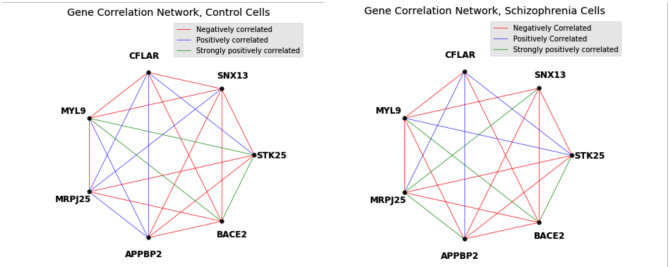

Fig. 4. Different colors represent different correlative strengths between genes. This visualization allows us to see the types of possible differences between HCS cells and SCZ cells.

V. Results

Since our zeroth order method was solely tested on synthetic data, the results we obtained were more indicative of the correctness of our model rather than of any interesting outcomes. Based on figure 5, we can see that our model outperforms the standard procedure for detecting anomalous gene expression levels. We plot the regions of a on synthetic truncated bi-modal distribution where our model detects statistical significance but SciPy Stats method `ttest_ind_from_stats` [9] does not. We demonstrate that, by modeling the data better, we can perform more accurate and informative hypothesis testing than by using a standard t-test.

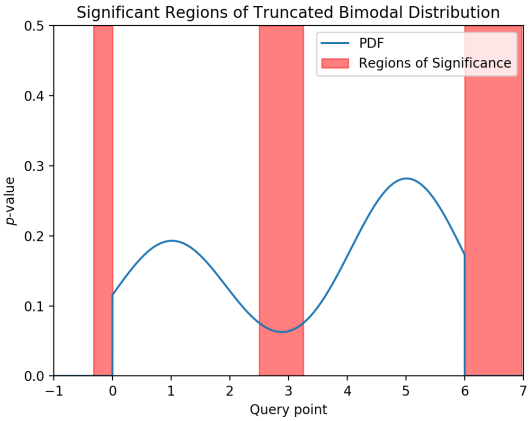

Fig. 5. Highlighted regions of this graph show areas that our p-value calculation determines to be statistically unlikely ($p < 0.05$) but a standard t-test does not detect as statistically significant.

We were able to obtain results from our first order method on real data from the experiment done at MGH using cells from healthy control subjects and patients with Schizophrenia. We constructed matrices of the strengths of pairwise correlations between genes and analyzed which gene-to-gene relationships differed the most in the two conditions. This data was then used to determine which single gene expression levels changed the most.

After establishing a baseline with a permutation test, we could evaluate the significance of the genes found through our correlative analysis. Using the results of our permutation test as our null distribution, we used a p-value of 0.05 to determine which genes were significantly differentially expressed between the two states, SCZ and HCS. We found four genes that were expressed significantly more in SCZ cells: MPZL1, ACLY, STXBP1, and FOXJ3. The p-values of these genes are shown in table I. All of these genes are related to neurons or neurotransmitters and MPZL1 has even been found to be related to Schizophrenia [7].

TABLE I
SIGNIFICANTLY DIFFERENTIALLY EXPRESSED GENES BETWEEN HCS AND SCZ CELLS

Gene	Significance
MPZL1	≤ 0.01
ACLY	0.02
STXBP1	0.02
FOXJ3	0.05

VI. DISCUSSION

A. Interpretation of Results

Zeroth Order Method: An improved method of hypothesis testing, as demonstrated by our zeroth order methods, can be extremely valuable for non-normal gene expression data. The p-value is commonly used to determine whether a given value is significantly different than what would be expected based on the null distribution. A p-value of 0.05 is typical, and means that there is a 5% chance that this value belongs to the null distribution. While the standard formulation of p-value can be valuable for many unimodal distributions, there are important exceptions that show the limited scope of this metric. For instance, the bi-modal data demonstrated that a standard p-value does not detect points in between the two modes as being statistically significant, even though points in that region are much less likely to belong to the bi-modal distribution. As well, points outside the truncated region will never occur and should thus be assigned a p-value of 0 since a value in that region would be very anomalous. However, due to the simplicity of the standard p-value, it is not sensitive to truncation or censoring. The modifications we made allow for more accurate and meaningful significance testing.

Genes Found Using First Order Method: Regarding our first order method, we found that the most differentially expressed genes seemed to be related to neurons or neural development. [7] shows that MPZL1 is related to Schizophrenia in a study done on a Han Chinese population. ACLY may be related to synthesis of acetylcholine, a neurotransmitter that operates at the neuromuscular junction [20]. STXBP1 is related to regulating the release of neurotransmitters at neural synapses [21]. FOXJ3 was found to be involved in neural development [11].

B. Limitations

We recognize several limitations in our work that are worth mentioning. First, our zeroth order methods were only tested on synthetic data. While our method clearly detects certain significant regions better than does a standard t-test, this improvement may be negligible in real data. Perhaps assuming a normal distribution works well enough, and does not merit the additional modeling. Second, we only had data from four patients in performing our first order analysis. This small patient sample size results in limited generalizability. Differences detected by our model could simply be individual differences rather than anything indicative of a disease state. We understand that the limited sample size reduces the significance of our discovered genes. However, we hope to apply the same methods to larger datasets in order to

validate our findings. Third, we are performing relatively naive correlational analysis methods. Ideally, we would be able to incorporate biological priors and perform higher order analysis on the relationships present in the network. Gene Set Enrichment Analysis [17] could have interesting connections to correlative work like ours since it groups genes by biological similarity.

C. Future Work

Although we have validated our hypothesis testing method on synthetic datasets, our next step is to run our model on real genetic data. As well, we hope to combine our zeroth and first order methods so as to perform a more accurate analysis of our data. We hope that these results will allow biological researchers to discover genes of interest more easily and more accurately. In addition, we hope to run our model on larger datasets, so as to correct for our second limitation of a small sample size. Finally, we want to package the software so that it is easily usable by biological researchers.

Differential gene extraction can often be challenging due to noisy, high-dimensional datasets. In this paper, we demonstrate that improved modeling of the underlying distributions, and correlative analysis may aid in analysis of gene expression datasets. Our results show that these methods could be promising in identifying significantly differentially expressed genes. As well, these methods could for improved modeling of non-normal distributions in a variety of fields. Our work has implications for both aiding our collaborators at MGH and for broader applications in improving analysis of gene expression datasets.

VII. Conclusion

In this paper, we offer three main contributions:

- We demonstrate a *zeroth order* method to model non-normal distributions, including bi-modal mixtures with truncated or censored components.
- We use a probabilistic p-value to better perform hypothesis testing for non-normal distributions.
- We use a *first order* method of correlative analysis to identify genes with significantly different patterns of expression in patients with Schizophrenia as opposed to healthy control subjects.

VIII. Acknowledgements

Many thanks to Matthew McDermott and Professor Peter Szolovits for supervising my research, to the 6.UAR staff, and to Roy Perlis, Steve Haggarty, Wen-Ning Zhao, Jennifer Wang, Ting Fu, Steve Sheridan, Isaac Kohane, Susanne Churchill, Kamber Hart, and the overall MGH and CEGS Teams; NIH (Grant P50-MH106933).

References

[1] Aliferis, Constantin F., Alexander Statnikov, and Ioannis Tsamardinos. *Challenges in the Analysis of Mass-Throughput Data: A Technical Commentary from the Statistical Machine Learning Perspective.* Cancer Informatics 2006: 133162. Print.

[2] Calcagno, Vincent & Claire de Mazancourt. *glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models.* Journal of Statistical Software [Online], 34.12 (2010): 1 - 29. Web. 22 Feb. 2018

[3] Clark, Neil R. et al. *The Characteristic Direction: A Geometrical Approach to Identify Differentially Expressed Genes*. BMC Bioinformatics 15.1 (2014): n. pag. BMC Bioinformatics. Web.

[4] Dempster, A. P., et al. *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, no. 1, 1977, pp. 138. JSTOR, JSTOR, www.jstor.org/stable/2984875.

[5] Greene, Casey S. et al. *Understanding Multicellular Function and Disease with Human Tissue-Specific Networks.* Nature genetics 47.6 (2015): 569576. PMC. Web. 17 May 2018.

[6] Gyemin Lee, Clayton Scott. *EM algorithms for multivariate Gaussian mixture models with truncated and censored data*. Computational Statistics & Data Analysis, Volume 56, Issue 9, 2012, Pages 2816-2829, ISSN 0167-9473, https://doi.org/10.1016/j.csda.2012.03.003.

[7] He G, Liu X, Qin W, Chen Q, Wang X, Yang Y, Zhou J, Xu Y, Gu N, Feng G, Sang H, Wang P, He L. *MPZL1/PZR, a novel candidate predisposing schizophrenia in Han Chinese.* Mol Psychiatry. 2006 Aug;11(8):748-51. Epub 2006 May 9. PubMed PMID: 16702974.

[8] Jeanmougin M, de Reynies A, Marisa L, Paccard C, Nuel G, Guedj M (2010) *Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies.* PLoS ONE 5(9): e12336. https://doi.org/10.1371/journal.pone.0012336

[9] Jones E, Oliphant E, Peterson P, et al. *SciPy: Open Source Scientific Tools for Python, 2001-*, http://www.scipy.org/ [Online; accessed 2018-05-14].

[10] Lai, Carmen et al. *A Comparison of Univariate and Multivariate Gene Selection Techniques for Classification of Cancer Datasets.* BMC Bioinformatics 7 (2006): 235. PMC. Web. 17 May 2018.

[11] Landgren, H. and Carlsson, P. (2004), *Foxj3, a novel mammalian forkhead gene expressed in neuroectoderm, neural crest, and myotome.* Dev. Dyn., 231: 396-401. doi:10.1002/dvdy.20131

[12] James W. Jawitz, *Moments of truncated continuous univariate distributions*, Advances in Water Resources, Volume 27, Issue 3, 2004, Pages 269-281, ISSN 0309-1708, https://doi.org/10.1016/j.advwatres.2003.12.002.

[13] Seemayer, Stefan. *mix'EM*. https://mixem.readthedocs.io/en/latest/

[14] Smyth G.K. (2005) *limma: Linear Models for Microarray Data.* In: Gentleman R., Carey V.J., Huber W., Irizarry R.A., Dudoit S. (eds) Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health. Springer, New York, NY

[15] Stuart, J et al. *A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules.* Science (2003): vol. 302 no. 5643 249-255.

[16] Subramanian, Aravind et al. *A next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles.* Cell 171.6 (2017): 14371452. Print.

[17] Subramanian, A. et al. *Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.* Proceedings of the National Academy of Sciences 102.43 (2005): 1554515550. Proceedings of the National Academy of Sciences. Web.

[18] Tusher, V. G., R. Tibshirani, and G. Chu. *Significance Analysis of Microarrays Applied to the Ionizing Radiation Response.* Proceedings of the National Academy of Sciences 98.9 (2001): 51165121. Proceedings of the National Academy of Sciences. Web.

[19] B. L. Welch. *THE GENERALIZATION OF STUDENT'S PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED*. Biometrika, Volume 34, Issue 1-2, 1 January 1947, Pages 2835, https://doi.org/10.1093/biomet/34.1-2.28

[20] https://www.genecards.org/cgi-bin/carddisp.pl?gene=ACLY

[21] https://www.genecards.org/cgi-bin/carddisp.pl?gene=STXBP1