Unsupervised Single-Cell Analysis in Triple-Negative Breast Cancer: A Case Study

Arjun P. Athreya, Alan J. Gaglio,
Zbigniew T. Kalbarczyk, and Ravishankar K. Iyer
University of Illinois at Urbana-Champaign, USA
Email:{athreya2, agaglio2, kalbarcz, rkiyer}@illinois.edu

Junmei Cairns, Krishna R. Kalari,
Richard M. Weinshilboum, and Liewei Wang
Mayo Clinic, Rochester, MN, USA
Email: {cairns.junmei, kalari.krishna, weinshilboum.richard, wang.liewei}@mayo.edu

***Abstract*—This paper demonstrates an unsupervised learning approach to identify genes with significant differential expression across single-cell subpopulations induced by therapeutic treatment. Identifying this set of genes makes it possible to use well-established bioinformatics approaches such as pathway analysis to establish their biological relevance. Then, a biologist can use his/her prior knowledge to investigate in the laboratory, a few particular candidates among the subset of genes overlapping with relevant pathways. Due to the large size of the human genome and limitations in cost and skilled resources, biologists benefit from analytical methods combined with pathway analysis to design laboratory experiments focusing on only a few significant genes. As an example, we show how model-based unsupervised methods can identify a small set of genes (1% of the genome) that have significant differential expression in single-cells and are also highly correlated to pathways (p-value $< 1E-7$) with anticancer effects driven by the antidiabetic drug metformin. Further analysis of genes on these relevant pathways reveal three candidate genes previously implicated in several anticancer mechanisms in other cancers, not driven by metformin. Identification of these genes can help biologists and clinicians design laboratory experiments to establish the molecular mechanisms of metformin in triple-negative breast cancer. In a domain where there is no prior knowledge of small biologically significant data, we demonstrate that careful data-driven methods can infer such significant small data to explain biological mechanisms.**

1. Introduction

Population studies have shown that an anti-diabetic drug, metformin, inhibits cancer growth in various types of cancer including triple-negative breast cancer [1], [2]. Triple-negative breast cancer is a molecular subtype of breast cancer that does not have any standard targeted therapies [3], [4]. Further, the molecular mechanisms of metformin's response in triple-negative breast cancer are not yet known.

The purpose of the overall project driving this paper is to infer the molecular mechanism by which metformin inhibits cancer growth in triple-negative breast cancer cells. The workflow of our analysis is illustrated in Fig. 1. In order to differentiate metformin's impact on the breast cancer cells, we used two identical breast cancer MDA-MB231 cell populations, 192 cells (2 assays, each comprising 96 cells) not treated with metformin (referred to as *baseline cells*), and an equal number of the same cells treated with metformin (referred to as *metformin-treated cells*). The cells were sequenced using single-cell RNA sequencing (scRNA-seq) technology, and the resulting data comprise the expression measure for each gene of the sequenced genome contained in each of the cells under study [5]. The data reflect $23,398$ genes and their associated gene expressions for baseline and metformin-treated cells. Thus, the overall data consist of 9M[1] gene expression values. Towards the said purpose of this project, the goal of the analytics in this work was to infer clusters of metformin-treated cells and then identify a small group of differentially expressed genes across clusters. These genes can then be used to identify associated diseases and pathways ("series of actions among molecules in a cell that leads to a certain product or a change in the cell" [2]) by performing pathway analysis. Combining differentially expressed genes overlapping with relevant pathways (obtained from our pathway analysis) and available data on these genes from existing literature in the context of metformin and anticancer mechanisms, we choose genes which have been implicated or studied in anticancer experiments. These few genes then become potential candidates for laboratory experiments to help establish molecular mechanisms of metformin in triple-negative breast cancer.

When we used a mixture model-based unsupervised learning approach, 310 of the $23,398$ genes were found to be significantly differentially expressed in six metformin-treated cells. That set of 310 genes is small enough to be manageable with well-understood bioinformatics approaches, such as pathway analysis. As a substantiation of our learning approach, pathway analysis of the differentially expressed genes showed strong correlations with three pathways: i) oxidative phosphorylation (p-value $3.81E-21$), ii) the citric acid (TCA) cycle, and the respiratory electron transport (p-value $2.10E-19$) and, iii) mitochondrial translation (p-value $1.41E-07$) pathways. All of these pathways have recently found to have anticancer properties, via both in-vivo and in-vitro experiments [6], [7], [8], [9]. Further, among the differentially expressed genes overlapping with

1. [192*2 cells]*$23,398 \approx$ 9M
2. https://www.genome.gov/27530687/biological-pathways-fact-sheet/

978-1-5090-1610-5/16/$31.00 ©2016 IEEE

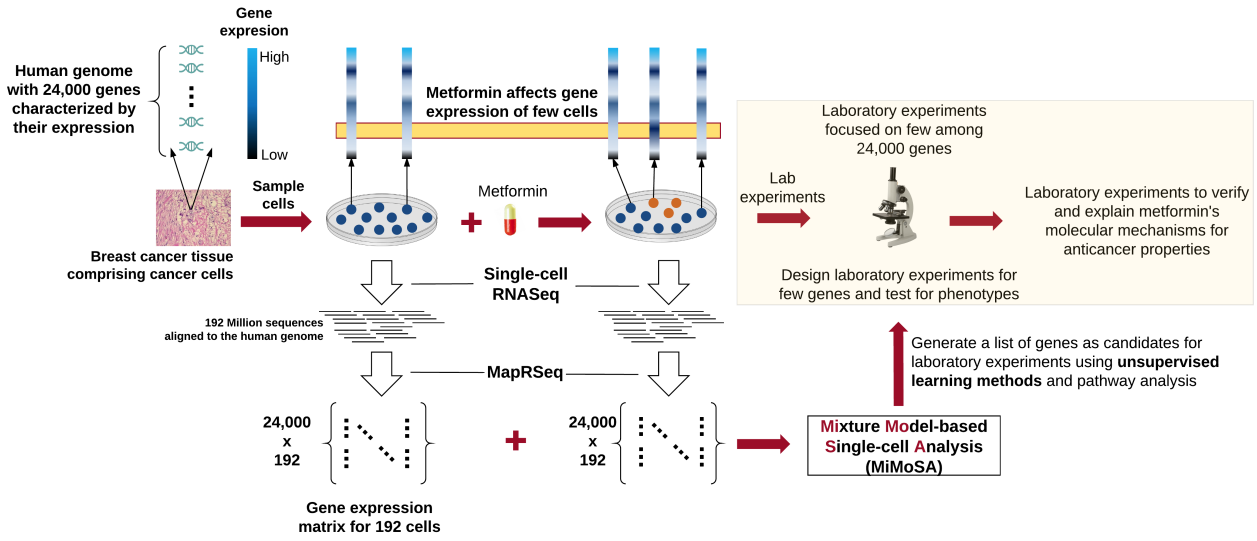


Figure 1: The single-cell RNAseq analysis workflow, starting from sequencing of cancer cells to generation of gene expression matrices and the use of unsupervised learning methods, combined with pathway analysis, to generate a list of few genes the become candidates for informing the design of focused laboratory experiments, which will be the basis of our future work.

those pathways, we have identified the NDUFB9, COX5B, and MRPS7 genes which have been implicated in other anticancer mechanisms for other cancers not driven by metformin. Hence, these three genes are now ideal candidates for laboratory experiments to establish metformin's mechanisms in triple-negative breast cancer.

Traditional bulk sequencing enabled the study of aggregate gene expressions in a tumor sample. However, with scRNA-seq, the amount of data is significantly larger and we have gained finer differentiation of the cells by using distributions of gene expression as opposed to the single aggregate value of gene expression provided by bulk sequencing. For example, scRNA-seq generates about 1 million RNA sequences per cell comprising roughly $24,000$ genes. With 96 cells comprising a sequencing panel, and for two panels analyzed, 192M sequences are generated (see Fig. 1 for the analysis workflow). The value of such fine granularity in data is that we can now cluster cells based on their associated distributions of gene expression.

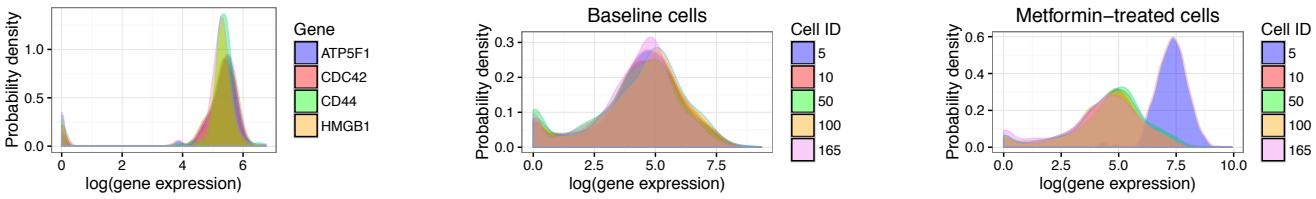
Key additional contributions of this work are as follows,

1) **Tool Development:** We use the mixture model-based clustering algorithm embodied in a tool, Mixture Model-based Single-cell Analysis (MiMoSA), to cluster cells based on the similarity of their gene expression distributions, by minimizing Kullback-Liebler divergence between distributions (Sec. 4.1.1). In our preliminary analysis of the data (Sec. 3), we observed that the probability density function (PDF) of the gene expression in the baseline cells was best explained as a mixture of Gaussians (Fig. 2(b)). Further, some metformin-treated cells had a component of the mixture in the distribution of gene expression significantly phase-shifted as shown in Fig. 2(c). This observation made the use of k-means clustering, which assumes normality in the data, and is therefore inappropriate for clustering cells with mixtures in their distributions.
2) **Validation using Alternative Mathematical Formulations:** Cluster assignment of metformin-treated cells did not change when a variety of hierarchical clustering algorithms were used, thereby validating MiMoSA's inference via methods with different underlying mathematics (Sec. 4.3). The choice of the hierarchical clustering algorithm was driven by the fact that the differentially expressed genes were major drivers of the clustering behavior when distributions were used to cluster cells. Hence, the six cells would be clustered together by algorithms that use pairwise euclidean distances as a similarity measure.
3) **Test Dataset and Tool Access:** We provide access to a test dataset and MiMoSA, which is compatible with the Linux, OS X and Windows operating system and the IBM's POWER8 and Intel's x86 architectures.

2. Related Work

The recently proposed methods for analyzing single-cell data have largely focused on finding subpopulations of cells in a population of cells [10], [11]. All of the proposed methods include two steps of processing, first reducing the number of genes being used to cluster cells, and then using a clustering method to find subpopulations of cells. Further, all these methods have found that only a few thousand genes are significantly differentially expressed in cell samples [10], [11], [12]. For the second step of these analyses, supervised, unsupervised, or graphical model approaches [12] are used.

The first step of the analysis tries to retain the genes that show variation in their expression levels across the samples. For example, in the transcriptome analysis of lung



(a) Gene (b) Baseline cells (c) Metformin-treated cells

Figure 2: The existence of mixtures in the feature space and samples is illustrated by the probability density functions (PDF) of gene expressions in a set of genes across (a) all baseline cells, (b) a set of baseline cells and (c) a set of metformin-treated cells.

adenocarcinoma [10], the method to reduce the dimensionality of the data (for genes in this context) was to start by looking at genes (expressed across all the samples) whose gene expressions were measured as greater than 0, thereby reducing the gene list to about $9,000$ genes. Then, the authors of [10] studied the correlation of gene expression among these genes by using Pearson's correlation analysis, and reduced the gene list to about $5,500$ genes by choosing genes with correlation coefficients greater than 0.9. In an analysis of cell-to-cell heterogeneity that revealed subpopulations [11], prior knowledge of cell-cycle genes was used, and only genes that showed significant correlation were chosen, bringing the gene list down from $23,398$ to $2,881$.

To reduce the number of genes needed to infer cell types (assuming that the data are normally distributed), shared-nearest neighbors (SNN) or k-means clustering is used [12]. In particular, when SNN is used on simulated and real cell data, it has been shown that it performs better than k-means clustering when no biological priors are used [12]. However, there remains an open problem on how to choose the optimal number of neighbors and k values; currently, we must perform an extensive search of values or use heuristics to estimate the best values. One particular approach that is different from the normal two-step process is the use of diffusion maps [13], a supervised approach. The authors of [13] assume the existence of known types of cells and then use a transition matrix for classifying cells based on the state they best match their signature to. To do so, the authors needed to define the Gaussian kernel and further approximate the transition probabilities, and those steps are hinged on the assumption that the cell types are known.

All the aforementioned methods for single-cell analysis are driven either by known correlations between genes and biological mechanisms, or by supervised methods that use cell signatures. However, in problems where either there are many mechanisms related to drug response or we do not know the mechanism by which the drug impacts the cells, we have to turn to data-driven methods that first study characteristics of the data and then choose a method/algorithm to apply on the data. Further interpretation of the chosen method's results requires interactions with domain experts to address the primary goal of the analysis. To the best of our knowledge, no methods exist that can use mixture-model distributions to infer clusters of cells, despite the observation that gene expression of cells is best described using mixture models. Hence, in this work, we propose a data-driven, mixture-model-based single-cell analysis (MiMoSA) to infer clusters of cells using gene expression distributions.

Multiple research gaps are addressed in our work that have been overlooked by previous analysis methods.

1) Our case study demonstrates a consistent method to go from data generation from drug intervention, to identifying major candidates for focused laboratory experimentation to establish a drug's molecular mechanisms.
2) Our work does not make prior assumptions on gene correlations with drug response to reduce the number of genes for analysis.
3) Our choice of method was driven completely by observations made in our preliminary analysis, which revealed that distributions of gene expression in cells were best described by mixtures of Gaussians; this observation was aided by the fine resolution of data provided by scRNA-seq. The observation meant that the k-means clustering algorithm, which assumes normality in the data, was not suitable inferring clusters of cells in this work.

3. Data and Data Characteristics

3.1. Data

The MDA-MB-231 breast cancer cell line (ATCC HTB-26) was cultured in Leibovitz's L-15 medium with 10% fetal bovine serum for 5 days with and without metformin. Duplicate cultures were processed for single-cell analysis. Single cells with and without metformin were captured on a large-sized ($17-25\mu m$ cell diameter) microfluidic mRNA-seq chip known as the C1 Single-Cell Auto Prep IFC, using the C^{TM} Single-Cell Auto Prep System (Fluidigm Corporation, South San Francisco, CA). Cells were loaded onto the chips at a concentration of 300 cells/μl, stained for viability with a LIVE/DEAD cell viability assay kit (Life Technologies, Carlsbad, CA), and imaged by phase contrast and fluorescence microscopy to assess the number and viability of cells per capture site. Only single, live cells were included in the analysis. cDNAs were prepared on chip using the SMARTer Ultra Low RNA kit for Illumina (Clontech Laboratories, Mountain View, CA). Single-cell cDNA

size distribution and concentration were measured with a Quant-iT Pico green dsDNA assay kit (Life Technologies). Illumina libraries were constructed in 96-well plates using the Illumina Nextera XT DNA Sample Preparation kit using the protocol supplied by Fluidigm. Libraries were quantified by Agilent BioAnalyzer, using a high-sensitivity DNA analysis kit. Single-cell Nextera XT (Illumina) libraries of one experiment were pooled and sequenced at 100 bp paired-end on Illumina Hiseq to a depth of about 1 million reads. Single-cell mRNA-Seq data were processed using MAP-Rseq pipeline [14].

3.2. Data Characteristics and Pre-processing

The data characteristics are as follows.

1) The data comprises 192 baseline cells, and an equal number of cells treated with metformin. 23,398 genes were sequenced in each cell, and MAP-Rseq [14] was used obtain gene expression from sequencing data.
2) The expression value for each gene was normalized by accounting for the sequencing depth (number of short DNA sequence strings from the sequencing platform aligned to a gene) and length of the gene. The measure of gene expression is Reads Per Kilo Million (RPKM). The range of the RPKM is between 0 and 2,000.
3) Only 20% of the 23,398 genes show expression levels greater than 32 on the RPKM scale, which is a heuristic that can be used to decide whether a gene is expressed or not as recommended by MAP-Rseq [14].
4) Roughly 10% of the baseline and metformin-treated cells had low sequencing coverage ($<$ 1M reads per cell), so we excluded those cells from our analysis.
5) The density functions of gene expression of a gene across all cells, and for all genes in a cell comprise mixtures as shown in Figures 2(a) and 2(b) respectively. We used mclust package to fit a distribution for each gene and the best fit was a Gaussian Mixture Model, where each component of the mixture was a Gaussian. Using the fitted distribution, we performed non-parametric tests (Kolmogorov-Smirnov and Wilcoxon-rank tests) against the distribution derived from the data. The null hypothesis in this test is that the two distributions have equal means. The p-value in these tests were greater than the significance level of 0.05, thereby failing to reject the null hypothesis, meaning that the model fit was statistically close to the actual data's distribution.
6) In some metformin-treated cells, at least one component of the mixture had phase-shifted significantly. Therefore, metformin was affecting these cells in a way differently from other cells, thereby making the drug's effect on the cells non-uniform as shown in Fig. 2(c).

We observed that 80% of the genes were considered inactive in the data and as our focus in this analysis was on the impact of metformin measured by changes in gene expression, only genes in the top 5% of variance across cells were considered. That narrowing of the data set reduced the number of genes (features) for our analysis from 23,398 to 1,170. We used the reduced set of genes and the samples as inputs to the unsupervised learning methods. This way of reducing feature space is common in bioinformatics practices, although there is no standard threshold for the amount of variance to consider in gene expression profiles. The general assumption is that only a few biological pathways, comprising $100 - 400$ genes, are affected by a treatment, and thus these genes would highly likely be present within the top 5% most variable genes.

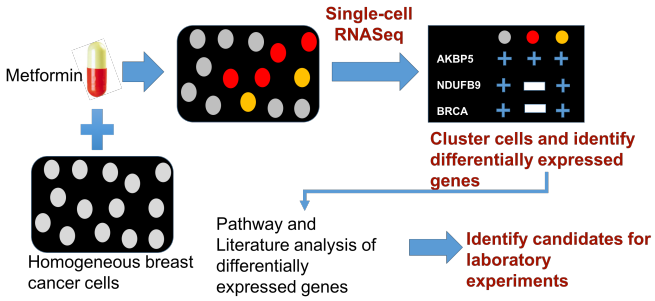


Figure 3: The unsupervised learning view of single-cell analysis is explained as follows. Cells exposed to a drug may exhibit differences in their gene expression behavior due to the molecular interactions with the drug, and these differences are not known. The computational problem is to cluster these cells (samples), and extract the genes (features) that are behaving differently compared to other clusters (referred to as differentially expressed genes). These differentially expressed genes are analyzed to study their biological significance in all known disease and molecular pathways.

4. Methods and Results

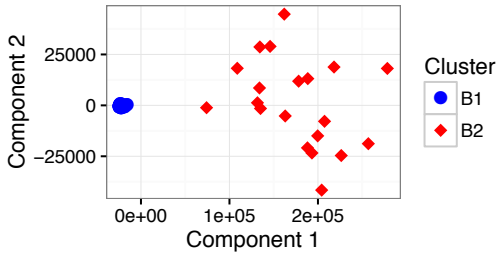
We describe the methods used to analyze single cells with an unsupervised learning approach as shown in Fig. 3, as cell-types induced by metformin are not yet known.

4.1. Inferring Cell Subpopulations

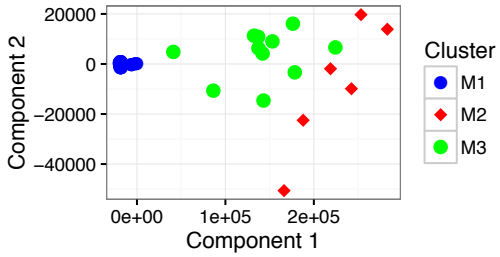
4.1.1. MiMoSA. Based on our knowledge of the presence of multiple distributions in our data as illustrated in Fig. 2, we chose to use probability distribution models to cluster the cells. Probability models cluster data according to a model (distribution) that best defines the data (such as Gaussian mixture model in this work), and treats each sample (cell) as being independent from other samples. The assumption of independence is acceptable, since each cell can behave differently and independently of other cells in response to metformin. We define the likelihood $\mathcal{L}_i$ of the mixture model for each cell y_i of N cells as shown in Equation 1, where there are K mixtures/components, f_k is the distribution of the component k, and θ_k is the distribution's parameters.

$$\mathcal{L}_i(\theta_1, ..., \theta_N; \tau_1, ...\tau_K|\mathbf{y}) = \prod_{i=1}^{N} \sum_{k=1}^{k=K} \tau_k . f_k(y_i, \theta_k). \quad (1)$$

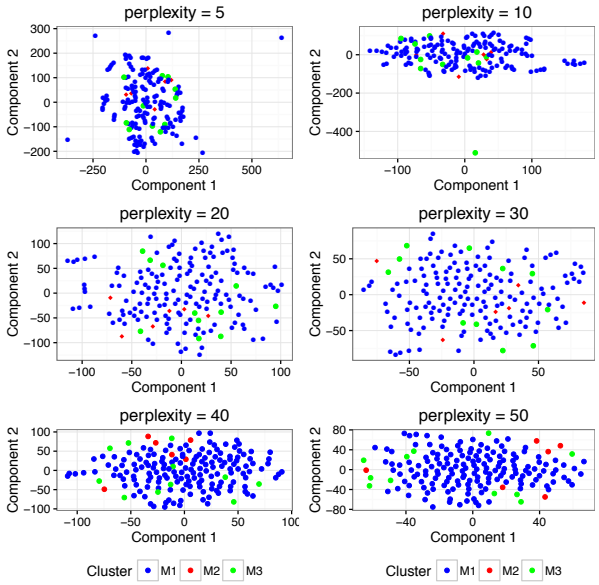
We chose to fit the data with Gaussian mixture models (GMM) with varying volume and finite mixtures, as this



(a) Baseline cell clusters projected using PCA



(b) Metformin-treated cell clusters projected using PCA



(c) Metformin-treated cell clusters projected using t-SNE

Figure 4: In Figs. (a) and (b), we project the baseline and metformin-treated cell clusters found by MiMoSA onto the first two principal components derived from principal component analysis (PCA). Fig. (c) shows that t-SNE was not as effective as PCA in helping to visualize the metformin-treated cell clusters.

provided the best likelihood score for fitting the data. The multivariate Gaussian distribution function is as defined in Equation 2, where $(\mu_k, \sum_k)$ is the mean and covariance of the component k. Because our model has mixtures of Gaussians, we need to compute the model parameters using the maximum likelihood (ML). For model estimation, there are two popular algorithms to choose from, the expectation maximization algorithm (EM) and the variational Bayesian EM algorithm (VBEM). Both algorithms are iterative and are known to have similar time complexities, and VBEM can be used to perform automatic model selection and is less prone to over-fitting when compared to EM. However, implementations of VBEM needed binning the gene expression measures, and with small sample space and large variation in the range of gene expression, the binning proved to be a challenge. Hence, we used the EM algorithm for ML to learn the model parameters. The EM algorithm is a two-step process. First, the E-step computes the conditional expectation of the observable data and current parameter estimate. Then, the $M-step$ maximizes the log-likelihood of the parameter estimates learned in the E-step.

$$\phi_k(y_i|\mu_k, \sum_k) = \frac{0.5exp\{(y_i-\mu_k)^T \sum_k^{-1}(y_i-\mu_k)\}}{\sqrt{det(2\pi \sum_k)}} \quad (2)$$

Once we have learned the model parameters, we then need to decide on an optimal number of clusters. One method that has historically provided a consistent estimator of the number of clusters is Bayesian Information Criteria (BIC) [15], where the value of K is chosen at which the BIC value asymptotically converges. We implemented the model-based clustering from the "mclust" package in R [16]. MiMoSA identified two clusters ($B1$,$B2$) in baseline cells, and three clusters ($M1, M2$,$M3$) in metformin-treated cells.

4.2. Visualizing Clusters

Since scRNA-seq allows for expression of the genome to be studied in each cell, it is also of interest to visualize the subpopulations inferred using single-cell analyses. Earlier works have either used linear methods such as the principal component analysis (PCA) [17] or non-linear methods such as t-distributed stochastic neighbor embedding (t-SNE) [18] to reduce the dimensionality of the data to two or three dimensions, and then project the clusters onto these lower dimensions [12], [19]. In this work we used both these methods to aid in visualizing the clusters inferred from MiMoSA in two-dimensional space (i.e. going from $1,170$ dimensions to two dimensions). The clustering visualizations for baseline and metformin-treated cells using PCA are shown in Figs. 4(a)- 4(b). We observe that metformin treatment induced very little variability in the gene expression across cells. Hence, majority of the cells are tightly clustered together in $M1$. Metformin-treated cells visualized using t-SNE are shown in Fig. 4(c). The authors of t-SNE suggest that there is no way to exactly know the optimal value of perplexity in the data, so values between 1 and 50 are suggested on a trial-and-error basis. Hence, we chose perplexity values of $[5, 10, 20, 30, 40, 50]$ and allowed for $100,000$ iterations with an epoch of 100 cycles; and in all these experiments on both datasets, the error in embedding

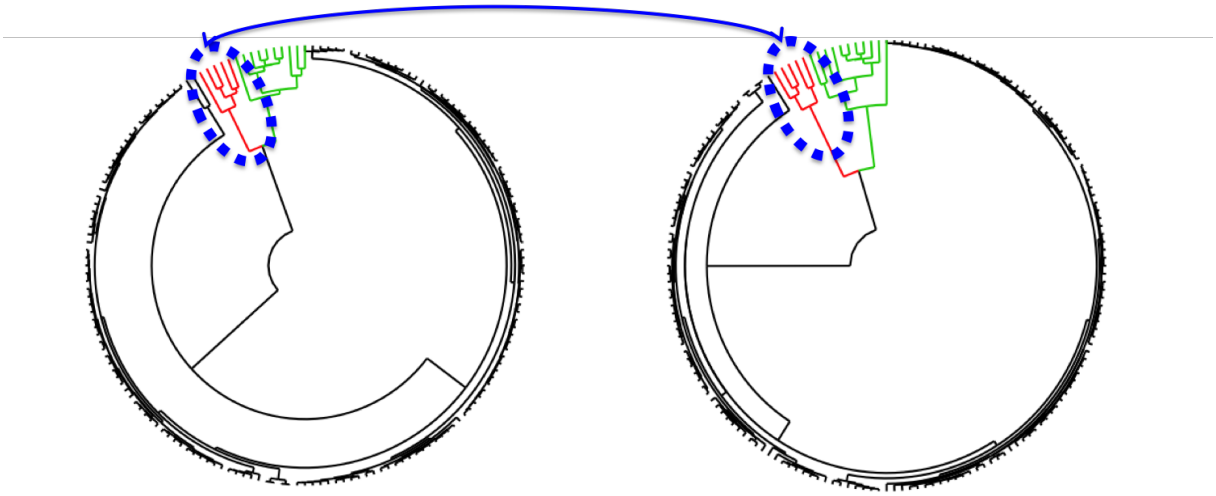


Figure 5: Dendrograms from the agglomerative (L) and divisive (R) hierarchical clustering methods. We show that the top levels of the clusters have the same numbers of cells (indicated by red lines). We also observed that these 5 cells were the same in both methods.

was less than 0.5%. From Fig. 4, we see that PCA provided better spatial separation of the cells.

4.3. Cluster validation: Hierarchical Clustering

We validate the clusters inferred by MiMoSA using hierarchical clustering methods. We performed both agglomerative and divisive hierarchical clustering on metformin-treated cells. The choice of hierarchical clustering was chosen because it is a method based on comparing pairwise similarity of features. Principal component analysis of metformin-treated cells showed that 97% of the variability was captured within the first two components. This mean't that it is likely that a few genes are probably significantly altered by metformin, while rest of the genes show little change in their expression. Hence, pairwise comparison of the cells would tend to cluster cells with similar changes in expression of the genome.

These analyses treat each of the N cells as a data point described by a set of M feature coordinates, where each individual feature coordinate is the expression of a gene. A relative measure of proximity between the cells was computed that encompassed all of the M gene expression levels. Agglomerative clustering was performed through successively merging of N total clusters, based on proximity, into a single, global cluster. Divisive clustering was performed in the inverse direction, beginning with a single cluster and successively splitting into N remaining clusters. Clustering was performed using the proximity matrix.

4.3.1. Proximity Matrix. The proximity matrix is formed using the features of the dataset, that is, the relative expression level of each gene for all cells. Since the expression measure has the same unit of measure for all genes in all the cells, we directly proceed to compute the distances. Commonly used distance metrics for hierarchical clustering include maximum (infinity norm), Euclidean (2 norm), and Manhattan (1 norm), among others. Given the exploratory nature of this work and the lack of community consensus on which distance metric best conveys proximity in the high dimensional space of SC-RNA data Euclidean distance was chosen because it can be easily interpreted. The transformation between the dataset and its associated distance matrix is as follows. For a given cell i has an expression level $E(i,k)$ associated with each gene $k \in [1:M]$, the Euclidean distance $D(i,j)$ between the i^{th} and j^{th} cells is given by

$$D(i,j) = \sqrt{\sum_{k=0}^{M} (E(j,k) - E(i,k))^2}. \tag{3}$$

Thus, the proximity matrix is symmetric ($D(i,j) = D(j,i)$), and we now have a measurement of proximity between cells with which we can use as inputs for clustering methods described next. A heuristic must be defined for how distance is measured between multi-child clusters, that is, clusters that contain multiple data points. Commonly used methods include single-linkage, complete-linkage, average-linkage, and centroid-linkage. Single- and complete-linkage methods report the distance from one cluster to another as the minimum and maximum distances between two individual data points of each cluster, respectively. In contrast, the average and centroid-linkage methods take the average euclidean distance of all child data points between clusters and the distance between centroid clusters respectively. The average and centroid-linkage methods act as a compromise between single and complete-linkage methods [20]. To encourage cluster compactness, hierarchical clustering was performed using complete-linkage as is commonly reported within the literature [21].

Direct comparison of the agglomerative and divisive hierarchical clustering methods reveals three main clusters within the gene expression data for metformin-treated cells, as shown in Fig. 5. The arrangement of branches on the

Method	Cluster 1	Cluster 2	Cluster 3
Model-based	160	6	12
Agglomerative	163	5	10
Divisive	162	5	11

TABLE 1: Heterogeneity inferred in metformin-treated cells

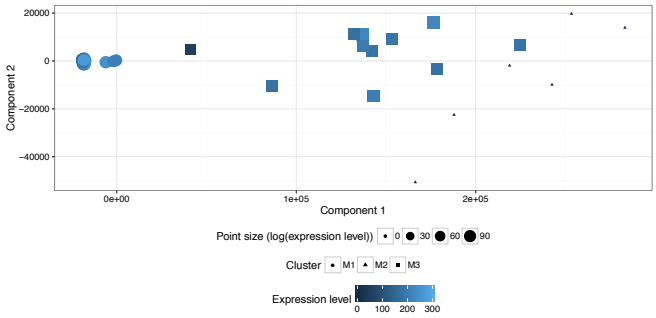


Figure 6: Differential expression of genes visualized in metformin-treated cells projected onto the first two principal components.

dendrograms indicates a level of similarity between cells within these branches. The radial length at which a branch occurs is a measure of similarity between cells within the branches. As seen in Fig. 5, the radial length of the red and green branches are far shorter than that of the black branch, indicating that the cells within the red and green branches are most dissimilar. Moving radially outward down the branches, there is similarity among cells until the leaf nodes, where the highest observed similarity exists. We note the similarity between the number of cells present in each of the three main clusters shown by Table 1. Further, we find that all the cells in cluster 2 of both hierarchical methods overlapped with $M2$ from MiMoSA, with $M2$ is having one additional cell. Going by the data-driven approach of this work, we decided to base all our further analysis on clustering results from MiMoSA.

5. Unsupervised Learning Informing Biology

The clusters inferred by MiMoSA in metformin-treated cells are characterized by 310 differentially expressed genes of which 200 are downregulated and about 100 are upregulated in cluster $M2$, compared to $M1$ and $M3$. Clusters $M1$ and $M3$ showed little variation in gene expression, and the cells comprising cluster $M2$ were the most affected by metformin as shown in Fig. 6. Specifically, for the downregulated genes, the size of the points (proportional to the expression of the genes) in Fig. 6 is comparable in clusters $M1$ and $M3$, and very small in cluster $M2$. Pathway analysis was performed to further underst the biological relevance of the differentially expressed genes. The top pathways (and the associated p-values) were oxidative phosphorylation ($3.8E{-}21$), the citric acid (TCA) cycle, and respiratory electron transport ($2.1E-19$) and mitochondrial translation ($1.4E-07$). These pathways are relevant in the context of metformin in several ways. 1) The possibility of chemo-prevention with metformin is being investigated by targeting the oxidaptive phosphorylation pathway [7]. 2) It has been shown that metformin inhibits cancer cell proliferation by regulating the TCA cycle [6]. 3) Metformin has been shown to target mitochondrial metabolism in cancer therapies [8], [9].

It is clear that the differentially expressed genes inferred from MiMoSA's clusters are on pathways known to have anticancer effects driven by metformin. Among these differentially expressed genes in the above listed pathways, we have identified the following genes which are implicated in anticancer mechanisms. 1) NDUFB9: an accessory subunit of the mitochondrial membrane respiratory chain NADH dehydrogenase (complex I), and loss of NDUFB9 promotes MDA-MB-231 cells proliferation, migration, and invasion; because of elevated levels of reactive oxygen species (ROS) [22]. 2) COX5B is a peripheral nuclear-encoded sub-unit of CcO (cytochrome c oxidase), and loss of COZ5B induces mitochondrial dysfunction and subsequently leads to suppression of cell growth and cell senescence [23]. 3) MRPS7 is a mitochondrial ribosomal protein, involved in mitochondrial translation, that is significantly elevated in human breast cancer cells, leading to amplified mitochondrial biogenesis and/or mitochondrial translation in epithelial breast cancer cells [24]. Therefore, mitochondrial biogenesis could be a potential target for anticancer agents and therefore could explain the retrospective success of metformin, which prevents the onset of nearly all types of cancer in diabetic patients, likely because it functions as a "weak" mitochondrial poison.

Future work based on current findings are the following, 1) investigate what makes the six metformin-treated "diagnostic" in terms of inferring metformin's response, as these cells seem to be more sensitive to metformin in comparison to other cells, and 2) conduct laboratory experiments for candidate genes identified in this work based on their differential expression after metformin treatment and their biological relevance from pathway analysis.

6. Tool Implementation and Availability

MiMoSA was developed in R, version 3.2.2. MiMoSA has been tested on and is compatible with the 32-bit and 64-bit Linux (Ubuntu 14 and later), OS X (Yosemite and later) and Windows (Windows 7 and later) operating systems. Further, in addition to being compatible with Intel's x86 architectures, MiMoSA is also compatible with the Power Architecture, and was tested on the IBM POWER8 processors, establishing that our tool is agnostic with respect to common, state-of-the-art high-performance computing platforms. The tool and sample data is available on request for now, and will be publicly available soon. The platforms and operating-system agnostic characteristics of MiMoSA means that it can easily be integrated into larger genomics and clinical workflows, both on stand-alone machines and in the cloud.

7. Conclusion

In this work, we proposed a model-based unsupervised learning method for using single-cells to help infer metformin's molecular mechanisms in triple-negative breast cancer, by inferring clusters in metformin-treated cells. Pathway analsysis of differentially expressed genes obtained from cluster analysis showed very significant correlations with pathways known to have anticancer effects driven by metformin. Further analysis of these pathways identified potential candidates for experiments in the laboratory, including NDUFB9, COX5B and MRPS7, upon which we will base our future work. Thus, we have now narrowed the search space for choosing genes as candidates for laboratory experiments. Based on the relevance of genes identified in the context of metformin and triple-negative breast cancer, a sub-type of breast cancer for which there are still no targeted drug therapies, in principle MiMoSA has the potential to accurately infer drug mechanisms in other diseases, if the cell's gene expression distribution is a mixture of Gaussians.

8. Acknowledgments

This material is based upon work partially supported by Mayo Clinic and Illinois Alliance Fellowship for Technology-Based Healthcare Research, CompGen Fellowship, IBM Faculty Award, National Science Foundation (NSF) under Grants CNS 13-37732, National Institutes of Health (NIH) under Grants RO1 GM28157, R01CA19664, U19 GM61388 (The Pharmacogenomics Research Network), Breast SPORE P50CA116201 and U19 GM61388 and Mayo Clinic Center for Individualized Medicine. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF and NIH. We thank Prof. Gene Robinson for helpful comments and Jenny Applequist in helping prepare the manuscript.

References

[1] B. Bao, Z. Wang, S. Ali, A. Ahmad, A. S. Azmi, S. H. Sarkar, S. Banerjee, D. Kong, Y. Li, S. Thakur *et al.*, "Metformin inhibits cell proliferation, migration and invasion by attenuating csc function mediated by deregulating mirnas in pancreatic cancer cells," *Cancer prevention research*, vol. 5, no. 3, pp. 355–364, 2012.

[2] H. A. Hirsch, D. Iliopoulos, and K. Struhl, "Metformin inhibits the inflammatory response associated with cellular transformation and cancer stem cell growth," *Proceedings of the National Academy of Sciences*, vol. 110, no. 3, pp. 972–977, 2013.

[3] C. A. Hudis and L. Gianni, "Triple-negative breast cancer: an unmet medical need," *The oncologist*, vol. 16, pp. 1–11, 2011.

[4] B. C. Litzenburger and P. H. Brown, "Advances in preventive therapy for estrogen-receptor-negative breast cancer," *Current breast cancer reports*, vol. 6, no. 2, pp. 96–109, 2014.

[5] C. Trapnell, "Defining cell types and states with single-cell genomics," *Genome research*, vol. 25, no. 10, pp. 1491–1498, 2015.

[6] T. Griss, E. E. Vincent, R. Egnatchik, J. Chen, E. H. Ma, B. Faubert, B. Viollet, R. J. DeBerardinis, and R. G. Jones, "Metformin antagonizes cancer cell proliferation by suppressing mitochondrial-dependent biosynthesis," *PLoS Biol*, vol. 13, no. 12, p. e1002309, 2015.

[7] M. Cazzaniga and B. Bonanni, "Breast cancer metabolism and mitochondrial activity: The possibility of chemoprevention with metformin," *BioMed research international*, vol. 2015, 2015.

[8] S. E. Weinberg and N. S. Chandel, "Targeting mitochondria metabolism for cancer therapy," *Nature chemical biology*, vol. 11, no. 1, pp. 9–15, 2015.

[9] M. Van Gisbergen, A. Voets, M. Starmans, I. de Coo, R. Yadak, R. Hoffmann, P. Boutros, H. Smeets, L. Dubois, and P. Lambin, "How do changes in the mtdna and mitochondrial dysfunction influence cancer and cancer therapy? challenges, opportunities and models," *Mutation Research/Reviews in Mutation Research*, vol. 764, pp. 16–30, 2015.

[10] J.-W. Min, W. J. Kim, J. A. Han, Y.-J. Jung, K.-T. Kim, W.-Y. Park, H.-O. Lee, and S. S. Choi, "Identification of distinct tumor subpopulations in lung adenocarcinoma via single-cell rna-seq," *PloS one*, vol. 10, no. 8, p. e0135817, 2015.

[11] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells," *Nature biotechnology*, vol. 33, no. 2, pp. 155–160, 2015.

[12] C. Xu and Z. Su, "Identification of cell types from single-cell transcriptomes using a novel clustering method," *Bioinformatics*, p. btv088, 2015.

[13] L. Haghverdi, F. Buettner, and F. J. Theis, "Diffusion maps for high-dimensional single-cell analysis of differentiation data," *Bioinformatics*, 2015.

[14] K. R. Kalari, A. A. Nair, J. D. Bhavsar, D. R. OBrien, J. I. Davila, M. A. Bockol, J. Nie, X. Tang, S. Baheti, J. B. Doughty *et al.*, "Mapr-seq: mayo analysis pipeline for rna sequencing," *BMC bioinformatics*, vol. 15, no. 1, p. 1, 2014.

[15] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American statistical Association*, pp. 611–631, 2002.

[16] ——, "Mclust: Software for model-based cluster analysis," *Journal of Classification*, vol. 16, no. 2, pp. 297–306, 1999.

[17] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, pp. 37–52, 1987.

[18] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, no. 2579-2605, p. 85, 2008.

[19] N. K. Wilson, D. G. Kent, F. Buettner, M. Shehata, I. C. Macaulay, F. J. Calero-Nieto, M. S. Castillo, C. A. Oedekoven, E. Diamanti, R. Schulte *et al.*, "Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations," *Cell stem cell*, 2015.

[20] S. Theodoridis and K. Koutroumbas, "Chapter 13 - clustering algorithms ii: Hierarchical algorithms," in *Pattern Recognition (Fourth Edition)*, fourth edition ed., S. Theodoridis and K. Koutroumbas, Eds. Boston: Academic Press, 2009.

[21] P. D'haeseleer, "How does gene expression clustering work?" *Nature biotechnology*, vol. 23, no. 12, pp. 1499–1501, 2005.

[22] L.-D. Li, H.-F. Sun, X.-X. Liu, S.-P. Gao, H.-L. Jiang, X. Hu, and W. Jin, "Down-regulation of ndufb9 promotes breast cancer cell proliferation, metastasis by mediating mitochondrial metabolism," *PloS one*, p. e0144441, 2015.

[23] S.-P. Gao, H.-F. Sun, H.-L. Jiang, L.-D. Li, X. Hu, X.-E. Xu, and W. Jin, "Loss of cox5b inhibits proliferation and promotes senescence via mitochondrial dysfunction in breast cancer," *Oncotarget*, vol. 6, no. 41, pp. 43 363–43 374, 2015.

[24] F. Sotgia, D. Whitaker-Menezes, U. E. Martinez-Outschoorn, A. F. Salem, A. Tsirigos, R. Lamb, S. Sneddon, J. Hulit, A. Howell, and M. P. Lisanti, "Mitochondria "fuel" breast cancer metabolism: fifteen markers of mitochondrial biogenesis label epithelial cancer cells, but are excluded from adjacent stromal cells," *Cell cycle*, vol. 11, no. 23, pp. 4390–4401, 2012.