Clustering, Biomarker and Cancer Model Selection Using Omics Data

by

Jian Zou

B.Sc. in Biotechnology and Applied Chemistry, Central China Normal University, 2017

M.Sc. in Biostatistics, Columbia University in the City of New York, 2019

Submitted to the Graduate Faculty of the

School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Jian Zou

It was defended on

April 24, 2023

and approved by

George C. Tseng, ScD, Professor, Department of Biostatistics, School of Public Health, University of Pittsburgh

Jiebiao Wang, PhD, Assistant Professor, Department of Biostatistics, School of Public Health, University of Pittsburgh

Wei Chen, PhD, Professor, Department of Pediatrics, School of Medicine, University of Pittsburgh

Adrian V. Lee, PhD, Professor, Department of Pharmacology & Chemical Biology, School of Medicine, University of Pittsburgh

Clustering, Biomarker and Cancer Model Selection Using Omics Data

Jian Zou, PhD

University of Pittsburgh, 2023

Central dogma reforms the biomedical science. Since then, biomedical researchers have focused mostly on the relationship between DNA, RNA, and protein. To quantify their sequence, structure, and abundance, numerous biotechnologies have been created. High-throughput technologies, which emerged since 2000s, offer researchers a fantastic opportunity to thoroughly grasp the mechanism of diseases and also bring many statistical challenges. This thesis focuses on constrained clustering (Chapter 2), multi-study multi-class concordant biomarker detection (Chapter 3), and cancer model selection (Chapter 4) in high-throughput omics data analysis.

In Chapter 2, we proposed Constrained Gaussian Mixture Model (CGMM) by extending the Gaussian mixture model (GMM) to solve empty or small cluster issue. We also generalized CGMM to sparse CGMM (SCGMM) using $L1$ penalty for gene selection. Extensive simulations and three real applications demonstrated the superior performance of our proposed method.

In Chapter 3, we proposed a two-step framework, Multi-Study Multi-Class Concordance (MSCC), to detect biomarkers in multi-class analysis across multiple studies from the aspect of information theory. We first detect biomarkers with partially shared concordant patterns across multiple studies and then identify the studies which contribute to such concordance. The simulation and real-world data analysis showed superiority over min-MCC, the only existing method for this problem so far.

In Chapter 4, we developed Congruence Analysis and Selection of CAncer Models (CASCAM), a statistical and machine learning framework for authenticating and selecting the most representative cancer models in pathway-specific and drug-relevant manner using transcriptomics data. CASCAM provides harmonization between tumor and cancer model omics data, interpretable machine learning for congruence quantification, mechanistic investigation, and pathway-based topological visualization to determine the most appropriate cancer model

selection. The workflow is presented using invasive lobular breast carcinoma (ILC) subtype, credentialing highly relevant models for ILC research. Our novel method is generalizable to any cancer subtype and will be impactful for furthering research in precision medicine.

Contribution to public health: The proposed clustering, biomarker and cancer model selection methods using omics data are crucial for disease mechanistic understanding that can lead to translational and clinical research. The related researches unravel knowledge towards precision medicine and benefit public health.

Table of Contents

List of Tables

List of Figures

Preface

As I reflect on my Ph.D. journey, I am overwhelmed with gratitude for the people who have supported and guided me through this transformative experience. Conducting research in the field of machine learning and oncology has been an exciting and meaningful journey, and I am incredibly grateful for the opportunity.

First, I would like to express my deepest gratitude to my advisor, Dr. George C. Tseng. His invaluable help, support, patience, and advice have been the driving force behind my success. I have learned so much from him, especially about scientific thinking, and he has been an excellent mentor in both my research and personal life.

I would also like to thank Drs. Steffi Oesterreich and Adrian V. Lee. Without their training, I wouldn't have a deep understanding of cancer research, and they guided me through the complexities of oncology research.

The members of our lab, Xiangning Xue, Yusi Fang, Wei Zong, Wenjia Wang, Rick Chang, Michael Gorczyca, Ruofei Yin, Danyang Li, Yujia Li, and Peng Liu, have been an essential source of constant support and joyful communication throughout my Ph.D. journey. Their presence and encouragement have made my research experience much more enjoyable.

I would also like to thank my dissertation committee members, Drs. Wei Chen and Jiebiao Wang, for their participation and valuable feedback during my defense. Their insights and suggestions were immensely helpful.

I am indebted to my parents, Yafan Zou and Wenyan Wang, for their unwavering support of every decision I have made and for always believing in me. Their love and encouragement have been a constant source of motivation.

I would like to express my heartfelt gratitude to my girlfriend, Shilei Liu, for her unwavering support and encouragement throughout this process. She has been my biggest cheerleader and I could not have done this without her.

Finally, I would like to thank the company of my domestic shorthair cat.

This has been an enjoyable and challenging journey, and I am grateful to all those who have made this an unforgettable experience.

1.0 Introduction

Central dogma, first proposed by Francis Crick in 1958, reforms the biomedical science [24]. Since then, the relationship between DNA, RNA, and protein has drawn the most attention of biomedical scientists. Numerous biotechnologies have been developed to measure their sequence, structure, and abundance. High-throughput technologies, arising from the 2000s, provide an excellent opportunity for researchers to comprehensively understand the mechanism of diseases [45] and bring multiple statistical challenges simultaneously. This chapter will introduce the high-throughput omics data (Section 1.1) and three related statistical modeling and learning issues (Sections 1.2, 1.3, and 1.4).

1.1 High-throughput omics data

Omics data analysis seeks to collectively characterize and quantify biological molecules (such as DNA, RNA, and proteins) to understand the structure, function, and dynamics of organisms, and the study subjects for these domains, such as genomics, proteomics, or metabolomics, are denoted by the suffix "-omics". By parallelizing the sequencing process, high-throughput technologies generate thousands or millions of sequences at once, giving researchers the excellent opportunity to conduct omics analysis on a larger scale. In this section, we introduce two commonly used omics data – genomics and transcriptomics, and the statistical challenges that accompany them.

1.1.1 Genomics

The entirety of an organism's DNA, including all of its genes and their interrelations and influence on the organism, is known as its genome. The human genome is the whole collection of nucleic acid sequences for humans, encoded as DNA in the 23 pairs of chromosomes found in cell nuclei and in small DNA molecules located in each mitochondria. It was originally

made public in February 2001. Both different non-coding DNA sequences and DNA that codes for proteins are present in the human genome.

Mutation detection is one of the main goals in genomics analysis [87]. The causality between some mutations and diseases has been constructed. For example, the cumulative breast cancer risk to age 80 years was 72% for $BRCA1$ and 69% for $BRCA2$ carriers [56]. Nowadays, several multigene panel testings were developed based on mutation detection research and widely implemented in medical practice [58].

Single nucleotide polymorphisms (SNPs) is another prevalent type of genetic variation that is utilized to link with diseases. Genome-wide association study (GWAS) is therefore proposed to find SNPs associated with clinical phenotypes [54]. Association differs from causality, though. The researchers are motivated to create more statistical methods as a result of the GWAS limitations [14]. For example, expression quantitative trait locus (eQTL) links the SNPs analysis with the gene expression under assumption that the SNPs with high correlation with gene expression are more likely to be functional, and methylation quantitative trait locus (mQTL) links SNPs with methylation level.

In addition to the previous two, copy number variation (CNV), defined as amplification or deletion of genetic materials [110], detected in DNA sequencing data is a useful tool for comprehending genetic variation [103], particularly in oncology. Different CNV status have been ensured to be correlated with cancer occurrences [63].

1.1.2 Transcriptomics

The study of all of the RNA transcripts, such as message RNA (mRNA) and micro RNA (miRNA), produced by the genome is known as transcriptomics. The abundance of mRNA shows the gene expression level related to the the level of activity of certain biological functions, since the mRNA is translated into peptide chains, which can then be folded to form proteins. DNA microarray and RNA sequencing (RNA-seq) are two main techniques to measure the transcriptomics.

DNA microarray measures the abundance of RNA based on known gene sequences. mRNA is first extracted from a control sample and an experimental sample, the latter of

which is typically representative of the disease. The target RNA is transformed into cDNA to boost stability and labeled with two fluorophores (red and green typically). A laser is used to scan the microarray after the cDNA has been dispersed across its surface and hybridized with oligonucleotides. We could then detect which of the samples exhibits higher amounts of mRNA based on the color of the fluorophores. The fluorescence intensity on each place of the microarray correlates to the degree of gene expression [41].

RNA sequencing (RNA-seq), the next-generation sequencing technology, allows for both qualitative and quantitative analysis of RNA transcripts with only a tiny amount of RNA and no prior knowledge of the genomes [47], which are gradually replacing the use of microarray. RNA samples are first extracted, transformed into cDNA libraries, sequenced, aligned to a reference, and quantified for further research. Single cell RNA sequencing (scRNA-seq) has recently opened up possibilities for the simultaneous measurement of gene expression in hundreds to thousands of individual cells, and it is now widely applied in multiple research areas such as understanding the heterogeneity of tumor samples [51].

1.1.3 Statistical modeling and learning issues in omics data analysis

The accumulation of omics data brings new statistical challenges and opportunities. This dissertation focuses on three issues: 1) The imbalanced group labels in real omics applications are not well handled by current clustering algorithms, which might even lead to empty cluster issues (i.e., one cluster is assigned no observations). 2) The biomarkers showing concordant expression patterns across multiple studies are believed to be valid disease indications. However, the statistical methods to identify these biomarkers are seldomly proposed. 3) The continuing increase in the number of cancer models, such as cell lines, patient-derived xenografts (PDX), and organoids (PDO), has led to an increasingly urgent need for statistical methods which could select the most appropriate ones for the specific research interest. However, such methods have not been developed yet.

In the following sections of this chapter, we will introduce the imbalance issues in clustering analysis (Section 1.2), the multi-study multi-class concordant biomarker detection (Section 1.3), and congruence and selection of representative cancer models (Section 1.4).

1.2 Constrained clustering

1.2.1 Overview of clustering algorithm

Clustering analysis is a set of practical data mining techniques widely applied in many fields to find groups of objects with similar patterns. It is a crucial tool for unsupervised machine learning and is widely applied in biomedical research, especially omics data exploration. Two major categories of clustering methods are distance-based and model-based methods.

Distance-based methods define pair-wise distances between observations at first, and the observations with modest distances are typically assigned to a single cluster. K-means, for example, is the most representative algorithm under this category. Given a set of observations $(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N})$, K-means aims to partition the N observations into K ($\leq N$) sets $S = \{\mathbf{S_1}, \mathbf{S_2}, ..., \mathbf{S_k}\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\text{argmin}_{\mathbf{S}} \sum_{k=1}^{K} \frac{1}{|\mathbf{S}_k|} \sum_{\mathbf{x},\mathbf{y} \in \mathbf{S_K}} \text{dist}(\mathbf{x}, \mathbf{y})$$

where $\text{dist}(\mathbf{x}, \mathbf{y})$ represents the distance between $\mathbf{x}$ and $\mathbf{y}$, which we usually use the Euclideadean distance.

Compared with distance-based methods, model-based clustering, considering the observations created from a finite combination of distributions, stands out for its precise inference, enhanced interpretation, and adaptability. Gaussian mixture model is a representative example of model-based approach (Section 2.2.1).

1.2.2 Constrained clustering

Constrained clustering is a series of methods extending the clustering by incorporating the prior knowledge into the unsupervised process for better results. Instance-level and cluster-level are two categories of constrained clustering. The prior one, which is commonly discussed, is about setting constraints on the objects such as *must-link* and *cannot-link* by the partially known labels or the expertise (semi-supervised learning). In contrast, the latter

focuses on the clusters themselves, such as cluster size and number, which is relatively less mentioned [37].

Most cluster size constrained methods are extended from K-means. In order to guarantee the lowest size of each cluster, Bradley et al. [15, 28] incorporate a linear transportation solver into the K-means cluster assignment stage. Pakhira [83] suggests modifying K-means to take the determined cluster center as an observation for the center calculation in the each iteration. The balanced K-means suggested by Malinen and Fränti [70] once more alters the assignment stage to pre-allocate the slots with an equal number of objects around the centroids and allocates the item to the slots rather than the centroids to ensure a balanced assignment. The initial and assignment steps are modified suggested by Ganganath et al. [38]. In the assignment step, a predetermined upper bound for the cluster size is predetermined by sorting the distance in ascending order. The updated beginning step still requires at least one object with a known group label for each cluster group. Another assignment step revision of K-means is called eXploratory K-Means with empty-cluster-reassignment (EXK-Means), which was developed by Hua et al. [46]. In order to replace the empty clusters, it finds the items that are the furthest from the centroids in each iteration and reassigns them as the new centroids.

A few other techniques fall within the aforementioned task in addition to the K-means based approaches. For instance, Zhu et al [136] seek to locate the clustering assignments that, when considered in the context of the size constraints, has the maximum degree of agreement with the known clustering assignment.

1.2.3 Problems in model-based clustering

Model-based clustering has the following three drawbacks: 1) one may have empty cluster issues when cluster sizes are uneven. 2) it would lose clustering accuracy owing to vulnerability to a local optimum. 3) it could fail to achieve the clustering objective when cluster size constraints are mandatory. All of three drawbacks can be solved by cluster size constrained clustering through narrowing down the searching space.

To our knowledge, no constrained model-based clustering method exists, and none of the

methods discussed above take into account the issue of feature selection, which is critical with high-dimensional data such as omics data. In Chapter 2, we proposed CGMM – a novel algorithm for constrained model-based clustering to bridge this gap.

1.3 Multi-study multi-class concordant biomarker detection

Biomarker detection is a critical component of biomedical research. Study integration is a typical strategy for enhancing the accuracy and potency of biomarker identification. If a gene exhibits consistent expression patterns in various studies, we could assume it is a reliable candidate for disease indication.

Two approaches to study integration include combining p-values and combining effect sizes. The former has received a lot of attention. In Fisher's approach, for instance, the log-transformed p-values are added up, and each p-value is presumptively distributed uniformly under the null hypothesis. The latter strategy splits the observed treatment effects of each study into two components: the actual effect size and the study-specific noise [31]. The effect size combination, on the other hand, is only available in the two-class scenario, and the p-value combination just considers the significance level without taking into account the gene expression pattern. The effect size combination is no longer valid when there are more than two categories, and the p-value combination cannot accurately identify the multi-class pattern.

To the best of our knowledge, min-MCC [68] is the sole strategy for identifying biomarkers with consistent multi-class patterns across several studies that makes use of the minimal correlation value for all study pairs. However, it is too strict to detect the biomarkers with consistent patterns in partial studies. In this thesis, we revisit this problem from the aspect of information theory, and re-design a new framework for biomarker detection in Chapter 3.

1.4 Selection of representative cancer model

1.4.1 Cancer models

Numerous cancer models, such as cell lines, patient-derived organoids (PDO), and xenografts (PDX), are created as a result of the growth of biotechnology. Cell lines originate from multicellular organisms and are immortalized and maintained in vitro. It is believed that cell lines retain the stability of specific phenotypes and functions. PDOs are created from isolated organ progenitors or pluripotent stem cells, which can develop into an organ-like tissue with a variety of cell types. PDOs have the ability to self-renew and self-organize, maintaining the physiological structure and function of their source tumor [130]. Surgery is used to remove tumor pieces from cancer patients, which are then transplanted directly into immunodeficient mice to create PDXs [131].

1.4.2 Current selection methods

It is necessary to evaluate and select appropriate cancer models prior to experiment. Cell lines could be mislabeled [133] and the genomic/epigenomic alterations or even contamination [39, 125, 8] may accumulate across passages in the culture. Similarly, PDO and PDX may also evolve due to the different microenvironment during time (Section 4.2). In practice, cancer models are typically chosen based on a select few important mutations or traits without thorough research. [32]

Congruence (correlation-based) analysis and authentication (machine-learning-based) analysis are the two main tool categories employed in current assessment research, which are typically focused on pan-cancer investigation. In the former congruence analysis, correlation/association metrics are typically used to evaluate the degree of genome-wide similarity between a cancer model and the target tumor cohort [125, 4, 64, 124, 5, 99]. In contrast, the later authentication analysis creates machine learning models for cancer model assignment to human cancer types, including suitability score [32], random forest [86, 133], ridge regression [98], and nearest template prediction [132].

However, four limitations exist in current approaches. 1) High prediction accuracy is

the goal of machine learning-based authentication approaches, but they are not intended to promote candidate cancer models that most closely resemble the target tumor cohort. 2) Correlation-based congruence approaches frequently yield poorer prediction accuracy though they are more suitable for prioritizing cancer models. 3) Current congruence or authentication methods are generally used at the genome-wide level and are unable to identify the pathways or molecular mechanisms that a cancer model most closely or least closely resembles, which is crucial for the development of precision medicine. 4) The current literature has not thoroughly studied and analyzed the data compatibility and harmonization between cancer model and tumor data, which is a crucial step to obtain high accuracy and prevent false mechanistic conclusions. In this thesis, we review the cancer model selection problem and provide a complete framework to select the appropriate cancer models using transcriptomics data.

1.5 Overview of this dissertation

This thesis focuses on the statistical concerns and challenges associated with omics data. In Chapter 1, we introduce the omics and three statistical issues (Section 1.2, 1.3, and 1.4).

In Chapter 2, we propose Constrained Gaussian Mixture Model (CGMM) by extending the Gaussian mixture model (GMM). We also generalize CGMM to sparse CGMM (SCGMM) using $L1$ penalty in high-dimensional data. Extensive simulations and three real applications demonstrate the superior performance of our proposed method.

In Chapter 3, we propose a two-step framework, Multi-Study multi-Class Concordance (MSCC), to detect the biomarkers in multi-class analysis across multiple studies from the aspect of information theory. We first detect the biomarkers with concordant patterns partially or entirely across multiple studies and then identify the studies which contribute to such concordance. The simulation and real-world data analysis demonstrated superiority over min-MCC [68], the only applicable method for this problem so far.

In Chapter 4, we develop Congruence Analysis and Selection of CAncer Models (CASCAM), a statistical and machine learning framework for authenticating and selecting the

most representative cancer models in pathway-specific and drug-relevant manner using transcriptomic data. CASCAM provides harmonization between tumor and cancer model omics data, interpretable machine learning for congruence quantification, mechanistic investigation, and pathway-based topological visualization to determine the most appropriate cancer model selection. The workflow is presented using invasive lobular breast carcinoma (ILC) subtype, credentialing highly relevant models, while questioning congruence of some frequently used models such as MDA-MB-134VI for ILC research. Our novel method is generalizable to any cancer subtype and will be impactful for furthering research in precision medicine.

2.0 CGMM: a novel algorithm for constrained model-based clustering

The contents of this Chapter are prepared and ready to be submitted to journal *Knowledge-Based Systems*. This work was awarded the 2022 Mihaela Serban Award for Best Poster Presentation from the ASA Pittsburgh Chapter.

2.1 Introduction

Clustering analysis, an essential tool for unsupervised machine learning, is a set of practical data mining techniques to identify groups of objects with similar patterns and has been widely used in many areas. For example, in biomedical applications, clustering patients into different subgroups is usually the first step to understanding the underlying mechanism of complex disease, followed by the development of precision medicine. Model-based clustering stands out among many clustering methods for its rigorous inference, better interpretation and extensibility. Recently, to address the rising challenges about "large p, small n" problem, many clustering methods constructed from conventional model-based clustering are also proposed for clustering objects and selecting features simultaneously, such as penalized Gaussian mixture model [84], sparse Poisson mixture model [129], and sparse negative binomial mixture model [62].

However, it encounters new challenges in modern data science. Unlike the conventional datasets, the group sizes in some particular datasets (such as omics data) are usually imbalanced (e.g., the rare subtypes of some diseases), and the sample sizes are moderate. Therefore, the small clusters are likely to be overlooked or ultimately merged into the larger groups, especially when the number of groups is significant. More specifically, one clustering group could have small probabilities among all the objects when applying the model-based clustering methods. This group would then "disappear" in the hard assignments. Besides the imbalanced and empty cluster issue, there are also needs to pre-specify the boundary of cluster size in practice, such as the job scheduling problem (similar jobs are clustered for

the same worker while the maximum number of jobs per worker is pre-specified) and the customer segmentation problem [136].

Clustering with size constraints, as a category of constrained clustering [37], is the approach to solve this issue. Most of the methods under this goal are extended from K-means. Bradley et al. [15, 28] implements a linear transportation solver into the cluster assignment step of K-means to ensure the minimal size of each cluster. Pakhira [83] proposes the modified K-means, focusing on the cluster updating step instead, to assume the calculated cluster center as an observation for the center calculation in the next iteration. The balanced k-means proposed by Malinen and Fränti [70] again modifies the assignment step to pre-allocate the slots with an equal number of objects around the centroids and assigns the object to the slots, rather than the centroids, to ensure a balanced assignment. Another K-means based approach proposed by Ganganath et al. [38] changes the initial step and assignment step. The modified initial step requires at least one object with a known group label for each cluster group; in the assignment step, an upper bound for the cluster size is pre-determined by sorting the distance in ascending order. eXploratory K-Means with empty-cluster-reassignment (EXK-Means) by Hua et al. [46] is also an assignment step revision of K-means. It works by detecting the most marginal objects (according to the distance toward the centroids) in every iteration and re-assigning them as the new centroids to replace the empty clusters. Besides the approaches based on K-means, a few other methods are also under the task mentioned above. For example, size constraints, developed by Zhu et al. [136], directly works on the clustering assignment and aims at identifying the clustering partition under the cluster size constraints with the highest agreement with the known clustering assignment.

However, to the best of our knowledge, none of them consider feature selection. The model-based clustering with cluster size constraints and details about when and how cluster size constraints work are still unsolved. In order to solve these problems, we introduce the cluster size constraints to the model-based clustering inspired by the idea of solving linear transportation problems from Bradley et al. [15, 28] and develop a new framework named Constrained Gaussian Mixture Model (CGMM), allowing the pre-determination of cluster size boundary. Our framework is a generalization of the Gaussian mixture model with

Expectation-Maximization algorithm [29]. We update the E-step by introducing a checkpoint to examine the constraint criteria for each iteration. If not met, a linear transportation solver is then implemented to re-arrange the assignment to ensure the minimal cluster size. We also extend CGMM to SCGMM (Sparse CGMM) using lasso penalty to allow feature selection in high-dimensional data, which adopts the model proposed by Pan and Shen [84].

The article is structured as follows. In Section 2.2, we showcase the conventional GMM and our proposed CGMM and SCGMM framework with their working mechanisms. We further illustrate the mechanism in high-dimensional and one-dimensional data simulation in Section 2.3. Section 2.4 includes the real applications under three different scenarios using our framework. Finally, the discussion and conclusion about this method and the following possible extensions are in section 2.5.

2.2 Methodology

We first introduce the model based clustering under the Gaussian assumption. Then we introduce our proposed cluster constrained framework and its extension for "large p, small n" scenario. The methods for parameter selection and results benchmark are included at last.

2.2.1 GMM and penalized GMM

We annotate $\boldsymbol{X_{n\times p}}$ as the standardized data matrix for clustering with n observations and p features. It is assumed that every observation $\mathbf{X_i}$ is generated from a Gaussian mixture distribution with K components $f(\mathbf{X_i};\Theta) = \sum_{k=1}^{K} \pi_k \phi_k(\mathbf{X_i}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma})$. As for the unknown parameters, π_k represents the prior probability with $\sum_{k=1}^{K} \pi_k = 1$ and $0 \leq \pi_k \leq 1$, $\boldsymbol{\mu_k}$ represents the p-dimensional vector of mean parameters for component k, and $\boldsymbol{\Sigma} = diag(\sigma_1^2, ..., \sigma_p^2)$ represents the co-variance matrix shared across K components with the feature independence

assumption in our study. The observed log-likelihood is

$$\log L(\Theta) = \sum_{i=1}^{n} \log f(\mathbf{X_i}; \Theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi_k(\mathbf{X_i}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma}) \tag{1}$$

Since it is difficult to directly maximize the observed log-likelihood, a latent variable as the group label indicator $\boldsymbol{Z_{n\times K}}$ is introduced with $z_{i,k} = 1$ if $\mathbf{X_i}$ comes from group k and $z_{i,k} = 0$ if not. The complete log-likelihood including the $\boldsymbol{Z_{n\times K}}$ is

$$\log L_c(\Theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{i,k} \log[\pi_k \phi(\mathbf{X_i}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma})] \tag{2}$$

The Expectation-Maximization (EM) algorithm [29] is then applied to obtain the maximized likelihood estimator (MLE). In every iteration, the expectation of $\boldsymbol{Z_{n\times K}}$ is updated in E-step and the estimator for Θ is updated in M-step. We denote $\boldsymbol{\Delta_{n\times K}} = \{\delta_{i,k}\}_{i\in 1:n, k\in 1:K}$ as the expectation of the latent variable $\boldsymbol{Z_{n\times K}}$ which is also the soft group assignments for the data matrix, and $map(\boldsymbol{\Delta_{n\times K}})$ is the function of mapping the soft assignment matrix to the hard assignment vector indicating the assigned label of each object.

To address the "large p, small n" issue, Pan and Shen [84] propose a penalized GMM (PGMM) and regularize the log-likelihood by a L_1 penalty term. The observed and complete log-likelihood become

$$\log L_P(\Theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi_k(\mathbf{X_i}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma}) - \lambda \sum_{k=1}^{K} \sum_{j=1}^{p} |\mu_{kj}| \tag{3}$$

$$\log L_{c,P}(\Theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{i,k} \log[\pi_k \phi(\mathbf{X_i}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma})] - \lambda \sum_{k=1}^{K} \sum_{j=1}^{p} |\mu_{kj}| \tag{4}$$

Under this setting, μ_{kj} is shrunken towards 0. If the estimated $\mu_{kj} = 0$ for all k, the j-th feature does not contribute to the clustering results and realizes the feature selection. Detailed derivations can be referred to [84].

2.2.2 CGMM framework

In the CGMM framework, besides the parameters in GMM, we introduce a new parameter τ requiring the minimal cluster size (i.e. each cluster should contain at least τ

observations). The framework is similar if the upper bound of cluster size needed.

At first, GMM (with random or k-means assignment as the initial) is performed until convergence, and we start the CGMM EM iterations if $\sum_{i=1}^{n}(map(\mathbf{\Delta}) == k) < \tau,\ \exists k \in \{1, 2, ..., K\}$. Otherwise, we directly report the GMM results since the cluster size has already satisfied the constraints.

For the E-step, the cluster size is checked every time. In iteration m, if $\sum_{i=1}^{n}(map(\mathbf{\Delta}) == k) \geq \tau\ \forall k \in \{1, 2, ..., K\}$, $\mathbf{\Delta}$ is updated in the usual way. For i-th observation in k-th group,

$$\delta_{i,k}^{(m)} = \frac{\pi_k^{(m-1)}\phi(\mathbf{X_i}; \boldsymbol{\mu_k}^{(m-1)}, \mathbf{\Sigma}^{(m-1)})}{\sum_{k=1}^{K}\pi_k^{(m-1)}\phi(\mathbf{X_i}; \boldsymbol{\mu_k}^{(m-1)}, \mathbf{\Sigma}^{(m-1)})} \tag{5}$$

Otherwise, we obtain the $\mathbf{\Delta}$ by maximizing the object function $h(\mathbf{\Delta}^{(m)})$

$$h(\mathbf{\Delta}^{(m)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{i,k}^{(m)}\log[\pi_k^{(m-1)}\phi(\mathbf{X_i}; \boldsymbol{\mu_k}^{(m-1)}, \mathbf{\Sigma}^{(m-1)})] \tag{6}$$

with the constraints satisfied,

$$\begin{cases} \sum_{i=1}^{n}\delta_{i,k}^{(m)} \geq \tau, k = 1, ..., K \\ \sum_{k=1}^{K}\delta_{i,k}^{(m)} = 1, i = 1, ..., n \\ \delta_{i,k}^{(m)} \in \{0, 1\}, i = 1, ..., n; k = 1, ..., K \end{cases} \tag{7}$$

If the constraints are applied, the goal of calculating the latent variable expectation is switched to maximizing the complete log-likelihood in the constrained searching space. The soft assignments then become hard assignments, which is similar to the idea of K-means. Though the continuity is sacrificed when the constraints are activated, the local optimum with unwanted small cluster size is avoided at the same time.

When the constraints are activated, it is equivalent to solving a linear transportation problem since the goal of maximizing the complete log-likelihood $h(\mathbf{\Delta}^{(m)})$ is equivalent to minimizing $-h(\mathbf{\Delta}^{(m)})$. A common example of a linear transportation problem is transporting goods from multiple factories to multiple warehouses. The goal is to minimize the transportation fee while satisfying the minimum number of goods requested by different warehouses. In our case, every observation can be seen as a factory with one piece of goods, and each

cluster can be seen as a warehouse. The cost of transporting observation $\mathbf{X_i}$ to cluster k is $-\log[\pi_k \phi(\mathbf{X_i}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma})]$, and the goal is to minimizing the total cost (maximizing the complete log-likelihood) while requiring each cluster contains at least τ observations. Therefore, the optimized transportation plan is the hard assignment results in every iteration. we use *lpSolve* R package [11] to solve it.

For the M-step in iteration m, the parameters are updated as usual [84],

$$\begin{aligned}
\boldsymbol{\mu_k}^{(m)} &= \frac{\sum_{i=1}^{n} \delta_{i,k}^{(m)} \mathbf{X_i}}{\sum_{i=1}^{n} z_{i,k}^{(m)}} \\
\sigma_p^{2,(m)} &= \sum_{k=1}^{K} \sum_{i=1}^{n} \frac{\delta_{i,k}^{(m)} (X_{i,p} - \mu_{k,p}^{(m)})^2}{n} \\
\pi_k^{(m)} &= \frac{\sum_{i=1}^{n} \delta_{i,k}^{(m)}}{n}
\end{aligned} \tag{8}$$

2.2.3 SCGMM framework

In order to solve the "large p, small n" issues in the high dimensional data with sparsity and realize the feature selection, we extend our method to SCGMM using the model proposed by Pan and Shen [84]. The included penalty term shrinks the group centers towards 0, and the less informative features are excluded for clustering assignment. The only difference lies in updating $\boldsymbol{\mu_k}$ in M-step. A further calculation is needed,

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{k}}^{(m)} = sgn(\boldsymbol{\mu}_{\boldsymbol{k}}^{(m)})(|\boldsymbol{\mu}_{\boldsymbol{k}}^{(m)}| - \frac{\lambda}{\sum_{i=1}^{n} \delta_{i,k}^{(m)}} \boldsymbol{\Sigma}^{(m)} \mathbf{1})_+ \tag{9}$$

where $\mathbf{1}$ is a vector of 1s, $sgn(x)$ is the sign function. $(x)_+ = x$ if $x > 0$, and $(x)_+ = 0$ otherwise. SCGMM becomes CGMM when $\lambda = 0$. Detailed derivations can be referred to [84]. Pseudo code is shown in Algorithm 1.

2.2.4 Parameter selection and methods for benchmarking

A modified version of Bayes Information Criterion (BIC) [84] is used in this study for λ selection in the Section 2.3, and it is defined as $BIC = -2 \log L(\hat{\Theta}) + \log(n) d_e$, where $\hat{\Theta}$ is the maximized penalized likelihood estimators (MPLE), and $d_e = K + p + Kp - 1 - q$ with

q for the number of MPLE mean components which are shrunken to 0. Comparing with the original BIC, d_e represents the degree of freedom for the penalized model through the introduction of q [115].

Adjusted Mutual Information (AMI), a measure of clustering results consistency adjusting for chance, is adopted for comparing the clustering results with the ground truth. AMI ranges in $[0, 1]$, and the larger value means better consistency. Compared with Adjusted Rand Index (ARI), another commonly used statistics, AMI emphasizes more on the imbalanced clustering results [96], which is more appropriate in our scenarios.

When the ground truth is not applicable in Section 2.4.1, we use the average silhouette method with Euclidean distance for the results evaluation [97]. It measures the cohesion of a data point to its cluster compared to the others and ranges from -1 to +1 with the larger value indicating better performance.

Furthermore, we reallocate the clustering labels by by maximizing consistency between the clustering results and the ground truth through the Hungarian method [42]. Confusion matrix, averaged confusion matrix across simulated replicates, and the cluster size table are used for results evaluation.

2.3 Simulation

This section demonstrates the performance of CGMM/SCGMM compared with GMM/PGMM under different scenarios. In simulation 1, we showcase the performance of SCGMM in high dimensions with sparsity, and the performance of CGMM in one dimension is shown in simulation 2. A random assignment is used as the initial if not specified.

2.3.1 Simulation 1: high dimension with sparsity

In simulation 1, we simulate 4 clusters with (40 - S, S, S, 40 - S) $\cdot N$ observations respectively, and the data are generated from D dimensional standard normal distribution centering at (-2, -1, 1, 2) $\cdot E$ for each group in the first 20% of the features and centering at

(0, 0, 0, 0) in the remaining features. Features are mutually independent. S, N, D, and E are variables changed to simulate different scenarios and evaluate the impact of imbalance, sample size, dimensionality and effect sizes respectively. Under each scenario, 100 datasets are generated. Details for each setting are outlined below:

- Simulation 1A: $S = 10$, $N = 1$, $D = (50, 100, 200, 500, 1000)$, and $E = 1$.
- Simulation 1B: $S = (5, 10, 15, 20)$, $N = 1$, $D = 500$, and $E = 1$.
- Simulation 1C: $S = 10$, $N = (1, 2, 5, 10)$, $D = 500$, and $E = 1$.
- Simulation 1D: $S = 10$, $N = 1$, $D = 500$, and $E =$ (0.8, 0.9, 1 , 1.2, 1.4, 1.8, 2, 50, 100).
- Simulation 1E: Same setting with 1A, but use 100 random initials and 1 K-means initial with modified BIC for initial selection.

We first demonstrate the impact of dimensionality in simulation 1A. In Figure 1A, we can observe the increasing trend of AMI and the difference in AMI between PGMM and SCGMM at the same time as the number of dimensions with signals increases. The averaged AMI of SCGMM is 0.93, while the average AMI of PGMM is 0.69 when $D = 1000$. This difference is caused by the group merging issue. Among 100 simulated datasets, there are 86 of them containing the empty cluster, which is concordant with the averaged confusion matrix (Table 1A), where PGMM tends to merge the small clusters and SCGMM can avoid these situations as such assignments are not in its searching space.

The effect of imbalance is then evaluated in simulation 1B. As the degree of imbalance decreases, the difference between SCGMM and PGMM decreases (Figure 1B). Table 1B shows the averaged confusion matrix when $S = 5$, and the small groups are nearly merged by the large groups.

We then evaluate the impact of sample size in simulation 1C. Not surprisingly, as the sample size becomes larger, AMI of SCGMM and PGMM increase simultaneously and the difference of them gets smaller (Figure 1C). Finally, both of them reach the perfect clustering results when $N = 10$.

The impact of effect sizes is also investigated. As the effect size increases, AMI of PGMM and SCGMM increases at the same time (Figure 1D). We can also see the AMI of PGMM bounces up and down at 0.9 even when the effect size is large enough, which is

because of the empty cluster in some simulated datasets. For group l as an instance, since $\tilde{\boldsymbol{\mu}}_{\boldsymbol{l}} = sgn(\boldsymbol{\mu}_{\boldsymbol{l}})(|\boldsymbol{\mu}_{\boldsymbol{l}}| - \frac{\lambda}{\sum_{i=1}^{n}\delta_{i,l}}\boldsymbol{\Sigma}\mathbf{1})_+$, which shrinks to 0, it is possible that $\tilde{\boldsymbol{\mu}}_{\boldsymbol{l}} = \mathbf{0}$ (i.e., $\mu_{lj} = 0$, $\forall j \in \{1, ..., p\}$) and $\pi_l \cdot \phi(\mathbf{X_i}; \mathbf{0}, \boldsymbol{\Sigma}) < \pi_k \cdot \phi(\mathbf{X_i}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma})$, $\forall k \neq l$ and $i \in \{1, ..., n\}$. Therefore, we see group l completely "disappears" under PGMM framework.

Figure 1E shows the impact of multiple initials including K-means to ensure a good initial is selected for every dataset. Compared with simulation 1A, the increasing number of initials can improve PGMM performance at the cost of a heavier computation burden. However, it still can not guarantee avoiding the empty cluster issue. For example, when $D = 100$, we still observe 12 empty cluster issues for GMM among 100 simulated datasets and the averaged confusion matrix table 1E also shows that large groups tend to merge the small groups in PGMM while SCGMM rescue this tendency, indicating the necessity of cluster size constraints.

To summarize, SCGMM works by limiting the searching space to avoid local optimum and the empty cluster issue. As a result, it has better performance for imbalanced datasets, small sample sizes, and small effect sizes. Multiple initials can improve the performance of PGMM, but they can not avoid empty cluster issues.

2.3.2 Simulation 2: one dimension

In this section, we focus on the performance of CGMM under the impact of different effect sizes, different sample sizes, and the different number of initials. Under the following scenarios, there are 4 groups, and data are generated from a standard normal distribution centered at (-2, -1, 1, 2) $\cdot E$ with (30, 10, 10, 30) $\cdot N$ observations, respectively. 100 $\cdot I$ random assignment initials are used, and the one with the largest log-likelihood is selected. Details for each setting are outlined below:

- Simulation 2A: $E = (0.5, 1, 2, 3, 5)$, $N = 10$, and $I = 1$.
- Simulation 2B: $E = 1$, $N = (1, 5, 10, 50, 100)$, and $I = 1$.
- Simulation 2C: $E = 1$, $N = 1$, and $I = (1, 2, 5, 10, 20)$.
- Simulation 2D: Same setting with Simulation 2A and 2B but using K-means clustering results as the initials.

Figure 2A shows that, in simulation 2A, as effect size increases, AMI of CGMM and GMM increase while the difference between them increases simultaneously. This difference is caused by the soft assignment property of GMM when the initial centers are not separable enough, and CGMM avoids the drawbacks by introducing the hard assignment. We can observe this case in Table 2 when $E = 5$.

Figure 2B demonstrates the simulation 2B for different sample sizes. As the sample size increases, the 100 random initials are less likely to include a good initial to have the centers separate enough. However, we could find that CGMM is more robust to these initials.

We also explore the effects of different number of initials in simulation 2C. As the number of initials increases, the AMI of GMM increases with better initials (Figure 2C).

However, the worse performance of GMM mentioned above is mainly caused by the random initials, which are not separative enough. If we perform the simulation 2A and 2B with K-means clustering results as the initials, the difference between GMM and CGMM diminishes (Figure 2D1-2).

2.4 Real application

In this section, we demonstrate the application of the (S)CGMM method in three different scenarios: 1) when the cluster size is pre-specified; 2) when the empty cluster issue exists; 3) when the rare group exists in different areas (business and biology). We clearly show that our proposed method can ensure the cluster size boundary while the normal (P)GMM cannot. Besides that, we show the superiority of our method to solve the empty cluster and rare group issues.

2.4.1 Real application 1: when the cluster size is pre-specified

Peer grouping is a tool about organizational learning, usually aiming to cluster similar peers within an organization for peer mentoring and communication. In order to design policy for differentiated groups and ensure a similar group size for better management, there

are always cluster size requirements for such tasks. In this section, we use the dataset from a case study [49] about clustering outlets within a large organization, and the variables include different measures of client demographics and organizational characteristics. The data have already been pre-processed with the 11 covariates carefully selected and de-identified. There are 200 outlets for clustering, 100 is the maximum cluster size pre-specified, and 3 clusters are recommended in the study. The true clustering labels are not available in the dataset. In the analysis, K-means is used as initials, and the constraint term in Equation 7 is modified to $\sum_{i=1}^{n} z_{i,k}^{(m)} \leq \tau$ for accommodating the maximum cluster size constraint.

The cluster sizes for each group obtained by GMM are 117, 65, and 18, respectively, violating the cluster size requirements. In contrast, the cluster sizes obtained by CGMM are 100, 82, and 18. t-SNE [122], a tool based on stochastic neighbor embedding for high dimensional data visualization, is used for showing the clustering results in Figure 3. We can observe that CGMM changes the labeling of 17 data points from group 1 to group 2, and the rearrangement is reasonable according to the t-SNE figure. The average silhouette is 0.24 for GMM and 0.22 for CGMM. We can find that constrained clustering can help with better assignments under the pre-specified requirement by sacrificing a little clustering performance in this task.

2.4.2 Real application 2: when the empty cluster issue exists

Empty cluster issue, which refers to the one or multiple clusters containing no data point after hard assignment, happens especially in the high dimensional data with many clusters. This section showcases the empty cluster issue and how SCGMM can help. Genotype-Tissue Expression (GTEx) project [66] is a public resource for tissue-specific gene expression and regulation, and we use the brain tissue gene expression (RNA-Seq) in 13 different brain tissue types as the example. The gene read counts and the sample annotations in GTEx Analysis v6p are downloaded on 12/15/2021 and filtered to contain brain tissues only. The dataset contains 56,238 genes and 1,259 tissue samples. It is preprocessed in two steps: 1) transforming the data to the log2 scaled normalized values; 2) selecting 2,000 genes with the highest interquartile ranges.

In the analysis, τ is set to be 60, the rounded number to the nearest tens of the smallest cluster size. λ is explored, ranging from 0 to 300, and loess regression with 95% confidence interval is fitted to show the trend of AMI with different λs in Figure 4. We find that SCGMM generally has better clustering results and is more robust to different λ.

We go further to analyze the detailed assignments when $\lambda = 200$, which is the turning point in Figure 4. When $\lambda = 200$, AMI(GMM) = 0.57 and AMI(CGMM) = 0.63. t-SNE figure (Figure 5) and the cluster size table (Table 3) show the details of clustering assignments after the class label reallocation. PGMM identifies only 6 regions compared to 13 regions detected by SCGMM, and we see that the results of SCGMM have better consistent with true labels according to Figure 5. For example, caudate nucleus, putamen, and nucleus accumbens are 3 important components of basal ganglia, and they are gathered together in Figure 5 annotated by actual labeling. We find that SCGMM successfully detects them, but PGMM fails to differentiate them and the groups of putamen and nucleus accumbens are completely merged into the group of caudate nucleus.

2.4.3 Real application 3: when the rare group exists

Data imbalance is a common issue in classification problems. Over-sampling and down-sampling are two possible solutions. However, we cannot take these approaches in the clustering problem as the group labels are unknown. Constrained clustering can help with this scenario. By introducing the constraints, the rare groups can be identified for further investigation. In this section, we use the single-cell gene expression data (scRNA-Seq) for illustration. *zhengmix4uneq* data, which is originally from [135] and pre-processed in [35], is used for analysis. The dataset consists of 4 pre-sorted cell types (B cells, naive cytotoxic T cells, CD14 monocytes, and regulatory T cells) from the human. There are 1,644 gene features and 3,830 cell samples. In order to have the imbalance and rare groups, the dataset is subsampled to have 2 groups (CD14 monocytes and regulatory T-cells) with 100 samples each and 2 groups (B-cells and naive cytotoxic T-cells) with 20 samples each.

We first do the subsampling 50 times to see the general performance of PGMM and SCGMM under this scenario. Then, for each subsampled dataset, the analysis is performed

among different λs ranging from 0 to 60, τ is set to be 20, and the one with the largest AMI (best performer) is selected for SCGMM and PGMM separately. The boxplot (Figure 6A) clearly shows that SCGMM outperforms PGMM in all 50 simulated datasets.

One subsampled dataset is then analyzed in detail. In general, we find SCGMM has better performance compared to PGMM in Figure 6B, and the larger likelihood does not indicate better clustering results due to the limited sample size (Figure 6C). When $\lambda = 24$, where SCGMM reaches the largest AMI (AMI(PGMM) = 0.61 and AMI(CGMM) = 0.78), GMM fails to identify the B cells due to its small sample size and B cells are nearly absorbed by the CD14 monocytes group, while SCGMM successfully detects them (Table 4 and Figure 7).

Since neither SCGMM nor PGMM finds cytotoxic T cells correctly as they mix with regulatory T cells according to the actual labeling in the t-SNE plot (7), we further perform another analysis by merging the cytotoxic T cells and regulatory T cells groups into one group and subsample 50 B cells, 200 T cells, and 200 CD14 monocytes 50 times. Similarly, SCGMM is not inferior to PGMM among 50 subsampled data (Figure 6D). In one simulated dataset, when λ ranges from 0 to 60, SCGMM has better clustering performance and does not guarantee a better log-likelihood (Figure 6E-F). When $\lambda = 29$, where SCGMM achieves the best AMI, We can see that B cells are merged with CD14 monocytes in PGMM (only 1 B cell identified), while SCGMM successfully identifies B cells (51 B cells identified) (Table 5 and Figure 8).

2.5 Discussion and conclusions

This study proposes CGMM and SCGMM to perform the model-based clustering with cluster size constraints. To the best of our knowledge, this is the first cluster size constrained method in model-based clustering, and the first study thoroughly analyzes the scenarios where constrained clustering is needed.

The application of cluster size constrained clustering mainly lies in two scenarios – when the cluster size boundary is required and when the data structure is exceptional. There are

several tasks in the real world that require the pre-determination of cluster size boundary, such as the job scheduling problem, the customer segmentation problem, and peer grouping mentioned in section 2.1 and section 2.4.1. Under these cases, the cluster size constraints are mandatory. However, we have not seen any algorithm solution based on model-based clustering except ours. As for the latter scenario, we demonstrate that (S)CGMM has better performance on clustering assignments and can avoid empty cluster problems for datasets with severe imbalance, small sample sizes, and small effect sizes in the simulations, especially when the data are high dimensional and sparse. Two real-world examples also showcase the performance of SCGMM when the empty cluster issues exist and when the dataset is imbalanced. We find that similar to the results in the simulation, SCGMM successfully identifies the ignored groups and achieves even more consistent clustering assignments with the actual labels.

Briefly, (S)CGMM works by limiting the searching space to avoid local optimum and guaranteeing a cluster size boundary. This method can be extended to all the possible distributional assumptions such as negative binomial and Poisson. The cluster size boundary can also be generalized to be different for each group. One potential limitation is that the cluster size boundary has to be pre-determined according to prior knowledge. There is no method available to decide the boundary automatically, which sometimes leads to cases when the cluster size equals the pre-specified τ.

Algorithm 1 Pseudo code for (S)CGMM

Perform conventional (P)GMM to obtain initial parameters $\boldsymbol{\mu_k}^{(0)}$, $\boldsymbol{Z}^{(0)}$, $\boldsymbol{\Sigma}^{(0)}$, $\boldsymbol{\pi}^{(0)}$, and $l(\boldsymbol{X}; \boldsymbol{\mu_k}^{(0)}, \boldsymbol{\Sigma}^{(0)}, \boldsymbol{\pi}^{(0)})$ (observed log likelihood). If $\sum_{i=1}^{n}(map(\boldsymbol{Z_{n\times p}}) == k) \geq \tau$, $\forall k \in \{1, 2, ..., K\}$, stop the algorithm and directly report the initialization results. Otherwise, activate (S)CGMM. $m = 1$ and $\delta = 100$.

while $m \leq 100$ and $\delta > 10^{-7}$ **do**

// E Step

if min cluster size from $\boldsymbol{Z}^{(0)} < \tau$ **then**

$\boldsymbol{Z}^{(m)} := \text{argmax}_{\boldsymbol{Z}^{(m)}}(\sum_{i=1}^{n}\sum_{k=1}^{K} z_{i,k}^{(m)} \log[\pi_k^{(m-1)}\phi(\mathbf{X_i}; \boldsymbol{\mu_k}^{(m-1)}, \boldsymbol{\Sigma}^{(m-1)})])$

subject to:

$$\begin{cases} \sum_{i=1}^{n} z_{i,k}^{(m)} \geq \tau, k = 1, ..., K \\ \sum_{k=1}^{K} z_{i,k}^{(m)} = 1, i = 1, ..., n \\ z_{i,k}^{(m)} \in \{0, 1\}, i = 1, ..., n; k = 1, ..., K \end{cases}$$

else if min cluster size from $\boldsymbol{Z}^{(0)} \geq \tau$ **then**

$z_{i,h}^{(m)} := \frac{\pi_k^{(m-1)}\phi_k(\mathbf{X_i}; \boldsymbol{\mu_k}^{(m-1)}, \boldsymbol{\Sigma}^{(m-1)})}{\sum_{k=1}^{K}\pi_k^{(m-1)}\phi_k(\mathbf{X_i}; \boldsymbol{\mu_k}^{(m-1)}, \boldsymbol{\Sigma}^{(m-1)})}$

end if

// M Step

$\boldsymbol{\mu_k}^{(m)} := \frac{\sum_{i=1}^{n} z_{i,k}^{(m)} \mathbf{X_i}}{\sum_{i=1}^{n} z_{i,k}^{(m)}}$

$\sigma_p^{2,(m)} := \sum_{k=1}^{K}\sum_{i=1}^{n} \frac{z_{i,k}^{(m)}(X_{i,p} - \mu_{k,p}^{(m)})^2}{n}$

$\pi_k^{(m)} := \frac{\sum_{i=1}^{n} z_{i,k}^{(m)}}{n}$

$\boldsymbol{\mu_k}^{(m)} := sgn(\boldsymbol{\mu_k}^{(m)})(|\boldsymbol{\mu_k}^{(m)}| - \frac{\lambda}{\sum_{i=1}^{n} z_{i,k}^{(m)}}\boldsymbol{\Sigma}^{(m)}\mathbf{1})_+$ ▷ $\lambda = 0$ for CGMM

$\delta := l(\boldsymbol{X}; \boldsymbol{\mu_k}^{(m)}, \boldsymbol{\Sigma}^{(m)}, \boldsymbol{\pi}^{(m)}) - l(\boldsymbol{X}; \boldsymbol{\mu_k}^{(m-1)}, \boldsymbol{\Sigma}^{(m-1)}, \boldsymbol{\pi}^{(m-1)})$

$m := m + 1$

end while

Figure 1: Trend of AMI in high dimension with sparsity.

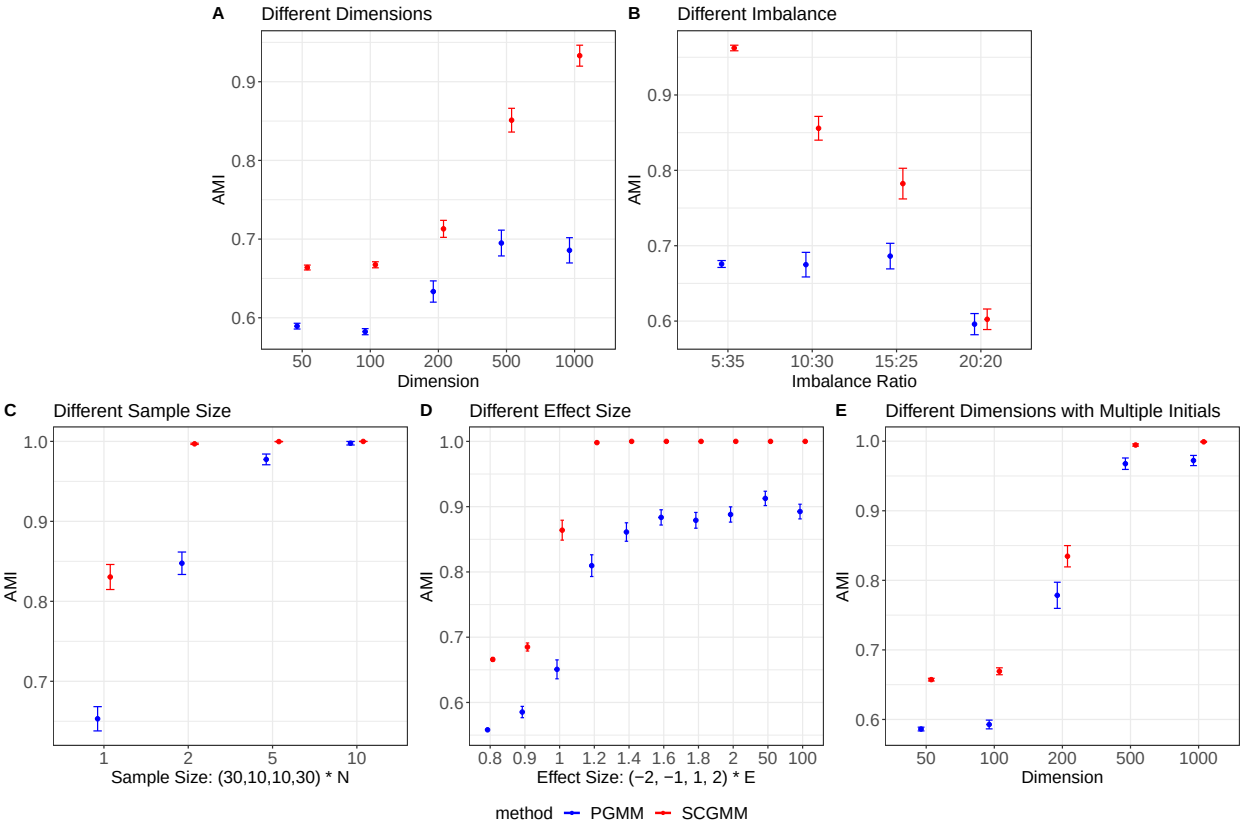


Figure 2: Trend of AMI in high dimension with sparsity.

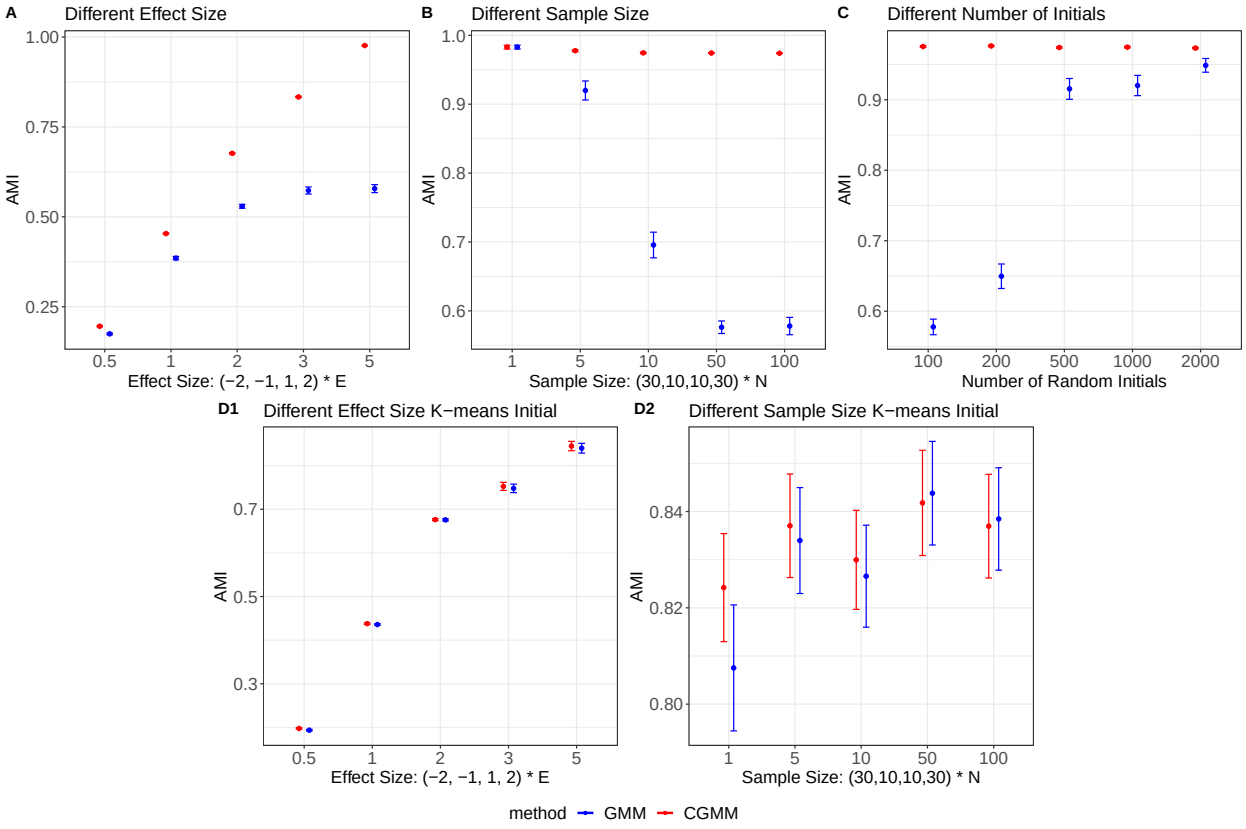


Figure 3: t-SNE for peer grouping data clustering with or without cluster size constraint.

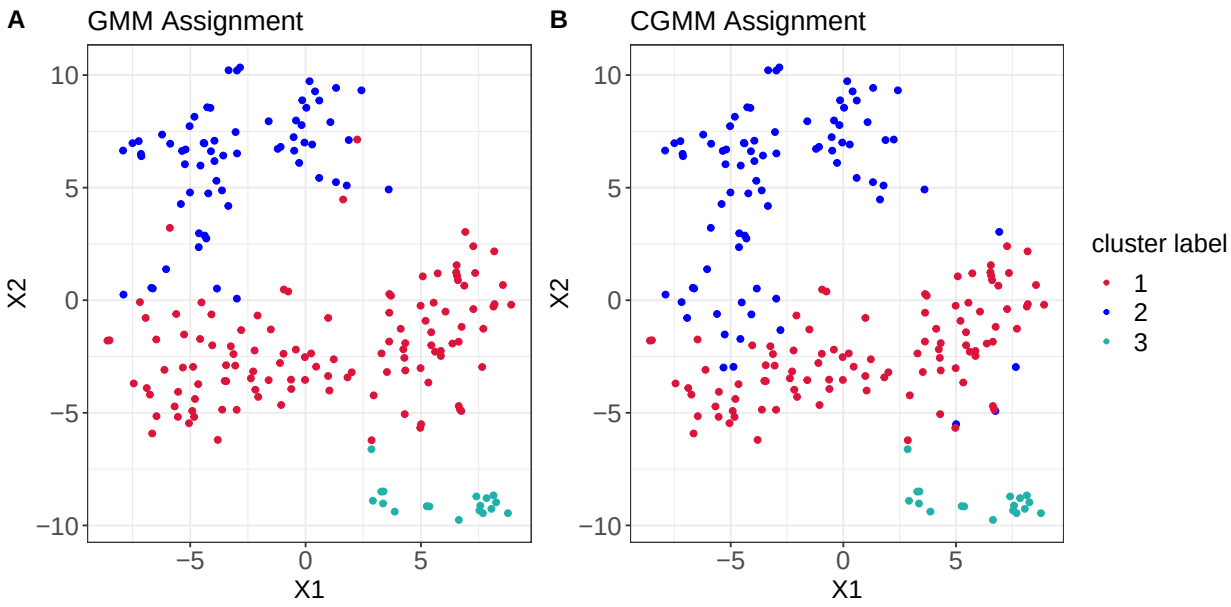


Figure 4: Trend of AMI for gene expression data in GTEx brain regions for different λs.

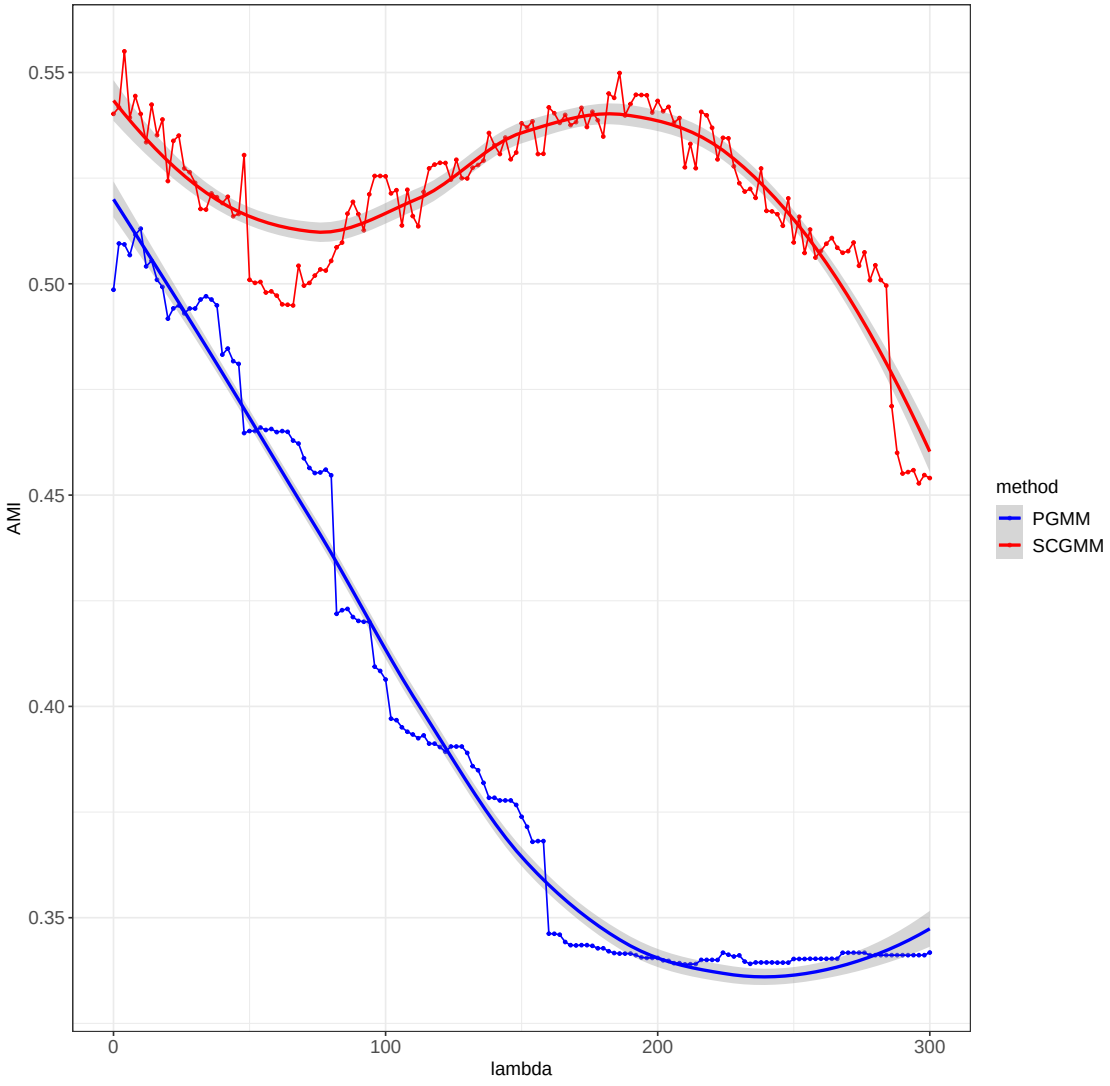


Figure 5: t-SNE plot for clustering assignments by actual labeling, PGMM assignments and SCGMM assignments in GTEx brain regions data.

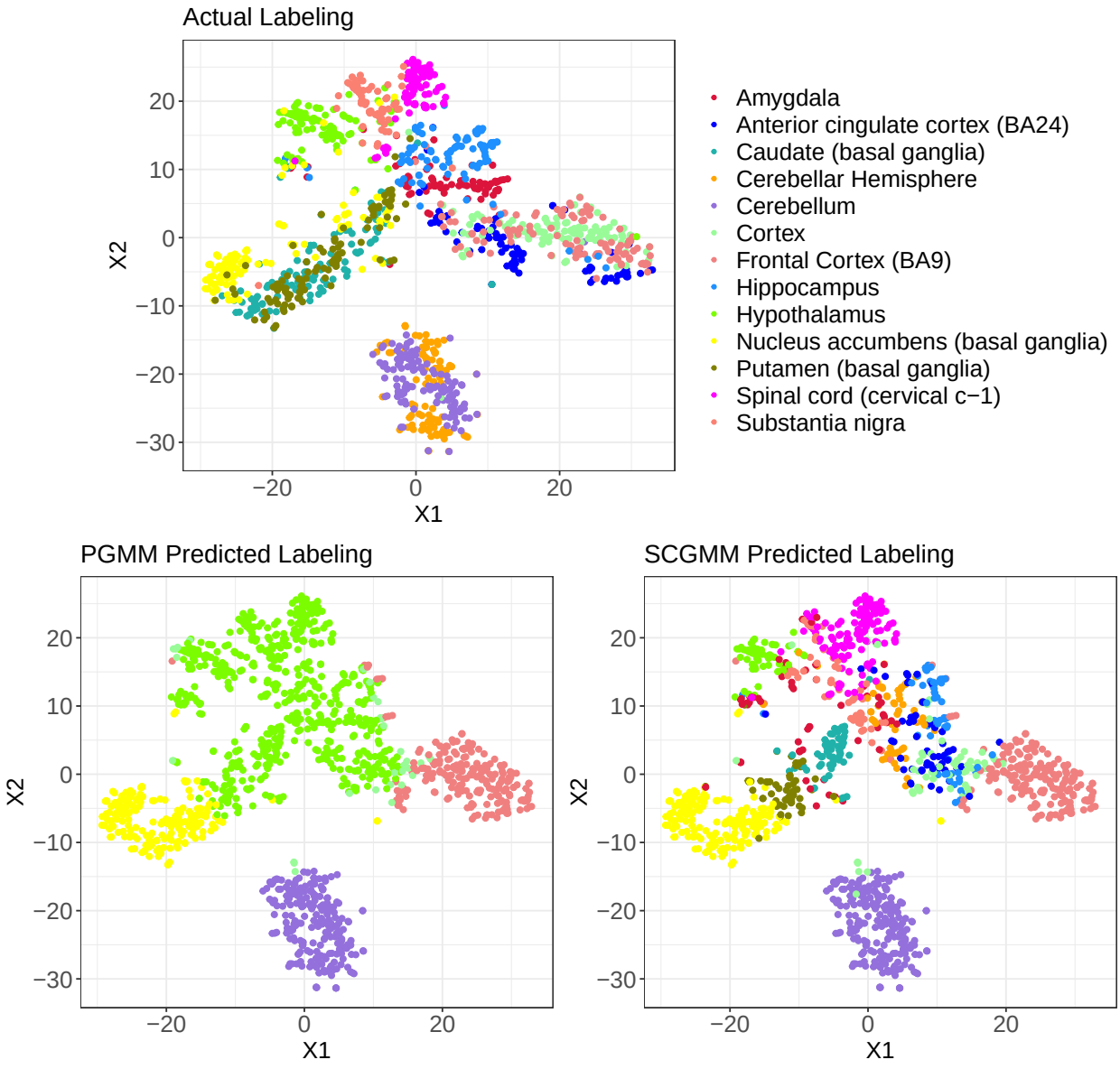


Figure 6: Trend of AMI in the 4-group and 3-group subsampled Zheng4uneq single cell gene expression data.

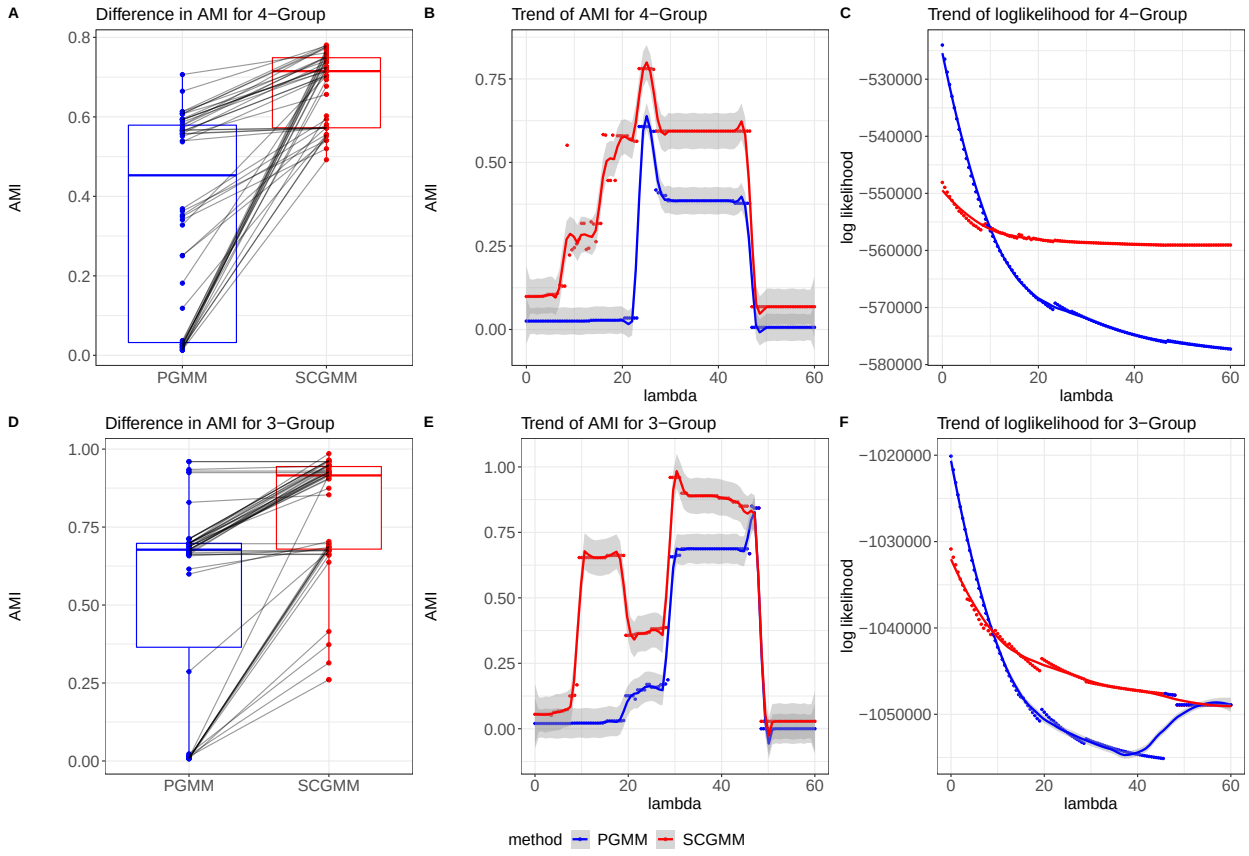


Figure 7: t-SNE plot for clustering assignments by actual labeling, PGMM and SCGMM in subsampled 4-group zhengmix4uneq single cell data.

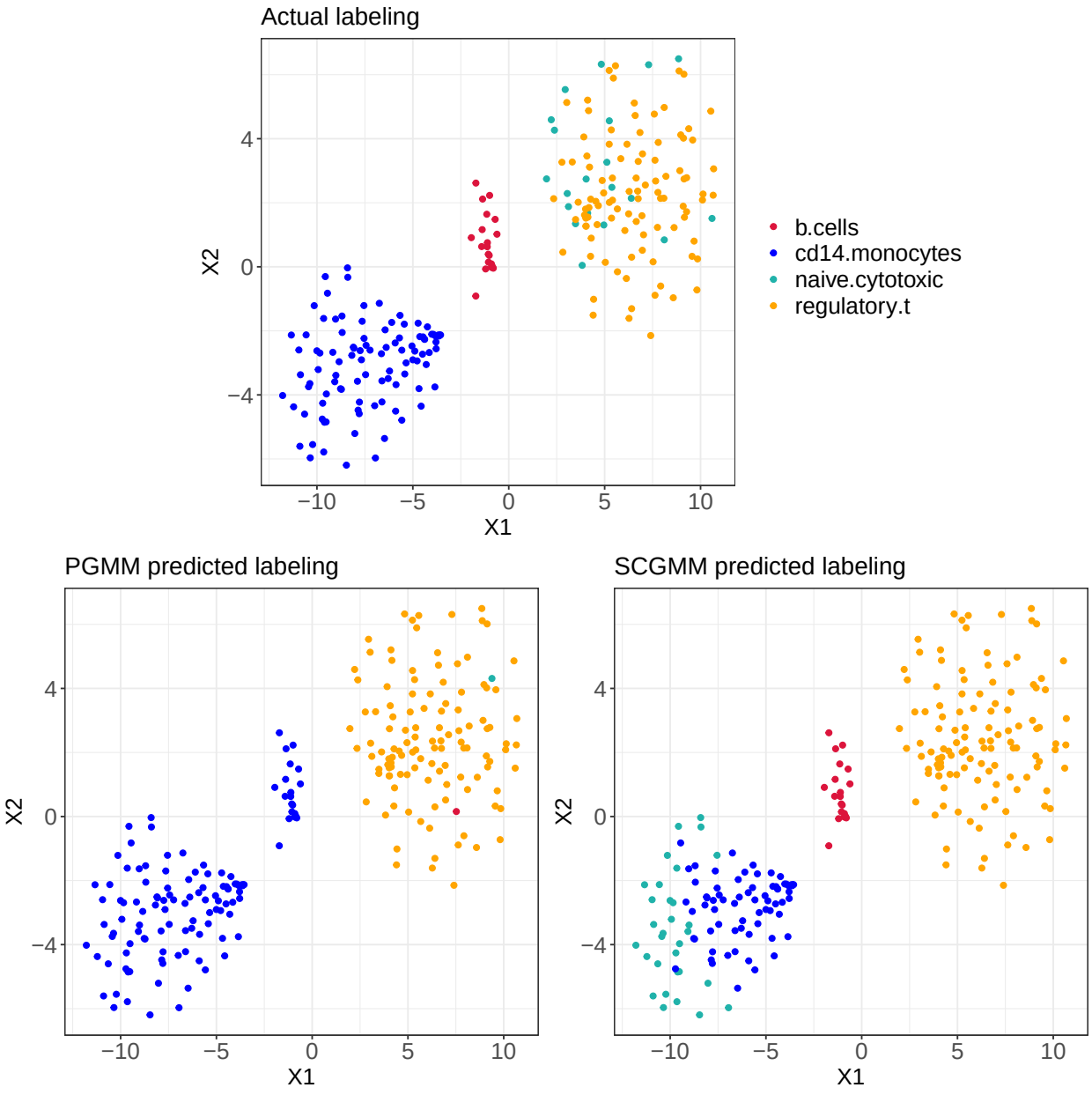


Figure 8: t-SNE plot for clustering assignments by actual labeling, PGMM and SCGMM in subsampled 3-group zhengmix4uneq single cell data.

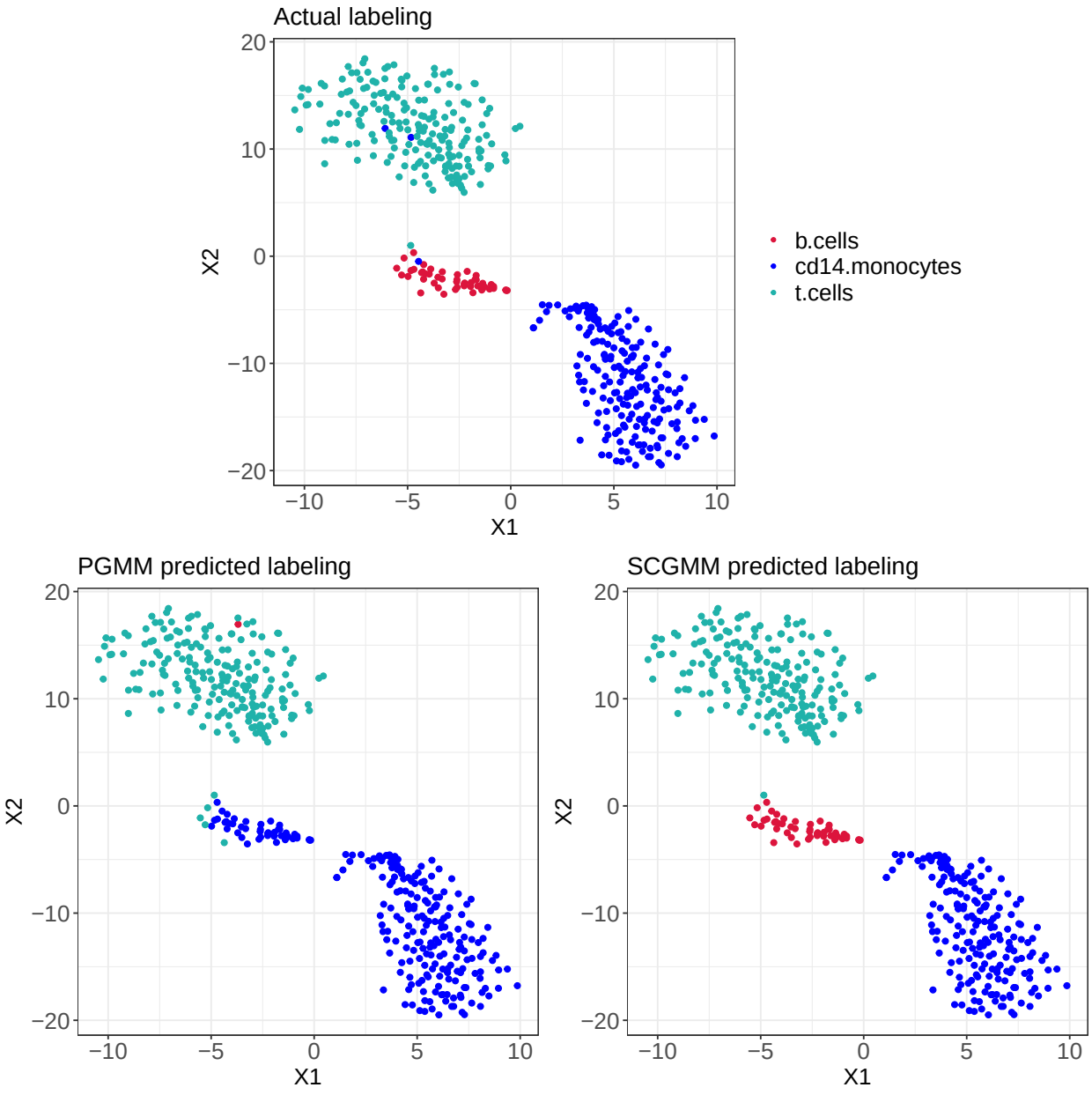


Table 1: Averaged confusion matrix when the difference of AMI reaches the maximum in each scenario of simulation 1. A: when $D = 1000$ in simulation 1A; B: when $S = 5$ in simulation 1B; C: when $N = 1$ in simulation 1C; D: when $E = 1$ in simulation 1D; E: when $D = 100$ in simulation 1E.

PGMM

A-1

Predicted \ Actual	1	2	3	4
1	30	6.71	0	0
2	0	3.29	0	0
3	0	0	2.7	0
4	0	0	7.3	30

B-1

Predicted \ Actual	1	2	3	4
1	35	4.39	0	0
2	0	0.5	0.08	0
3	0	0.11	0.44	0
4	0	0	4.48	35

C-1

Predicted \ Actual	1	2	3	4
1	30	7.24	0	0
2	0	2.74	0.01	0
3	0	0.02	1.88	0
4	0	0	8.11	30

D-1

Predicted \ Actual	1	2	3	4
1	30	8.08	0	0
2	0	1.91	0	0
3	0	0.01	2.56	0
4	0	0	7.44	30

E-1

Predicted \ Actual	1	2	3	4
1	30	8.62	0	0
2	0	1.08	0.34	0
3	0	0.3	1.47	0
4	0	0	8.19	30

SCGMM

A-2

Predicted \ Actual	1	2	3	4
1	30	0.97	0	0
2	0	8.69	0.34	0
3	0	0.34	8.67	0
4	0	0	0.99	30

B-2

Predicted \ Actual	1	2	3	4
1	35	0	0	0
2	0	4.12	0.88	0
3	0	0.88	4.12	0
4	0	0	0	35

C-2

Predicted \ Actual	1	2	3	4
1	30	2.35	0	0
2	0	6.97	0.64	0
3	0	0.68	6.65	0
4	0	0	2.71	30

D-2

Predicted \ Actual	1	2	3	4
1	30	1.9	0	0
2	0	7.54	0.49	0
3	0	0.56	7.36	0
4	0	0	2.15	30

E-2

Predicted \ Actual	1	2	3	4
1	30	5.05	0	0
2	0	3.44	1.72	0
3	0	1.51	3.61	0
4	0	0	4.67	30

Table 2: Averaged confusion matrix when the difference of AMI reaches the maximum in each scenario of simulation 2. A: when $E = 5$ in simulation 2A; B: when $N = 100$ in simulation 2B; C: when $I = 1$ in simulation 2C; D1: when $E = 5$ in simulation 2D-1; D2: when $N = 1$ in simulation 2D-2.

GMM

A1

Predicted \ Actual	1	2	3	4
1	298.19	91.93	1.58	0
2	1.81	7.92	0.22	0
3	0	0	6.9	0.43
4	0	0.15	91.3	299.57

B1

Predicted \ Actual	1	2	3	4
1	2868.49	763.57	66.45	25.47
2	84.67	166.45	24.45	5.96
3	7.6	19.91	152.12	33.63
4	39.24	50.07	756.98	2934.94

C1

Predicted \ Actual	1	2	3	4
1	299.94	93.08	1.57	0
2	0.06	6.92	0	0
3	0	0	6.89	0.07
4	0	0	91.54	299.93

D1-1

Predicted \ Actual	1	2	3	4
1	298.43	32.64	0	0
2	0.59	67.35	0.02	0
3	0.98	0.01	68.66	0.55
4	0	0	31.32	299.45

D2-1

Predicted \ Actual	1	2	3	4
1	29.94	3.68	0	0
2	0.06	6.31	0	0
3	0	0.01	5.74	0.07
4	0	0	4.26	29.93

CGMM

A2

Predicted \ Actual	1	2	3	4
1	299.09	1.32	0	0
2	0.91	98.68	0	0
3	0	0	99.01	0.87
4	0	0	0.99	299.13

B2

Predicted \ Actual	1	2	3	4
1	2990.3	11.51	0	0
2	9.71	988.49	0	0
3	0	0	988.75	10.48
4	0	0	11.22	2989.5

C2

Predicted \ Actual	1	2	3	4
1	299.1	1.29	0	0
2	0.9	98.71	0	0
3	0	0	98.84	0.95
4	0	0	1.16	299.05

D1-2

Predicted \ Actual	1	2	3	4
1	291.82	32.07	0	0
2	0.56	60.6	6.9	8.61
3	7.62	7.33	62.69	0.51
4	0	0	30.41	290.88

D2-2

Predicted \ Actual	1	2	3	4
1	28.49	3.52	0	0
2	0.06	6.12	0.72	0.83
3	1.45	0.36	5.52	0.06
4	0	0	3.76	29.11

Table 3: Cluster assignment in GTEx Brain Region (the numbers of correctly selected are shown in the parentheses).

	PGMM	**SCGMM**	**Actual**
Hypothalamus	590 (90)	60	96
Cerebellum	227 (123)	225	125
Frontal Cortex (BA9)	217 (77)	192	108
Nucleus accumbens (basal ganglia)	189 (74)	178	113
Cortex	36 (9)	60	114
Caudate (basal ganglia)		60	117
Hippocampus		60	94
Spinal cord (cervical c-1)		124	71
Amygdala		60	72
Anterior cingulate cortex (BA24)		60	84
Cerebellar Hemisphere		60	105
Putamen (basal ganglia)		60	97
Substantia nigra		60	63

Table 4: Cluster assignments in 4 group subsampled Zhengmix4uneq single cell gene expression data (the numbers of correctly selected are shown in the parentheses).

	PGMM	**SCGMM**	**Actual**
B cells	1 (0)	20 (20)	20
CD14 monocytes	120 (100)	69 (69)	100
naive cytotoxic T cells	1 (0)	31 (0)	20
regulatory T cells	118 (98)	120 (100)	100

Table 5: Cluster assignments in 3 group subsampled Zhengmix4uneq single cell gene expression data (the numbers of correctly selected are shown in the parentheses).

	PGMM	**SCGMM**	**Actual**
B cells	1 (0)	51 (50)	50
CD14 monocytes	244 (198)	197 (197)	200
T cells	205 (199)	202 (200)	200

3.0 Mutual information for multi-study multi-class concordant biomarker detection

3.1 Introduction

Biomarker detection, which provides accurate biological information for early disease diagnosis, is a critical element in biomedical research [65]. Study integration is a common approach to improve the reliability and power of biomarker detection. If a biomarker shows similar patterns across multiple studies, we could assume that it is a robust choice for disease indication.

Combining p-values and combing effect sizes are two leading solutions for study integration. The first has been widely discussed. For example, Fisher's method sums up the log-transformed p-values, and each p-value is assumed to follow standard uniform distribution under the null hypothesis. Besides Fisher's method, the Stouffer's method [109], the minimum p-value method [116], the higher criticism method [33], and the Berk-Jones method [10] are all constructed under this category and are widely used in the omics study integration, such as GWAS [6], transcriptomics [118], and methylation [107]. Random effects models [31] are an example of the latter approach, which decompose the observed treatment effects of each study into two parts: the actual effect size and the study-specific noise.

These methods have their limitations. The p-value combination focuses only on the significance level without considering the data pattern, and the effect size combination is only available in the two-class scenario (usually the disease vs. normal). When there are more than two categories, the effect size combination is no longer applicable, while the p-value combination cannot precisely decipher the multi-class pattern.

The min-MCC [68] is the only known method to detect the biomarkers with concordant multi-class patterns across multiple studies. It uses the minimum value of the correlations for all pairs of studies. The hypothesis test HS_A for min-MCC is H_0: $\exists\ \rho_{ij} \leq 0$ vs.H_A: $\forall\ \rho_{ij} > 0$, where ρ_{ij} represents the measurement of concordance in the multi-class pattern between study i and j. However, it has two drawbacks. First, it ignores the partially shared strong

signal due to its strict requirement that all the studies should contain a consistent pattern simultaneously. Second, a significant min-MCC does not imply a high degree of concordance between each study pair, because only the minimum pairwise concordance across multiple studies is considered. In other words, none of the study pairs need to have high concordance for the min-MCC to be significantly large.

Based on the previous two drawbacks, we revisited this problem from the perspective of information theory. We proposed a new method called Multi-Study multi-Class Concordance (MSCC), a two-step method for informative biomarker identification. Multi-Study multi-Class Association (MSCA) is the first step based on the concept of total correlation with the corresponding hypothesis HS_B: H_0: $\forall\ \rho_{ij} \leq 0$ vs. H_A: $\exists\ \rho_{ij} > 0$, which solves the above two drawbacks. To identify the studies which share the concordant expression pattern, a post-hoc Multi-Class Mutual Information (MCMI) is then computed in the second step.

In this article, we focused on the gene expression data and aimed to identify the informative genes that showed concordant expression patterns across studies. A visual illustration is provided in Fig. 1 using toy examples (see simulation settings in Supplement Table 9). MSCA first identified Gene 1-3 as those with concordant expression patterns (p-value = 0 for all three genes, enclosed by red triangle). Pairwise MCMI then determined the studies (p-values = 0, enclosed by yellow triangle) that contribute to such concordance for the genes identified in the first step.

The article is organized as follows. In Section 3.2, we started from the problem statement and reviewed MCC and min-MCC [68], followed by a problem reappraisal from an information theory perspective, where we demonstrated the better properties and extensions of MSCC framework. A simulation study and three real-world data applications (Section 3.3) were conducted to compare min-MCC and MSCC.

3.2 Methods

We assume that there are S studies, G genes, and K classes ($K \geq 2$). x^g_{ski} represents the gene expression for gene g ($1 \leq g \leq G$), study s ($1 \leq s \leq S$), class k ($1 \leq k \leq K$), and

sample i ($1 \leq i \leq n_{sk}$).

3.2.1 A brief introduction of MCC and min-MCC

For simplicity, we start from the scenario of two studies ($S = 2$), X and Y, for one gene. For study X, the observed gene expression x_{kj} from sample j class k is assumed to be obtained from $X_k \sim N(\mu_{X_k}, \sigma^2_{X_k})$, where $X_k \perp\!\!\!\perp X'_k$ ($\forall\ k \neq k'$). Therefore, X can be naturally defined as a mixture distribution of X_k ($k = 1 : K$), where

$$f_X(x) = \sum_{k=1}^{K} w_k f_{X_k}(x)$$

$$E(X) = \mu_X = \sum_{k=1}^{K} w_k \mu_{X_k}$$

$$Var(X) = \sigma^2_X = = \sum_{k=1}^{K} w_k(\sigma^2_{X_k} + \mu^2_{X_k}) - \mu^2_X$$

Study Y is similarly defined, and Y_k is independent with X_k. w_k represents the class weight, which is assumed to be $1/K$ in the previous study [68]. To gain the flexibility, we allow it to be estimated from the data. Besides that, the above-mentioned parameters can all be directly estimated from the data.

$$\hat{w}_k = (n_{X_k} + n_{Y_k})/n$$

$$\hat{\mu}_{X_k} = \sum_{j=1}^{n_{X_k}} x_{kj}/n_{X_k}$$

$$\hat{\sigma}^2_{X_k} = \sum_{j=1}^{n_{X_k}} (x_{kj} - \hat{\mu}_{X_k})^2/n_{X_k}$$

Multi-class correlation (MCC) is therefore defined as

$$MCC = \rho = \frac{E(XY) - EX \cdot EY}{\sqrt{Var(X) \cdot Var(Y)}} = \frac{(\sum_{k=1}^{K} w_k \mu_{X_k} \mu_{Y_k} - \mu_X \cdot \mu_Y)}{\sigma_X \cdot \sigma_Y}$$

For multiple S studies, min-MCC is then defined as the minimum value of MCC statistics

across all the pair-wise study combinations:

$$min - MCC = min_{1 \leq u < v \leq S}(MCC_{(u),(v)})$$

3.2.2 MCMI and MSCA

We revisit this problem from the aspect of information theory. We assumed X and Y to be jointly bivariate normal and annotate Z and $Z^{\perp\!\!\!\perp}$ as the bivariate random variables when X and Y are correlated or not respectively.

$$Z \sim N\left(\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}, \begin{bmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y\\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{bmatrix}\right)$$
$$Z^{\perp\!\!\!\perp} \sim N\left(\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}, \begin{bmatrix}\sigma_X^2 & 0\\ 0 & \sigma_Y^2\end{bmatrix}\right)$$

Therefore, we can define the mutual information between X and Y as

$$MI = D_{KL}(Z||Z^{\perp\!\!\!\perp}) = -\frac{1}{2}\log\left(1-\rho^2\right)$$

D_{KL} means the Kullback-Leibler divergence, and ρ is exactly the MCC between X and Y. To be consistent with MCC and limits to the positive correlation, we define multi-class mutual information (MCMI) as

$$MCMI = -\frac{1}{2}(1-\rho_+^2)$$

where

$$\rho_+ = \begin{cases}\rho & \text{if } \rho > 0\\ 0 & \text{if } \rho \leq 0\end{cases}$$

In this case, we can find that MCMI same with MCC, but it has better potential to be generalized to multiple studies. For S studies, we have $Z \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $Z^{\perp\!\!\!\perp} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{\perp\!\!\!\perp})$,

where

$$\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_S)^T$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \cdots & \rho_{1,S_+} \\ \vdots & \ddots & \vdots \\ \rho_{1,S_+} & \cdots & \sigma_S^2 \end{bmatrix}$$

Therefore, we can define the measurement for multiple studies, which is a generalized form of mutual information and known as total correlation [126].

$$MSCA = D_{KL}(Z||Z^{\perp\!\!\!\perp}) = -\frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}^{\perp\!\!\!\perp}|}\right) = -\frac{1}{2}\left(\log|\boldsymbol{\Sigma}| - \sum_{s=1}^{S}\log\sigma_s^2\right)$$

3.2.3 Permutation test for the four statistics

Permutation test is designed to obtain the significance levels for the above four statistics (MCC, min-MCC, MSCA, MCMI) since the analytical solution is not available. We use θ to denote them, and permutation steps are as follows.

1. Compute statistics θ_g for gene g.
2. Permutate the group label B times and calculate the permutated statistics $\theta_g^{(b)}$, where $1 \leq b \leq B$.
3. Calculate the p-value of θ_g,

$$p(\theta_g) = \frac{\sum_{b=1}^{B}\sum_{g'=1}^{G} I(\theta_{g'}^{(b)} \geq \theta_g)}{G \cdot B}$$

4. Obtain the p-values $p(\theta_g)$ for each gene where $1 \leq g \leq G$, and estimate q-values for G genes using Benjamin-Hochberg procedure. ($p_{(i)}$ is ordered i-th p-value)

$$q_{(i)} = min\{min_{j \geq i}\{\frac{Gp_{(j)}}{j}\}, 1\}$$

3.3 Results

In this section, we applied the methods to the simulated datasets and three real-world scenarios, mouse metabolism [68](Lu et al., 2010), Estro-Gene (https://estrogene.org/), and three leukemia datasets [62].

3.3.1 Simulation

We performed the same simulation with the MCC study [68] to identify the genes showing concordant patterns for three classes among three studies. 2,000 genes from four expression patterns were simulated for each study. Among 2,000 genes, 300 genes (category I) have concordant expression across three studies, 100 genes (category II) have discordant expression across three studies, 100 genes (category III) have concordant expression in study 1 and 2 only, and the remaining 1500 genes (category Null) contain no signals (Supplement Table 10). A gene with a $q - value < 0.05$ is considered as informative in the concordant expression pattern. The number of detected genes is shown in Table 6.

The category Null is used for quality control, and both methods show the expected results. The false discovery rates (FDR) are 0.59%, 0.76%, and 0.83% for min-MCC and 0.75%, 0.95% and 1.04% for MSCA in three different effect sizes respectively. MSCA provides higher FDR due to its less stringent null hypothesis.

Categories I and II represent the scenarios when all three studies share the same expression pattern or not simultaneously. Similarly, compared with min-MCC, MSCA shows less stringent results with more genes detected. In category I, with all genes concordant, both methods can successfully detect concordant genes with the false negative rate (FNR) of 30.42% for min-MCC and 21.13% for MSCA when the effect size is 0.5, and the false negative rates decrease when the effect size increases. In contrast, in category II, where genes are discordant across all the studies, both methods fail to detect the concordant genes with a false discovery rate of 0.01% for min-MCC and 13.26% for MSCA when effect size is 0.7.

The main difference between min-MCC and MSCA lies in category III, where Study 1 and Study 2 have the same pattern, while Study 3 contains noise. MSCA tends to identify

the biomarkers as informative ones (detection rate is 81.90% when effect size is 0.7), while min-MCC tends not to (detection rate is 22.06% when effect size is 0.7). In real-world data, it usually happens that only part of the datasets contains the signals, while the others do not due to poor data quality or limited sample size. Therefore, it is more reasonable to detect the genes with concordant expression patterns in the subset of studies and identify which dataset exhibits such consistency, rather than directly detecting the genes with the concordant pattern in all the datasets.

3.3.2 Mouse metabolism data analysis

In this section, we applied MSCC to the study analyzed in the min-MCC paper [68]. A dataset with samples from three genotypes of mice (wild-type, LCAD knock-out, and VLCAD knock-out) was analyzed. LCAD deficiency is associated with impaired fatty acid oxidation, and VLCAD deficiency is associated with energy metabolism disorders in children. Microarray experiments were conducted on tissues from 12 mice (four mice per genotype) including brown fat, liver, heart, and skeletal. The expression changes across genotypes were studied, and genes with little information content were filtered out to have 4288 genes remaining for downstream analysis. Four arrays were identified with quality defects and excluded from further analysis.

A total of 1,394 concordant genes were identified through MSCA analysis ($q - value < 0.05$). To gain further insights of these concordant genes, we implemented QIAGEN Ingenuity Pathway Analysis (IPA) [55] on the MSCA q-values (Supplement Table 11). The top three pathways associated with the MSCA results were mito-chondrial dysfunction, Sirtuin signaling pathway, and oxidative phos-phorylation ($p - values < 0.01$), which have been shown to correlate with LCAD and VLCAD knock-outs [134, 82]. These findings confirm the roles of LCAD and VLCAD and validate the efficacy of MSCA.

Compared to MSCA, min-MCC only detected 393 concordant genes, suggesting tissue heterogeneity. To assess the necessity of MSCA, we classified genes into three subsets: genes identified by min-MCC only (V), genes detected by min-MCC and MSCA simultaneously (M1), and genes identified only by MSCA (M2-M11). In subset 3, we classified genes into 10

categories based on post-hoc MCMI results and clustered genes within same category using k-means. The number of clusters was determined using the NbClust R package [19]. Figure 10 displays the expression patterns for each gene category. Ambiguous expression patterns were observed for genes in V. Genes in M1 were divided into three clusters and showed high concordance across all four tissues. Partial concordance was observed in categories M2-M11, which were not detected by the min-MCC method and are highlighted in red panels. Supplement Figure 19 further illustrates the expression patterns for each gene category using boxplots.

It is crucial to identify the genes with concordant expressions in partial tissues. For example, *Blvrb* showed the largest MSCA statistic (stat = 2.323, q-value = 0), while min-MCC failed to detect it (stat = -0.711, q-value = 1) (Supplement Figure 20). *Blvrb* demonstrated lower expression in LCAD knock-out samples in brown fat, heart, and skeletal tissues, but higher expression in the liver. Despite lacking the reported direct relation with LCAD and VLCAD, *Blvrb* is related to metabolism and converts biliverdin to bilirubin in the liver [1]. Notably, *Blvrb* exhibits the highest gene expression in the liver among multiple tissues, according to the Human Protein Atlas (proteinatlas.org), in the GTEx database [66, 120], suggesting unique liver-specific functions compared to the other three tissues.

3.3.3 EstroGene data analysis

The EstroGene project (related paper submitted for publication) focuses on improving the understanding of the estrogen receptor and its role in the development of breast cancer. It aims to document and integrate the publicly available estrogen-related datasets, including RNA-seq, microarray, ChIP-seq, ATAC-seq, DNase-seq, ChIA-PET, Hi-C, GRO-seq and others, to establish a comprehensive database that allows for customized data search and visualization.

In this section, we only considered studies that included gene expression data (microarray and RNA-seq) and limited our analysis to the samples with estrogen receptor positive (ER+) treated with estradiol (E2) doses greater than 1nM for varying duration. We first combined the samples by cell line and sequencing technology. To further analyze the data, we then

classified the treatment duration into three categories: short (< 6 hours), medium (≥ 6 hours and ≤ 24 hours), and long (> 24 hours). Finally, we normalized the data for the newly pooled studies using trimmed mean of M values (TMM) [16] followed by ComBat [48] with the study indication as a batch covariate. These steps resulted in three pooled studies: MCF7 microarray (25 samples in short treatment, 34 in medium treatment, and 7 in long treatment), MCF7 RNA-seq (49 in short treatment, 62 in medium treatment, and 10 in long treatment), and T47D RNA-seq (3 in short treatment, 22 in medium treatment, and 11 in long treatment). 1,983 genes were intersected across multiple platforms for downstream analysis.

We first validated the two well-established benchmark genes, $GREB1$ and $IL1R1$, which have been widely reported as E2 activated and repressed genes [21, 89, 102, 60]. Figure 11 revealed the up- and downregulation of $GREB1$ and $IL1R1$ in MCF7 microarray and MCF7 RNA-seq studies. However, these trends were not observed in the T47D RNA-seq study, which may be due to the limited sample size. As a result, MSCA identified both genes as concordant with q-values of 0.03 and 0, while the min-MCC failed to detect them with q-values of 0.07 and 0.11, respectively.

In addition to validation, we are also able to detect novel biomarkers. For example, $MECOM$ was the only gene identified by MSCA and min-MCC with q-values = 0 simultaneously (Figure 11). Prior to our study, $MECOM$ was not recognized as a biomarker for E2 treatment. Our analysis revealed that $MECOM$ is a gene that is repressed by E2, indicating lower E2 response responsiveness and potentially poorer response to endocrine treatment. Therefore, we could hypothesize a worse survival outcome if a patient has higher $MECOM$ gene expression.

We tested our hypothesis using the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database [26] and extracted 1,459 patients with ER+ breast cancer for the analysis. We normalized the microarray gene expression data using TMM and analyzed the overall survival (OS) and relapse free survival (RFS) outcomes. By fitting Cox proportional hazards regression models [2], we observed that higher MECOM gene expression was associated with worse hazard ration (HR) in terms of OS (HR = 2.27, p-value = 0.048) and RFS (HR = 3.34, p-value = 0.015). The potential mechanism of the clinical prognosis

could partially be explained by the regulation of estrogen receptor, as we observed several consistent ER binding sites at transcription start sites (TSS) proximity from ChIP-seq data in EstroGene web, and $MECOM$ may also play a role in immune suppression [72] which could not be reflected in these cell culture experiments. Future investigation is still needed.

In total, MSCA identified concordant 281 genes ($q - value < 0.05$). To gain a deeper understanding of the upstream transcription factors associated with these genes, we applied LISA, an algorithm that uses chromatin profile and H3K27ac ChIP-seq data to determine the transcrip-tion factors (TF) and chromatin regulators related to a given gene set [88]. Among the top-ranked TFs (Supplement Table 11), $ESR1$ and $FOXA1$ are the TFs that have previously been reported to be associated with E2 [20, 114]. In addition, the presence of $SMC1A$ and $CTCF$ among the top 3 candidates suggests a potential role of topologically associating domain (TAD) in the regulation of these gene [95, 27]. These findings revealed that the E2 response may involve gene regulation through chromatin looping mechanisms. Further experimental studies are needed to fully elucidate the underlying mechanisms.

3.3.4 Three leukemia datasets analysis

Following Li [62], we analyzed three leukemia transcriptomic studies with 3 pre-detected chromosome translocation subtypes: inv(16), t(15;17), and t(8;21). The microarray datasets were directly obtained from NCBI GEO with GSE6891 [123], GSE17855 [3], and GSE13159 [52]. We preprocessed the data by removing probesets with missing values and selecting probesets with the largest interquartile range if multiple probes were mapped to the same gene. The remaining 20,192 genes were used in the analysis.

We identified 9,889 concordant genes by MSCA and compared the results from min-MCC, which identified 5,834 genes. Similar to section 3.3.1, we divided the genes into three subsets: genes identified by both MSCA and min-MCC (M1), genes identified exclusively by MSCA (M2-M4), and genes identified only by min-MCC (V). We used K-means to cluster the genes within each category (Figure 12). Although the three studies were conducted similarly, we observed 2,838 genes that are only concordant in partial studies (M3-M5). We also prepared boxplots of averaged gene expressions to visualize the features of each

category (Supplement Figure 24). Weak signals were observed in V, suggesting that min-MCC compromises signaling strength when requiring the same expression across studies.

3.4 Discussion and conclusions

Meta-analysis is an efficient tool for biomarker detection by increasing the statistical power [25, 117]. To date, min-MCC is the only available method to detect the biomarkers with concordant multi-study multi-class expression patterns [68]. However, since min-MCC cannot identify the partially concordant biomarkers and is insensitive to the pairwise high concordance, we revisited this problem from the aspect of information theory. We proposed a two-step framework MSCC (multi-study multi-class concordance), including MSCA (multi-study multi-class association) and pairwise MCMI (multi-class mutual information). Both the simulation and real application results disclose the superiority of the MSCC framework in selecting more informative biomarkers and detecting the datasets that exhibit such concordance.

Through the simulation, we aimed to investigate the differences between MSCA and min-MCC. Our results showed that MSCA tends to select more biomarkers than min-MCC due to different hypothesis testing. Specifically, in category III of the simulation, where gene expressions are concordant in a subset of studies, 81.90% of the genes are identified as informative by MSCA, and it is 22.06% for min-MCC when the effect size is 0.7, suggesting that MSCA can detect the partially shared signals while min-MCC cannot.

In the analysis of mouse metabolic data, 1,394 concordant genes were identified by MSCA, while min-MCC detected only 393 genes, indicating tissue heterogeneity. Genes were classified into multiple categories based on the results of both methods, and 371 genes were concordant in only a subset of tissues. It is crucial to identify such genes as they may have unique tissue-specific functions. One such example is *Blvrb*, which was downregulated in brown fat, heart and skeletal tissues in LCAD knock-out samples but upregulated in liver. *Blvrb* is related to metabolism and has the highest expression in liver compared to other tissues, suggesting possible unique liver-specific functions compared to the other three

tissues.

The EstroGene data analysis also provides a compelling illustration of the efficacy of MSCC. The detection of $GREB1$ and $IL1R1$ highlighted the utility of MSCC for biomarkers identification, even when some studies fail to provide useful information due to limited sample size or poor data quality. In addition, the identification of $MECOM$ provided a potential biomarker to predict the clinical prognosis of E2 treatment. Finally, using the 281 MSCC identified genes in LISA, we found the involvement of $ESR1$, $FOXA1$ and chromatin looping mechanisms in E2 response.

Similarly, in the analysis of three leukemia datasets, we observed 2,838 genes that are concordant only in partial studies (M2-M4), and weak signals were found in the genes unique to min-MCC (V), indicating that min-MCC compromises signal strength when requiring the same expression across studies.

There are two possible extensions to our method. First, a non-parametric approach could be achieved by changing the definition of pairwise correlation ($\rho_{i,j}$) to a rank-based formula. Second, we assume gene-wise independence in this study, which could be generalized to the dependency structure considering the possible relationship among different genes.

Figure 9: The illustration of MSCC framework. The Gene 1-3 that show concordant patterns across studies are first identified by MSCA (enclosed by red triangle). The studies which share the concordance for each gene are later detected by MCMI (enclosed by yellow triangle).

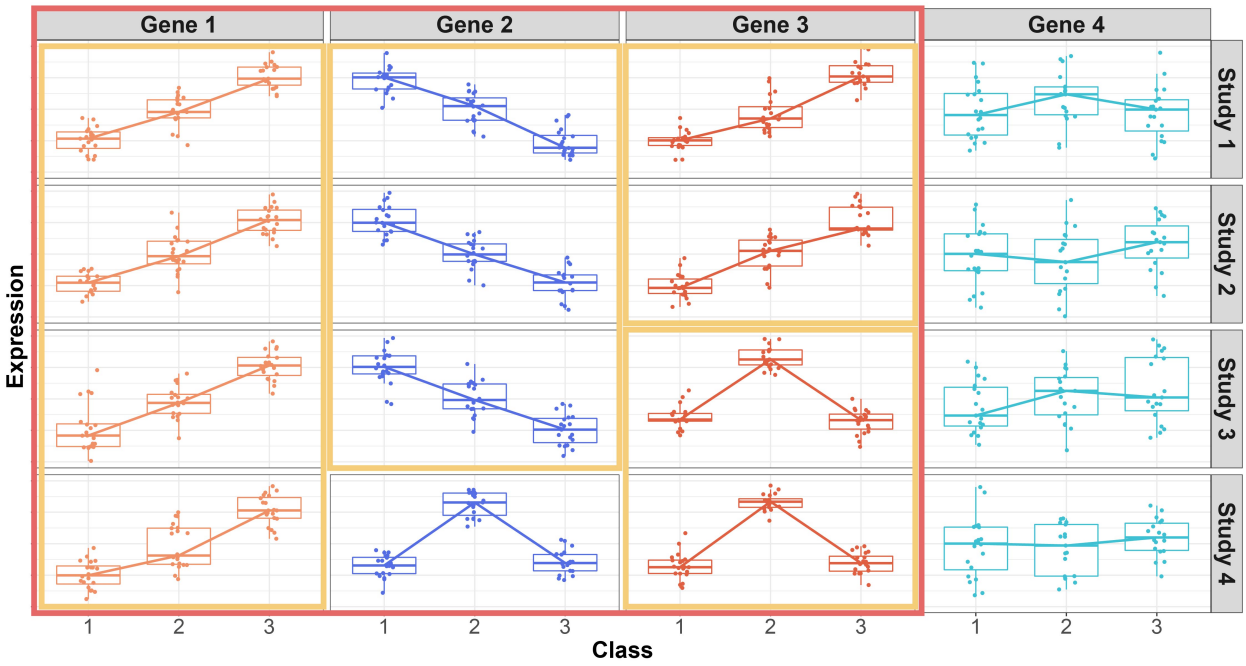


Figure 10: The heatmap of the gene expression patterns of different gene categories across four tissues in mouse metabolism data analysis. The rows represent for the genes and the columns represents for the samples. V includes genes detected by min-MCC only, while M1 includes genes detected by min-MCC and MSCA at the same time. The genes in M2-M11 were identified by MSCA alone and categorized by the contributing studies using MCMI post-hoc analysis. Studies that contributed to the concordance are shown in red panel, while those that did not are shown in gray. More stringent threshold (q-value < 0.01) for concordant gene identification was applied for visualization.

Figure 11: The expression patterns of $GREB1$, $IL1R1$, and $MECOM$ across three data sources. $GREB1$ and $IL1R1$ are widely reported as E2 activated and repressed genes and were detected by MSCA while failed to be identified by min-MCC. $MECOM$ was the only gene detected by MSCA and min-MCC simultaneously.

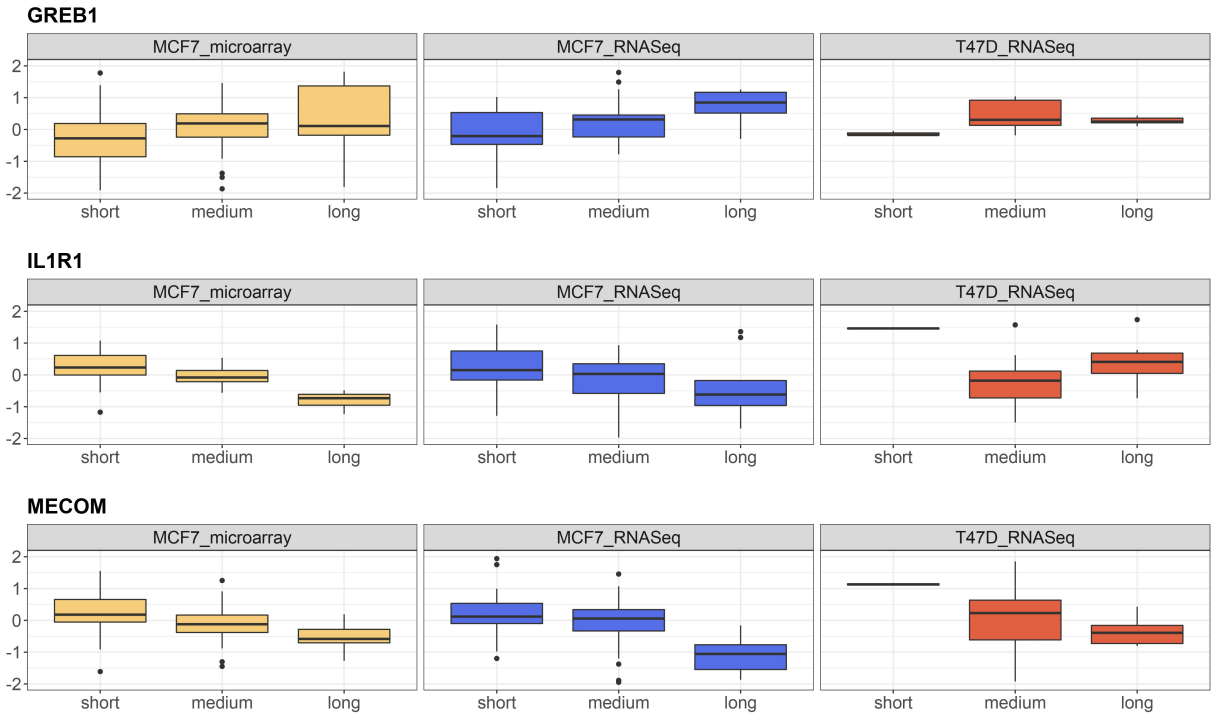


Figure 12: Heatmap of the gene expression pattern of different gene categories across three studies in leukemia data analysis. V includes the genes identified by min-MCC alone, and M1 includes the genes identified by min-MCC and MSCA together. Genes in M2-M4 were detected by MSCA alone and categorized by contributing studies using MCMI post-hoc analysis. Studies that contributed to the concordance are shown in red and those that did not are shown in gray. A stricter threshold (q-value < 0.01) for concordant gene identification was used for visualization.

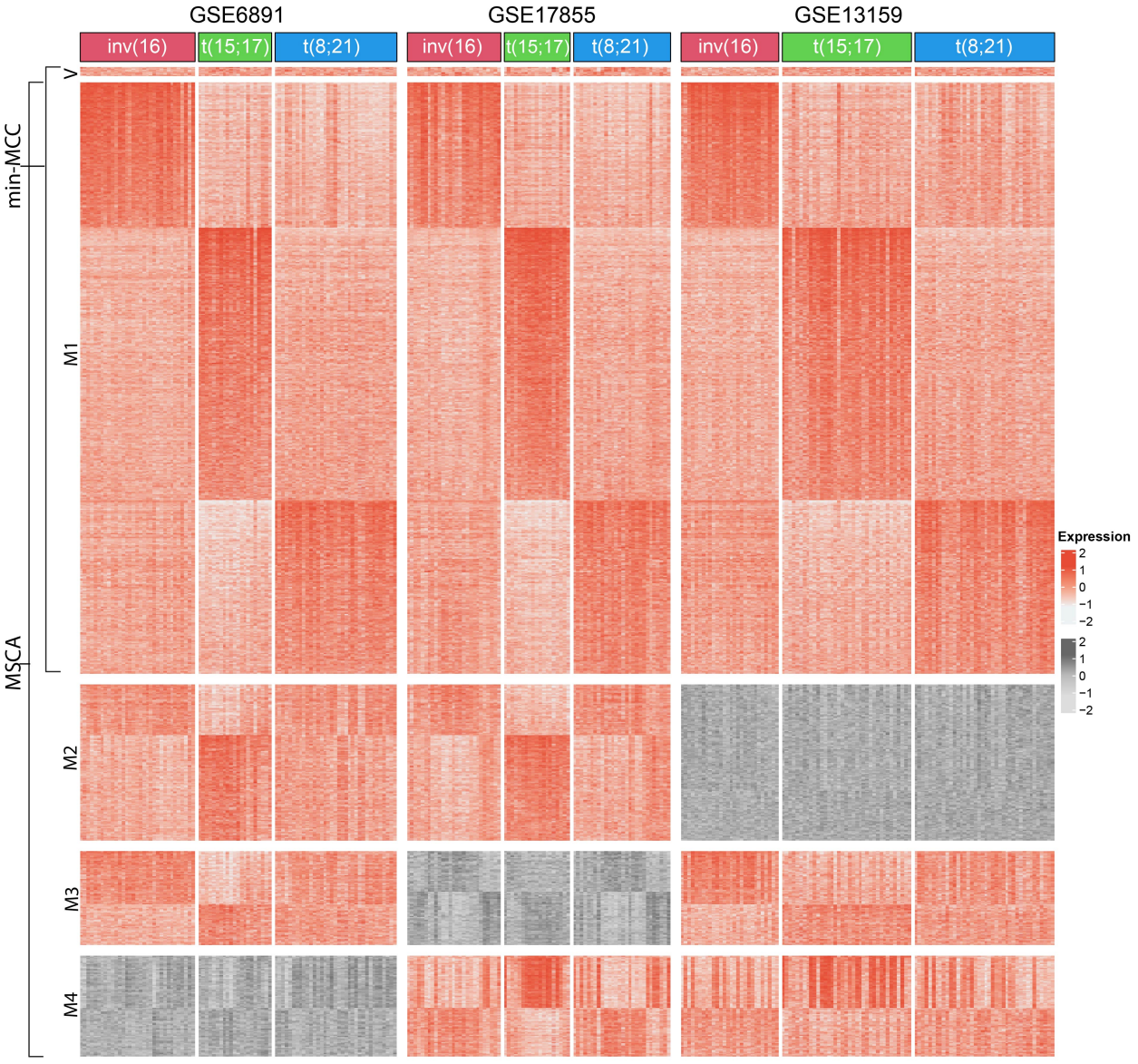


Table 6: The average number of detected genes which show the concordant expression pattern. MSCA is less stringent and detects more genes, especially when the signals are only present in part of the datasets.

Effect size	Methods	I (300)	II (100)	III (100)	Null (1500)
0.5	min- MCC	209.48	0.18	12.94	10.08
	MC-TC	237.48	7.02	38.22	13.12
0.6	min- MCC	266.96	0.10	18.14	12.14
	MC-TC	284.36	10.92	61.72	16.00
0.7	min- MCC	290.58	0.04	22.28	13.62
	MC-TC	297.50	13.74	81.24	17.24

4.0 Transcriptomic congruence and selection of representative cancer models towards precision medicine

The contents of this Chapter are prepared and ready for submission.

4.1 Introduction

Cancer models, inheriting genetic properties of the tumors of origin, are essential tools in cancer research for exploring carcinogenesis and developing drugs in basic, translational and clinical studies. For a given cancer subtype, a wide selection of models, such as cell lines, patient-derived xenografts (PDX), patient-derived organoids (PDO), and genetically modified murine models, are often available to researchers. Specifically, patient-derived cancer models, such as PDO, are increasingly available with molecular profiling and are expected to play a heightened role in disease understanding, drug response prediction and precision medicine [50, 127]. For example, the NCI-funded PDCM Finder (Patient Derived Cancer Model Finder) provides an open catalog of patient-derived cancer models with an established "minimal information standard" for researchers to upload new cancer models [74], which currently includes 4,661 xenograft models, 1547 cell lines and 108 PDO as of 10/20/2022.

Despite advances in technology and reduced cost, cancer models can be mislabeled [133] and genomic/epigenomic alterations may accumulate across passages in culture. Many cancer models may be potentially mis-annotated from their origins or the quality of congruence may vary or decay over time [125, 133, 8]. Due to increasing availability of new cancer models and associated comprehensive omics data, evaluation and comparison of cancer models with human tumors using transcriptomic and multi-omics data have drawn increasing attention in recent years [125, 86, 78, 98, 133, 4, 64, 132, 94, 124, 5, 32, 100, 99]. However, existing evaluation tools mostly belong to two major categories, congruence (correlation-based) analysis and authentication (machine-learning-based) analysis, and do not sufficiently serve the

purpose of identifying appropriate models for precision medicine. In congruence analysis, correlation/association measures are usually applied to quantify similarity of a cancer model to the target tumor cohort in a genome-wide scale [125, 4, 64, 124, 5, 99]. In contrast, authentication analysis develops machine learning models, such as suitability score [32], random forest, ridge regression and nearest template prediction, for accurate assignment of cancer models to human cancer types. Appendix B.1 outlines features and shortcomings of the existing methods. Overall, these tools have significant limitations in the following four areas: (1) Machine-learning-based authentication methods focus on predication accuracy but are not designed to prioritize candidate cancer models that best mimic the target tumor cohort; (2) On the other hand, correlation-based congruence methods can prioritize cancer models but they often produce lower prediction accuracy; (3) Current congruence or authentication methods cannot characterize pathways or molecular mechanisms that are most or least mimicked by a cancer model, which is essential in precision medicine development; (4) Data compatibility and harmonization between cancer model and human tumor data have not been systematically considered and evaluated in the current literature, which is a critical step to achieve high accuracy and avoid misleading mechanistic conclusions.

To this end, we developed CASCAM with three modules to overcome the aforementioned shortcomings of existing methods (see Figure 13). In the first "data harmonization" module, we applied the recently developed Celligner method to correct for batch effects and obvious variations between cancer models and tumors that prevent analysis of congruence. In the second "interpretable machine learning pre-selection" module, we developed an interpretable machine learning approach, integrating prediction assignment probability from sparse linear discriminant analysis (SDA) and deviance score derived from the SDA projected space. The integrative framework combines advantages of high classification accuracy by machine-learning-based authentication analysis and prioritization by correlation-based congruence analysis to pre-select, say, the top 10 promising cancer models from up to hundreds of initial candidates. The pre-selected cancer models then enter the final "pathway and mechanistic-based selection" module. By integrating pathway and regulatory network information, multiple bioinformatic and visualization tools, including differential expression, pathway enrichment analysis, heatmaps, violin plots and topological network plots, itera-

tively investigate disease-relevant biological mechanisms that are best or least mimicked by each cancer model. We note that the two-stage selection by global (genome-wide congruence) pre-selection in Module 2 and then targeted (pathway- and gene-based congruence) evaluation in Module 3 is an essential and innovative aspect of CASCAM. We demonstrate that the highest genome-wide congruent cancer models selected from Module 2 may not harbor critical pathways and genes relevant to the target tumor subtype and thus show a lower score in essential pathways. On the other hand, pre-selection in Module 2 is necessary to reduce the number of cancer model candidates for allowing detailed mechanistic investigation in Module 3.

For demonstration purposes, both case studies in this paper focused on invasive lobular breast carcinoma (ILC), a histological subtype containing 10-15% of all breast cancers and with a hallmark genomic feature consisting of CDH1 gene (E-cadherin) mutation and subsequent loss of cell-cell adherent junctions. There is a compelling need to develop and identify representative cancer models for ILC since previous breast cancer models mostly focus on the more prevalent ($\sim$80%) invasive ductal carcinoma (IDC) subtype (also known as no special type (NST)). Indeed, there are very few ILC annotated cell lines publicly available; however, a previous study identified numerous breast cancer (BC) cell lines which lack ILC annotation but harbor CDH1 mutations – they were named 'ILC-like' and these potentially could serve as representative models of human-ILC disease [75]. Beyond ILC, we note that CASCAM is applicable in general cancer research by quantifying congruence and identifying the most appropriate cancer model for any given tumor (sub)type.

4.2 Results

4.2.1 Case study 1: Selection of cell line for ILC

4.2.1.1 Data harmonization between cancer model and tumor transcriptomic data

The critical first step for quantifying congruence and selection of cancer model(s) is to ensure omics data harmonization between cancer models and human tumors. This is important as cell lines do not contain many genes expressed in the tumor microenvironment. We accessed bulk transcriptomic data of 9,264 pan-cancer tumor samples across 24 cancer types from TCGA (960 samples are breast cancer, BC), and 1,257 pan-cancer cell lines from CCLE and ICLE (65 annotated as BC cell lines) (see Section 4.4). We then evaluated performance of normalization using five approaches – A) no normalization; B) quantile normalization [13] to normalize BC tumors and BC cell lines ; C) ComBat [48] to normalize BC tumors and BC cell lines; D) Celligner to normalize BC tumors and BC cell lines; E) Celligner to normalize pan-cancer tumors and pan-cancer cell lines. Figure 14 shows UMAP [73] plots of BC tumors and BC cell lines when different normalization approaches were applied. Biased separation of tumors and cell lines was clearly found when no normalization or conventional quantile normalization were implemented. Combat and Celligner using BC tumors and BC cell lines (approaches C and D) produced improved normalization although systematic bias was still observed from small clusters of cell lines, showing insufficient quality of data harmonization. In contrast, Celligner using pan-cancer tumors and pan-cancer cell lines (approach E) best eliminated batch effects between BC tumors and BC cell lines.

To further examine the quality of Celligner normalization in approach E, we investigated eight cell lines each with three experimental replicates from different sources (see Section 4.4) and confirmed their high reproducibility in the UMAP plot (Figure 15A). From breast cancer subtype annotation [85] , we confirmed that TCGA tumors in the lower-right cluster were mostly annotated as basal-like (118 out of 160) (Figure 15B). Reassuringly, 26 of 28 CCLE cell lines in that cluster were also annotated as basal-like. As a result, we performed Approach E Celligner normalization before all down-stream analyses in this paper.

4.2.1.2 Interpretable machine learning pre-selection

We next extended and applied an interpretable machine learning (ML) method, namely sparse discriminant analysis (SDA), and combined with a deviance score (DS) derived from SDA to pre-select from up to hundreds of candidate cancer models and narrow down to < 10 of the most promising cancer models. The machine learning (ML) setting here was similar to existing literature, where a prediction model was constructed using human tumor data (e.g., TCGA) as the training set and then was used to classify cancer models to the targeted group (ILC) versus comparison group (IDC). To justify application of SDA, Table 7 shows performance of 16 popular machine learning methods, six of which were used to classify cancer models according to TCGA cancer types in the literature. Detailed description of these machine learning methods can be found in B.1 and Method section. In existing publications, machine learning analyses aimed to classify cancer models into major cancer types, such as the 24 cancer types in TCGA. We note that since the two subtypes we focus on (ILC and IDC) are two histological subtypes within breast cancer, the differences are more subtle. The machine learning and congruence analysis tasks are expected to be more difficult but biologically more impactful.

Table 7 shows evaluation result of the 16 machine learning methods in BC machine learning tasks from three different aspects (tumor type, histological subtype, and molecular subtype). Convolutional neural network (CNN) is a category of deep learning methods commonly designed for classification problems [81, 76, 91]. In this study, we included three CNN models initially optimized for pan-cancer classification [76]. Columns 2-4 contain prediction accuracy results: 5-fold cross validation of ILC versus IDC using TCGA BC data, ER+ versus ER- classification using TCGA as training data and CCLE as test data, and BC versus other cancer types using TCGA as training data and CCLE as test data. The result shows SDA and elastic net to have the highest average accuracy, followed by 2D-Hybrid-CNN and ridge regression methods. Specifically, SDA achieved 91% accuracy for ILC vs IDC cross-validated tumor classification, 91% for ER+ vs ER- cell line prediction and 86% for BRCA vs other cancers in cell line prediction. The CNN methods produced reasonably high accuracy in the three tasks but not among the best, possibly due to limited sample

size. 2D-Hybrid-CNN was proposed to benefit from having two-dimensional inputs with simple one-dimensional convolution operations and had better performance than 1D-CNN and 2D-Vanilla-CNN, consistent with previous results in the cancer subtype classification [76].

In addition to binary prediction accuracy performance, Table 7 lists three machine-learning relevant properties that are critical for evaluation and selection towards precision medicine: feature (gene) selection, prediction assignment probability and deviance measure. Explicit gene selection identifies gene signatures involved in the prediction model and provides interpretable machine learning. Assignment (prediction) probability reports prediction confidence and ranking for cancer models predicted into the target tumor subtype. Finally, deviance measure (e.g., dissimilarity measure or lack-of-association measure) provides supplemental information to prediction assignment probability for cancer model suitability. Of the 16 methods in Table 7, only SDA and a robust variant, RSDA, can be extended for all three interpretable machine learning properties. Taken together, SDA was among the most accurate machine learning methods and provided three essential properties of gene selection, assignment probability (denoted as P_{SDA}) and deviance score (denoted as DS_{SDA}); it was chosen to be the core machine learning method in CASCAM. Particularly, we defined deviance score DS_{SDA} as the (signed) standardized distance between a cell line to the center of the target tumor cohort on the SDA projected space. Bootstrap analysis was then performed in the tumor data to calculate the confidence interval and two-sided p-value, denoted as pval(DS_{SDA}) (see Section 4.4 for details).

Since nearly all ILC cases are luminal ER-positive (i.e. basal negative or non-basal [93]), we focused on the large luminal non-basal cluster in Fig. 2B, which contains 798 BC tumors and 37 BC cell lines. DU4475 [59] was manually included to explore the performance of a basal positive cell line. Of the 38 cell lines, we pre-selected 14 candidate cell lines using Module 2 by high prediction assignment probability ($P_{SDA} > 0.5$) and small deviance score such that the corresponding p-value is not statistically significant (i.e., $pval(DS_{SDA}) > 0.05$) (Supplement Table 13). There were the 153 genes selected by SDA for constructing machine learning model. The pathognomonic feature of ILC is mutation of $CDH1$ and a subsequent reduction $CDH1$ mRNA expression. Thus, as expected the weight for $CDH1$ was -10.428

and was at least 5-10 fold greater than all the other predictive genes.

The necessity of using P_{SDA} and DS_{SDA} simultaneously can be seen in the SDA projected scatter plot of the 38 cell lines in Figure 16A. If we only used P_{SDA} information (marked by red color), cell lines such as UACC812 and ZR751 were predicted to be ILC with $P_{SDA} = 100\%$, disregarding the fact that these cell lines' expression patterns were highly dissimilar to the averaged expression pattern (center) of ILC tumor cohort on SDA projection (large $DS_{SDA} = 2.991$ and 2.970 to ILC, respectively) (Supplement Table 13). In contrast, if we only used DS_{SDA} (marked by round shape), cell lines such as OCUBM and UACC893 had a relatively small deviance score to ILC ($DS_{SDA} = 1.597$ and 1.667 to ILC, respectively), but they were also close to the center of IDC tumor cohort. By applying the combined criteria of DS_{SDA} and P_{SDA}, 14 of the 38 cell lines were identified as well-resemblance to the ILC subtype ($P_{SDA} = 54.8 - 100\%$ and $DS_{SDA} = 0.024$–1.892). Specifically, CASCAM identified SUM44PE ($DS_{SDA} = 0.024$ and $P_{SDA} = 98.7\%$) and DU4475 ($DS_{SDA} = 0.188$ and $P_{SDA} = 99.2\%$) as the two most genome-wide representative cell lines for ILC. UACC3133 was ranked the third with small deviance $DS_{SDA} = 0.452$ but had a wide 95% confidence interval [0.067, 3.040]. The congruent finding of SUM44PE is consistent with literature, as it has been reported to have anchorage-independence and limited migration and invasion ability, which are unique properties to the ILC-like cell lines [113] and are widely studied in ILC [92, 77]. To avoid the ambiguous assignment probabilities, we further restricted the selection criteria to $P_{SDA} > 0.8$ (enclosed by green dashed rectangle) and $pval(DS_{SDA}) > 0.1$ (enclosed by orange dashed rectangle). ZR7530 ($P_{SDA} = 0.764$), MDAMB453 ($P_{SDA} = 0.670$), SKBR3 ($DS_{SDA} = 2.615$, p-value = 0.052) and AU565 ($DS_{SDA} = 2.405$, p-value = 0.062) were filtered out, and the 9 cell lines that met the criteria were used for further investigation. We note that MDA-MB-134VI ($P_{SDA} = 0.548$) was manually included for further evaluation as it is widely used in ILC research [105, 108]. In Figure 16B, the 9 unbiased-selected and 1 manually-included cell lines were ranked by DS_{SDA} with 95% confidence interval provided.

4.2.1.3 Pathway and mechanistic-based selection of cancer model(s)

Next, we applied Module 3 with pathway-specific and gene-specific evaluation for further prioritization of the 10 pre-selected breast cancer cell lines. Using a similar definition of DS_{SDA}, we calculated gene- and pathway-specific deviance scores, DS_{gene} and DS_{path}, for characterizing congruence of each candidate cell line. Differential expression analysis on 769 IDC versus 191 ILC samples in TCGA identified 3,065 DE genes. For pathway investigation, 236 pathways in Hallmark and KEGG from MSigDB [111] were first identified, and 53 pathways with more than 20 DE genes were used for GSEA pathway analysis [53].

Supplement Figure 22 shows pathway-specific deviance scores (DS_{path}) for the 53 selected pathways (rows) and 10 selected cell lines (columns) in the heatmap. The side-bar on the top shows genome-wide congruence DS_{SDA} for each cell line and the side-bar on the left margin shows size and normalized enrichment score (NES) for each pathway. In general, genome-wide resemblance does not guarantee similar performance in specific pathways. SUM44PE, for example, was the most congruent ILC cell line with the smallest DS_{SDA} . However, it was second to worst congruent cell line in Hallmark heme metabolism, where heme is an iron-containing porphyrin with multifaceted roles in cancer ($DS_{Path} = 1.029$). When users have prior knowledge of known relevant pathways, the most congruent cell line can be selected by the smallest averaged DS_{path} of the pre-selected pathways. If no prior biological knowledge is used, we recommend using pathways with adequate pathway size (e.g., $30 < size < 200$) and enrichment (e.g., $|NES| > 1.5$) for final cell line decision. This criterion selected 14 pathways and the heatmap of their pathway-specific deviance score was shown in Figure 17A (detailed values available in Supplement Table 14). Among the 14 pathways, the majority of pathways were cancer related (marked star in Figure 17A). For example, Hallmark E2F Targets (Supplement Figure 23) has the most significant NES (NES = -2.18, adjusted p-value < 0.0001, 79 DE genes), which includes genes encoding cell cycle related targets of E2F transcription factors. Related to the loss of E-cadherin, E2F was reported to show difference in ILC compared with IDC [92, 34]. KEGG PPAR Signaling Pathway, including genes related to peroxisome proliferator-activated receptors (PPARs) signaling, is significantly enriched (NES = 1.55, adjusted p-value = 0.048, 22 DE genes, Supplement Figure 23), and is also

widely reported for its upregulation in ILC tumors in multi-omics studies [112, 106].

Given that loss of $CDH1$ [22] and subsequent dysfunction of cell-cell adhesion [128] is the hallmark of ILC we manually included "KEGG Cell Adhesion Molecules" pathway for analyses shown in Figure 17 and Supplement Table 14, in addition to the 14 unbiased selected pathways. The pathway was not selected because its $|NES| = 0.854$ did not meet the prespecified criterion. The second to the last row in Figure 17A shows average DS_{path} of the 14 pathways for each cell line, in which CAMA1 had the smallest average deviation. CAMA1 was also congruent to ILC in the "KEGG Cell Adhesion Molecules" pathway (22 DE genes, $DS_{path} = 0.468$, p-value = 0.634). Although SUM44PE, DU4475 and UACC3133 outperformed CAMA1 in the genome-wide SDA-based deviance score (Figure 16), each of them did not mimic well in at least part of the 14 pathways (one circle: $p < 0.1$; two concentric circles: $p < 0.05$; three concentric circles: $p < 0.01$, showing non-congruence) while CAMA1 had uniformly high congruence. For example, DU4475 did not mimic ILC in several important cancer and ILC-related pathways, such as "Hallmark TNFA Signaling Via NFKB", "Hallmark glycolysis", "Hallmark MTORC1 Signaling", etc.

For a pathway of interest, CASCAM further generated a gene-specific deviance score (DS_{gene}) heatmap. Figure 17B shows DS_{gene} heatmap of 22 DE genes in the "KEGG Cell Adhesion Molecules" pathway, giving gene-level resolution of congruence information. BCK4 appeared to be the least congruent cell line in this pathway with 10 genes having large deviance scores ($|DS_{gene}| > 2$; Figure 17B). Next, we utilized KEGG topological regulatory network information [69] to investigate gene-specific congruence to ILC in KEGG Cell Adhesion Molecules for selected cell lines. The well-known ILC hallmark gene $CDH1$ only showed congruence in CAMA1 and DU4475 ($CDH1$ highlighted in Figure 17B). Although the MDA-MB-134VI cell line has been widely used in ILC research, Figure 17A and Supplement Table 13 (P_{SDA} to ILC = 0.548) show that it has similar congruence to both ILC and IDC. Furthermore, although MDA-MB-134VI was congruent to ILC in many of the 14 pathways, it was not congruent in the "KEGG Cell Adhesion Molecules" pathway and many ILC-relevant genes in this pathway (Figure 17A and B). Among the 10 cell lines, the BCK4 cell line ($DS_{path} = 1.323$, p-value = 0.005) had the largest DS_{path}, indicating worst genome-wide congruence (Figure 17B and Supplement Figure 24). Figure 17C shows part of KEGG

PathView plot of BCK4 in the Cell Adhesion Molecules pathway. BCK4 had many discordant genes to ILC: 2 genes highly up-regulated to ILC ($DS_{gene} > 2$; $CLDN11$ and $NRCAM$) and 8 genes highly down-regulated to ILC ($DS_{gene} < -2$; $CADM1$, $CDH1$, PVR, $L1CAM$, $CLDN1$, $CLDN16$, $CDH4$, and $JAM2$). Of these genes, cadherin genes ($CDH1$, $CDH4$ and $CDH15$) were cell adhesion molecules that are critical in the formation of adhesion junctions for cells to adhere to each other [43]. Similarly, claudin genes ($CLDN1$, $CLDN11$ and $CLDN16$) are proteins essential for the formation of tight junctions in epithelial and endothelial cells [119].

As shown in this ILC representative cell line selection example, CASCAM provided multiple visualization tools and interactive software functions, including violin plot (Supplement Figure 24), pathway-specific congruence heatmap (DS_{path}; Figure 17A), gene-specific congruence heatmap (DS_{gene}; Figure 17B), and KEGG topological network plot (Figure 17C), to allow researchers to iteratively investigate concordance and discordance of cell lines with the target tumor cohort. In conclusion, 5 of the 10 cell lines are determined as ILC-like in the "KEGG Cell Adhesion" pathway ($pval(DS_{path}) > 0.05$), and we recommend them as appropriate ILC cell lines in the order of average DS_{path} of the 14 selected pathways: CAMA1 ($DS_{path} = 0.505$), UACC3133 ($DS_{path} = 0.667$), SUM44PE ($DS_{path} = 0.689$), HCC2218 ($DS_{path} = 0.748$), IPH926 ($DS_{path} = 0.754$).

4.2.2 Case study 2: selection of PDO and PDX for ILC

To extend the algorithm to PDO and PDX, we applied CASCAM to 11 PDO and 136 PDX breast cancer models from the PDMR [79] database to select congruent cancer models for ILC versus IDC. These 147 cancer models were first normalized by the data harmonization module with the 9,264 TCGA pan-cancer tumor samples, and 960 TCGA BC samples were used for further investigation after normalization. UMAP in Supplement Figure 25A showed three distinct clusters. Except for the basal and non-basal group observed before (Figure 14), there was a small third cluster with 15 samples (2 PDO models, 12 PDX models, and 1 TCGA sample) from four patients. All four samples were annotated with triple-negative IDC, and two patients from PDMR have a metaplastic squamous cell carcinoma diagnosis. Due to the

rare and unique subtype features of these tumors, we excluded these samples from further analysis and reproduced UMAP in Supplement Figure 25B. Similarly, we also excluded the samples in basal cluster, and 4 PDO and 25 PDX models were then kept for downstream analysis. The normalization result demonstrated excellent performance of Celligner for PDO and PDX.

We next applied the criteria ($P_{SDA} > 0.5$ and $pval(DS_{SDA}) > 0.05$) in the "interpretable machine learning pre-selection" module and identified four candidate cancer models (3 PDX and 1 PDO) to represent ILC tumors (Supplement Table 15). Cross-referencing with the PDMR database revealed that all four cancer models originated from the same patient (PRMR ID:171881-019-R). Table 8 showed 5 PDX and 1 PDO (denoted as PDO.1) originate from this patient. The 5 PDX samples contained one sample with passage 0 (denoted as PDX.0), two samples with passage 1 (denoted as PDX.1A and PDX.1B), and two samples with passage 2 (denoted as PDX.2A and PDX.2B). Intriguingly, the three highly congruent ILC PDX models were of passage 0 and 1 (PDX.0, PDX.1A and PDX.1B) while two PDX models with passage 2 (PDX.2A and PDX.2B) were not selected. Figure 18A showed a clear pattern that PDX.0 has almost perfect DS_{SDA} congruence to represent ILC but the deviance score increased with increasing passage numbers (also see Table 15 Column 5), indicating that the xenografts may evolve and be affected by the microenvironments in mice and deviate from the original tumor over time. When we investigated information of this patient, the specific histological subtype was not annotated but insertion frameshift mutation (p.T115Nfs*53) for $CDH1$ was detected. Since loss of $CDH1$ is a key determinant of ILC, it suggests that the original cancer is likely ILC, and cancer models derived from this patient's tumor are representative of the ILC tumor cohort.

We then applied the default pathway selection criteria, adequate size ($30 < size < 200$) and $|NES| > 1.5$ and selected 14 pathways as in the first cell line case study. As rationalized before, we again manually included the KEGG Cell Adhesion Molecules in addition to the 14 pathways. The pathway-specific congruence heatmap revealed performance of the six cancer models originating from the same patient (171881-019-R) (Figure 18B). Variations in performances were seen in the models even though they were developed from the same patient. Of the four pre-selected cancer models, PDO.1 had the largest pathway deviance

score ($DS_{path} = 0.455$) while PDX.1B has the smallest ($DS_{path} = 0.294$) in "KEGG Cell Adhesion Molecules" pathway (Figure 18B) although none of them was statistically significant in lack of congruence. We next investigated KEGG topological network plot (Supplement Figure 27) for the "KEGG Cell Adhesion Molecules" pathway comparing PDO.1 and PDX.1B. Supplement Figure 27B showed that expression of $CADM1$, $CADM3$, $CDH2$ was down-regulated ($DS_{gene} < -1.5$) while expression of three other genes ($CDH15$, $NRXN2$, $L1CAM$) was up-regulated ($DS_{gene} > 1.5$) in PDO.1 compared with average expression of ILC tumor, while we observed better congruence in PDX.1B (Supplement Figure 27A). Violin plot (Figure 18C) further elucidated the comparison between PDX.1B and PDO with the gene expression distribution in IDC and ILC tumors as reference.

4.3 Discussion

Cancer models play a crucial role in cancer research for understanding carcinogenesis and drug development. However, how to best select the most congruent cancer model to faithfully represent a specific tumor subtype remains mostly unsolved, which is an urgent gap to fill given the increasing number of cell lines and PDOs being generated. In contrast to pure machine-learning-based methods in the literature, we developed a pipeline, CASCAM, to progressively select the most representative cancer model(s) by genome-wide pre-selection and pathway-specific mechanistic investigation using transcriptomics data. First, tumor and cancer model data are harmonized by Celligner (Module 1). The congruence evaluation combines merits of both machine learning and correlation-based approaches to pre-select cancer models (Module 2). In-depth bioinformatic tools provide iterative exploration of the most and least mimicked biological mechanisms of selected cancer models (Module 3).

The first example of this framework used ILC breast cancer data to select the most representative cell line(s), and it is demonstrated that CASCAM is suitable either in a supervised manner with prior knowledge of disease mechanism (e.g., cell adhesion pathway in ILC) or drug targeted pathways, or in an unsupervised manner when no prior knowledge is given. 14 cell lines were credentialed as ILC cell lines on the genome-wide evaluation by Module

2, and 10 of them (including user-specified MDA-MB-134VI) were used for pathway-specific analysis in Module 3. Though widely used in ILC research [105, 108], MDA-MB-134VI was not congruent with ILC tumors on the genome wide and in the "KEGG cell adhesion molecules" pathway. All results combined together indicated that CAMA1, UACC3133, SUM44PE, HCC2218, and IPH926 were recommended in order as appropriate cell lines for ILC research.

DU4475 is an example of the necessity of pathway-specific analysis (Module 3). As this cell line is E-cadherin positive, estrogen receptor positive [121], and without $CDH1$ mutation detected [40], DU4475 does not exhibit features of the classic ILC subtype. However, as epithelial-mesenchymal transition (EMT) preferentially occurs in basal cell lines [101], it often accounts for reduced $CDH1$ and $CDH2$ expression, which is also the key features of ILC tumors. Therefore, DU4475 was genome-wide classified as ILC. Importantly, pathway-specific analysis provided higher resolution to differentiate IDC and ILC, with DU4475 being dissimilar to ILC on average of the 14 selected pathways ($DS_{path} = 0.734$, p-value = 0.093) and in the "KEGG cell adhesion molecules" pathway ($DS_{path} = 0.917$, p-value = 0.026) and finally was not selected as a representative ILC cell line by CASCAM.

In practice, researchers tend to credential cell lines according to the annotation of their origin or pre-specified mutations (e.g., $CDH1$ for ILC) if available. However, the origins might be mislabeled, and the cell line evolution in culture has uncovered the possibility of genetic diversification, weakening the credibility of the original annotation. On the other hand, selected mutations cannot guarantee eligibility for a cell line. In our study; for example, we observed large genomic differences between SUM44PE ($P_{SDA} = 0.987$, $DS_{SDA} = 0.024$) and 600MPE ($P_{SDA} = 0.125$, $DS_{SDA} = 2.013$) although both have $CDH1$ mutation and are ER+, which are essential features of ILC. Therefore, the proposed CASCAM captures systems information in pathways, topological networks and genes and provides a thorough congruent investigation of the cell lines.

We also extended our framework to examine congruence of PDO and PDX cancer models to ILC tumors in the second case study. Of 11 PDOs and 136 PDXs in the PDMR database, only four from the same patient were credentialed as ILC in Module 2 evaluation. Strikingly, this tumor and model while not annotated as ILC has a $CDH1$ mutation, suggesting that

CASCAM authenticated a new model of ILC. Aside from offering a "yes" or "no" answer, CASCAM can score the cancer models according to how similar they are to the targeted tumor cohort. We therefore observed a progressive deviation trend for PDX samples over passages, which is consistent with recent reports that PDX often undergo murine-specific tumor evolution and congruence decays over passages [7, 104]. In fact, due to discrepancies in drug response for late-passage PDXs, recent studies have suggested design to use early-passage PDX models [80]. In addition, we found that PDO is not guaranteed to be better than PDX, although PDO is widely believed to be a highly conserved cancer model promising for precision medicine development and superior to PDX [86, 71]. Its discordance in the six coding genes related to cell adhesion gives a cautious sign of using it to represent ILC.

The current CASCAM has limitations and multiple directions of development are on-going. The methodologies are now developed for transcriptomic data evaluation. As multi-level omics data (e.g., mutation, copy number variation, methylation and miRNA expression) are becoming affordable and prevalent, an extended congruence framework for evaluating cancer models with multi-omics data will provide deeper insight. Secondly, congruence analysis using single cell RNA-seq or single cell multi-omics data will provide a high-resolution understanding of clonal and micro-environment information for selecting the most representative cancer model, which is also an on-going work. Thirdly, the current framework is built upon binary contrast (i.e., ILC versus IDC). An extension to evaluating multi-class (i.e., three or more tumor subtypes) scenario is also a future direction. Currently, the molecular congruence we focus is on transcriptomic resemblance in genome-wide, pathway or gene level. The method can be extended to incorporate additional information, such as drug response, when available. Finally, the goal of CASCAM is to identify the most congruent cancer model from a long list of candidates to represent a target tumor cohort. For precision medicine, one may be interested in quantifying congruence of a PDO compared to the tumor from the patient origin. CASCAM can be easily extended for that purpose.

Collectively, we demonstrated CASCAM as a comprehensive and effective congruence evaluation tool for selecting the most representative cancer model for investigating cancer pathways and ultimately for precision medicine. CASCAM provides harmonization between human tumor and cancer model omics data, interpretable machine learning for congru-

ence quantification, mechanistic investigation, and pathway-based topological visualization to determine the most appropriate cancer model selection. The workflow is presented using invasive lobular breast carcinoma (ILC) subtype, credentialing highly relevant models, and suggesting CAMA1 followed by UACC3133 as the most representative cell lines for ILC research. Our novel method is generalizable to any cancer subtype and will be impactful for furthering research in precision medicine. An R package, CASCAM, with an interactive app is publicly available (https://github.com/jianzou75/CASCAM.) to facilitate the use of our proposed framework.

4.4 Method

4.4.1 Gene expression data

Gene expression matrices in raw read count and transcripts per million (TPM) versions for 9,264 The Cancer Genome Atlas (TCGA) pan-cancer tumor samples were downloaded from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo) with query ID GSM1536837 [90], and there were 960 breast cancer primary tumor samples with histology annotated as IDC or ILC. Log2-TPM gene expression data for 1,248 Cancer Cell Line Encyclopedia (CCLE) pan-cancer cell line samples were taken from DepMap Public 19Q4 file [30] (https://depmap.org/portal/ccle), and there were 65 breast cancer cell lines. Due to the limited representation of ILC cell lines in the CCLE project, we further included seventeen cell lines from an ongoing project (R01CA252378), namely Invasive Lobular Cancer Cell Line Encyclopedia (ICLE). The following eight cell lines were overlapping in ICLE and CCLE datasets: CAMA1, HCC1187, HCC2218, MDA-MB-134, MDA-MB-453, MDA-MB-468, SKBR3, and ZR7530. Those from ICLE were annotated as I, those from CCLE (sequencing data from Sequence Read Archive (SRA) under accession number PRJNA523380) were annotated as C, and the processed CCLE data directly from DepMap were not annotated. Gene expression data of the breast cancer PDO and PDX models in TPM were obtained from NCI Patient-Derived Models Repository (PDMR) database

(https://pdmr.cancer.gov/), and was log transformed for downstream evaluation. The genetic variants (e.g. mutations) in PDMR was extracted from whole genome sequence and annotated through oncoKB annotation pipeline version 1.1.0 [18].

4.4.2 Gene expression normalization between tumor and cell lines

The gene expression matrices from tumors and cell lines are not directly comparable. We evaluated three different approaches for normalization. Quantile normalization is a widely used method to achieve equal quantiles across all the samples ("normalize.quantiles" function in preprocessCore [12] package). ComBat [48] is method for batch effect correction under empirical Bayes frameworks, where we treated tumor and cell lines as two different batches ("ComBat" function in sva [61] package). Celligner is a two-step machine learning method specifically developed for tumor and cell line normalization. The first step is to remove systemic differences, such as normal cell contamination, between tumor and cell lines using contrastive principal component analysis (cPCA). The second step is to perform further normalization using mutual nearest neighbors (MNN) [44]. We used the default parameters in Celligner implementation (celligner package), using either 960 breast cancer tumor samples or all 9,264 pan-cancer samples to harmonize the datasets.

4.4.3 Differential expression analysis and gene set enrichment analysis

We applied DESeq2 [67] R package using TCGA tumor read count data for differential expression analysis (IDC vs. ILC). A gene with absolute fold change > 1.5 and two-sided Benjamini-Hochberg adjusted p-value [9] < 0.05 was defined as "differentially expressed (DE)". For gene set enrichment analysis, we used fgsea R package. Kyoto Encyclopedia of Genes and Genomes (KEGG) and Hallmark gene sets in the Molecular Signatures Database (MSigDB) were analyzed, and log2 fold changes from the differential expression analysis were used for gene ranking.

4.4.4 Machine learning methods

We compared 16 machine learning methods, including sparse discriminant analysis (SDA), random forest on pre-filtered transformed data* (CancerCellNet), robust sparse discriminant analysis (RSDA), logistic regression with elastic net (ElasticNet), logistic regression with ridge penalty* (RidgeRegress), K nearest neighbors (KNN), majority voting according to 25 highest Pearson correlated tumor samples* (Pearson25), linear discriminant analysis (LDA), random forest* (RandomForest), nearest template prediction* (NTP), subtype assignment according to the median of within subtype Spearman correlations* (SpearmanMed), subtype assignment according to the median of within subtype Pearson correlations* (PearsonMed), logistic regression (Logistic), and three convolutional neural networks which were originally optimized on pan-cancer datasets (1D-CNN, 2D-Vanilla-CNN, and 2D-Hybrid-CNN) [76]. Six of these methods (marked with asterisk in Table 7) have been extended and used in publications for cancer model prediction analysis. The following three prediction evaluations were performed: (1) Five-fold cross-validation on breast cancer histology (769 IDC vs. 191 ILC) using 960 TCGA BC samples. (2) Construction of prediction model on Celligner aligned TCGA BC samples (training set, 712 ER+ and 205 ER-) and validated on Celligner aligned CCLE BC cell line samples (testing set, 19 ER+ and 37 ER- cell lines). (3) Construction of prediction model on Celligner normalized TCGA pan-cancer samples (training set, 960 BC and 960 non-BC) and validated on Celligner normalized CCLE cell line samples (testing set, 56 BC and 56 non-BC cell lines). To avoid accuracy calculation issue of imbalanced sample sizes, 960 TCGA non-BC tumor samples and 56 CCLE non-BC cell lines were randomly subsampled from 8,304 TCGA pan-cancer non-BC samples and 1,192 CCLE pan-cancer non-BC cell lines.

4.4.5 SDA projected deviance score

The SDA projected deviance score, DS_{SDA}, was designed based on the sparse discriminant analysis (SDA) method [23] to quantify genome-wide dissimilarity between a cancer model and the targeted tumor subtype. We denote the tumor gene expression $N \times G$ matrix as X with N samples and G DE genes, the $N \times 2$ class indicator matrix as Y with

$Y_{ik} = 1_{(i \in C_k)}$ for tumor sample i belonging to targeted tumor subtype, and the cancer model gene expression as C. SDA extends linear discriminant analysis with elastic net to identify ($\boldsymbol{\theta}$, $\boldsymbol{\beta}$) by

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\beta},\boldsymbol{\theta}} \left\{ \|\boldsymbol{Y\theta} - \boldsymbol{X\beta}\|^2 + \gamma \boldsymbol{\beta}^T \boldsymbol{I\beta} + \lambda \|\boldsymbol{\beta}\|_1 \right\} \\ &\text{subject to } \frac{1}{n} \boldsymbol{\theta}^T \boldsymbol{Y}^T \boldsymbol{Y\theta} = 1 \end{aligned}$$

where $\boldsymbol{\theta}$ is the optimal scores, $\boldsymbol{\beta}$ is the sparse discriminant vector, $\boldsymbol{I}$ is the identity matrix, γ and λ are nonnegative tuning parameters selected by cross-validation. An iterative algorithm is applied to solve the pair ($\boldsymbol{\theta}$, $\boldsymbol{\beta}$). The tumor and cancer model gene expressions are then projected to the direction of estimated $\boldsymbol{\beta}$–$\boldsymbol{X\beta}$ and $\boldsymbol{C\beta}$. The assignment probability, P_{SDA}, was calculated from the standard LDA on the reduced data matrix $\boldsymbol{X\beta}$ and $\boldsymbol{C\beta}$ in selected gene features.

To simplify annotation, we use c_i to denote the projected value for cancer model i and t_k to denote the projected tumor sample vector for subtype k. The SDA projected deviance score for cancer model i in class k, is defined as $DS_{SDA}^{(i,k)} = |c_i - \hat{\mu}_k|/\hat{s}$ where $\hat{\mu}_k = median_k(\boldsymbol{t_k})$, $\hat{s} = mad_k(\boldsymbol{t_k} - \hat{\mu}_k)$, and mad is abbreviation for scaled median absolute deviation. Intuitively, $\hat{\mu}_k$ and $\hat{s}$ are robust forms of mean and standard deviation, and DS_{SDA} can be seen as a robust form of absolute t-statistics as the standardized distance of the cancer model to the center of tumor cohort on the SDA projected space. Smaller deviance score indicates higher congruence of the cancer model to the desired tumor subtype cohort. By setting the null hypothesis as $c_i = \mu_k$, the p-value of $DS_{SDA}^{(i,k)}$, denoted as $pval(DS_{SDA}^{(i,k)})$, is obtained from the distribution of tumor $\boldsymbol{t_k} \sim N(\mu_k, \sigma)$, where (μ_k, σ) are estimeated by $(\hat{\mu_k}, \hat{\sigma})$. The ordinary bootstrap [36] with 1,000 times on the tumor projected data is performed to obtain the 95% confidence interval of $DS_{SDA}^{(i,k)}$ on the log2 scale. The implementation of this method is based on sparseLDA [23], caret [57] and boot [17] R package.

4.4.6 Gene and pathway specific deviance score

We denoted the Celligner aligned gene expression for cancer model i and gene g as $c_{g,i}$ and for tumor samples in subtype k and gene g as $t_{g,k}$.Similar to SDA-projected deviance score, we defined the gene specific deviance score (DS_{Gene}) for model i and subtype k in gene g as $DS_{Gene}^{(g,i,k)} = (c_{g,i} - \hat{mu}_{g,k}/\hat{\sigma}_g$, where $\hat{\mu}_{g,k} = median_k(\boldsymbol{t_{g,k}})$ and $\hat{\sigma}_g = mad_k(\boldsymbol{t_{g,k}})$. The pathway specific deviance score (DS_{path}) for pathway p, cancer model i, and tumor subtype k is then defined, based on the DS_{Gene}, as $DS_{path}^{(p,i,k)} = \text{geometric mean}_{g \in P^{(DE)}}(|DS_{gene}^{(g,i,k)}|)$, where $P^{(DE)}$ is the set of DE genes in pathway p. The geometric mean is proposed to reduce the effects of outliers. The significance levels of DS_{path} (one-sided p-values) were defined similar to $pval(DS_{SDA})$, which were obtained from the null distribution empirically constructed by DS_{path} of the tumor samples.

Figure 13: Flowchart of CASCAM for congruence quantification and selection. Tumor and cancer model gene expression data are first harmonized (Module 1). Interpretable machine learning by sparse discriminant analysis (SDA) is applied by combining predication accuracy and SDA-based deviance score for pre-selecting candidate cancer models (Module 2). Pathway-specific mechanistic explorations are iteratively investigated to conclude the final representative cancer model (Module 3). Blue frames represent input data, orange frames for essential output results, parallelogram frames for intermediate results, rectangular frames for analysis process, bullet-shaped frames for visualization, and rhombus frames for decision making.

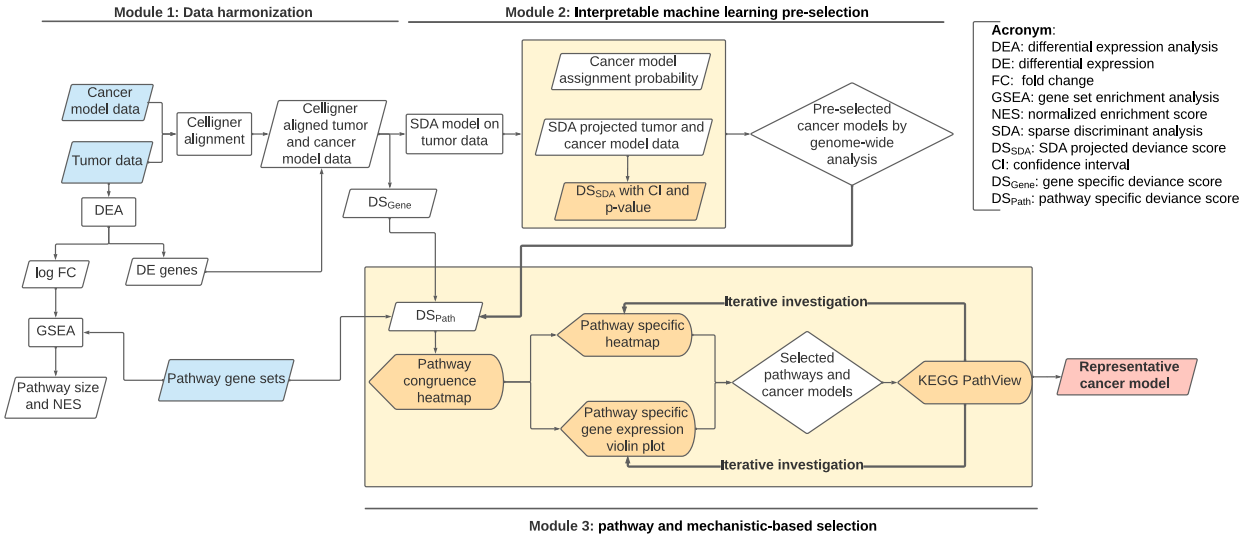


Figure 14: UMAP for comparison of multiple data harmonization approaches. UMAP for normalized BC tumors (n=960) and BC cell lines (n=65) to compare five normalization approaches: (A) no correction and (B) quantile normalization (C) ComBat and (D) Celligner utilizing BC tumors and BC cell lines (E) Celligner utilizing pan-cancer tumors and pan-cancer cell lines. The final approach best eliminates batch effects by mixing well the BC tumors and BC cell lines.

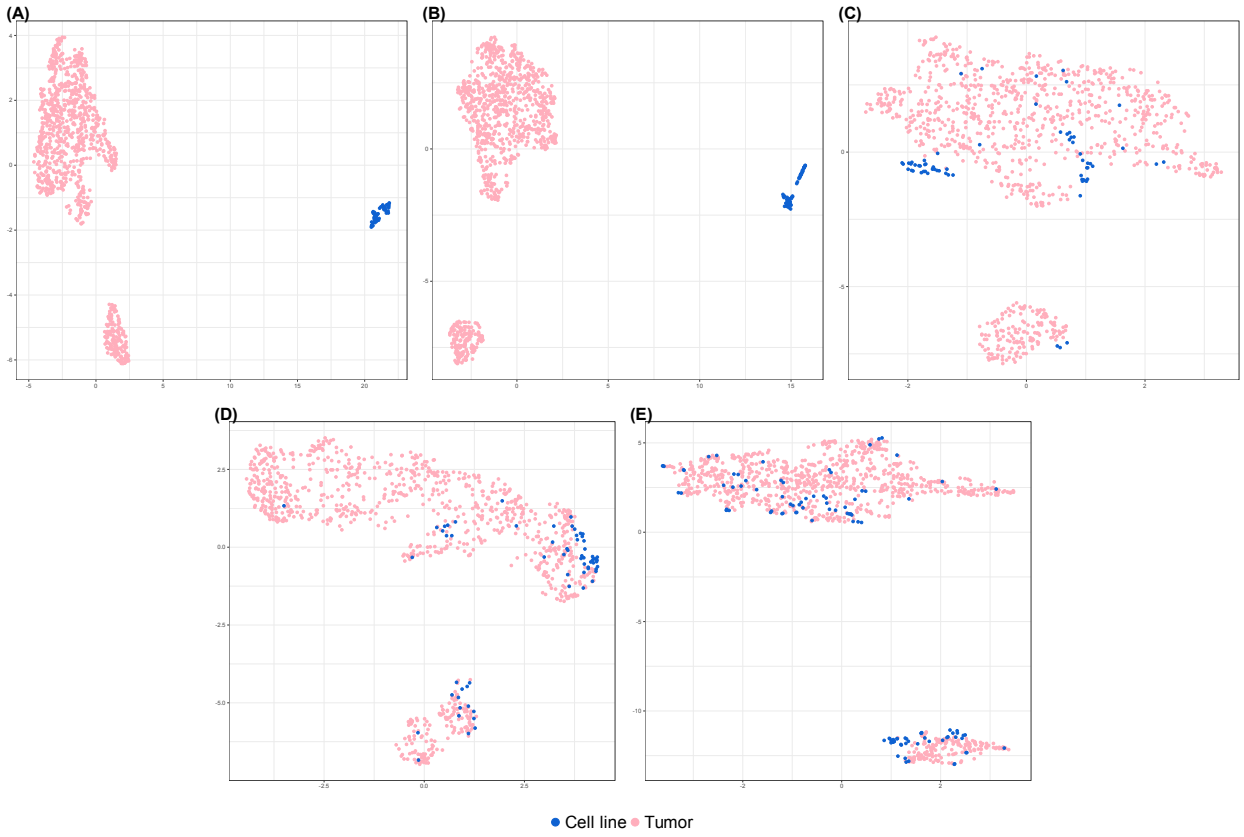


Figure 15: UMAP after data harmonization with replicates and basal subtype information. (A) Three replicates (cell line; cell line_C; cell line_I) for each of the eight cell lines are highly reproducible. (B) The lower-right cluster contains dominantly tumors and cell lines annotated as basal-like (118/160 tumors and 26/28 cell lines).

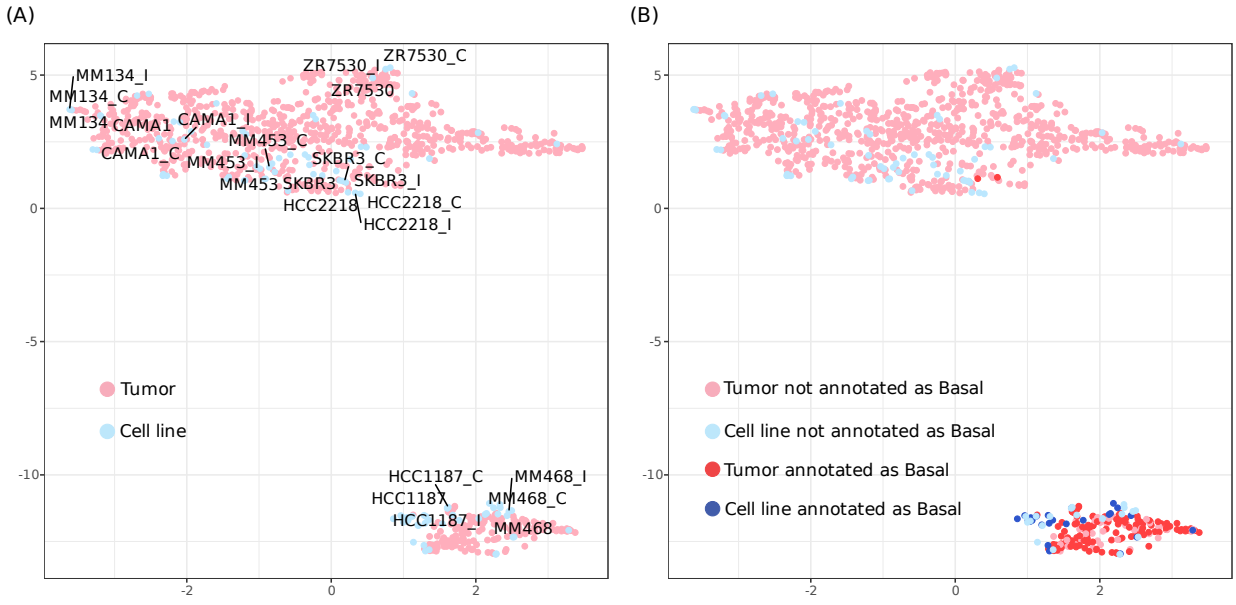


Figure 16: Genome-wide cell line congruence and pre-selection. (A) SDA projected scatter plot. y-axis represents the projected values for 38 cell lines, the red and blue horizontal lines represent the median projected value (center) of ILC and IDC tumor samples respectively. The density plots on the right shows distributions of 769 IDC (blue) and 191 ILC (red) tumors. Red color of the dots represents SDA classification to ILC (threshold $P_{SDA} > 50\%$), and the solid dots represent small SDA-based deviance scores (threshold $pval(DS_{SDA}) > 0.05$). More stringent criteria were indicated by the dashed rectangle. Cell lines with $P_{SDA} > 0.8$ were enclosed by green dashed rectangle and the ones with $pval(DS_{SDA}) > 0.1$ were enclosed by orange dashed rectangle. (B) SDA projected deviance score (absolute value) with 95% confidence interval. 9 unbiased-selected and 1 manually-included (marked with asterisk) cell lines are ranked based on $|DS_{SDA}|$, and 95% confidence intervals are obtained by bootstrap analysis on log-scale.

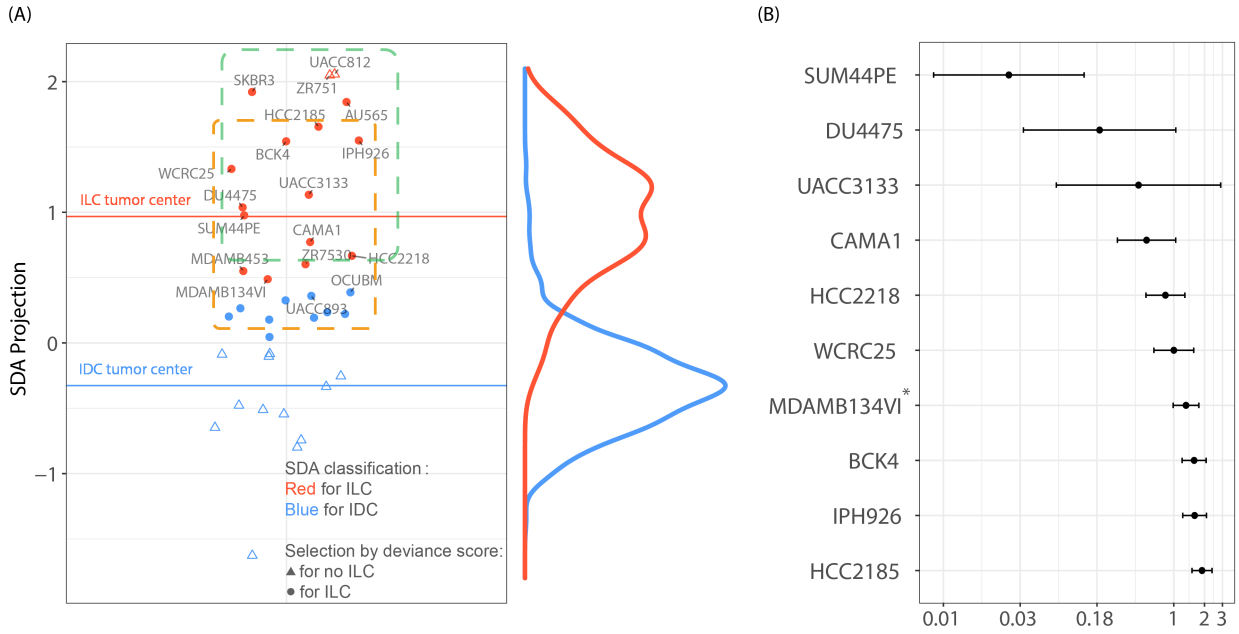


Figure 17: Pathway- and gene-specific analysis for selection of representative cell line(s). (A) Heatmap of pathway-specific deviance scores (DS_{path}) with 14 unbiased-selected and 1 manually-included pathways ($30 < size < 200$, $|NES| > 1.5$; shown on the rows) and 9 unbiased-selected and 1 manually-included cell lines (columns). The genome-wide SDA projected deviance score (DS_{SDA}) is shown on the top side-bar and the pathway size and normalized enrichment score (NES) are on the left. Positive (negative) NES indicates up-regulation (down-regulation) in ILC compared to IDC. Average of the 14 pathways and the pre-selected "KEGG Cell Adhesion Molecules" pathway are shown at the bottom. The p-values of DS_{path} are annotated in the heatmap (one circle: $p - value < 0.1$; two concentric circles: $p - value < 0.05$; three concentric circles: $p - value < 0.01$), and smaller p-values indicate worse congruence. (B) Gene-specific heatmap shows DS_{gene} for the 10 selected cell lines and 22 DE genes in "KEGG Cell Adhesion Molecules" pathway. (C) Part of KEGG PathView topological networks for BCK4 ($DS_{path} = 1.323$) for the "KEGG Cell Adhesion Molecules" pathway. The result shows discordance of 10 genes in BCK4 (orange stars showing up-regulation compared to ILC tumors and blue start showing down-regulation).

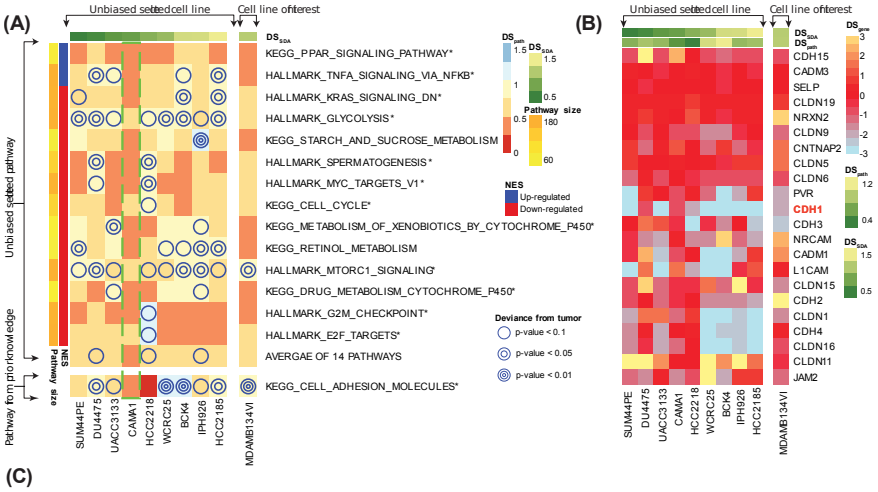


Figure 18: Selecting representative PDO/PDX for ILC. (A) SDA projected positions for PDO and PDX models from PDMR. Four models (three PDXs and one PDO; red circles) from the same patient (171881-019-R) were identified as candidate ILC models. Six models from this patient are labeled with the sample ID. High consistency was observed between SDA deviance scores and passages among PDX models. (B) Six models originated from the same patient were used for pathway-specific analysis. Six models show high congruence in the majority of 14 pathways and the Cell Adhesion pathway. (C) Violin plot shows the position of PDO.1 and PDX.1B on the six genes on which PDO.1 is discordant with.

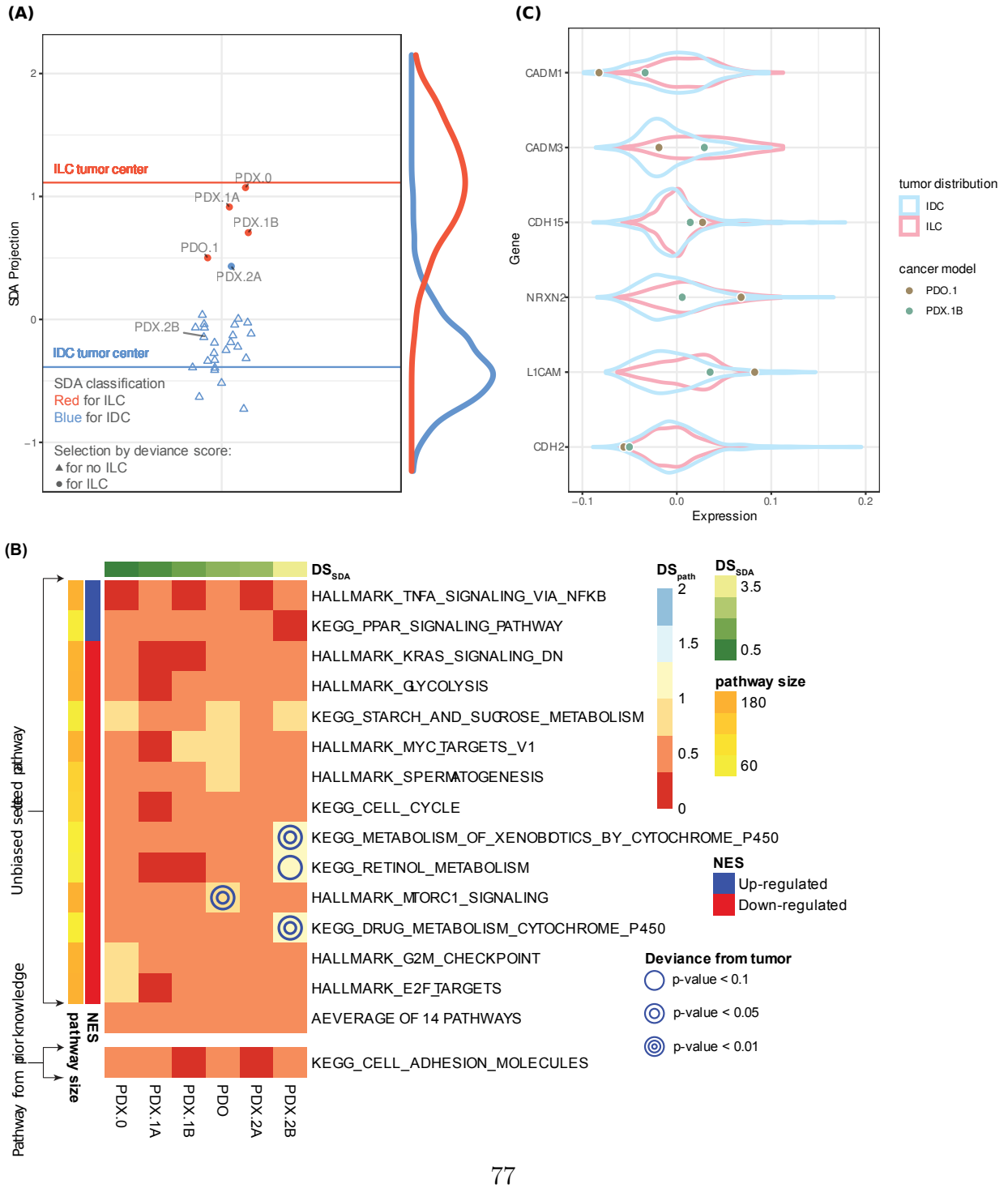


Table 7: Evaluation and properties of 13 popular machine learning methods. Six methods applied for cancer model prediction in previous papers are highlighted (*). Prediction accuracies are shown in three machine learning evaluation examples. Parentheses in the second column are standard deviations of accuracies in five repeats of five-fold cross-validation.

	Machine learning evaluation			**Machine learning relevant properties**		
	ILC vs IDC	ER+ vs ER-	BRCA vs other cancers	Gene selection	Assignment probability	Deviance score
	TCGA; 5-fold CV	Training data: TCGA; Test data: CCLE	Training data: TCGA; Test data: CCLE			
SDA	0.91 (0.02)	0.91	0.86	Yes	Yes	Yes
ElasticNet	0.90 (0.03)	0.93	0.85	Yes	Yes	No
2D-Hybrid-CNN	0.87 (0.03)	0.93	0.86	No	No	No
RidgeRegress*	0.88 (0.02)	0.91	0.84	Yes	Yes	No
Pearson25*	0.86 (0.01)	0.86	0.9	No	No	No
KNN	0.85 (0.03)	0.86	0.91	No	Yes	No
2D-Vanilla-CNN	0.86 (0.04)	0.88	0.85	No	No	No
1D-CNN	0.86 (0.03)	0.86	0.86	No	No	No
RandomForest*	0.85 (0.01)	0.91	0.82	Yes	Yes	No
RSLDA	0.81 (0.11)	0.77	0.86	Yes	Yes	Yes
CancerCellNet*	0.79 (0.03)	0.82	0.79	Yes	Yes	No
LDA	0.80 (0.03)	0.68	0.82	No	Yes	Yes
NTP	0.61 (0.03)	0.86	0.82	No	No	Yes
SpearmanMed*	0.40 (0.03)	0.84	0.61	No	No	Yes
PearsonMed*	0.38 (0.04)	0.84	0.62	No	No	Yes
Logistic	0.52 (0.04)	0.43	0.65	No	Yes	No

Table 8: SDA-based genome-wide congruence summary for six models from patient 171881-019-R. Later passages of PDX models have worse congruence (i.e., larger deviance scores).

Sample ID	**Label name**	**Model type**	**Passage**	$\boldsymbol{DS}_{SDA}^{(ILC)}$	$\boldsymbol{P}_{SDA}^{(ILC)}$	**Identified as ILC**
APW-DS2	PDX.0	PDX	0	0.12	1.00	Yes
APYF68	PDX.1A	PDX	1	0.56	0.99	Yes
APWG05	PDX.1B	PDX	1	1.15	0.90	Yes
APWG05PF7	PDX.2A	PDX	2	1.93	0.34	No
APVG40_RG-G15	PDX.2B	PDX	2	3.56	0.00	No
V1-organoid	PDO	PDO		1.73	0.51	Yes

Appendix A Chapter 3

A.1 Supplement tables and figures

Table 9: Simulation settings for the toy example. Gene 1: the gene has concordant expression across all 4 studies; Gene 2: the gene has concordant expression in study 1, 2, and 3; Gene 3: The gene has concordant expression between study 1 and 2 and between 3 and 4; Gene 4: gene without any concordant signals.

	Study 1	Study 2	Study 3	Study 4
	$(n_{11}, n_{12}, n_{13}) =$ $(20, 20, 20)$	$(n_{21}, n_{22}, n_{23}) =$ $(20, 20, 20)$	$(n_{31}, n_{32}, n_{33}) =$ $(20, 20, 20)$	$(n_{41}, n_{42}, n_{43}) =$ $(20, 20, 20)$
	$(\mu_{11}, \mu_{12}, \mu_{13}), \sigma_1$	$(\mu_{21}, \mu_{22}, \mu_{23}), \sigma_2$	$(\mu_{31}, \mu_{32}, \mu_{33}), \sigma_3$	$(\mu_{41}, \mu_{42}, \mu_{43}), \sigma_4$
Gene 1	$(1, 3, 5), 1$	$(1, 3, 5), 1$	$(1, 3, 5), 1$	$(1, 3, 5), 1$
Gene 2	$(5, 3, 1), 1$	$(5, 3, 1), 1$	$(5, 3, 1), 1$	$(1, 7, 1), 1$
Gene 3	$(1, 3, 5), 1$	$(1, 3, 5), 1$	$(1, 7, 1), 1$	$(1, 7, 1), 1$
Gene 4	$(0, 0, 0), 1$	$(0, 0, 0), 1$	$(0, 0, 0), 1$	$(0, 0, 0), 1$

Table 10: Simulation settings for different effect sizes. Category I: genes with concordant expression patterns across three studies; Category II: genes with discordant expression patterns across three studies; Category III: genes with concordant expression patterns between study 1 and 2 only; Category Null: genes without any signals.

Effect size		Study 1	Study 2	Study 3
		$(n_{11}, n_{12}, n_{13}) =$ $(10, 5, 8)$	$(n_{21}, n_{22}, n_{23}) =$ $(5, 8, 10)$	$(n_{31}, n_{32}, n_{33}) =$ $(8, 10, 5)$
		$(\mu_{11}, \mu_{12}, \mu_{13}), \sigma_1$	$(\mu_{21}, \mu_{22}, \mu_{23}), \sigma_2$	$(\mu_{31}, \mu_{32}, \mu_{33}), \sigma_3$
0.5	I (n = 300)	$(1, 3, 5), 3.5$	$(2, 4, 6), 3.1$	$(1, 4, 7), 4.4$
	II (n = 100)	$(1, 3, 5), 3.5$	$(6, 4, 2), 3.1$	$(1, 7, 1), 5.9$
	III (n = 100)	$(1, 3, 5), 3.5$	$(2, 4, 6), 3.1$	$(0, 0, 0), 4.4$
	Null (n = 1500)	$(0, 0, 0), 3.5$	$(0, 0, 0), 3.1$	$(0, 0, 0), 4.4$
0.6	I (n = 300)	$(1, 3, 5), 2.9$	$(2, 4, 6), 2.6$	$(1, 4, 7), 3.7$
	II (n = 100)	$(1, 3, 5), 2.9$	$(6, 4, 2), 2.6$	$(1, 7, 1), 4.8$
	III (n = 100)	$(1, 3, 5), 2.9$	$(2, 4, 6), 2.6$	$(0, 0, 0), 3.7$
	Null (n = 1500)	$(0, 0, 0), 2.9$	$(0, 0, 0), 2.6$	$(0, 0, 0), 3.7$
0.6	I (n = 300)	$(1, 3, 5), 2.5$	$(2, 4, 6), 2.2$	$(1, 4, 7), 3.2$
	II (n = 100)	$(1, 3, 5), 2.5$	$(6, 4, 2), 2.2$	$(1, 7, 1), 4.3$
	III (n = 100)	$(1, 3, 5), 2.5$	$(2, 4, 6), 2.2$	$(0, 0, 0), 3.2$
	Null (n = 1500)	$(0, 0, 0), 2.5$	$(0, 0, 0), 2.2$	$(0, 0, 0), 3.2$

Table 11: IPA canonical pathway analysis using the q-values from the MSCA analysis on the mouse metabolism data. Top 50 pathways (sorted by the p-value) are listed.

Pathway	**-log(p-value)**	**Ratio**
Mitochondrial Dysfunction	43.2	0.493
Sirtuin Signaling Pathway	42.9	0.522
Oxidative Phosphorylation	35.7	0.712
Estrogen Receptor Signaling	24.1	0.381
NRF2-mediated Oxidative Stress Response	18.2	0.414
Acute Phase Response Signaling	15.8	0.427
Integrin Signaling	15.5	0.406
Neutrophil Extracellular Trap Signaling Pathway	15.2	0.332
Granzyme A Signaling	14.5	0.573
Unfolded protein response	14.4	0.533
ILK Signaling	13.9	0.398
Huntington's Disease Signaling	13.2	0.353
CLEAR Signaling Pathway	13	0.351
FXR/RXR Activation	12.9	0.452
Protein Kinase A Signaling	12.7	0.316
Glucocorticoid Receptor Signaling	12.3	0.289
PPARα/RXRα Activation	12.2	0.385
Actin Cytoskeleton Signaling	11.8	0.357
LXR/RXR Activation	11.6	0.439
Protein Ubiquitination Pathway	11.3	0.341
Ferroptosis Signaling Pathway	11.3	0.424
Valine Degradation I	11	0.857
Epithelial Adherens Junction Signaling	10.8	0.395
Fatty Acid β-oxidation I	10.7	0.686
Insulin Receptor Signaling	10.6	0.407
Germ Cell-Sertoli Cell Junction Signaling	10.5	0.382
Necroptosis Signaling Pathway	10.4	0.391
Superpathway of Cholesterol Biosynthesis	10.1	0.515

Table 12: LISA results for top 30 ranked transcription factors. The ranking is obtained by combining Peak-RP method, H3K27ac, DNase-seq in silico deletion of TF ChIP-seq peaks.

Transcription Factor	p-value	Transcription Factor	p-value	Transcription Factor	p-value
SMC1A	3.25E-67	MYC	1.13E-22	TCF7L1	4.45E-20
DPF1	7.35E-64	TERC	1.22E-21	E2F1	4.55E-20
CTCF	3.90E-56	EGR3	2.05E-21	MAX	5.50E-20
ZMYM3	4.50E-54	ERG	4.75E-21	KDM5B	5.65E-20
NFIA	1.01E-51	SP1	5.87E-21	SP2	6.73E-20
ESR1	1.27E-48	SP140	8.85E-21	NRF1	8.23E-20
BATF3	6.89E-44	HIF1A	9.49E-21	YY1	1.44E-19
MED1	2.46E-43	NR2F2	1.20E-20	RUNX1	1.83E-19
T	1.80E-31	TFAP2C	3.20E-20	ZNF143	4.87E-19
FOXA1	5.99E-23	TFAP2A	3.71E-20	TAF1	9.18E-19

Figure 19: The boxplots for the averaged gene expression patterns of all the different gene categories across four tissues in the mouse metabolism study. V consists of genes detected by min-MCC only, while M1 represents the intersection of genes detected by both min-MCC and MSCA. M2-M11 represent gene categories with concordance shared between different tissue pairs: M2 in brown fat and liver, M3 in brown fat and heart, M4 in brown fat and skeletal, M5 in liver and heart, M6 in liver and skeletal, M7 in heart and skeletal, M8 in brown fat, liver, and heart, M9 in brown fat, liver, and skeletal, M10 in brown fat, heart, and skeletal, and M11 in liver, heart, and skeletal. More stringent threshold ($q - value < 0.01$) for concordant gene identification was applied for visualization.

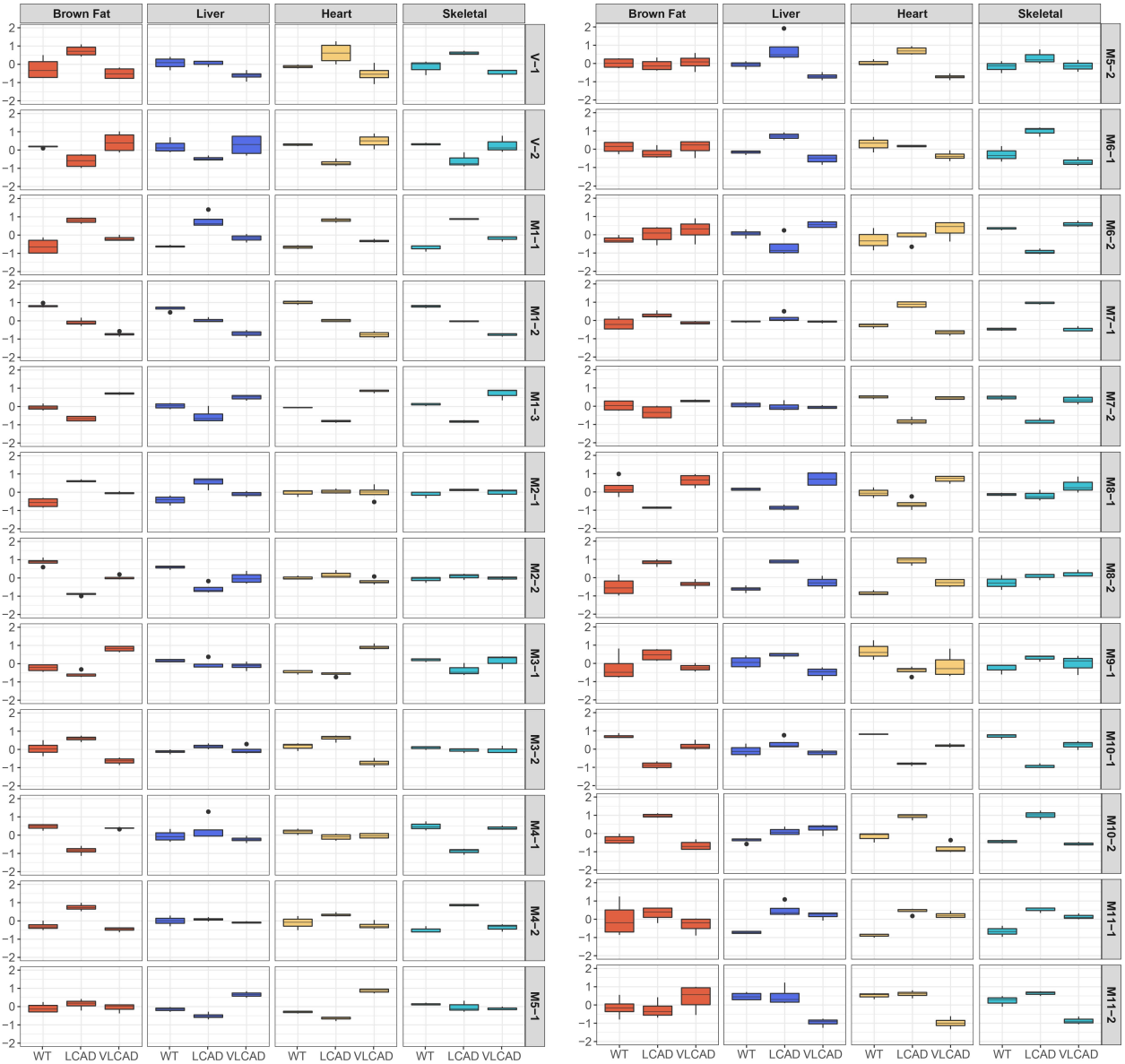


Figure 20: The boxplot for the gene expression patterns of *Blvrb*. Concordance gene expression is in brown fat, heart and skeletal tissues.

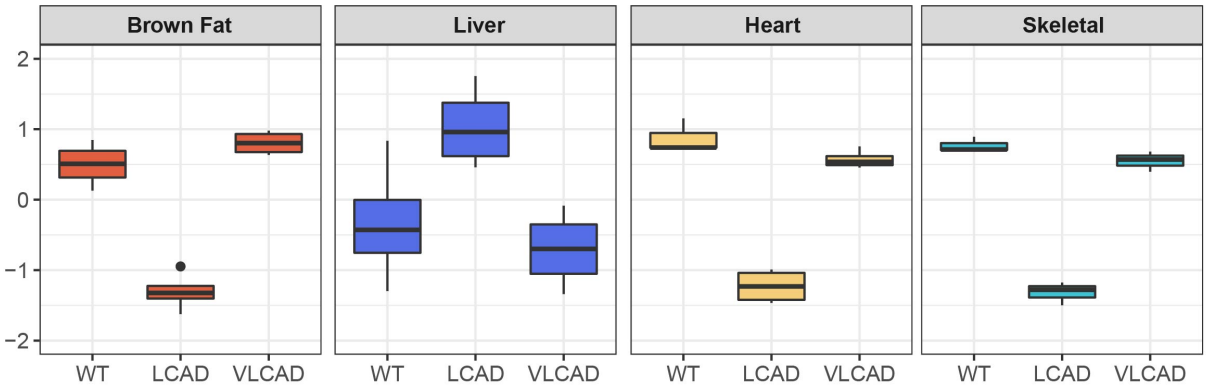
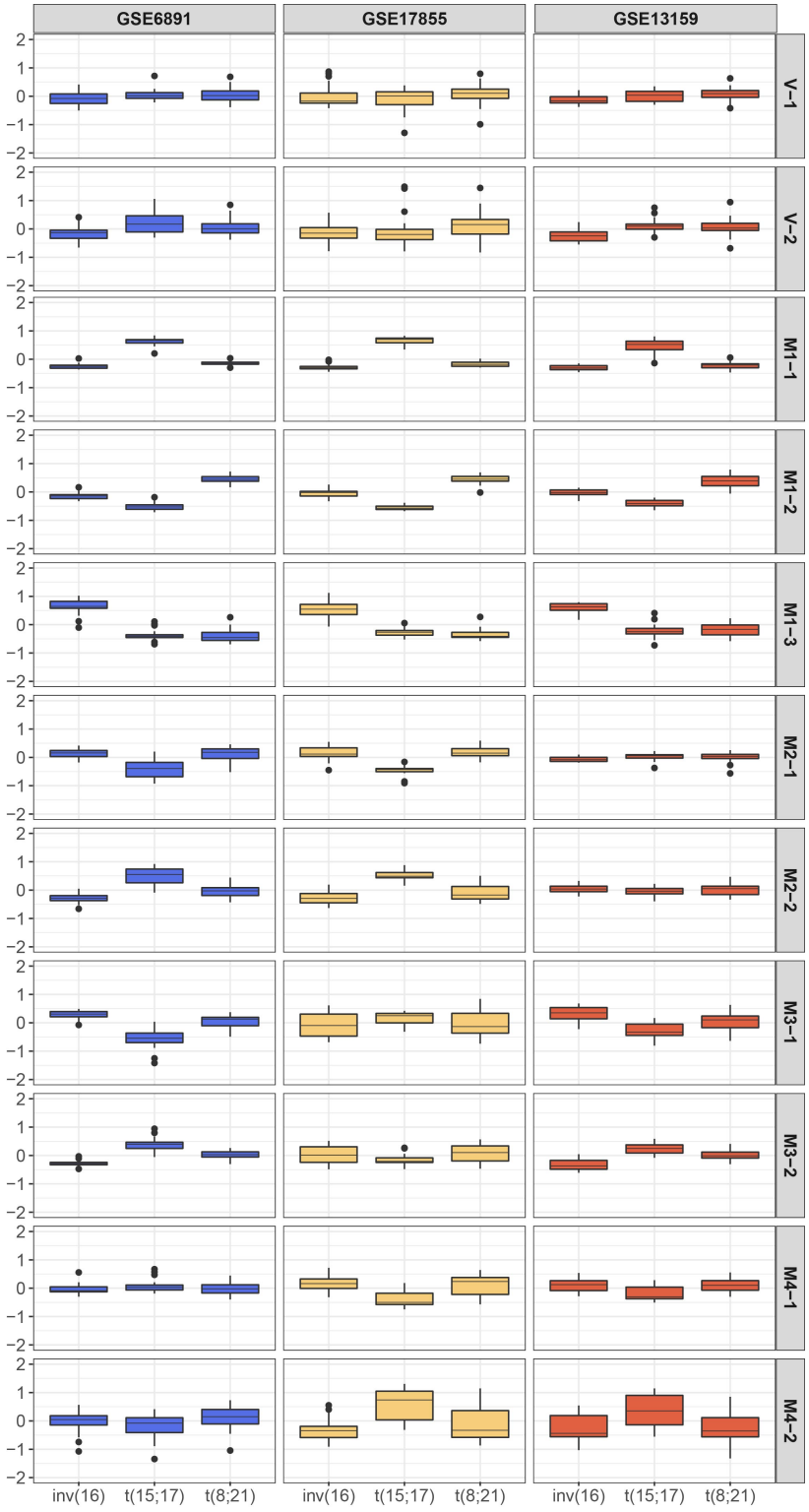


Figure 21: The boxplots for the averaged gene expression patterns of all the different gene categories across three leukemia studies. The gene categories include the genes identified by min-MCC alone (V), the intersected genes identified by min-MCC and MSCA (M1), genes identified only by MSCA and the partial shared concordance detected in GSE6891 and GSE17855 (M2), concordance between GSE6891 and GSE13159 (M3), and concordance between GSE17855 and GSE13159 (M4). More stringent threshold ($q - value < 0.01$) for concordant gene identification was applied for visualization.



Appendix B Chapter 4

B.1 Literature review

Evaluating the transcriptional fidelity of cancer models, *Genome Medicine*, April 2021

Dataset: TCGA, CCLE, ICGC, etc.

Method Evaluation:

1. Identify upregulated, downregulated, and invariant genes in each tumor type by template vector and Pearson correlation
2. Select the most discriminative gene pairs for each tumor type from the above identified genes
3. Train the random forest model using above selected gene pairs
4. Evaluate the cancer models on the 22 tumor types and 36 sub-types
5. Evaluate the similarity in cancer cell lines, xenografts, mouse models, and tumoroids

Pros:

1. The method is platform- and species - agnostic because of ranked-based design
2. Many cancer models are studied

Cons:

1. Cannot measure absolute distance between cell line and tumor
2. The comparability between models and tumors is not considered

Global computational alignment of tumor and cell line transcriptional profiles, *Nature Communications*, 04 January 2021

Dataset: Treehouse, CCLE

Method Evaluation: Propose a method to perform an unsupervised global alignment of tumor and cell line gene expressions, allowing for direct comparisons of their transcriptional profiles.

1. Calculate the Pearson correlation between aligned tumor and cell line
2. Cell lines are classified by identifying the most frequently occurring tumor type within each cell line's 25 highest correlated tumors
3. Show the information transformation between cell and tumor
4. Validate the method using known truth

Pros:

1. New method to make the cell line and tumor data comparable

Cons:

1. The classification method does not provide enough information

CCLA: an accurate method and web server for cancer cell line authentication using gene expression profiles, *Briefings in Bioinformatics*, 08 June 2020

Dataset: CCLE, GDSC, CHCC

Method Evaluation:

1. Apply *single sample gene set enrichment analysis* to the reference set for obtaining the reference score matrix
2. Cluster the reference score matrix into 3 groups by t-SNE
3. Apply random forest using the above group label
4. Obtain the group label for the new sample, and calculate the Pearson correlation with the reference samples within that group
5. Use independent data source
6. Compare the distribution of expressed signature genes in the query samples and resulting reference

Pros:

1. User-friendly web sever
2. Gene signatures for each cell line are carefully selected

Cons:

1. Not consider the tumor

Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns, *Science Advances*, 01 Jul 2020

Dataset: CCLE, TCGA

Method Evaluation:

1. Construct the one vs. rest ridge regression model for each cancer type using TCGA methylation and gene expression data respectively
2. Select the cell lines which have high precision score in both models but reported from the other origin
3. Construct the one (origin) vs. one (suspected) model for the above selected cell lines
4. Verify the misclassification using other mutant data
5. Validate the misclassification using cancer type-specific drugs and specific mutation signatures
6. Use UV-linked signature 7 and sensitivity (IC_{50}) for mutant target drugs to evaluate the 6 cell lines which are consistently reassigned to skin cancer
7. Subtype the cell lines and validate the results using breast cancer cell line subtyping labels
8. Perform association study using different set of cell lines on drug sensitivity and gene dependency screenings

Pros:

1. Apply multi-omics data, especially drug sensitivity data
2. Classification model works well

Cons:

1. Cannot measure absolute distance between cell line and tumor

Evaluating cell lines as tumor models by comparison of genomic profiles, *Nature Communications*, 09 July 2013

Dataset: CCLE, TCGA

Method Evaluation: Suitability Score: $S = A + B – 2 \times C – D/7$

1. A: Correlation with mean CNA of HGSOC tumors
2. B: 1 or 0, TP53 mutation
3. C: 1 or 0, hypermutated
4. D: number of genes mutated among 7 'non-HGSOC' genes

Pros:

1. Design a new score including the important factors

Cons:

1. The score is specific for this one case, and how this score designed (the weights) is not fully illustrated

Integrated analyses of murine breast cancer models reveal critical parallels with human disease, *Nature Communications*, 22 July 2019

Dataset: Lab data

Method Evaluation:

1. Filter the resulting genes based on human data and cluster on gene expression
2. Identify genes highly altered in human, and analyzing CNV in mouse

Pros:

1. The evaluation of mouse model is based on human information

Cons:

1. The correlation between human and mouse model is not fully compared

Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types, *Nature Communications*, 08 August 2019

Dataset: TCGA, CCLE

Method Evaluation:

1. Correlation analysis (remove the tumor purity genes) and gene set enrichment analysis
2. *Nearest Template Prediction* for subtype prediction of cell lines

Pros:

1. Remove the tumor purity genes to make cell lines and tumors more comparable
2. Generate subtype templates using specific genes for NTP

Cons:

1. Cannot measure absolute distance between cell line and tumor

Evaluating cell lines as models for metastatic breast cancer through integrative analysis of genomic data, *Nature Communications*, 15 May 2019

Dataset: MET500, CCLE

Method Evaluation:

1. Compare genomic profiles (Genes highly mutated in metastatic breast cancer & differentially mutated between metastatic and primary breast cancer)
2. Spearman correlation across 1,000 most-varied genes

Pros:

1. Focus on metastasis instead of primary tumor
2. Important gene selection part is thought-provoking

Cons:

1. Cannot measure absolute distance between cell line and tumor

Analysis of Transcriptomic Similarity between Osteosarcoma Cell Lines and Primary Tumors, *Oncology*, 23 Jul 2020; Assessing alveolar rhabdomyosarcoma cell lines as tumor models by comparison of mRNA expression profiles, *Gene*, 15 November 2020

Dataset: TCGA, CCLE, GEO

Method Evaluation:

1. Calculate spearman correlation using 5,000 top genes (by IQR) from primary tumor
2. Differential expression analysis for tumor versus cell lines with purity score and sequencing platform as covariate

3. Gene ontology enrichment analysis based on DEA results

Pros:

1. It is interesting to use DEA to identify genes that are differentially expressed between cell lines and primary tumors

Cons:

1. Cannot measure absolute distance between cell line and tumor
2. Cannot perform the cell line selection

Investigating the utility of human melanoma cell lines as tumour models, *Oncotarget*, 7 Feb 2017

Dataset: TCGA, GEO

Method Evaluation:

1. PCA on cell lines and tumors by top 5,000 genes
2. DEA on cell lines versus tumors
3. Calculate the Pearson association between RNA-seq of cell lines and single cells
4. Subtype the cell lines by clustering using 2 gene sets identified in tumors
5. Detect the UV-induced mutational signatures
6. Prepare a panel for selection based on average properties and genetic events from tumor study

Pros:

1. Use PCA to evaluate the performance of batch correction method
2. A relatively complete analysis
3. Use important signatures for validation
4. Use a panel to summarize the results

Cons:

1. The correlation calculation is too simple
2. The resulting panel cannot directly provide a selection suggestion

B.2 Supplement tables and figures

Table 13: Summary table of the 38 candidate BC cell lines.

	Projected Position	P_{SDA}^{ILC}	DS_{SDA}^{ILC}	Classification
SUM44PE	-0.977	0.987	0.024	ILC
DU4475	-1.036	0.992	0.188	ILC
UACC3133	-1.132	0.996	0.452	ILC
CAMA1	-0.771	0.929	0.541	ILC
HCC2218	-0.666	0.845	0.828	ILC
ZR7530	-0.604	0.764	0.999	No ILC
WCRC25	-1.333	0.999	1.002	ILC
MDAMB453	-0.549	0.67	1.15	No ILC
MDAMB134VI	-0.488	0.548	1.318	No ILC
BCK4	-1.544	1	1.583	ILC
OCUBM	-0.386	0.339	1.597	No ILC
IPH926	-1.551	1	1.601	ILC
UACC893	-0.361	0.293	1.667	No ILC
MDAMB175VII	-0.326	0.236	1.762	No ILC
HCC2185	-1.657	1	1.892	ILC
T47D	-0.264	0.155	1.934	No ILC
MPE600	-0.235	0.125	2.013	No ILC
HCC1428	-0.222	0.114	2.05	No ILC
CAL148	-0.2	0.097	2.108	No ILC
SUM185PE	-0.193	0.091	2.129	No ILC
HCC1419	-0.179	0.082	2.166	No ILC
AU565	-1.844	1	2.405	No ILC
SUM52PE	-0.044	0.028	2.538	No ILC

SKBR3	-1.92	1	2.615	No ILC
BT483	0.083	0.01	2.885	No ILC
MDAMB415	0.088	0.009	2.899	No ILC
MM330	0.103	0.008	2.942	No ILC
ZR751	-2.05	1	2.97	No ILC
UACC812	-2.057	1	2.991	No ILC
HCC1500	0.254	0.002	3.356	No ILC
MFM223	0.334	0.001	3.577	No ILC
MDAMB361	0.479	0	3.974	No ILC
HCC202	0.512	0	4.066	No ILC
KPL1	0.543	0	4.15	No ILC
EFM19	0.649	0	4.439	No ILC
MCF7	0.745	0	4.705	No ILC
EFM192A	0.8	0	4.856	No ILC
BT474	1.627	0	7.127	No ILC

Table 14: Summary table of the pathway specific analysis (DS_{path}) for 9 unbiased-selected + 1 manually-included cell lines and 14 unbiased-selected + 1 manually-included pathways.

	SUM44PE	DU4475	UACC3133	CAMA1	HCC2218	WCRC25	BCK4	IPH926	HCC2185
KEGG_PPAR_SIGNALING_PATHWAY	0.405	0.572	0.42	0.493	0.349	0.358	0.72	0.615	0.395
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.589	0.984	0.877	0.369	0.644	0.635	0.875	0.599	0.932
HALLMARK_KRAS_SIGNALING_DN	0.73	0.667	0.558	0.456	0.572	0.645	0.813	0.553	0.863
HALLMARK_GLYCOLYSIS	0.958	0.97	0.816	0.591	0.8	0.99	1.009	0.791	0.861
KEGG_STARCH_AND_SUCROSE_METABOLISM	0.868	0.581	0.62	0.425	0.623	0.915	0.785	1.284	0.669
HALLMARK_SPERMATOGENESIS	0.43	0.875	0.513	0.465	0.904	0.535	0.507	0.624	0.438
HALLMARK_MYC_TARGETS_V1	0.525	0.849	0.497	0.488	0.991	0.456	0.529	0.551	0.556
KEGG_CELL_CYCLE	0.572	0.755	0.363	0.579	0.992	0.548	0.45	0.655	0.582
KEGG_METABOLISM_OF_XENOBIOTICS_BY_CYTOCHROME_P450	0.822	0.543	1.05	0.528	0.508	0.837	0.911	0.972	0.638
KEGG_RETINOL_METABOLISM	1.051	0.54	0.806	0.453	0.541	0.981	0.932	1.062	0.997
HALLMARK_MTORC1_SIGNALING	0.88	1.04	0.689	0.619	0.778	0.768	0.898	0.951	0.715
KEGG_DRUG_METABOLISM_CYTOCHROME_P450	0.768	0.499	1.035	0.498	0.482	0.823	0.832	0.991	0.611
HALLMARK_G2M_CHECKPOINT	0.478	0.609	0.458	0.525	1.138	0.451	0.376	0.391	0.425
HALLMARK_E2F_TARGETS	0.569	0.799	0.629	0.576	1.156	0.389	0.44	0.52	0.47
AVERAGE OF 14 PATHWAYS	0.689	0.734	0.667	0.505	0.748	0.666	0.72	0.754	0.654
KEGG_CELL_ADHESION_MOLECULES	0.726	0.917	0.796	0.468	0.264	1.155	1.323	0.795	0.869

Table 15: Summary table of the 11 PDO and 136 PDX BC models.

Patient ID	Specimen ID	Sample ID	Projected Position	P_{SDA}^{ILC}	DS_{SDA}^{ILC}	Classification
171881	019-R	APW-DS2	1.072	0.998	0.117	ILC
171881	019-R	APYF68	0.915	0.987	0.562	ILC
171881	019-R	APWG05	0.706	0.898	1.154	ILC
171881	019-R	V1-organoid	0.501	0.508	1.733	ILC
171881	019-R	APWG05PF7	0.432	0.335	1.928	No ILC
337426	197-R	AL-F5Y_AL-A80	0.036	0.008	3.05	No ILC
755229	096-R	AL-VNC_AL-C53_AL-J67	0.004	0.006	3.141	No ILC
755229	096-R	AL-VNC_AL-C53_AL-J67_AL-Q60	-0.026	0.004	3.227	No ILC
337426	197-R	AL-F5Y	-0.039	0.004	3.261	No ILC
397859	316-R	P0POOL_OT-Q25	-0.044	0.003	3.275	No ILC
337426	197-R	AL-F5W_AL-A70	-0.067	0.003	3.34	No ILC
755229	096-R	AL-VNC_AL-C54_AL-Q07	-0.068	0.003	3.343	No ILC
397859	316-R	P0POOL_OT-Q25_RG-NP9	-0.116	0.002	3.48	No ILC
337426	197-R	AL-F5Y_AL-A81_AL-C56_AL-E24_AL-F39	-0.13	0.001	3.521	No ILC
171881	019-R	APVG40_RG-G15	-0.145	0.001	3.562	No ILC
755229	096-R	AL-VNC_AL-C55	-0.186	0.001	3.68	No ILC
755229	096-R	AL-VNC	-0.191	0.001	3.691	No ILC
397859	316-R	P0POOL_OT-Q23N59	-0.223	0.001	3.783	No ILC
755229	096-R	V1-organoid	-0.25	0	3.861	No ILC
337426	197-R	V2-organoid	-0.274	0	3.929	No ILC
913291	066-R	V1-organoid	-0.316	0	4.047	No ILC
397859	316-R	P0POOL_OT-Q23	-0.331	0	4.09	No ILC
913291	066-R	UJH	-0.338	0	4.109	No ILC
913291	066-R	UJHG08	-0.391	0	4.259	No ILC
913291	066-R	UJHG07K01	-0.393	0	4.266	No ILC
913291	066-R	UJHG08J25	-0.412	0	4.319	No ILC
397859	316-R	P0POOL_OT-Q23N60KY7W19	-0.517	0	4.617	No ILC
913291	066-R	UJF	-0.631	0	4.939	No ILC
913291	066-R	UJHG08J26_AL-KX9	-0.728	0	5.214	No ILC

Figure 22: Heatmap of pathway-specific deviance scores (DS_{path}) with 53 pathways (rows) and 9 genome-wide pre-selected cell lines + 1 manually selected cell line (MDA-MB-134VI) (columns). The genome-wide SDA projected deviance score (DS_{SDA}) is shown on the top sidebar and the pathway size and normalized enrichment score (NES) are on the left.

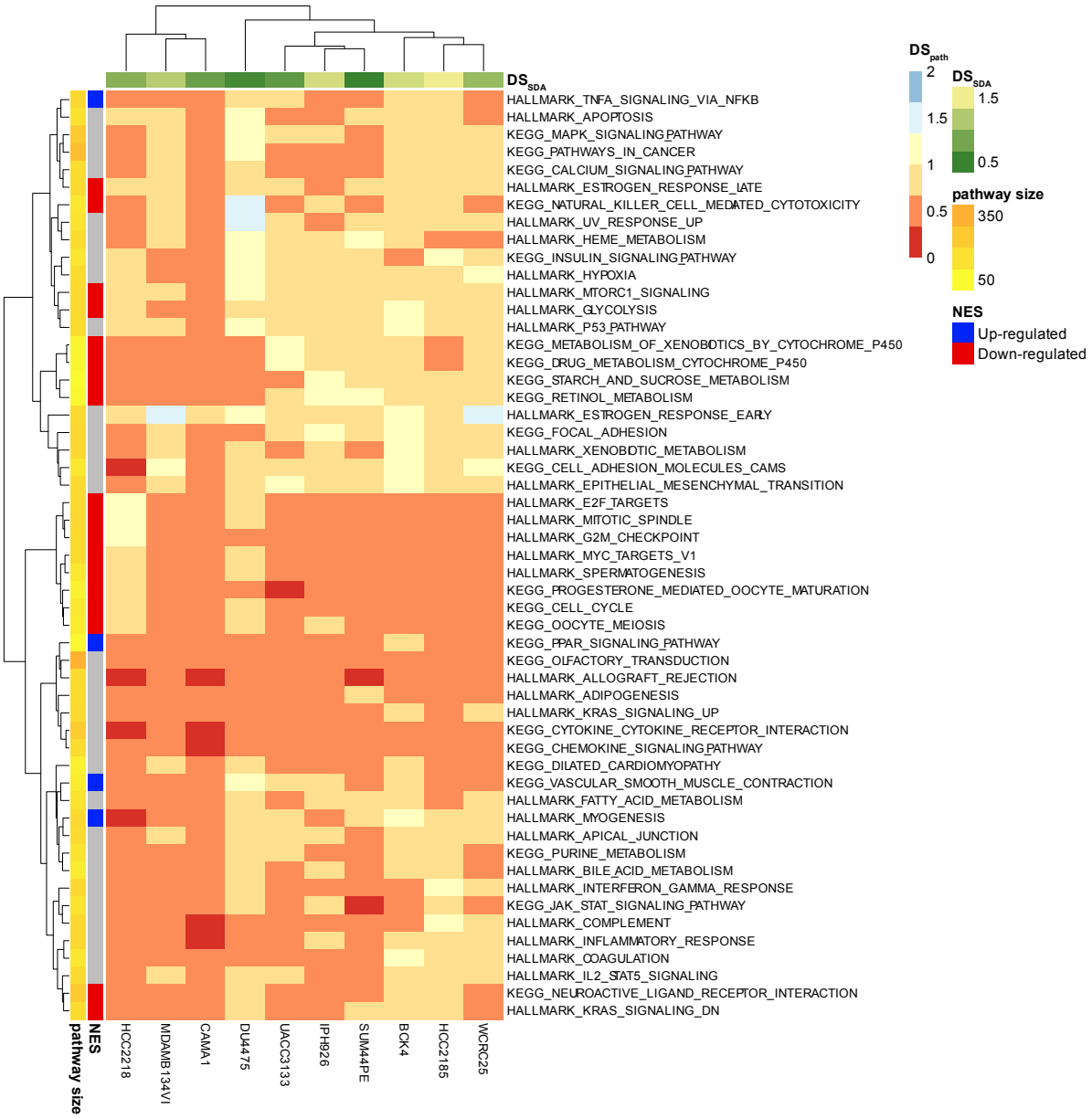


Figure 23: Enrichment plots for Hallmark E2F Targets and KEGG PPAR Signaling Pathway.

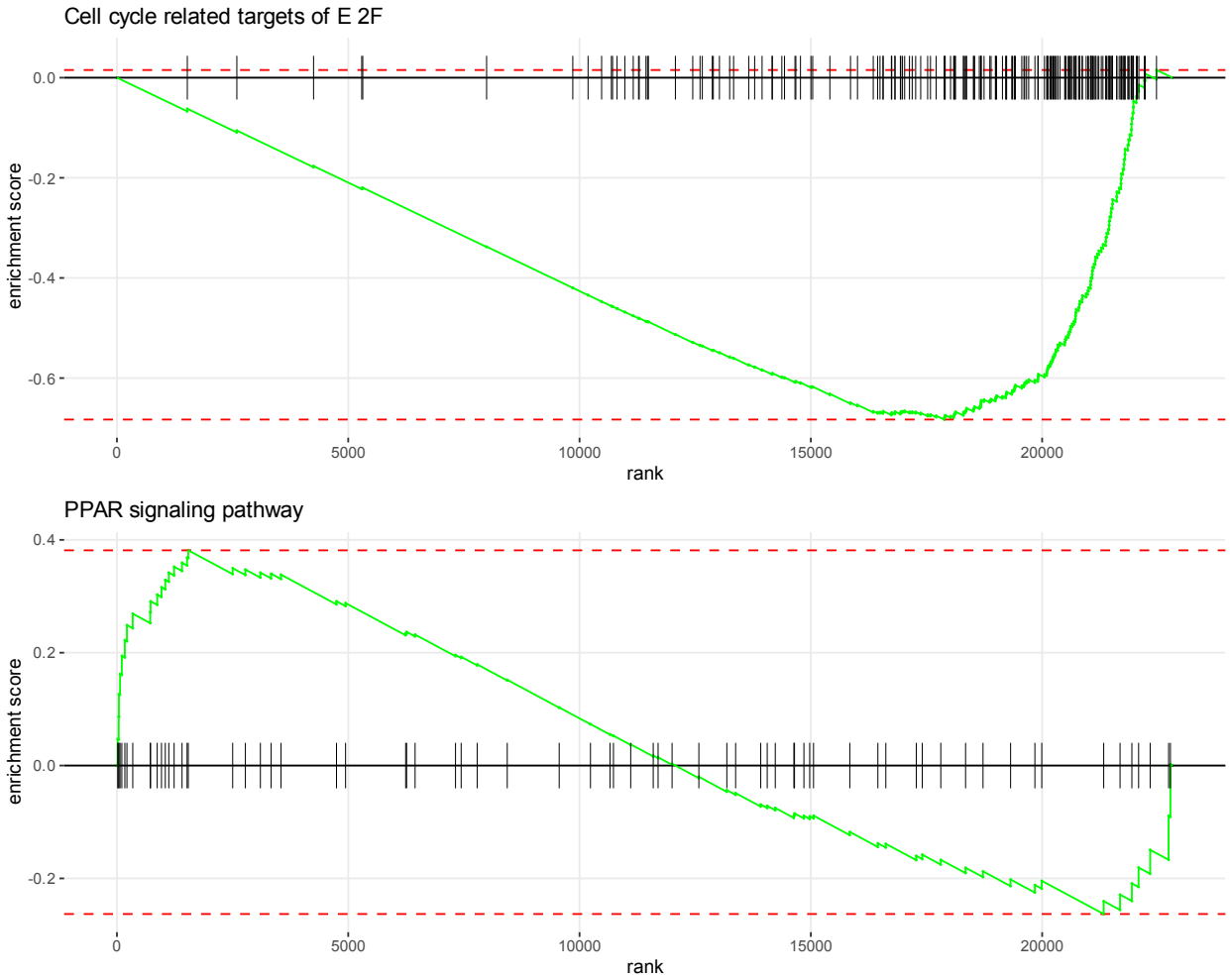


Figure 24: Violin plots for CAMA1 and BCK4 in KEGG Cell Adhesion Molecules.

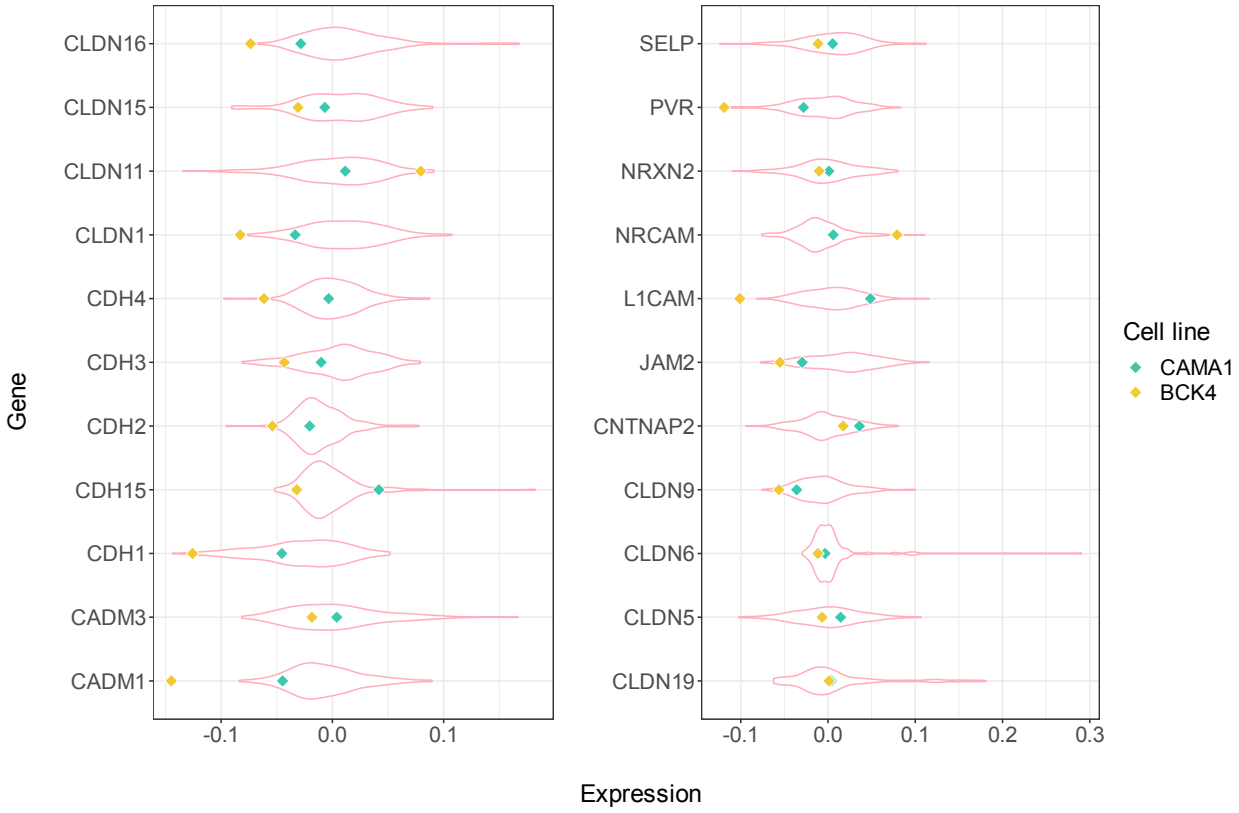


Figure 25: UMAP of Celligner alignment between tumors and PDX/PDO models. (A) Three distinct clusters were observed. The small cluster on the left consists of a seemingly rare breast cancer subtype, the upper-right cluster includes mostly non-basal samples, and the lower-right cluster includes mostly basal samples. (B) UMAP is redrawn when the small cluster in (A) is removed.

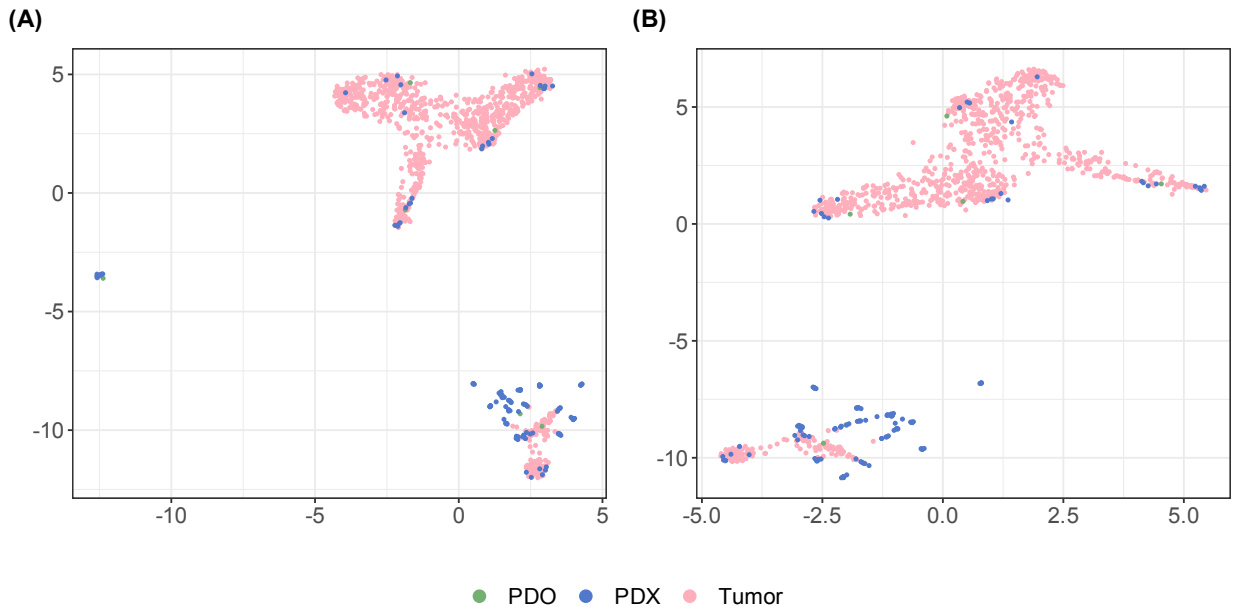


Figure 26: Violin plots for PDO.1 and PDX.1B in KEGG Cell Adhesion Molecules.

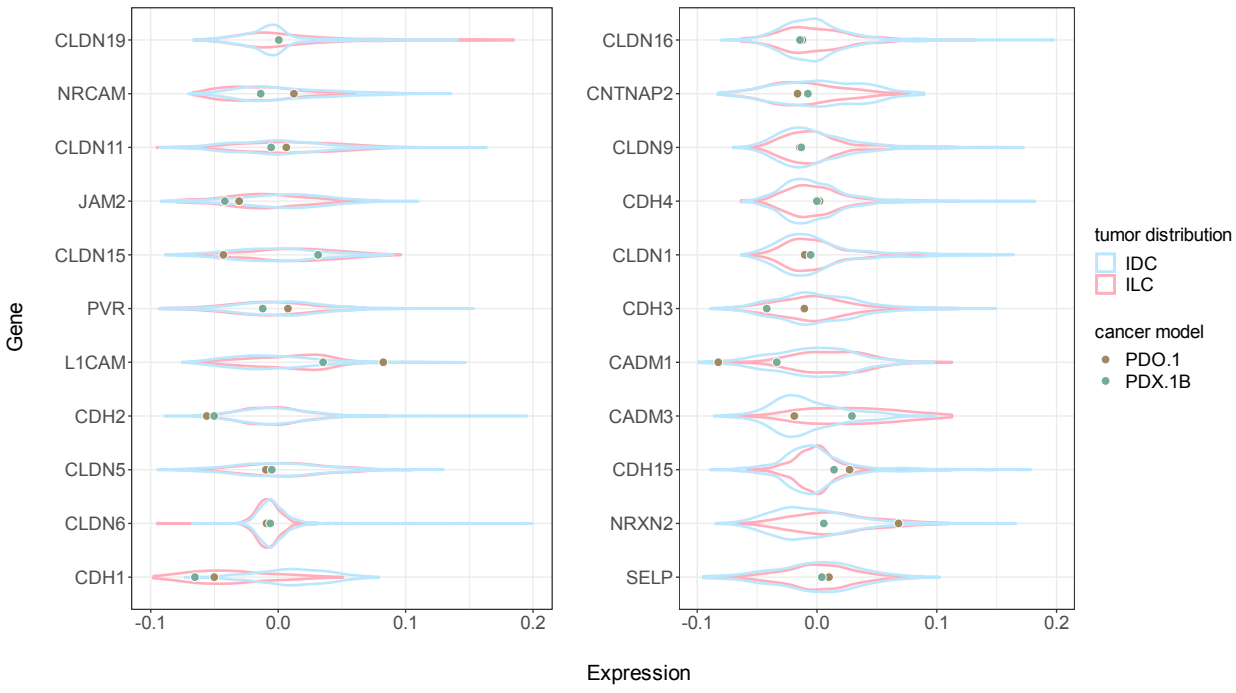
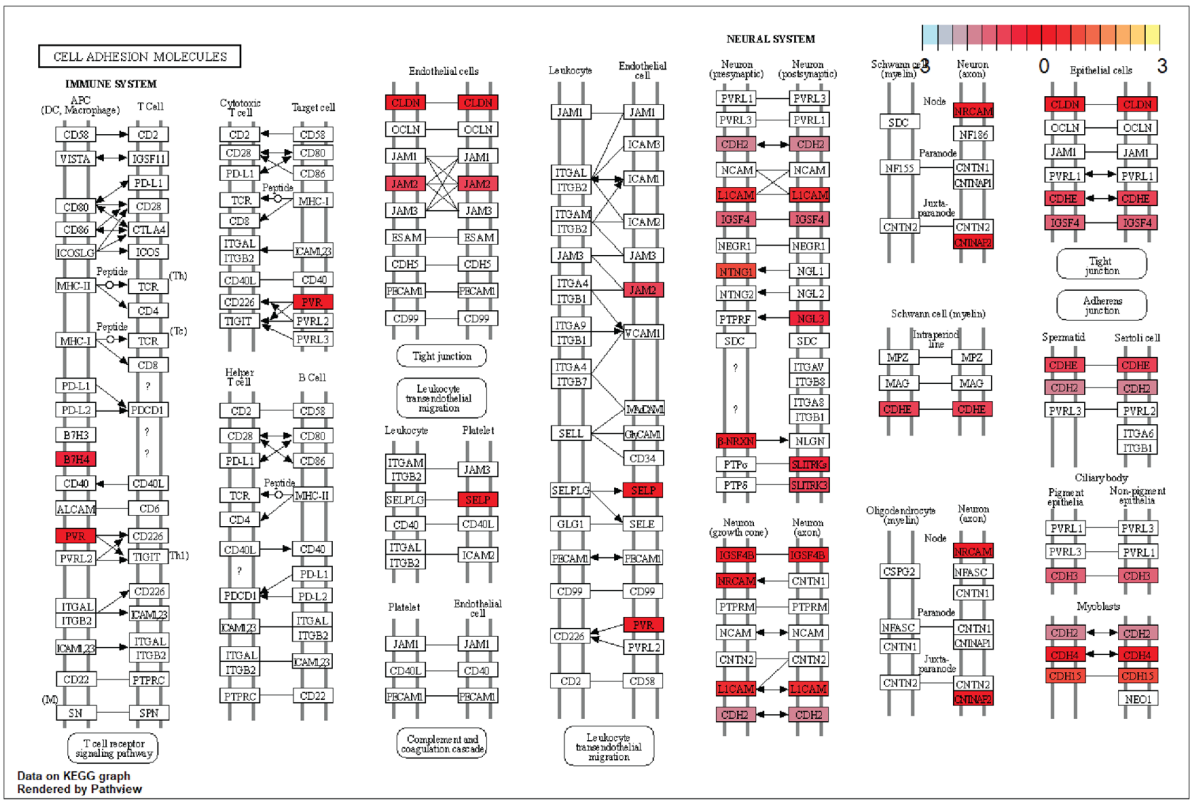
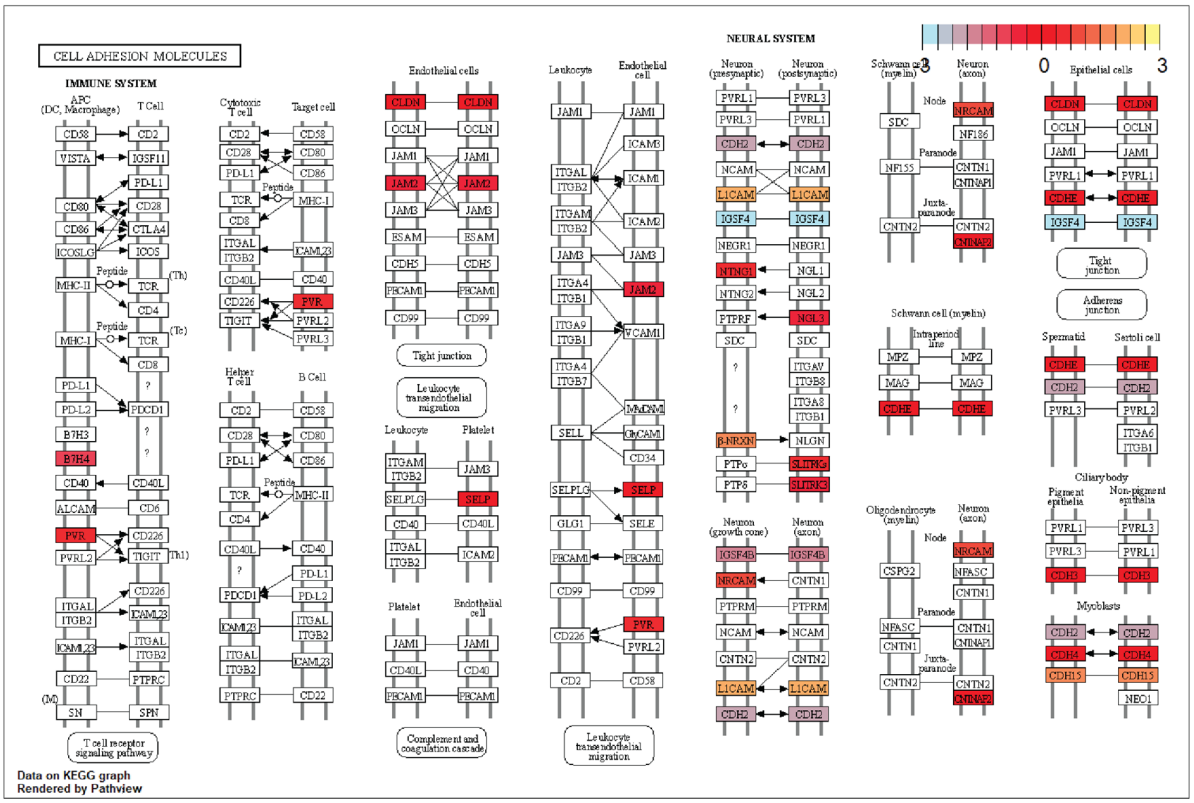


Figure 27: Topological plots for APWG05PF7 (PDX.1B) and V1-organoid (PDO) from the same patient (171881-09-R) in KEGG Cell Adhesion Molecules. (A) APWG05 (PDX.1B). (B) V1-organoid (PDO.1).

(A)



(B)



Bibliography

[1] Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017.

[2] Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, pages 1100–1120, 1982.

[3] Brian Balgobind, Marry van den Heuvel-Eibrink, Renee de Menezes, Dirk Reinhardt, Iris Hollink, Susan Arentsen-Peters, Elisabeth van Wering, Gertjan Kaspers, Jacqueline Cloos, Eveline de Bont, et al. Evaluation of gene expression signatures predictive of cytogenetic and molecular subtypes of pediatric acute myeloid leukemia. *Haematologica*, 96(2):221–230, 2011.

[4] Sai Batchu and Justin Lee Gold. Analysis of transcriptomic similarity between osteosarcoma cell lines and primary tumors. *Oncology*, 98(11):814–816, 2020.

[5] Sai Batchu, Alec S Kellish, and Abraham A Hakim. Assessing alveolar rhabdomyosarcoma cell lines as tumor models by comparison of mrna expression profiles. *Gene*, 760:145025, 2020.

[6] Ferdouse Begum, Debashis Ghosh, George C Tseng, and Eleanor Feingold. Comprehensive literature review and statistical considerations for gwas meta-analysis. *Nucleic Acids Research*, 40(9):3777–3784, 2012.

[7] Uri Ben-David, Gavin Ha, Yuen-Yi Tseng, Noah F Greenwald, Coyin Oh, Juliann Shih, James M McFarland, Bang Wong, Jesse S Boehm, Rameen Beroukhim, et al. Patient-derived xenografts undergo mouse-specific tumor evolution. *Nature Genetics*, 49(11):1567–1575, 2017.

[8] Uri Ben-David, Benjamin Siranosian, Gavin Ha, Helen Tang, Yaara Oren, Kunihiko Hinohara, Craig A Strathdee, Joshua Dempster, Nicholas J Lyons, Robert Burns, et al. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*, 560(7718):325–330, 2018.

[9] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[10] Robert H Berk and Douglas H Jones. Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(1):47–59, 1979.

[11] Michel Berkelaar et al. *lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs*, 2020. R package version 5.6.15.

[12] Ben Bolstad. preprocesscore: A collection of pre-processing functions, 2021.

[13] Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[14] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[15] Paul S Bradley, Kristin P Bennett, and Ayhan Demiriz. Constrained k-means clustering. *Microsoft Research, Redmond*, 20(0):0, 2000.

[16] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11(1):1–13, 2010.

[17] Angelo J Canty. Resampling methods in r: the boot package. *The Newsletter of the R Project Volume*, 2(3):2–7, 2002.

[18] Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E Rudolph, Rona Yaeger, Tara Soumerai, Moriah H Nissan, et al. Oncokb: a precision oncology knowledge base. *JCO Precision Oncology*, 1:1–16, 2017.

[19] Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. Nbclust: an r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61:1–36, 2014.

[20] Sanjib Chaudhary, B Madhu Krishna, and Sandip K Mishra. A novel foxa1/esr1 interacting pathway: A study of oncomine™ breast cancer microarrays. *Oncology Letters*, 14(2):1247–1264, 2017.

[21] Meng Cheng, Stephanie Michalski, and Ramakrishna Kommagani. Role for growth regulation by estrogen in breast cancer 1 (greb1) in hormone-dependent cancers. *International Journal of Molecular Sciences*, 19(9):2543, 2018.

[22] Giovanni Ciriello, Michael L Gatza, Andrew H Beck, Matthew D Wilkerson, Suhn K Rhie, Alessandro Pastore, Hailei Zhang, Michael McLellan, Christina Yau, Cyriac Kandoth, et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519, 2015.

[23] Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4):406–413, 2011.

[24] Matthew Cobb. 60 years ago, francis crick changed the logic of biology. *PLOS Biology*, 15(9):e2003243, 2017.

[25] Lawrence D Cohn and Betsy J Becker. How meta-analysis increases statistical power. *Psychological Methods*, 8(3):243, 2003.

[26] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[27] Laura E DeMare, Jing Leng, Justin Cotney, Steven K Reilly, Jun Yin, Richard Sarro, and James P Noonan. The genomic landscape of cohesin-associated chromatin interactions. *Genome Research*, 23(8):1224–1234, 2013.

[28] Ayhan Demiriz, Kristin P Bennett, and Paul S Bradley. Using assignment constraints to avoid empty clusters in k-means clustering. In *Constrained clustering: advances in algorithms, theory, and applications*, pages 201–219. Chapman & Hall/CRC Boca Raton, FL, 2008.

[29] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[30] Broad DepMap, Steven Corsello, Mustafa Kocak, and Todd Golub. PRISM repurposing 19Q4 dataset, December 2019.

[31] Rebecca DerSimonian and Raghu Kacker. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials*, 28(2):105–114, 2007.

[32] Silvia Domcke, Rileen Sinha, Douglas A Levine, Chris Sander, and Nikolaus Schultz. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature Communications*, 4(1):1–10, 2013.

[33] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

[34] Tian Du, Matthew J Sikora, Kevin M Levine, Nilgun Tasdemir, Rebecca B Riggins, Stacy G Wendell, Bennett Van Houten, and Steffi Oesterreich. Key regulators of lipid metabolism drive endocrine resistance in invasive lobular breast cancer. *Breast Cancer Research*, 20(1):1–15, 2018.

[35] Angelo Duò, Mark D Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, 7, 2018.

[36] Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987.

[37] Pierre Gançarski, Thi-Bich-Hanh Dao, Bruno Crémilleux, Germain Forestier, and Thomas Lampert. Constrained clustering: Current and new trends. In *A guided tour of artificial intelligence research*, pages 447–484. Springer, 2020.

[38] Nuwan Ganganath, Chi-Tsun Cheng, and K Tse Chi. Data clustering with cluster size constraints using a modified k-means algorithm. In *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 158–161. IEEE, 2014.

[39] RJ Geraghty, A Capes-Davis, JM Davis, J Downward, RI Freshney, I Knezevic, R Lovell-Badge, JRW Masters, J Meredith, GN Stacey, et al. Guidelines for the use of cell lines in biomedical research. *British Journal of Cancer*, 111(6):1021–1046, 2014.

[40] Mahmoud Ghandi, Franklin W Huang, Judit Jané-Valbuena, Gregory V Kryukov, Christopher C Lo, E Robert McDonald III, Jordi Barretina, Ellen T Gelfand, Craig M Bielski, Haoxin Li, et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757):503–508, 2019.

[41] Rajeshwar Govindarajan, Jeyapradha Duraiyan, Karunakaran Kaliyappan, and Murugesan Palanisamy. Microarray and its applications. *Journal of Pharmacy and Bioallied Sciences*, 4(Suppl 2):S310, 2012.

[42] Zuguang Gu, Matthias Schlesner, and Daniel Hübschmann. cola: an r/bioconductor package for consensus partitioning through a general framework. *Nucleic Acids Research*, 49(3):e15, 2021.

[43] Barry M Gumbiner. Regulation of cadherin adhesive activity. *The Journal of Cell Biology*, 148(3):399–404, 2000.

[44] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

[45] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):1–15, 2017.

[46] Chun Hua, Feng Li, Chao Zhang, Jie Yang, and Wei Wu. A genetic xk-means algorithm with empty cluster reassignment. *Symmetry*, 11(6):744, 2019.

[47] Josep C Jiménez-Chillarón, Rubén Díaz, and Marta Ramón-Krauel. Omics tools for the genome-wide analysis of methylation and histone modifications. In *Comprehensive Analytical Chemistry*, volume 63, pages 81–110. Elsevier, 2014.

[48] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[49] DW Kennedy, J Cameron, PP-Y Wu, and K Mengersen. Peer groups for organisational learning: Clustering with practical constraints. *PLOS ONE*, 16(6):e0251723, 2021.

[50] Jihoon Kim, Bon-Kyoung Koo, and Juergen A Knoblich. Human organoids: model systems for human biology and medicine. *Nature Reviews Molecular Cell Biology*, 21(10):571–584, 2020.

[51] Gabriela S Kinker, Alissa C Greenwald, Rotem Tal, Zhanna Orlova, Michael S Cuoco, James M McFarland, Allison Warren, Christopher Rodman, Jennifer A Roth, Samantha A Bender, et al. Pan-cancer single-cell rna-seq identifies recurring programs of cellular heterogeneity. *Nature Genetics*, 52(11):1208–1218, 2020.

[52] Alexander Kohlmann, Thomas J Kipps, Laura Z Rassenti, James R Downing, Sheila A Shurtleff, Ken I Mills, Amanda F Gilkes, Wolf-Karsten Hofmann, Giuseppe Basso, Marta Campo Dell'Orto, et al. An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the microarray innovations in leukemia study prephase. *British Journal of Haematology*, 142(5):802–807, 2008.

[53] Gennady Korotkevich, Vladimir Sukhov, Nikolay Budin, Boris Shpak, Maxim N Artyomov, and Alexey Sergushichev. Fast gene set enrichment analysis. *BioRxiv*, page 060012, 2021.

[54] Arthur Korte and Ashley Farlow. The advantages and limitations of trait analysis with gwas: a review. *Plant Methods*, 9(1):1–9, 2013.

[55] Andreas Krämer, Jeff Green, Jack Pollard Jr, and Stuart Tugendreich. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4):523–530, 2014.

[56] Karoline B Kuchenbaecker, John L Hopper, Daniel R Barnes, Kelly-Anne Phillips, Thea M Mooij, Marie-José Roos-Blom, Sarah Jervis, Flora E Van Leeuwen, Roger L Milne, Nadine Andrieu, et al. Risks of breast, ovarian, and contralateral breast cancer for brca1 and brca2 mutation carriers. *JAMA*, 317(23):2402–2416, 2017.

[57] Max Kuhn. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28:1–26, 2008.

[58] Allison W Kurian and James M Ford. Multigene panel testing in oncology practice: how should we respond? *JAMA Oncology*, 1(3):277–278, 2015.

[59] AJ Langlois, WD Holder Jr, JD Iglehart, WA Nelson-Rees, SA Wells Jr, and DP Bolognesi. Morphological and biochemical properties of a new human breast cancer cell line. *Cancer Research*, 39(7_Part_1):2604–2613, 1979.

[60] Jackie A Lavigne, Yoko Takahashi, Gadisetti VR Chandramouli, Huaitian Liu, Susan N Perkins, Stephen D Hursting, and Thomas TY Wang. Concentration-dependent effects of genistein on global gene expression in mcf-7 breast cancer cells: an oligo microarray study. *Breast Cancer Research and Treatment*, 110:85–98, 2008.

[61] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

[62] Yujia Li, Tanbin Rahman, Tianzhou Ma, Lu Tang, and George C Tseng. A sparse negative binomial mixture model for clustering rna-seq count data. *Biostatistics*, 2021.

[63] Biao Liu, Carl D Morrison, Candace S Johnson, Donald L Trump, Maochun Qin, Jeffrey C Conroy, Jianmin Wang, and Song Liu. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. *Oncotarget*, 4(11):1868–1881, 2013.

[64] Ke Liu, Patrick A Newbury, Benjamin S Glicksberg, William ZD Zeng, Shreya Paithankar, Eran R Andrechek, and Bin Chen. Evaluating cell lines as models for metastatic breast cancer through integrative analysis of genomic data. *Nature Communications*, 10(1):1–12, 2019.

[65] Ruitao Liu, Xiongying Ye, and Tianhong Cui. Recent progress of biomarker detection sensors. *Research*, 2020, 2020.

[66] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The Genotype-Tissue expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

[67] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):1–21, 2014.

[68] Shuya Lu, Jia Li, Chi Song, Kui Shen, and George C Tseng. Biomarker detection in the integration of multiple multi-class genomic studies. *Bioinformatics*, 26(3):333–340, 2010.

[69] Weijun Luo and Cory Brouwer. Pathview: an r/bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1831, 2013.

[70] Mikko I Malinen and Pasi Fränti. Balanced k-means for clustering. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 32–41. Springer, 2014.

[71] B Mao, X Xu, G Sheng, W Qian, and HQ Li. Transcriptome comparison among patients, pdx, pdo, pdxo, pdxc and cell lines. *European Journal of Cancer*, 138:S31, 2020.

[72] Yosuke Masamoto, Akira Chiba, Hideaki Mizuno, Toshiya Hino, Hiroki Hayashida, Tomohiko Sato, Masashige Bando, Katsuhiko Shirahige, and Mineo Kurokawa. Evi1 exerts distinct roles in aml via erg and cyclin d1 promoting a chemoresistance and immune-suppressive environment. *Blood Advances*, 2022.

[73] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[74] Terrence F Meehan, Nathalie Conte, Theodore Goldstein, Giorgio Inghirami, Mark A Murakami, Sebastian Brabetz, Zhiping Gu, Jeffrey A Wiser, Patrick Dunn, Dale A Begley, et al. Pdx-mi: minimal information for patient-derived tumor xenograft models. *Cancer Research*, 77(21):e62–e66, 2017.

[75] Magali Michaut, Suet-Feung Chin, Ian Majewski, Tesa M Severson, Tycho Bismeijer, Leanne De Koning, Justine K Peeters, Philip C Schouten, Oscar M Rueda, Astrid J Bosma, et al. Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. *Scientific Reports*, 6(1):1–13, 2016.

[76] Milad Mostavi, Yu-Chiao Chiu, Yufei Huang, and Yidong Chen. Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics*, 13:1–13, 2020.

[77] Alison M Nagle, Kevin M Levine, Nilgun Tasdemir, Julie A Scott, Kara Burlbaugh, Justin Kehm, Tiffany A Katz, David N Boone, Britta M Jacobsen, Jennifer M Atkinson, et al. Loss of e-cadherin enhances igf1–igf1r pathway activation and sensitizes breast cancers to anti-igf1r/insr inhibitors. *Clinical Cancer Research*, 24(20):5165–5177, 2018.

[78] Hanna Najgebauer, Mi Yang, Hayley E Francies, Clare Pacini, Euan A Stronach, Mathew J Garnett, Julio Saez-Rodriguez, and Francesco Iorio. Cellector: genomics-guided selection of cancer in vitro models. *Cell Systems*, 10(5):424–432, 2020.

[79] NCI-Frederick. The NCI Patient-Derived models repository (PDMR).

[80] Samuel Y Ng, Noriaki Yoshida, Amanda L Christie, Mahmoud Ghandi, Neekesh V Dharia, Joshua Dempster, Mark Murakami, Kay Shigemori, Sara N Morrow, Alexandria Van Scoyk, et al. Targetable vulnerabilities in t-and nk-cell lymphomas identified through preclinical models. *Nature Communications*, 9(1):2024, 2018.

[81] Hao Nie, Xianpei Han, Ben He, Le Sun, Bo Chen, Wei Zhang, Suhui Wu, and Hao Kong. Deep sequence-to-sequence entity matching for heterogeneous entity resolution. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 629–638, 2019.

[82] Abena Nsiah-Sefaa and Matthew McKenzie. Combined defects in oxidative phosphorylation and fatty acid β-oxidation in mitochondrial disease. *Bioscience Reports*, 36(2), 2016.

[83] Malay K Pakhira. A modified k-means algorithm to avoid empty clusters. *International Journal of Recent Trends in Engineering*, 1(1):220, 2009.

[84] Wei Pan and Xiaotong Shen. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8(5), 2007.

[85] Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160, 2009.

[86] Da Peng, Rachel Gleyzer, Wen-Hsin Tai, Pavithra Kumar, Qin Bian, Bradley Isaacs, Edroaldo Lummertz da Rocha, Stephanie Cai, Kathleen DiNapoli, Franklin W Huang, et al. Evaluating the transcriptional fidelity of cancer models. *Genome Medicine*, 13(1):1–27, 2021.

[87] Orsolya Pipek, Dezso Ribli, József Molnár, Á Póti, Marcin Krzystanek, András Bodor, Gabor E Tusnady, Zoltán Szállási, Istvan Csabai, and Dávid Szüts. Fast and accurate mutation detection in whole genome sequences of multiple isogenic samples with isomut. *BMC Bioinformatics*, 18(1):1–11, 2017.

[88] Qian Qin, Jingyu Fan, Rongbin Zheng, Changxin Wan, Shenglin Mei, Qiu Wu, Hanfei Sun, Myles Brown, Jing Zhang, Clifford A Meyer, et al. Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and chip-seq data. *Genome Biology*, 21(1):1–14, 2020.

[89] James M Rae, Michael D Johnson, Joshua O Scheys, Kevin E Cordero, José M Larios, and Marc E Lippman. Greb1 is a critical regulator of hormone dependent breast cancer growth. *Breast Cancer Research and Treatment*, 92:141–149, 2005.

[90] Mumtahena Rahman, Laurie K Jackson, W Evan Johnson, Dean Y Li, Andrea H Bild, and Stephen R Piccolo. Alternative preprocessing of rna-sequencing data in the cancer genome atlas leads to improved analysis results. *Bioinformatics*, 31(22):3666–3672, 2015.

[91] Ricardo Ramirez, Yu-Chiao Chiu, Allen Hererra, Milad Mostavi, Joshua Ramirez, Yidong Chen, Yufei Huang, and Yu-Fang Jin. Classification of cancer types using graph convolutional neural networks. *Frontiers in Physics*, 8:203, 2020.

[92] Max AK Rätze, Thijs Koorman, Thijmen Sijnesael, Blessing Bassey-Archibong, Robert van de Ven, Lotte Enserink, Daan Visser, Sridevi Jaksani, Ignacio Viciano, Elvira RM Bakker, et al. Loss of e-cadherin leads to id2-dependent inhibition of cell cycle progression in metastatic lobular breast cancer. *Oncogene*, 41(21):2932–2944, 2022.

[93] Amy E McCart Reed, Jamie R Kutasovic, Sunil R Lakhani, and Peter T Simpson. Invasive lobular carcinoma of the breast: morphology, biomarkers and'omics. *Breast Cancer Research*, 17(1):1–11, 2015.

[94] Jonathan P Rennhack, Briana To, Matthew Swiatnicki, Caleb Dulak, Martin P Ogrodzinski, Yueqi Zhang, Caralynn Li, Evan Bylett, Christina Ross, Karol Szczepanek, et al. Integrated analyses of murine breast cancer models reveal critical parallels with human disease. *Nature Communications*, 10(1):1–12, 2019.

[95] Niels J Rinzema, Konstantinos Sofiadis, Sjoerd JD Tjalsma, Marjon JAM Verstegen, Yuva Oz, Christian Valdes-Quezada, Anna-Karina Felder, Teodora Filipovska, Stefan van der Elst, Zaria de Andrade dos Ramos, et al. Building regulatory landscapes reveals that an enhancer can recruit cohesin to create contact domains, engage ctcf sites and activate distant genes. *Nature Structural & Molecular Biology*, 29(6):563–574, 2022.

[96] Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. Adjusting for chance clustering comparison measures. *The Journal of Machine Learning Research*, 17(1):4635–4666, 2016.

[97] Peter Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.

[98] Marina Salvadores, Francisco Fuster-Tormo, and Fran Supek. Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. *Science Advances*, 6(27):eaba1862, 2020.

[99] Rickard Sandberg and Ingemar Ernberg. Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (tsi). *Proceedings of the National Academy of Sciences*, 102(6):2052–2057, 2005.

[100] Vishesh Sarin, Katharine Yu, Ian D Ferguson, Olivia Gugliemini, Matthew A Nix, Byron Hann, Marina Sirota, and Arun P Wiita. Evaluating the efficacy of multiple myeloma cell lines as models for patient tumors via transcriptomic correlation analysis. *Leukemia*, 34(10):2754–2765, 2020.

[101] David Sarrió, Socorro María Rodriguez-Pinilla, David Hardisson, Amparo Cano, Gema Moreno-Bueno, and José Palacios. Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer Research*, 68(4):989–997, 2008.

[102] Todd M Schaefer, Jacqueline A Wright, Patricia A Pioli, and Charles R Wira. Il-1β-mediated proinflammatory responses are inhibited by estradiol via down-regulation of il-1 receptor type i in uterine epithelial cells. *The Journal of Immunology*, 175(10):6509–6516, 2005.

[103] Andrew J Sharp, Devin P Locke, Sean D McGrath, Ze Cheng, Jeffrey A Bailey, Rhea U Vallente, Lisa M Pertz, Royden A Clark, Stuart Schwartz, Rick Segraves, et al. Segmental duplications and copy-number variation in the human genome. *The American Journal of Human Genetics*, 77(1):78–88, 2005.

[104] Jiahao Shi, Yongyun Li, Renbing Jia, and Xianqun Fan. The fidelity of cancer cells in pdx models: Characteristics, mechanism and clinical significance. *International Journal of Cancer*, 146(8):2078–2088, 2020.

[105] Matthew J Sikora, Kristine L Cooper, Amir Bahreini, Soumya Luthra, Guoying Wang, Uma R Chandran, Nancy E Davidson, David J Dabbs, Alana L Welm, and Steffi Oesterreich. Invasive lobular carcinoma cell lines are characterized by unique estrogen-mediated gene expression patterns and altered tamoxifen responseendocrine response and resistance in ilc. *Cancer Research*, 74(5):1463–1474, 2014.

[106] Ambily Sivadas, Victor C Kok, and Ka-Lok Ng. Multi-omics analyses provide novel biological insights to distinguish lobular ductal types of invasive breast cancers. *Breast Cancer Research and Treatment*, 193(2):361–379, 2022.

[107] Rebecca G Smith, Eilis Hannon, Philip L De Jager, Lori Chibnik, Simon J Lott, Daniel Condliffe, Adam R Smith, Vahram Haroutunian, Claire Troakes, Safa Al-Sarraj, et al. Elevated dna methylation across a 48-kb region spanning the hoxa gene cluster is associated with alzheimer's disease neuropathology. *Alzheimer's & Dementia*, 14(12):1580–1588, 2018.

[108] Joseph L Sottnik, Evelyn K Bordeaux, Sanjana Mehrotra, Sarah E Ferrara, Andrew E Goodspeed, James C Costello, and Matthew J Sikora. Mediator of dna damage checkpoint 1 (mdc1) is a novel estrogen receptor coregulator in invasive lobular carcinoma of the breast. *Molecular Cancer Research*, 19(8):1270–1282, 2021.

[109] Samuel A Stouffer, Edward A Suchman, Leland C DeVinney, Shirley A Star, and Robin M Williams Jr. *The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1.* Princeton Univ. Press, 1949.

[110] Barbara E Stranger, Matthew S Forrest, Mark Dunning, Catherine E Ingle, Claude Beazley, Natalie Thorne, Richard Redon, Christine P Bird, Anna De Grassi, Charles Lee, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, 2007.

[111] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, 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):15545–15550, 2005.

[112] Ghazala Sultan, Swaleha Zubair, Iftikhar Aslam Tayubi, Hans-Uwe Dahms, and Inamul Hasan Madar. Towards the early detection of ductal carcinoma (a common type of breast cancer) using biomarkers linked to the ppar (γ) signaling pathway. *Bioinformation*, 15(11):799, 2019.

[113] Nilgun Tasdemir, Emily A Bossart, Zheqi Li, Li Zhu, Matthew J Sikora, Kevin M Levine, Britta M Jacobsen, George C Tseng, Nancy E Davidson, and Steffi Oesterre-

ich. Comprehensive phenotypic characterization of human invasive lobular carcinoma cell lines in 2d and 3d culturescharacterizing human invasive lobular carcinoma cell lines. *Cancer Research*, 78(21):6209–6222, 2018.

[114] Vasiliki Theodorou, Rory Stark, Suraj Menon, and Jason S Carroll. Gata3 acts upstream of foxa1 in mediating esr1 binding by shaping enhancer accessibility. *Genome Research*, 23(1):12–22, 2013.

[115] Ryan J Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

[116] Leonard Henry Caleb Tippett et al. The methods of statistics. *The Methods of Statistics*, 1931.

[117] Thomas A Trikalinos, Georgia Salanti, Elias Zintzaras, and John PA Ioannidis. Meta-analysis methods. *Advances in Genetics*, 60:311–334, 2008.

[118] George C Tseng, Debashis Ghosh, and Eleanor Feingold. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research*, 40(9):3785–3799, 2012.

[119] Shoichiro Tsukita and Mikio Furuse. Pores in the wall: claudins constitute tight junction strands containing aqueous pores. *The Journal of Cell Biology*, 149(1):13–16, 2000.

[120] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[121] Marc van de Wetering, Nick Barker, I Clara Harkes, Marcel van der Heyden, Nicolette J Dijk, Antoinette Hollestelle, Jan GM Klijn, Hans Clevers, and Mieke Schutte. Mutant e-cadherin breast cancer cells do not display constitutive wnt signaling. *Cancer Research*, 61(1):278–284, 2001.

[122] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

[123] Roel GW Verhaak, Bas J Wouters, Claudia AJ Erpelinck, Saman Abbas, H Berna Beverloo, Sanne Lugthart, Bob Löwenberg, Ruud Delwel, and Peter JM Valk. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica*, 94(1):131, 2009.

[124] Krista Marie Vincent and Lynne-Marie Postovit. Investigating the utility of human melanoma cell lines as tumour models. *Oncotarget*, 8(6):10498, 2017.

[125] Allison Warren, Yejia Chen, Andrew Jones, Tsukasa Shibue, William C Hahn, Jesse S Boehm, Francisca Vazquez, Aviad Tsherniak, and James M McFarland. Global com-

putational alignment of tumor and cell line transcriptional profiles. *Nature Communications*, 12(1):1–12, 2021.

[126] Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960.

[127] G Emerens Wensink, Sjoerd G Elias, Jasper Mullenders, Miriam Koopman, Sylvia F Boj, Onno W Kranenburg, and Jeanine ML Roodhart. Patient-derived organoids as a predictive biomarker for treatment response in cancer patients. *NPJ Precision Oncology*, 5(1):1–13, 2021.

[128] Natalie Wilson, Alastair Ironside, Anna Diana, and Olga Oikonomidou. Lobular breast cancer: a review. *Frontiers in Oncology*, 10:591399, 2021.

[129] Daniela M Witten. Classification and clustering of sequencing data using a poisson model. *The Annals of Applied Statistics*, 5(4):2493–2518, 2011.

[130] Huayu Yang, Lejia Sun, Meixi Liu, and Yilei Mao. Patient-derived organoids: A promising model for personalized cancer treatment, 2018.

[131] Go J Yoshida. Applications of patient-derived tumor xenograft models and tumor organoids. *Journal of Hematology and Oncology*, 13(1):1–16, 2020.

[132] K Yu, B Chen, D Aran, J Charalel, C Yau, DM Wolf, LJ Van'T Veer, AJ Butte, T Goldstein, and M Sirota. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nature Communications*, 10(1):1–11, 2019.

[133] Qiong Zhang, Mei Luo, Chun-Jie Liu, and An-Yuan Guo. Ccla: an accurate method and web server for cancer cell line authentication using gene expression profiles. *Briefings in Bioinformatics*, 22(3):bbaa093, 2021.

[134] Yuxun Zhang, Sivakama S Bharathi, Matthew J Rardin, Radha Uppala, Eric Verdin, Bradford W Gibson, and Eric S Goetzman. Sirt3 and sirt5 regulate the enzyme activity and cardiolipin binding of very long-chain acyl-coa dehydrogenase. *PLOS ONE*, 10(3):e0122297, 2015.

[135] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):1–12, 2017.

[136] Shunzhi Zhu, Dingding Wang, and Tao Li. Data clustering with size constraints. *Knowledge-Based Systems*, 23(8):883–889, 2010.