STATISTICAL LEARNING METHODS FOR MULTI-OMICS DATA INTEGRATION IN DIMENSION REDUCTION, SUPERVISED AND UNSUPERVISED MACHINE LEARNING

by

SungHwan Kim

M.S., Statistics, Korea University, South Korea, 2010

B.A., Education, Korea University, South Korea, 2007

Submitted to the Graduate Faculty of

the Department of Biostatistics of the

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF BIOSTATISTICS OF THE

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

SungHwan Kim

It was defended on

April 20, 2015

and approved by

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

YongSeok Park, PhD, Assistant Professor Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Daniel E. Weeks, PhD, Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

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

Jing Lei, PhD, Assistant Professor, Department of Statistics, Carnegie Mellon University

STATISTICAL LEARNING METHODS FOR MULTI-OMICS DATA INTEGRATION IN DIMENSION REDUCTION, SUPERVISED AND UNSUPERVISED MACHINE LEARNING

SungHwan Kim, PhD

University of Pittsburgh, 2015

Abstract

Over the decades, many statistical learning techniques such as supervised learning, unsupervised learning, dimension reduction technique have played ground breaking roles for important tasks in biomedical research. More recently, multi-omics data integration analysis has become increasingly popular to answer to many intractable biomedical questions, to improve statistical power by exploiting large size samples and different types omics data, and to replicate individual experiments for validation. This dissertation covers the several analytic methods and frameworks to tackle with practical problems in multi-omics data integration analysis.

Supervised prediction rules have been widely applied to high-throughput omics data to predict disease diagnosis, prognosis or survival risk. The top scoring pair (TSP) algorithm is a supervised discriminant rule that applies a robust simple rank-based algorithm to identify rank-altered gene pairs in case/control classes. TSP usually generates greatly reduced accuracy in inter-study prediction (i.e., the prediction model is established in the training study and applied to an independent test study). In the first part, we introduce a MetaTSP algorithm that combines multiple transcriptomic studies and generates a robust prediction model applicable to independent test studies.

One important objective of omics data analysis is clustering unlabeled patients in order to identify meaningful disease subtypes. In the second part, we propose a group structured

integrative clustering method to incorporate a sparse overlapping group lasso technique and a tight clustering via regularization to integrate inter-omics regulation flow, and to encourage outlier samples scattering away from tight clusters. We show by two real examples and simulated data that our proposed methods improve the existing integrative clustering in clustering accuracy, biological interpretation, and are able to generate coherent tight clusters.

Principal component analysis (PCA) is commonly used for projection to low-dimensional space for visualization. In the third part, we introduce two meta-analysis frameworks of PCA (Meta-PCA) for analyzing multiple high-dimensional studies in common principal component space. Theoretically, Meta-PCA specializes to identify meta principal component (Meta-PC) space; (1) by decomposing the sum of variances and (2) by minimizing the sum of squared cosines. Applications to various simulated data shows that Meta-PCAs outstandingly identify true principal component space, and retain robustness to noise features and outlier samples. We also propose sparse Meta-PCAs that penalize principal components in order to selectively accommodate significant principal component projections. With several simulated and real data applications, we found Meta-PCA efficient to detect significant transcriptomic features, and to recognize visual patterns for multi-omics data sets.

In the future, the success of data integration analysis will play an important role in revealing the molecular and cellular process inside multiple data, and will facilitate disease subtype discovery and characterization that improve hypothesis generation towards precision medicine, and potentially advance public health research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 OVERVIEW OF HIGH-THROUGHPUT OMICS DATA

1.1.1 High-throughput data analysis

The system biological information flow is fundamentally rooted upon the central dogma paradigm from DNA to RNA, and RNA to Protein. This principle applies to all living creatures with exception of some simple organisms. DNA, encoding for genetic instructions, contains all necessary information for transcribing RNA transcripts, and hence functions as the structural sketch for cellular and bio-molecular mechanisms. The ground-breaking genome projects expedite deciphering genetic information of molecular organisms. Particularly, whole genome sequencing has become a bifurcation toward the biomedical research history. Such advances in genomic technologies have spurred technological orchestration among various disciplines of biomedical, physiological, and bio-chemical sciences, and facilitated the understanding of organismic functions, evolution and disease psychophysiology.

With the advance of modern bioinformatics, high-throughput technologies such as DNA microarrays, next generation sequencing (NGS) or mass spectrometry progressively produce abundant large genome-scale data with hundreds of thousands of features and large sample sizes. Over the past decades, genome profiling techniques have become viable at different levels of molecular cells and cellular organisms, including epigenome, transcriptome, metabolome, proteome, and interactome (Joyce et al., 2006). In addition, many other bioinformatics technologies, such as proteomics assays and imaging techniques (e.g. fMRI and PET scan) have been actively applied to support the biomedical research community. Since each type of omics data has its unique characteristics, techniques that can accommodate

differences of multiple types of omic data have been heavily sought beyond traditional bioinformatics analysis. Recently, multi-omics data integration analysis has been highlighted by considering to elucidate inter-regulatory flows and whole bio-molecular systems. It has also revolutionized the understanding of the complex molecular biology process and disease developments (Cancer Genome Atlas Research Network, 2012). Furthermore, with the advances of hight-throughput technologies and rapid drop of the genomic experimental cost, generation of genomic data has been exponentially increased. Several large-scale data depositories have been constructed for public access such as Gene Expression Omnibus (GEO), ArrayExpress and Sequence Read Archive (SRA), and The Cancer Genome Atlas (TCGA). The emerging of large scale multi-omic data provides great opportunities for data integration analysis in the future.

1.1.2 High-throughput omics data technologies

1.1.2.1 Microarrays have been widely utilized in most of biological research domain and produced various real applications for translational research (**?**). Particularly, microarray data have been analyzed, together with computational algorithms and machine learning techniques, in applications for drug discovery, biomarker detection, pharmacology, toxicogenomics, prognostic testing, population genomics and disease subtype identifications. In the microarray technology, tens of thousands of microarray probes are immobilized on a solid support, such as a microscope glass slide or silicon chips. Labeled target sequences bind to probes for identifying unknown sequences. mRNA is reversely transcribed, amplified and hybridized to cDNA templates. Several levels of mRNA bound to different sites on the array for expression profiles of thousands of genes, and/or even the whole genome. The microarray technique can also be applied to detect single nucleotide polymorphisms (SNPs), copy number variation (CNVs), DNA methylation, and protein-DNA binding.

Since bulk microarray data sets have been generated, public access of those datasets have been required in hope of routinely storing and openly sharing of the data resources in the public domain. The National Center for Biotechnology Information (NCBI) has managed Gene Expression Omnibus (GEO), where a multitude of gene expression data

sets are available. The Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov/tcga/) provides large-scale microarray datasets that can be downloaded via a public access. Up to date, TCGA has accumulated over 600 miroarray samples that are clinically annotated to primary breast cancer specimens. In this dissertation, several public microarray datasets are used to demonstrate our novel bioinformatics analysis methods such as robust prediction rules, coherent disease subtype identification, feature discovery, and dimension reduction for data visualization.

1.1.2.2 Next-generation sequencing (NGS) has been introduced in recent years, and mostly used in many biomedical applications (e.g., mutation discovery, meta-genomics, defining DNA-protein interactions, noncoding RNAs, and de-novo assembly of transcriptomic sequences, (Mardis, 2008). Next-generation sequencing is so-called ultra deep high-throughput that processes millions of sequence reads simultaneously. The workflow is to bind specific adapter oligos to both ends of each DNA fragment and then sequence the DNA fragments. Next-generation sequencers can generate short sequencing reads while read lengths can vary depending on user preference, technologies or platforms (e.g., Illumina1, SOLiD2, Roche). The generated short sequencing reads are aligned to a reference genome or transcriptome to quantify the expression levels of genes or transcripts by counting mapped short reads.

RNA-Seq is the most popular next-generation sequencing technology to quantify gene expression. RNA-Seq is an efficient way to produce gene-expression profiles, transcriptional structures of genes, and post-transcriptional modifications. Compared to the microarray technology, RNA-Seq has quite a few better properties such as high resolution, novel exons and genes detection, higher specificity and sensitivity with low background noise, no need for reference sequence, distinguishing isoforms and allelic expression (Wang et al., 2009), and accurately measuring the amounts of transcripts and their isoforms (alternatively spliced transcripts from the same gene). RNA-Seq can be flexibly extended to different types of analyses, for example, single nucleotide polymorphism discovery, alternative transcript identification, and gene expression profiling. TCGA projects also include thousands of primary tumor samples from more than 30 different tumor types in order to study underlying mechanism of malignant transformation and progression (http://tcga-data.nci.nih.gov/tcga).

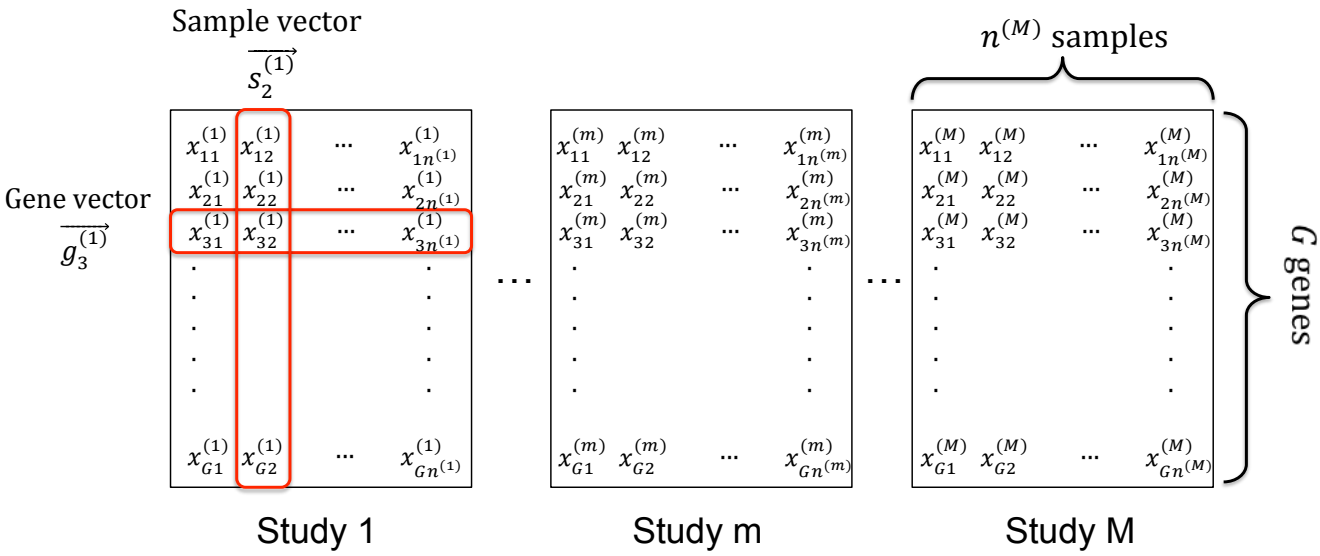

Figure 1: Data structure of multiple genomic studies

1.1.3 Data structure of omics study

In genomic data analysis, large-scale genomic data are generated under various conditions for different tissue samples. As shown in Figure 1, the data after proper pre-processing are in an expression matrix $D = \{x_{ij}^{(m)}\}$ $(1 \leq i \leq G, 1 \leq j \leq n^{(m)}, 1 \leq m \leq M)$, where the rows refer to expression features, the columns represent sample profiles, and $x_{ij}^{(m)}$ is the expression level for gene i in sample j of study m. $\overrightarrow{g_i}^{(m)} = \{x_{i1}^{(m)}, \ldots, x_{in^{(m)}}^{(m)}\}$ is the i^{th} gene vector that contains expression levels across all samples of study m. $\overrightarrow{s_j}^{(m)} = \{x_{1j}^{(m)}; \ldots, x_{Gj}^{(m)}\}$ is the j^{th} sample vector of expression levels across all gene features in study m. In microarray data, $x_{ij}^{(m)}$ is a log2 transformed of raw intensity or an intensity ratio as continuous values. For RNA-Seq data, $x_{ij}^{(m)}$ is the read count of gene i of subject j in study m. In this dissertation, we use these notations and the dataset structure unless explicitly described.

1.2 MACHINE LEARNING ANALYSIS ON HIGH-THROUGHPUT OMICS DATA

1.2.1 Major aims of statistical analysis in bioinformatics

The analyses of high-throughput data can be classified into two based on its major objectives. The first aim is to decipher underlying biological or disease development system through various omics data from different patient cohorts or/and various treatments. Exploitative and analytic methods such as differential expression (DE) analysis, clustering analysis, pathway analysis, and network analysis (Hawkins et al., 2010; Quackenbush, 2001) have played crucial roles in the identification of relations between bio-molecular units and clinical phenotype patterns (e.g., candidate biomarker detection, disease subtype identification and associated biological pathways) (Figure 2). This analytic trend has also revolutionized the target drug development, preventive disease procedures (Zografos et al., 2013) that will ultimately lead to "translational medicine" (Winslow *et al.*, 2012). For example, breast cancer developments has diverse patterns that depend on expressed marker genes related to Estrogen Receptor (ER)-positive or negative. Accurate biomarker detection closely links to the success of relevant clinical treatments and/or radio- or chemotherapy.

The second objective is to identify novel biomarker classifiers for clinical trial design and decision theory in many biomedical applications (Baek et al., 2009). The advent of prediction rules applicable to high-throughput omics data facilitates novel translational products such as disease diagnosis, prognosis prediction, treatment selection, preventative intervention, and precision medicine. However, high-throughput genomic, proteomic and metabolomic data brings new challenges in constructing robust prediction models. Model building with effective feature selection closely links to success in the development of a biomarker classifier. For this reason, the effort to detect biomarkers of disease development (translational products in the aim above) can closely related to developing accurate and feasible prediction models (Figure 2).

In Chapter 2, we propose "meta top scoring pairs (Meta-TSP)", a robust prediction model using the rank-order of paired genes. Meta-TSP can successfully function as a disease

prediction model. In Chapter 3, we develop "group structured tight integrative clustering (GST-iCluster)". We show that GST-iCluster can efficiently identify biologically relevant genes related to disease development mechanisms, and discover coherent disease subtypes. We expect these machine learning methods will significantly contribute to the community of the high-throughput omic data analysis.

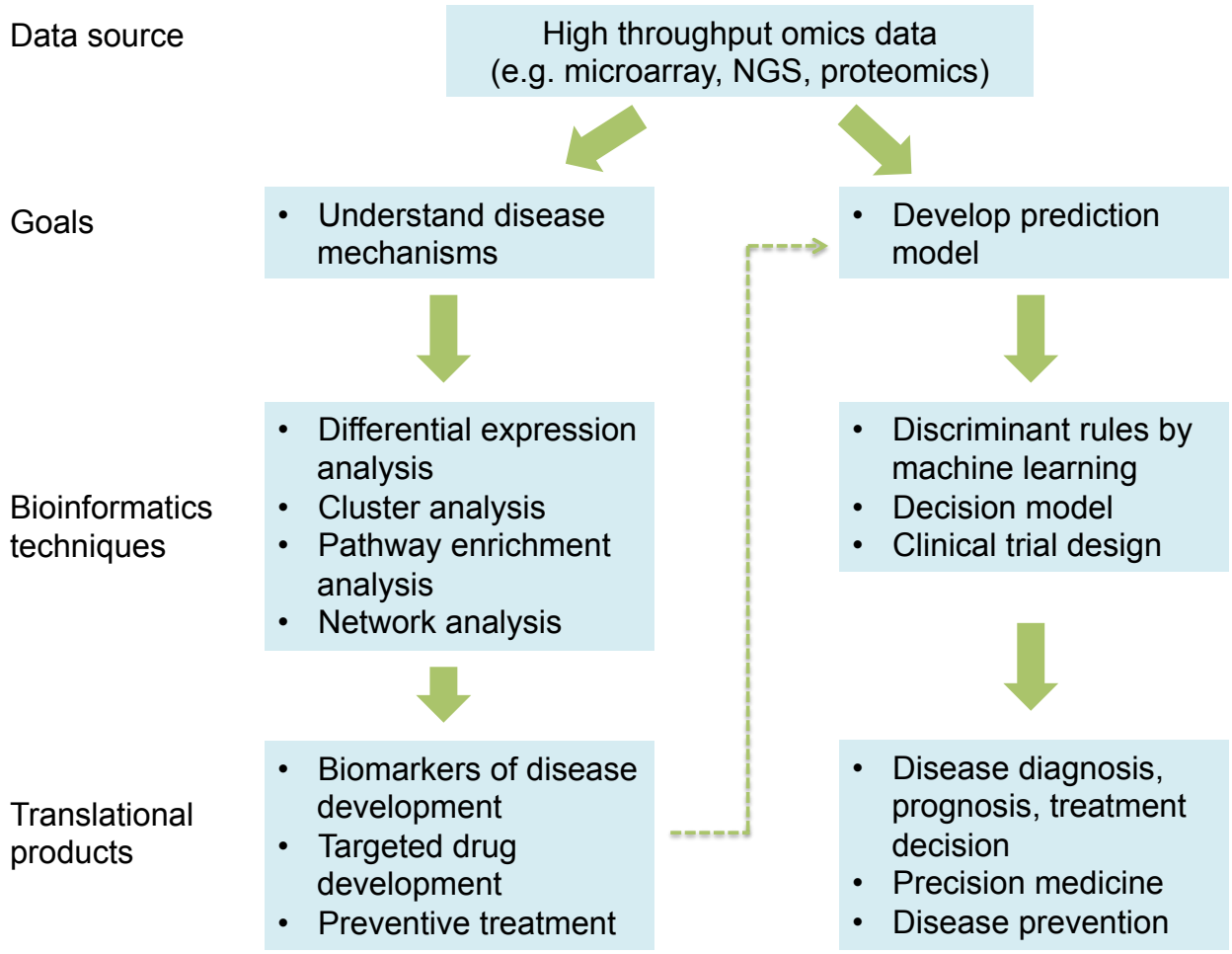

Figure 2: Overview of high-throughput data analysis.

1.2.2 Unsupervised learning on omics data

Unsupervised machine learning, aka clustering analysis, is a set of methods that do not rely on class label information, and separate samples into clusters under a predefined distance measure. By nature of unsupervised learning, it is intractable to evaluate its statistical property and performance due to the absence of so-called "gold standard". When performing clustering analysis, a distance or dissimilarity matrix measures the degree of closeness or separation of each pair of observations, and thereby a clustering algorithm can assign samples in proximity to each cluster. Many classical algorithms such as hierarchical clustering (Defays, 1977), K-means (Hartigan et al., 1979), self-organizing maps (Kohonen, 1982), Gaussian mixture model-based clustering (Banfield et al., 1993) and Bayesian clustering (Laua et al., 2007) have been developed. In addition, to control separation degrees of estimated clusters, a few cutting-edge methods have been proposed such as tight clustering (Tseng and Wong, 2005), penalized K-means (Tseng, 2007), and consensus clustering (Monti et al., 2003).

Consider a gene expression matrix of p genes and n samples. The data matrix can be viewed from row-wise (clustering genes) or column-wise (clustering samples) perspectives. When performing gene clustering, it is believed that highly correlated genes have a high chance of belonging to the same co-regulated systems and similar biological functions. Such gene cluster analysis generates gene modules that reveals relevant biological functions and evidences. In contrast, the problem of disease subtype discovery (clustering samples) has also been received wide attention. The purpose of subtype discovery is to cluster samples based on expression profiles in hope of identifying patient clusters with biologically (e.g. different pathway activation and disease progression mechanisms) and clinically (e.g. different drug response or survival) meaningful disease subtypes. For example, breast cancer was once thought as one type of disease. However, the well-known paper from Perous lab (Perou et al., 2010) applied hierarchical clustering in their microarray dataset and successfully identified five molecular breast cancer subtypes (Luminal A, Luminal B, Basal, Her2, and Normal-like) and demonstrated their biological and clinical relevance. This finding of novel disease subtypes will eventually become the fundamental basis for precision medicine.

1.2.3 Supervised discriminant analyses on omics data

Supervised machine learning has contributed to the advance of biomedical and clinical applications. In general, the task is to learn a classification model from training high-throughput data and predict the disease status or prognosis for incoming new patients. For instance, MammaPrint (Cardoso et al., 2007) established via supervised learning is a diagnostic tool to assess the metastatic risk of breast tumor based on the Amsterdam 70-gene breast cancer signature. MammaPrint measures a dichotomous risk using microarray profiles and samples from lymph node-negative breast cancers.

Here we introduce the generic data structure for supervised machine learning. Let G dimensional random vector $\overrightarrow{X} = (\overrightarrow{g_1}, \ldots, \overrightarrow{g_G})$ be the input data of population as covariates (e.g. the gene expression of G genes) and a random variable Y of values on $\{1, 2, \ldots, K\}$ as the class labels. In biomedical research, X can be high-throughput data that contain G features of clinical variables, gene expression levels, miRNA expression levels, protein expression levels, methylation intensities, SNPs/mutations. Y represents labels for different groups such as "disease vs control", "metastatic vs non-metastatic", "short patient survival vs long patient survival", "drug respondents vs non-respondents" or "multiple disease subtypes". The observed data as a whole comprise n patients: $D = ((y_1, \overrightarrow{s_1}), \ldots, (y_n, \overrightarrow{s_n}))$ where $(y_j, \overrightarrow{s_j}) \sim (Y, \overrightarrow{X})$ for $1 \leq j \leq n$. Using supervised learning techniques, a model is learned from the observed data D (including label information), and predicts new labels for future patients.

When applying machine learning techniques, it is essential to understand the motivations and details of each algorithm (e.g. data distribution assumption) to achieve accurate and interpretable results. The true distribution of data is typically unknown and is impossible to precisely estimate under high-dimensional settings due to "Curse of dimensionality" (refer to Section 1.3) . This problem has led to development of various machine learning methods based on different assumptions and types of data structure. To analyze high-throughput data, many popular machine learning methods have been proposed and applied, such as logistic regression, linear (quadratic) discriminant analysis, classification and regression tree (CART), random forest and support vector machines. There are also many fundamental

issues for better fitting machine learning models, e.g. cross-validation, feature selection and avoiding overfitting.

1.3 DIMENSION REDUCTION ON HIGH-THROUGHPUT OMICS DATA

With the advances in technology, high-dimensional data are now commonly generated in a wide range of research fields including genomics, signal processing, and financial risk management. The data analysis methods to deal with high dimensionality have been received increasing attentions as high-dimensional and large size data are accumulated over the years. The term of "Curse of dimensionality" introduced by Richard Bellman (Richard et al., 1957) refers to problems that occur under high-dimension of state variables in optimization problems (e.g. the computing complexity increases exponentially as the dimension increases). As a solution to this, he proposed a dynamic programming method for particular optimization problems. When fitting statistical models, the "curse of dimensionality" causes estimation to converge at a very slow rate. For example, the required sample sizes are only 4 and 19 for one or two dimensional space, whereas the required sample size rises to 842,000 if the dimension increases to 10. Another example is "concentration of measure" that influences the shape of a standard multivariate (d-dimensional) normal distribution. When d=1 or 2, the density concentrates to the origin, but when d is large, the distribution is concentrated on a d-dimensional sphere/shell with radius equals $\sqrt{d}$.

To circumvent high dimensionality problems, many dimension reduction techniques have been developed (e.g. Principal component analysis (PCA), multidimensional scaling (MDS), non-negative matrix factorization (NMF), etc.). In particular, the dimension reduction is suitable for high-throughput genomic data analysis, in which the signals of interest to differentiate groups tend to be in lower dimension subspace. Nevertheless, there are still many practical challenges of PCA method to deal with high-dimensional data. For example, noise features contained in most of large-scale microarray data often cause potential failure of dimension reduction (Hubert et al., 2005).

1.3.1 Principal component analysis (PCA)

Principal component analysis (PCA) has been one of the most popular data-processing and dimension reduction technique in multivariate analysis. It is particularly suitable to discover low-dimensional signals for high-dimensional data. In the setting of small-p and large-n, the estimated principal components of the covariance matrix are shown to be consistent as the sample size n increases when p fixed. For high-throughput data analysis, PCA has been applied to gene expression data (Alter et al., 2000). For example, the "gene shaving" technique (Hastie et al., 2000) uses PCA to cluster highly variable and coherent genes in microarray datasets. In spite of its advantages, PCA has several fundamental flaws.

Several experimental studies (Baik et al., 2006) show that the sample principal component is inconsistent with the principal component of whole population. For example, the high-dimensional setting of large-p and small-n causes very poor estimates. In addition, sample principal eigenvectors generally have nonzero loading values for each coordinate component. This drawback results in low interpretability as the dimension p increases. To overcome this issue, we introduce a meta analytic framework for principal components (Meta-PCA) in Chapter 4. Meta-PCA is designed to discover the best common eigenvector space, and is less sensitive to the effect of noise samples and features than single PCA method.

1.3.2 Regularized principal component analysis (Sparse PCA)

Sparse PCA is designed to overcome the aforementioned shortcomings of PCA, especially, the variable selection problem in high dimensional eigenvectors. In theory, PCA holds two beneficial properties: (1) the leading principal components minimize information loss (maximized variability);(2) Principal components are projected into perpendicular subspaces. However, small but non-zero loadings from many features in eigenvectors often act as a major barrier to interpret estimated principal components. One potential way to increase the interpretability of principal component (PC) is to apply regularization (i.e., penalization) over leading eigenvector components. This approach is commonly called "sparse PCA", and various sparse PCA methods have been proposed in the literature (Hoyle et al., 2004; dAspremont et al., 2007; Journ ee et al., 2010; Shen et al., 2008; Ulfarsson et al., 2008; Jolliffe et al.,

2003; Witten et al., 2009; Zou et al., 2006). Jolliffe et al. (2003) introduced SCoTLASS to estimate principal components with possible zero loadings. Zou et al. (2006) proposed sparse PCA (SPCA) based on a regression-type optimization via the elastic net incorporating the regression technique. Similar to SPCA, Witten et al. (2009) developed sparsePCA that exploits the penalized matrix decomposition (PMD) using SVD approximated matrix to minimize errors to the original observed matrix. In Chapter 5, we develop several sparse Meta-PCAs to improve the proposed Meta-PCA's interpretability. Various numerical examples have shown the sparse Meta-PCAs outperform Meta-PCA in favor of efficient and distinctive visualization in low dimension space.

1.4 MULTI-OMICS DATA INTEGRATION ANALYSIS

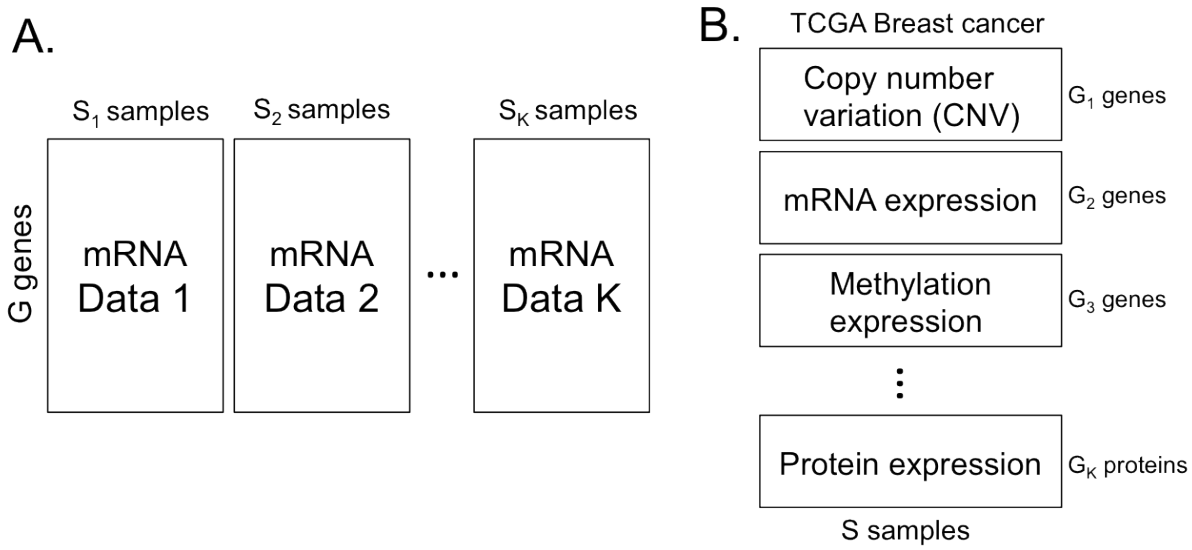

Figure 3: Two major types of omics data integration (A) Horizontal omics meta-analysis to combine K transcriptomic datasets (B) Vertical omics integrative analysis to combine different omics data in a given cohort.

1.4.1 Horizontal omics data integration (Meta-analysis)

In the past two decades, high-throughput experimental techniques have revolutionized biomedical research with large genome-scale data. The fruitful successes of this research are paving the way towards better drug targets and precision medicine. Such "big" datasets are routinely generated now that the cost has been greatly decreased. As a result, abundant datasets are available in the public domain. For example, as of 9/16/2014, GEO contains 1,234,880 samples and SRA has >2,826 terabases of sequencing data. Effective integrative methods are urgently required to decipher the biological information inside these data, leading to a better understanding of disease mechanisms. Omics integrative methods are commonly divided into two major categories. Due to high experimental cost, and/or limitation of clinical tissue access, individual labs usually generate omics datasets with small to moderate sample sizes (e.g. n=40-100). Statistical power and reproducibility of such small studies has long been a concern in this field (Simon et al., 2003; Simon, 2005; Domany, 2014). An increasingly popular solution is to search the literature, seek similar datasets (of similar design and biological hypothesis) and perform data integration. In this context, the analytic questions and methods are analogous to traditional meta-analysis (Ramasamy et al., 2008; Tseng *et al.*, 2012). Since the microarray boom of the late 90s, a convention has been developed in which genes on the rows and samples on the columns. As a result, multi-study data integration is often called "horizontal omics meta-analysis" since datasets are laid out horizontally (Figure 3A). The horizontal meta-analysis methods can conceptually be applied to other types of omics data, such as GWAS, mRNA expression, methylation, miRNA, copy number variation and protein expression. In particular, horizontal meta-analysis methods are useful and practical for individual labs mostly generating data of moderate sample size. Successful development of horizontal meta-analysis methods for subtype discovery and characterization can greatly enhance knowledge of finding and generating hypotheses towards precision medicine. In this dissertation, we introduce two methods on horizontal data integration analysis. In Chapter 2, we introduce "meta top scoring pairs (Meta-TSP)", a robust prediction model of paired genes. Meta-TSP integrates multiple studies when the prediction model is fitted, so Meta-TSP generates high accuracy especially in inter-study prediction. In Chapter 4, we propose

a meta-analytic framework for principal component analysis (Meta-PCA). Single PCA study tends to be sensitive to the effects of noise samples and features in high-dimensional data. On the contrary, Meta-PCA aims to form common eigenvector space that can capture variations of multiple studies in parallel. Due to nature of common eigenvector space, Meta-PCA is robust to noise features and outlier samples.

1.4.2 Vertical omics data integration

In contrast to horizontal meta-analysis, many large consortia (e.g. the Cancer Genome Atlas (TCGA) and the Lung Genomics Research Consortium (LGRC)) have started to generate multiple different types of -omics data using samples in a single cohort, including SNP genotyping, mutation, copy number variation, mRNA expression, miRNA expression and protein expression. The integration of the multi-omics data for understanding the inter-omics interaction mechanisms is challenging in statistical problem. The datasets are aligned vertically (Figure 3B) and thus, the integration of such multi-omics data is called "vertical omics integrative analysis". While integration of multiple omics data sources on the same cohort provides great insight into the molecular and cellular processes of the disease and has become popular, the problem brings new analytical challenges and it is in need to develop of statistical methods in this field.

Similar to traditional microarray data analysis, vertical omics integration can target on the following biological objectives: (i) candidate marker detection (Wang et al., 2008); (ii) gene set or pathway analysis (Hu et al., 2014); (iii) dimension reduction (Lock et al., 2013A; Li et al., 2012; Zhang et al., 2011); (iv) classification (Setty et al., 2014); and finally (v) clustering analysis. Using multi-omics data sources, several methods for disease subtype discovery using vertical omics integration have been proposed. Rey and Roth (2012) introduced a copula mixture model for dependency-seeking clustering of multi-omics data. Lock et al. (2013B) proposed a Bayesian consensus clustering to account for consensus and source-specific information in the cluster formation.

Shen et al. (2009, 2013) developed an integrative clustering approach (iCluster) via a Gaussian latent regression model. iCluster has several advantages that brought it pop-

ularity in many, particularly cancer, applications. Multi-omics integrative clustering has the focus to identify disease subtypes. For n subjects, suppose we have M different omics datasets. Let $X^{(m)}$ be the m^{th} dataset with p_m features, where each column of $X^{(m)}$ consists of mean-centered features of n subjects ($1 \leq m \leq M$). The combined dataset is $X = \left(X^{(1)^T}, X^{(2)^T}, \ldots, X^{(M)^T}\right)^T$, where X is a $\sum_{m=1}^{M} p^{(m)} \times n$ matrix, and $X^{(m)}$ is a $p^{(m)} \times n$ matrix. The joint latent regression model is

$$X^{(m)} = B^{(m)} Z + E^{(m)} \quad \text{for } 1 \leq m \leq M,$$

where Z is a $\ell \times n$ matrix whose rows are latent variables and columns are samples. The matrix $B^{(m)}$ is used to control the degree of relation between feature intensities and latent variables (usually $l << p^{(m)}, \forall m$). To achieve the sparse estimation of $B^{(m)}$, the expectation-maximization (EM) algorithm (Dempster et al., 1977) is applied to estimate $\hat{B}$ and $\hat{\psi}$, together with a L_1-lasso penalty (Tibshirani, 1996). Once we estimate the latent variable matrix Z, the standard k-means clustering is applied to Z with respect to samples to produce the integrative clusters.

1.5 OVERVIEW OF THE DISSERTATION

This dissertation covers three major data integration analyses: (1) robust prediction in a meta-analytic framework (horizontal integration); (2) coherent and tight integrative clustering specialized in feature gene discovery (vertical integration); (3) meta-analytic framework of dimension reduction for visualization (horizontal integration). In Chapter 2, we introduce a MetaTSP algorithm that combines multiple transcriptomic studies and generates a robust prediction model applicable to independent test studies. The top scoring pair (TSP) algorithm is a supervised discriminant rule by applying a robust simple rank-based algorithm. TSP exhaustively explores rank-altered gene pairs in case/control classes but often suffers from low accuracy in inter-study prediction (i.e. the prediction model is established in the training study and applied to an independent test study). With comprehensive applications and simulated data, the performance of MetaTSP is shown to outperform single study TSP.

In Chapter 3, we propose a group structured iCluster together with a sparse overlapping group lasso technique via regularization to incorporate information of inter-omics regulation flow, and also applying a tight clustering concept to form tight clusters by scattering outlier samples away. This integrative clustering (unsupervised) method can identify meaningful disease subtypes and biologically associated gene modules. We show by two real examples and simulated data that our proposed methods improve the original iCluster in clustering accuracy, biological interpretation, and are able to generate coherent tight clusters. In Chapter 4, we introduce two meta-analysis frameworks of PCA (Meta PCA) for analyzing multiple high-dimensional studies in common principal component space. Meta PCA aims to identify the best common PC space. Applications to various simulated data show that Meta PCA is able to find the true principal component space, and retains robustness on noise features and outlier samples. In Chapter 5, we further propose several sparse Meta PCA methods that can regularize principal components, and facilitate feature identifications and visual pattern recognition for multiple omics datasets.

2.0 METAKTSP: A META-ANALYTIC TOP SCORING PAIR METHOD FOR ROBUST CROSS-STUDY VALIDATION OF OMICS PREDICTION ANALYSIS

2.1 INTRODUCTION

High-throughput experimental techniques, including microarray and massively parallel sequencing, have been widely applied to discover underlying biological processes and to predict the multi-causes of complex diseases (e.g., cancer diagnosis, (Ramaswamy *et al.*, 2001), prognosis (van de Vijver *et al.*, 2002), and therapeutic outcomes (Ma *et al.*, 2004)). The associated data analysis has brought new statistical and bioinformatics challenges and many new methods have been developed in the past 15 years. In particular, methods for classification and prediction analysis (a.k.a. supervised machine learning) are probably the most relevant tools towards translational and clinical applications. Take breast cancer as an example, many expression-based biomarker panels have been developed (e.g. MammaPrint (van 't Veer *et al.*, 2002), Oncotype DX (Paik *et al.*, 2004), Breast Cancer Index BCI (Zhang *et al.*, 2013) and PAM50 (Parker *et al.*, 2009)) for classification/prediction of survival, recurrence, drug response and disease subtype. Reproducibility analysis of these markers and classification models has been a major concern and has drawn significant attention to ensure clinical applicability of these panels (Garrett-Mayer *et al.*, 2008; Kuo *et al.*, 2006; MAQC Consortium *et al.*, 2006; Mitchell *et al.*, 2004; Sato *et al.*, 2009). Many papers have focused on normalization, reproducibility of marker detection, inter-lab or inter-platform correlation concordance. For direct clinical utilities, more attention have shifted towards cross-study situation or inter-study prediction (i.e. a prediction model is established in one study and validated independently in a test study (Xu *et al.*, 2008; Cheng *et al.*, 2009; Mi et al., 2010;

Bernau *et al.*, 2014)). Such an issue is critical for translating models from transcriptomic studies into a practical clinical tool. For example, the training cohort may have utilized an old Affymetrix U133 platform. A biomarker panel and a model are constructed and a test study from a different medical center using an RNA-seq platform is available. A successful machine learning model should retain high prediction accuracy in such inter-lab and inter-platform validation. We note that many normalization methods have been developed to adjust for systematic biases across studies, including distance weighted discrimination (DWD, (Benito *et al.*, 2004)), cross-platform normalization (XPN, (Shabalin *et al.*, 2008)) and Knorm correlation (Teng *et al.*, 2007). But the normalization performance largely depends on whether the observed data structure fits the model assumptions. In most applications, researchers have often applied meta-analysis methods instead of normalization and data merging (Tseng *et al.*, 2012). Similarly, we will not consider normalization and data merging approach (a.k.a. mega-analysis).

In addition to the issue of cross-study validation, selection of a robust and accurate machine learning method is also critical. In the literature, many supervised machine learning methods have been proposed and applied to high-throughput experimental data. For example, the CMA package allows easy implementation of 21 popular classification methods such as linear or quadratic discriminant analysis, lasso, elastic net, support vector machines, random forest, PAM, etc (Slawski *et al.*, 2008). Most of these parametric and model-based methods potentially can suffer from heterogeneity across platforms and limit the feasibility and reproducibility in the cross-study validation. In addition to these popular methods, the top scoring pair (TSP) method (Geman *et al.*, 2004; Tan *et al.*, 2005; Afsari *et al.*, 2014) is a straightforward prediction rule utilizing building blocks of rank-altered gene pairs in case and control comparison (see Section 2.1 for more details). The method is rank-based without any model parameter. It is invariant to monotone data transformation that relieves from normalization necessity, and the feature selection and the model are more transparent for biological interpretation. Although TSP and its variant are robust methods that do not require normalization in cross-study validation, we have found that some of the selected TSPs from the training study may not reproduce in the test study and appear to be false positives.

Here we consider four Idiopathic pulmonary fibrosis(IPF) studies (see Table 2). Figure 1A illustrates the expression levels of a good TSP gene pair, CBS and MOXD1, identified from the first IPF training study Emblom XBP1 is more over-expressed than ITGAX in control samples but under-expressed in cases. If we use this TSP to validate in the test study Konishi, we find that XBP1 is over-expressed than ITGAX in both cases and controls and we obtain 0% sensitivity and 100% specificity (i.e. Youden index = sensitivity + specificity - 1 = 0). We find similar poor performance in two other studies Tedrow B and Pardo, showing that the TSP is likely a false positive. In Figure 1B, GPR160 is over-expressed than COMP in controls and under-expressed in cases for all three studies Emblom, Tedrow B and Pardo. It is a more reliable TSP across three studies and conceptually is less likely a false positive. Indeed, the cross-study validation in Konishi shows good performance with 80% Youden index. The two real examples in Figure 1 argue the potential of a meta-analytic approach by combining multiple training transcritomic studies to identify reliable TSPs so the resulting model has enhanced cross-study validation performance.

2.2 METHODS

2.2.1 Top scoring pair algorithm (TSP) and *k*TSP

The original TSP algorithm was first proposed by Geman *et al.* (2004). Denote by data matrix $X = \{x_{gn}\}$ the gene expression intensity of gene g ($1 \leq g \leq G$) in sample n ($1 \leq n \leq N$) and y_n the class label of sample n. Particularly, we consider $y_n \in \{0, 1\}$, representing controls and cases for binary classification. For any gene pair i and j ($1 \leq i, j \leq G$), define the conditional ordering probability score $T_{ij}(C) = Pr(X_i < X_j | Y = C)$ for $C \in \{0, 1\}$, where X_i and X_j are gene expression intensities of gene i and j. Intuitively, $T_{ij}(0)$ is the probability in controls that gene j has larger expression intensity than that of gene i and similarly $T_{ij}(1)$ is for cases. Given observed expression profile data matrix X, the probability scores can be estimated as $\hat{T}_{ij}(C) = \Big(\sum_{n=1}^{N} I(x_{in} < x_{jn}) \cdot I(y_n = C) \Big) / \Big(\sum_{n=1}^{N} I(y_n = C) \Big)$, where $I(\cdot)$ is an indicator function that is one if the statement inside the parenthesis is true and zero oth-

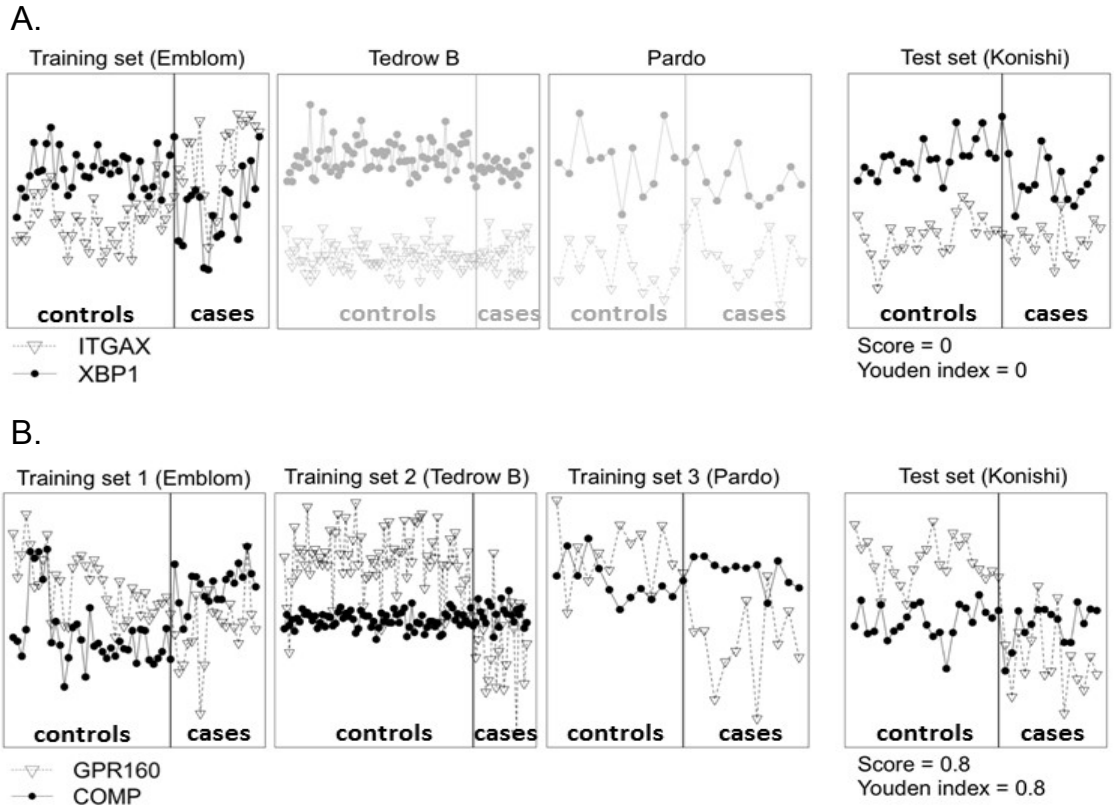

Figure 4: Two TSP examples from real data to show advantage of MetaTSP. X-axis and Y-axis refer to sample indices and gene expression levels, respectively. (A) Gene pair ITGAX/XBP1 has high TSP score (XBP1>ITGAX in controls but ITGAX>XBP1 in cases) in the training 'Emblom' study but fail to replicate in the testing 'Konishi' study as well as the other two Tedrow B and Pardo studies. (B) Gene pair GPR160/COMP has high TSP scores (GPR160>COMP in controls and COMP>GPR160 in cases) in all three training studies 'Emblom', 'Tedrow B' and 'Pardo'. The gene pair is successfully validated in the testing 'Konishi' study.

erwise. The discriminant score of the gene pair is defined as $S_{ij} = \hat{T}_{ij}(1) - \hat{T}_{ij}(0)$. Note that $-1 \leq S_{ij} \leq 1$ always holds. When $S_{ij} = 1$, expression of gene j is always greater than that of gene i in cases and expression of gene j is always smaller than that in gene i among controls. As a result, the ordering of gene i and gene j expression is predictive to the class label. On the contrary, if $S_{ij} = -1$, gene j always has smaller expression than gene i in cases and the relation is reversed in controls. In summary, the absolute value of S_{ij} reflects the predictive value of the gene pair. The TSP algorithm seeks the best gene pair $(i^*, j^*) = \arg\max_{i \neq j} |S_{ij}|$ as the classifier. When multiple gene pairs give the same highest absolute score, the best pair that gives the largest differential magnitude D_{ij} is chosen, where $D_{ij} = |d_{ij}(1) - d_{ij}(0)|$ and $d_{ij}(C) = \Big(\sum_{n=1}^{N} (x_{in} - x_{jn}) \cdot I(y_n = C) \Big) / \Big(\sum_{n=1}^{N} I(y_n = C) \Big)$. When a new test sample $\overrightarrow{x}^{(test)} = \left(x_1^{(test)}, \cdots, x_G^{(test)}\right)$ is entered in the future, the class prediction is determined by

$$\hat{C}_{i^*j^*}(\overrightarrow{x}^{(test)}) = \begin{cases} 1, & if \quad S_{i^*j^*} \cdot \left(x_{i^*}^{(test)} - x_{j^*}^{(test)} \right) \leq 0 \\ 0, & if \quad S_{i^*j^*} \cdot \left(x_{i^*}^{(test)} - x_{j^*}^{(test)} \right) > 0 \end{cases}$$

TSP classifier above is based on only one top scoring pair (two genes) and so the method can be very sensitive to slight noise perturbations (Geman *et al.*, 2004). To circumvent this issue, Tan *et al.* (2005) introduced kTSP to combine multiple TSPs for a more stable algorithm. The method identified the sorted TSPs similar to above. Instead of choosing only the best TSP, it selected the top K (where K is a parameter to be tuned) TSPs to construct the model. The TSPs were selected from the sorted list such that the genes in the TSPs had no overlap otherwise the latter TSPs containing overlapping genes would be skipped and the next TSP in the sorted list would be considered. In other words, the selected top K TSPs always contain $2K$ distinct genes. Suppose $\left\{(i_1^*, j_1^*), \cdots, (i_K^*, j_K^*)\right\}$ represents the K selected TSPs. The kTSP algorithm makes a prediction for a new test sample $\overrightarrow{x}^{(test)}$ by $\hat{C}(\overrightarrow{x}^{(test)}) = \arg\max_C \sum_{k=1}^{K} I\big(\hat{C}_{i_k^* j_k^*}(\overrightarrow{x}^{(test)}) = C\big)$. In a sense, the k-TSP is an ensemble classifier that aggregates multiple weak classifiers by majority vote (Opitz *et al.*, 1999). To avoid ties, we usually select odd numbers for K.

The TSP algorithms have the following advantages for omics prediction analysis: (1) The method is non-parametric and thus robust since the method is constructed based on the

relative ranking of gene pairs. Since different transcriptomic studies are usually conducted in different labs and in different platforms, the robust nonparametric nature is more likely to succeed in cross-study validation that we aim in this dissertation. (2) The method is based on one or a few gene pairs. The biological interpretation of the model and the translational application are more straightforward. It is more likely to succeed by designing a reproducible commercial assay for wider clinical applications, such as the 21-gene RT-PCR-based Oncotype DX test for breast cancer (Paik *et al.*, 2004). (3) Researchers have repeatedly found that the family of TSP algorithms provides good prediction performance in many transcriptomic data (Xu *et al.*, 2005; Raponi *et al.*, 2001; Price *et al.*, 2007).

2.2.2 Estimate K for kTSP

To estimate the best K in the kTSP algorithm, we can apply and compare the following two methods.

Cross-validation In Tan *et al.* (2005), leave-one-out cross validation was used to determine K in kTSP. In each iteration, one sample was left out as the test sample. The remaining samples were used to construct a prediction model and apply to the test sample. The procedure was repeated until each sample was left out as the test sample once. The cross-validated error rates were then calculated for different selections of K and the best K that produced the smallest cross validation error rate was chosen.

Variance optimization Afsari *et al.* (2014) recently developed a variance optimization method to estimate K in kTSP. Recall that $S_{ij} = Pr(X_i < X_j|Y = 1) - Pr(X_i < X_j|Y = 0)$. The kTSP algorithm searches for the optimized top scoring pairs without overlapping genes:

$$\left\{(i_1^*, j_1^*), \cdots, (i_K^*, j_K^*)\right\} = \arg\max_{\{(i_1,j_1),\cdots,(i_K,j_K)\}} \sum_{k=1}^{K} S_{i_k j_k}.$$

Define the t-statistics of the target function:

$$t_{kTSP}(K) = \frac{\sum_{k=1}^{K} S_{i_k^* j_k^*}}{\sqrt{Var\left(\sum_{k=1}^{K} I(X_{i_k^*} < X_{j_k^*})|Y=0\right) + Var\left(\sum_{k=1}^{K} I(X_{i_k^*} < X_{j_k^*})|Y=1\right)}}.$$

K is chosen by the value that maximizes t_{kTSP} $\left(\text{i.e. } K^* = \arg\max_K t_{kTSP}(K)\right)$. The variance optimization procedure greatly reduced high computational demand in cross validation.

2.2.3 Meta-kTSP algorithms

As mentioned in the introduction section, cross-study validation via Mega-kTSP (i.e. naively combine multiple normalized data sets and apply kTSP) may not be suitable to identify a robust prediction gene pair. Alternatively, we propose a Meta-kTSP framework below. Denote by $X^{(m)} = \left\{x_{gn}^{(m)}\right\}$ the expression profile of study m, where $x_{gn}^{(m)}$ represents the gene expression intensity of gene g $(1 \leq g \leq G)$, sample n $(1 \leq n \leq N^{(m)})$ in study m $(1 \leq m \leq M)$. The discriminant score $S_{ij}^{(m)}$ for gene i and j in study m takes the difference of two summations of Bernoulli random variables:

$$\begin{aligned} S_{ij}^{(m)} =& \frac{\left(\sum_{n=1}^{N^{(m)}} I(x_{in}^{(m)} < x_{jn}^{(m)}) \cdot I(y_n^{(m)} = 1)\right)}{N_1^{(m)}} \\ & - \frac{\left(\sum_{n=1}^{N^{(m)}} I(x_{in}^{(m)} < x_{jn}^{(m)}) \cdot I(y_n^{(m)} = 0)\right)}{N_0^{(m)}}, \end{aligned}$$

where $N_1^{(m)} = \sum_{n=1}^{N^{(m)}} I(y_n^{(m)} = 1)$ and $N_0^{(m)} = \sum_{n=1}^{N^{(m)}} I(y_n^{(m)} = 0)$ are the number of case and control samples in study m. We first develop three meta-analytic approaches (by Fisher score, Stouffer score and mean score) to choose the K non-overlapping top scoring pairs (TSPs) for prediction model construction (denoted as $\left\{(i_1^*, j_1^*), \cdots, (i_K^*, j_K^*)\right\}$). When a new test sample, $\overrightarrow{x}^{(test)} = \left(x_1^{(test)}, \cdots, x_G^{(test)}\right)$ is entered in the future, the class prediction by the k^{th} TSP and study m is:

$$\hat{C}_{i_k^*, j_k^*}^{(m)}(\overrightarrow{x}^{(test)}) = \begin{cases} 1, & if \quad S_{i_k^*, j_k^*}^{(m)} \cdot \left(x_{i_k^*}^{(test)} - x_{j_k^*}^{(test)}\right) \leq 0 \\ 0, & if \quad S_{i_k^*, j_k^*}^{(m)} \cdot \left(x_{i_k^*}^{(test)} - x_{j_k^*}^{(test)}\right) > 0. \end{cases}$$

The final meta-analyzed class prediction is determined by

$$\hat{C}(\overrightarrow{x}^{(test)}) = \arg\max_C \sum_{m=1}^{M} \sum_{k=1}^{K} I\left(\hat{C}_{i_k^*, j_k^*}^{(m)}(\overrightarrow{x}^{(test)}) = C\right).$$

Below we introduce the three meta-analytic approaches to select the top K TSPs. In meta-analysis, test statistics (e.g. t-statistics) across studies are not comparable and combining p-values has become a popular practice. Under the null hypothesis that gene i and j are not discriminant, $S_{i,j}^{(m)}$ can be well-approximated by Gaussian distribution $S_{i,j}^{(m)} \Big/ \sqrt{\frac{0.25}{N_1^{(m)}} + \frac{0.25}{N_0^{(m)}}} \sim$

$N(0,1)$ since $S_{ij}^{(m)}$ is the difference of two summations of independent Bernoulli trials. The two-sided p-value of $S_{ij}^{(m)}$ is calculated as $P_{ij}^{(m)} = 2 \times \left(1 - \Phi\left(\left|S_{ij}^{(m)}\right| \Big/ \sqrt{\frac{0.25}{N_1^{(m)}} + \frac{0.25}{N_0^{(m)}}}\right)\right)$. Alternatively, one-sided p-values can be calculated as $P_{ij}^{(m);L} = \Phi\left(S_{ij}^{(m)} \Big/ \sqrt{\frac{0.25}{N_1^{(m)}} + \frac{0.25}{N_0^{(m)}}}\right)$ for left-sided p-values and $P_{ij}^{(m);R} = 1 - \Phi\left(S_{ij}^{(m)} \Big/ \sqrt{\frac{0.25}{N_1^{(m)}} + \frac{0.25}{N_0^{(m)}}}\right)$ for right-sided p-values.

Select K TSPs by Fisher's method

The Fisher's method combines p-values across studies by $T_{ij}^{(Fisher)} = -2 \times \sum_{m=1}^{M} log\left(P_{ij}^{(m)}\right)$, where $P_{ij}^{(m)}$ is the two-sided p-value of the discriminant score $S_{ij}^{(m)}$ of gene i and j in study m. Under null hypothesis that gene i and j have no discriminant power in all studies, $T_{ij}^{(Fisher)} \sim \chi_{2M}^2$. This classical p-value combination procedure has a well-known problem that the discriminant scores across studies may have discordant signs but all with small two-sided p-values that generate a significant meta-analyzed p-value. To circumvent this discordant problem, we apply a one-sided test modification technique discussed in Owen (2009). Define $T_{ij}^{(Fisher);L} = -2 \times \sum_{m=1}^{M} log\left(P_{ij}^{(m);L}\right)$ and $T_{ij}^{(Fisher);R} = -2 \times \sum_{m=1}^{M} log\left(P_{ij}^{(m);R}\right)$, where $P_{ij}^{(m);L}$ and $P_{ij}^{(m);R}$ are the left and right one-sided p-values of discriminant score $S_{ij}^{(m)}$ of gene i and j in study m. The modified one-sided corrected Fisher's statistic is $T_{ij}^{(Fisher);OC} = max\left(T_{ij}^{(Fisher);L}, T_{ij}^{(Fisher);R}\right)$. The top K gene pairs with the largest meta-analyzed Fisher score (i.e. $T_{ij}^{(Fisher);OC}$) and with no overlapping genes are selected.

Select K TSPs by Stouffer's method

Instead of using log-transformation in Fisher's method, Stouffer's method applies an inverse normal transformation by $T_{ij}^{(Stouffer)} = \sum_{m=1}^{M} \Phi^{-1}\left(P_{ij}^{(m);L}\right) / \sqrt{M}$ Under null hypothesis that gene i and j have no discriminant power in all studies, $T_{ij} \sim N(0,1)$. The top K gene pairs with the smallest meta-analyzed two-sided p-values and with no overlapping genes are selected for prediction. Note that Stouffer's method has an advantage over Fisher's method that one-sided concordance correction is not necessary if one-sided p-values are input in the inverse normal transformation.

Select K TSPs by mean score Since the discriminant score is difference of two conditional probabilities, the scores are directly comparable across studies and can be directly combined. We define the mean score $T_{ij}^{(mean)} = \sum_{m=1}^{M} S_{ij}^{(m)} / M$ to combine M studies. The top K gene pairs with the largest absolute value of the meta-analyzed scores (i.e. $\left|T_{ij}^{(mean)}\right|$) and with no overlapping genes are selected for prediction model construction.

2.2.4 Estimate K for Meta-kTSP

Similar to Section 2.2, cross-validation and variance optimization methods can be extended to estimate K for Meta-kTSP.

Cross-validation We perform V-fold cross-validation for Meta-kTSP. Each of the M studies are firstly split into V equal-sized subgroups. In each cross-validation, one subgroup of samples in each study is left out as the testing samples. The remaining $(V-1)$ subgroups are used as training samples to construct the classifier and then apply to the test sample. The procedure is repeated for V times until all samples are left out and tested. We choose the optimal K such that the highest average Youden index over M studies is obtained. We adopted 5-fold cross-validation.

Variance optimization Similar to single study kTSP algorithm in Afsari *et al.* (2014), we define the following target function:

$$t_{kTSP}^{(meta)}(K) = \frac{\sum_{m=1}^{M}\sum_{k=1}^{K} S_{i_k^* j_k^*}^{(m)}}{\sqrt{Var\Big(\sum_{m=1}^{M}\sum_{k=1}^{K} I(X_{i_k^*}^{(m)} < X_{j_k^*}^{(m)}) | Y=0\Big) + Var\Big(\sum_{m=1}^{M}\sum_{k=1}^{K} I(X_{i_k^*}^{(m)} < X_{j_k^*}^{(m)}) | Y=1\Big)}}.$$

K is chosen by the value that maximizes $t_{kTSP}^{(meta)}(K)$ (i.e. $K^* = \arg\max_K t_{kTSP}^{(meta)}(K)$). The variance optimization procedure greatly reduced computational complexity in cross validation. We will show its equal or slightly improved performance compared to cross validation in our proposed meta-analytic scheme and this estimation method will be recommended in practice.

2.3 RESULTS

2.3.1 Simulations

We hypothesize that if gene pairs are consistently identified with strong TSP scores over multiple training studies, such gene pairs outperform original TSPs from a single study. We tested this hypothetical argument using simulated data sets. Below we describe simulated expression profiles under correlated gene structures to mimic real data sets. We performed a smaller scale of simulation with $G = 200$ genes and $M = 4$ transcriptomic studies, where

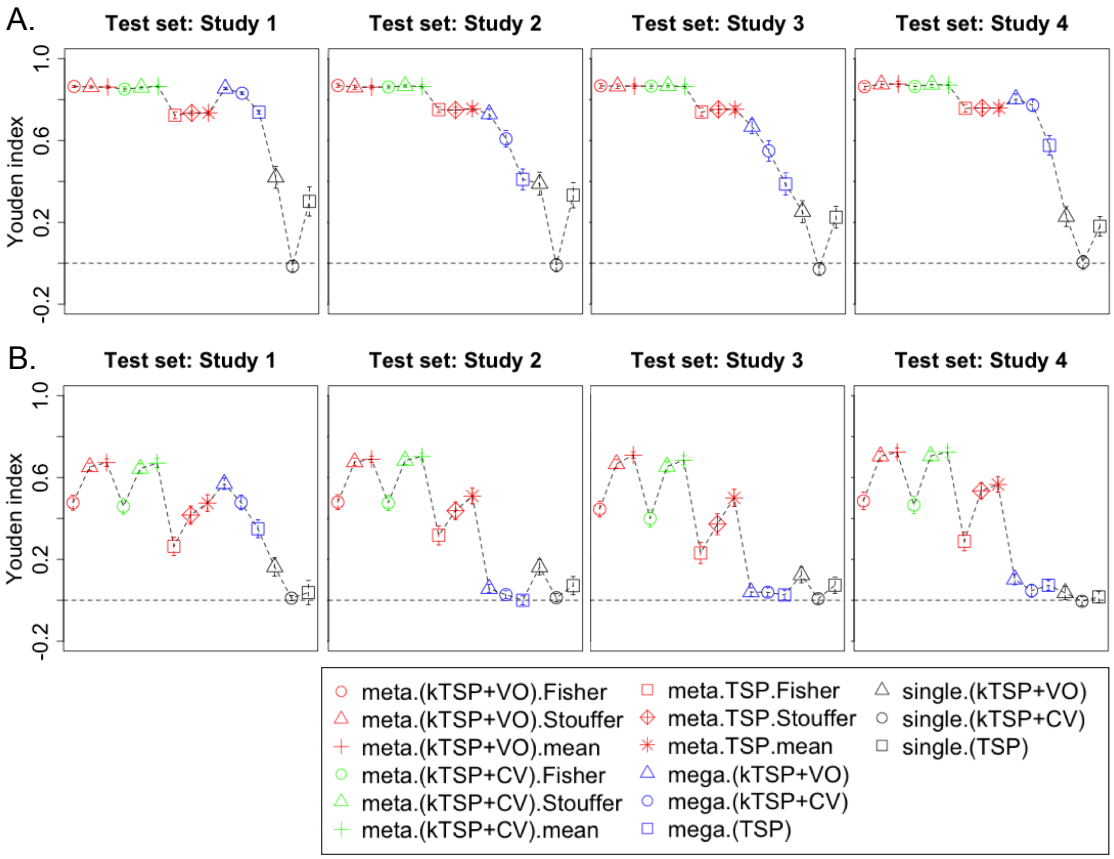

Figure 5: Results of inter-study prediction using four simulated data sets (A: $\mu_a = 1$, B: $\mu_a = 0.8$; $n_1^1 = n_1^2 = 100$). Y-axis represents the average Youden index. The bar plots indicate the standard error of estimated Youden index.

the number of samples is $n_j^{(m)}$ ($n_1^{(1)} = n_2^{(1)} = 80, 100, 200, n_1^{(2)} = n_2^{(2)} = 30, n_1^{(3)} = n_2^{(3)} = 20$, and $n_1^{(4)} = n_2^{(4)} = 15$) for study m ($1 \leq m \leq M = 4$) of sample subgroup j (i.e. $j = 1$ for controls and $j = 2$ for cases). Denote expression data matrix by $X^{(m)} = \{x_{g,u}^{(m)}\}$ for gene $1 \leq g \leq G = 200$, $1 \leq u \leq n_1^{(m)} + n_2^{(m)}$ and $1 \leq m \leq M = 4$.

Step 1. Simulate consensus predictive genes

(1) For each of the two consensus clusters c ($1 \leq c \leq 2$) in study $m(1 \leq m \leq M)$, sample gene correlation structure $\Sigma_{cjm}^* \sim W^{-1}(\Psi, 60)$ for every gene cluster c and sample subgroup j of study m, where $\Psi = 0.5I_{20\times 20} + 0.5J_{20\times 20}$, W^{-1} denotes inverse Wishart distribution, I is the identity matrix, and J is the matrix with all the entries being 1. Set vector σ_{cjm} as the square roots of the diagonal elements in Σ_{cjm}^*. Calculate Σ_{cjm} such that $\sigma_{cjm}\Sigma_{cjm}\sigma_{cjm}^T = \Sigma_{cjm}^*$.

(2) We simulate two clusters of consensus predictive genes, each containing 20 genes. The first down-regulated gene cluster is generated by $(x_{1,u}^{(m)}, \cdots, x_{20,u}^{(m)}) \sim MVN(\mu_a, \Sigma_{1jm})$, where sample u belongs to class j in study m and $\mu_a = 0.8$ for $j = 1$ (controls) and $\mu_a = -0.8$ for $j = 2$ (cases). This is a smaller effect size simulation. We also simulate a strong effect size simulation by $\mu_a = 1$ or -1 for controls and cases. Similarly, the second up-regulated gene cluster is simulated by $(x_{21,u}^{(m)}, \cdots, x_{40,u}^{(m)}) \sim MVN(\mu_a, \Sigma_{2jm})$, where $\mu_a = -0.8$ and 0.8 for controls and cases in weak signal scenario and $\mu_a = -1$ and 1 in strong signal scenario. These 40 consensus predictive genes are the basis to aggregate predictive power across studies.

Step 2. Simulate study-specific predictive genes

We next simulate four clusters ($m' = 1, 2, 3, 4$) of study specific genes, each containing 20 genes. Each gene cluster has specific predictive power to the corresponding study m. The down-regulated genes are simulated by $(x_{40+(m'-1)\cdot 20+1,u}^{(m)}, \cdots, x_{40+(m'-1)\cdot 20+10,u}^{(m)}) \sim MVN(\mu_b, \Sigma_{2+m',j,m})$, where $m' = m$, $\Sigma_{2+m,j,m}$ ($1 \leq m \leq 4$) are simulated similar to (1) of Step 1 and $\mu_b = 4$ or -4 for controls and cases. For up-regulated predictive genes, $(x_{40+(m'-1)\cdot 20+11,u}^{(m)}, \cdots, x_{40+m'\cdot 20,u}^{(m)}) \sim MVN(\mu_b, \Sigma_{6+m',j,m})$ and $\mu_b = -4$ or 4 for controls and cases. When $m' \neq m$, the gene cluster m' has no predictive power in study m and $(x_{40+(m'-1)\cdot 20+1,u}^{(m)}, \cdots, x_{40+m'\cdot 20,u}^{(m)}) \sim MVN(0, I)$. These study-specific genes are a main source of errors in cross-study validation.

Step 3. simulate non-informative genes

Finally, the remaining 80 non-informative genes are simulated by $x_{g,u}^{(m)} \sim N(0, 1)$ for $121 \leq g \leq 200$.

We repeated simulations for 50 times, and the results are benchmarked by averaged Youden index. Figure 5 shows the simulation evaluation for different methods using Youden index. For meta-analysis methods, we tested three meta-analyzed approaches for selected TSPs (Fisher, Stouffer, mean) and TSP/kTSP options. In kTSP, we have two further options (cross validation CV and variance optimization VO) to determine K, the number of TSPs for model construction. This gives a total of 9 meta-analysis methods to compare. In each meta-analysis evaluation, we take one study out as the test study, combine the remaining three studies to select the TSPs and construct the model, and finally use the model to predict samples in the test study. The result of Figure 2A in a stronger signal setting ($\mu_a = 1$ or -1) shows that all six meta-analysis methods by kTSP performed well (Youden Index = 0.851 - 0.876). The three meta-analysis methods by TSP performed slightly worse (Youden Index = 0.723 - 0.759). In contrast, we also compared three mega-analysis and three single study analysis approaches. In mega-analysis approaches, the three training studies are normalized and combined into one study to construct the prediction model and evaluate in the test study. In single study analysis, the accuracy was evaluated by averaging inter-study accuracy from each of the three training studies to the test study. The result clearly shows inferior performance of the three mega-analysis approaches and poor performance of single study prediction. This confirms our hypothesis that prediction model from a single study may not be robust and accurate. Proper meta-analysis by combining multiple training studies improves the stability and accuracy of the model to predict an independent test study. Figure 9 and 10 shows results of different parameter settings varied by the mean of consensus genes and the number sample of the first study. In the weaker signal case in Figure 2B ($\mu_a = 0.8$ or -0.8), we found that Meta-kTSP using Fisher's selecting approach sometimes has inferior performance than Stouffer and mean methods. This is probably because of the nature of heavy tail log-transformation in the Fisher's method. A p-value close to 0 (e.g. 1E-20) can contribute a very large score in Fisher's method and can easily dominate the analysis. The inverse transformation in Stouffer's method and the mean score approach somewhat alleviated the problem. From this aspect, we will only compare Stouffer and mean score approaches in the real data applications.

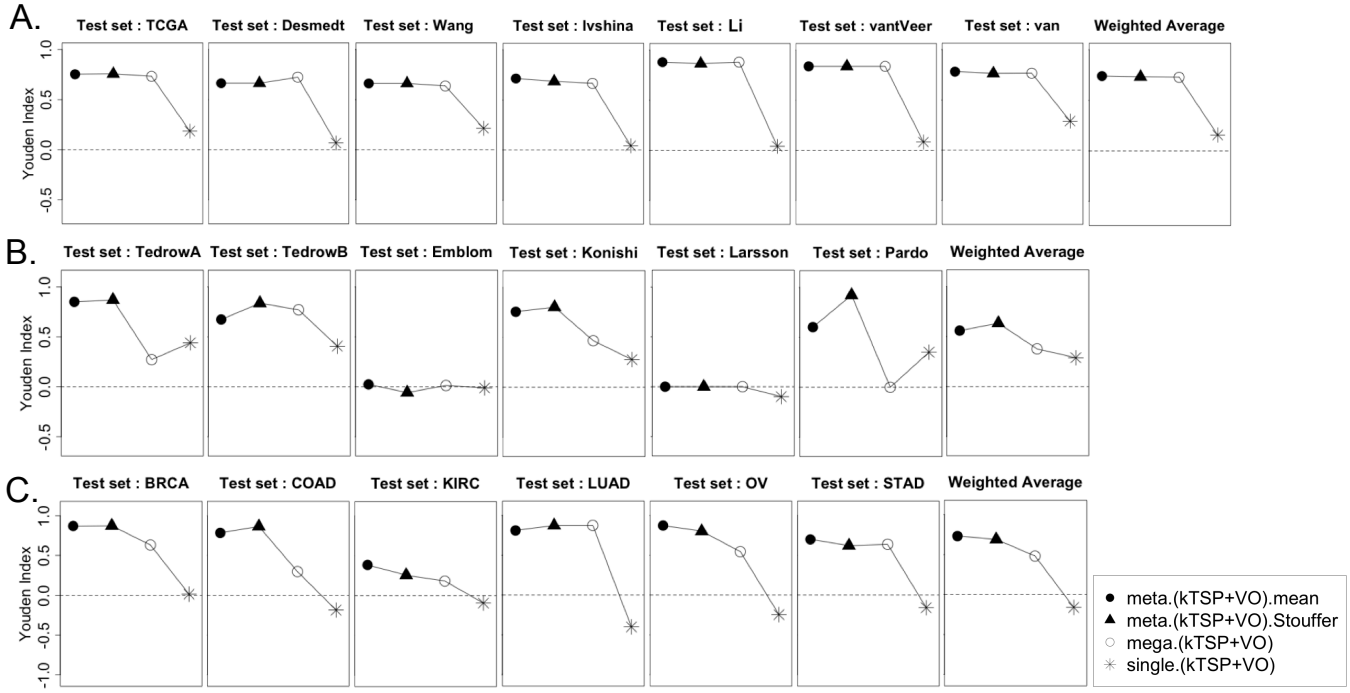

Figure 6: Three examples of Inter-study prediction with applications to real data sets (A. Breast Cancer: ER+ vs ER-, B. Idiopathic pulmonary fibrosis, B. Six differents cancers in TCGA). Y-axis represents the average Youden index.

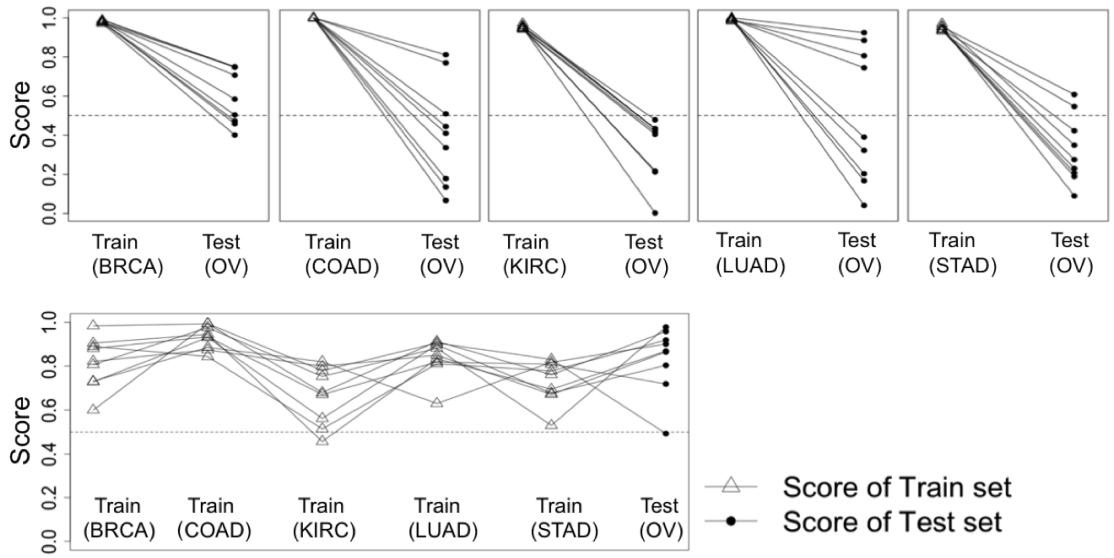

Figure 7: Comparison between single TSP scores and Meta-TSP scores using TCGA cancer data sets. The upper panels illustrate the scores of single study TSP to test data set (Ovarian; OV), whereas the bottom panel shows the Meta-TSP scores of multiple the train studies to test data set (Ovarian; OV).

Table 1: The list of nine identified gene pairs of Average Meta-KTSP and the existing breast cancer gene signatures.

Label	Gene1	Gene2	Averaged scores	References
Pair 1	PI3 (ER-)	GATA3 (ER+)	-0.698	Usary et al. (2004, GATA3, ER+)
Pair 2	ODC1 (ER-)	DNALI1 (ER+)	-0.644	Parris et al. (2010, DNALI1, ER+)
Pair 3	LAD1 (ER-)	SCCPDH (ER+)	-0.632	Dvorkin-Gheva et al. (2011, SCCPDH, ER+), Smith et al. (2008, LAD1, ER-)
Pair 4	FOXC1 (ER-)	MYB (ER+)	-0.623	Ray et al. (2010, FOXC1, ER-)
Pair 5	PSME4 (ER-)	DACH1 (ER+)	-0.620	Powe et al. (2014, DACH1, ER+)
Pair 6	WARS (ER-)	FBP1 (ER+)	-0.609	van't Veer et al. (2002, FBP1, ER+)
Pair 7	CDCA8 (ER-)	AFF3 (ER+)	-0.608	Thakkar et al. (2010, AFF3, ER+)
Pair 8	MRFAP1L1 (ER+)	KCMF1 (ER-)	0.599	Symmans et al. (2010, KCMF1, ER-)
Pair 9	RNASEH1 (ER-)	MAGED2 (ER+)	-0.598	Thakkar et al. (2010, MAGED2, ER+)

2.3.2 Application to genomic data sets

Below we demonstrate application of Meta-kTSP methods to three real omics examples of breast cancer expression profiles (1,658 samples in 7 studies), idiopathic pulmonary fibrosis expression profiles (IPF; 291 samples in 6 studies) and The Cancer Genome Atlas multi-cancer methylation profiles (TCGA, http://cancergenome.nih.gov/; 1,785 samples in 6 studies). Table 2 provides detailed data description of all 19 studies and their data sources. Genes and methylation probes were matched across studies. Non-expressed and/or non-informative genes were filtered according to the rank sum of mean intensities or standard deviations across studies (Wang et al., 2012). This generated 3,035 genes in breast cancer, 3,010 genes in IPF and 3,061 methylation probes in TCGA for down-stream prediction analysis.

Figure 7 shows the inter-study prediction performance of three examples (A: breast cancer ER+ versus ER- prediction by expression profiles; B: IPF versus controls prediction by expression profiles; C: cancer versus adjacent normal prediction by methylation profiles). In each example, we plotted performance of metaKTSP methods (using OV feature selection method to determine K) when each study was chosen as the test study and the remaining studies were used as training studies. For single study analysis, we performed all pairs of

cross-study validation and averaged the performance. For mega-analysis, each sample was standardized to mean zero and unit variance and multiple studies were merged for analysis. Finally, we aggregated Youden indexes of all studies using weighted average by sample size (last plot in each row). The result clearly showed best performance of the two Meta-kTSP methods (TSP selection by mean score or Stouffer's method). Mega-analysis methods had worse performance and single study analysis without combining information across studies performed the worst. In the final example of pan-cancer analysis, the performance of single study analysis was below random guess (Youden index < 0). This suggests that prediction models from single study analysis mostly reflected study-specific (cancer-specific) signature that could not be generalized to other cancers. By meta-analysis, pan-cancer methylation features were successfully selected to facilitate successful inter-study prediction. Figure 4 provides further insight on this concept. In Figure 4A, 9 TSPs were selected in individual training studies (BRCA, COAD, KIRC, LUAD and STAD), respectively. When these TSPs were evaluated in the ovarian cancer (OV) study, the absolute discriminant scores dropped significantly, many of which dropped from close to 1 to below 0.5. On the contrary, the 9 TSPs selected by meta-analysis shared universally large discriminant scores for all five training studies (Figure 4B) and the discriminant scores were mostly maintained in the test OV study. Figure 11-13 provides the full results of all 15 methods comparison in the three examples. We note that Figure 3 did not present results of Fishers method and CV model selection method since Fishers method performed almost identical to Stouffers method and CV and OV also had similar performance.

It is interesting to note that Emblom and Larsson studies in the IPF examples had almost no predictive value (Youden index near 0) while the other four studies performed well. This argues that the two studies might have heterogeneous cohorts from the other four studies or they may have worse experimental quality (Kang *et al.*, 2012). In practice, one may perform such cross validation to exclude potential 'outlier' studies before implementing Meta-kTSP.

Below we explore biological validation of detected gene pairs from Meta-kTSP using existing literature. We first applied MetaKTSP in all the seven breast cancer studies and identified 9 TSPs using mean score method and OV model selection. For the 18 genes in the 9 detected pairs, 10 of them were found to associate with ER expression in previous

publications and all of them had consistent differential expression direction compared to the microarray data (Table 1). For the pan-cancer methylation result, we identified 10 genes from the top 5 TSPs (mean scores and six data sets in Table 2) The PCDH8 gene from the second gene pair was previously confirmed as a candidate tumor suppressor regulated by methylation in multiple cancers (1) Kidney cancer: frequent promoter region methylation (58%) in primary renal cell carcinoma tumour samples (Morris *et al.*, 2011). (2) Breast cancer: either mutation or epigenetic silencing in a high fraction of breast carcinomas inactivates PCDH8 that leads to oncogenesis in cancers (Yu *et al.*, 2008) (3) Stomach cancer: tumor suppressor function in gastric cancer (Zhang *et al.*, 2012).

2.4 DISCUSSION

As high-throughput experimental data become more and more prevalent and publicly available, integrative methods to fully utilize information from the abundant multi-lab data sets have become critical. Generating predictive biomarkers and classification model from a single study often suffer from limited sample size and possibly study-specific biases. The resulting models are often found with poor performance in cross-study validation (Reid *et al.*, 2005; Correa *et al.*, 2009; McShane *et al.*, 2013; Kern *et al.*, 2012). To improve translational and clinical utility of the biomarker discovery and classification model construction, combining information from multiple studies provide a promising opportunity. We seek to improve a top scoring pair (TSP) method that is a non-parametric, accurate and easily interpretable model that likely will succeed in cross-study validation for clinical applications. We developed three MetaTSP approaches that combine multiple omics data sets to improve the credibility of TSP biomarker selection. Using simulations and real transcriptome and methylome data sets, we demonstrate its improved performance on cross-study validation. We compared two methods, cross validation (CV) and variance optimization (VO), to decide the number of TSPs used in the model construction. The result showed similar performance of the two model selection methods. Since VO does not involve repeated subsampling and is computationally faster, we recommend to use VO for future applications.

Table 2: Shown are the brief descriptions of the nineteen microarray datasets of disease-related binary phenotypes (e.g., case and control or ER+/-). All datasets are publicly available.

Name	Study	Type	# of samples	Control	Case	# of matched genes	Reference
TCGA	Breast cancer	mRNA	406	319 (+)	87 (-)	9,024	The Cancer Genome Atlas (TCGA)
Desmedt	Breast cancer	mRNA	198	134 (+)	64 (-)	9,024	Desmedt et al. (2007), GSE7390
Wang	Breast cancer	mRNA	286	209 (+)	77 (-)	9,024	Wang et al. (2005), GSE2034
Ivshina	Breast cancer	mRNA	245	211 (+)	34 (-)	9,024	Ivshina et al. (2006), GSE4922
Li	Breast cancer	mRNA	111	66 (+)	45 (-)	9,024	Li et al. (2010), GSE19615
vant	Breast cancer	mRNA	117	75 (+)	42 (-)	9,024	van de Vijver et al. (2002), Bioconductor
van	Breast cancer	mRNA	295	226 (+)	69 (-)	9,024	van't Veer et al. (2002), Bioconductor
Tedrow A	IPF	mRNA	63	11	52	5,807	GSE47460
Tedrow B	IPF	mRNA	96	21	75	5,807	GSE47460
Emblom	IPF	mRNA	58	20	38	5,807	Emblom et al. (2010) GSE17978
Konishi	IPF	mRNA	38	15	23	5,807	Konishi et al. (2009), GSE10667
Pardo	IPF	mRNA	24	11	13	5,807	Pardo et al. (2005), GSE2052
Larsson	IPF	mRNA	12	6	6	5,807	Larsson et al. (2008), GSE11196
BRCA	Breast cancer	Methylation	343	27	316	10,121	The Cancer Genome Atlas (TCGA)
COAD	Colon cancer	Methylation	204	37	167	10,121	The Cancer Genome Atlas (TCGA)
KIRC	Kidney cancer	Methylation	418	199	219	10,121	The Cancer Genome Atlas (TCGA)
LUAD	Lung cancer	Methylation	151	24	127	10,121	The Cancer Genome Atlas (TCGA)
OV	Ovarian cancer	Methylation	560	4	556	10,121	The Cancer Genome Atlas (TCGA)
STAD	Stomach cancer	Methylation	109	43	66	10,121	The Cancer Genome Atlas (TCGA)

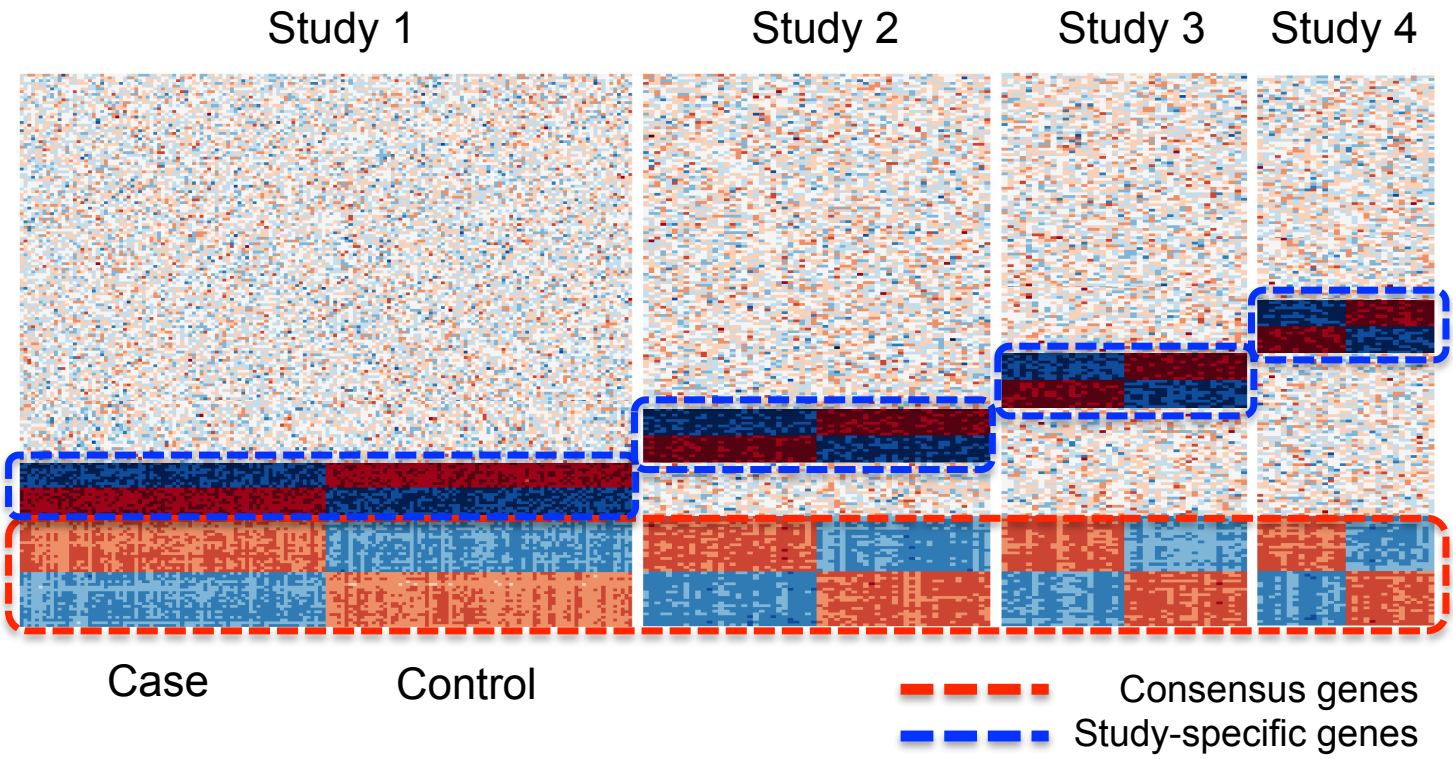

Figure 8: Heatmap of the four simulated data. Genes encircled by red dotted line refer to correlated consensus genes. Study-specific genes are encircled by the blue dotted line.

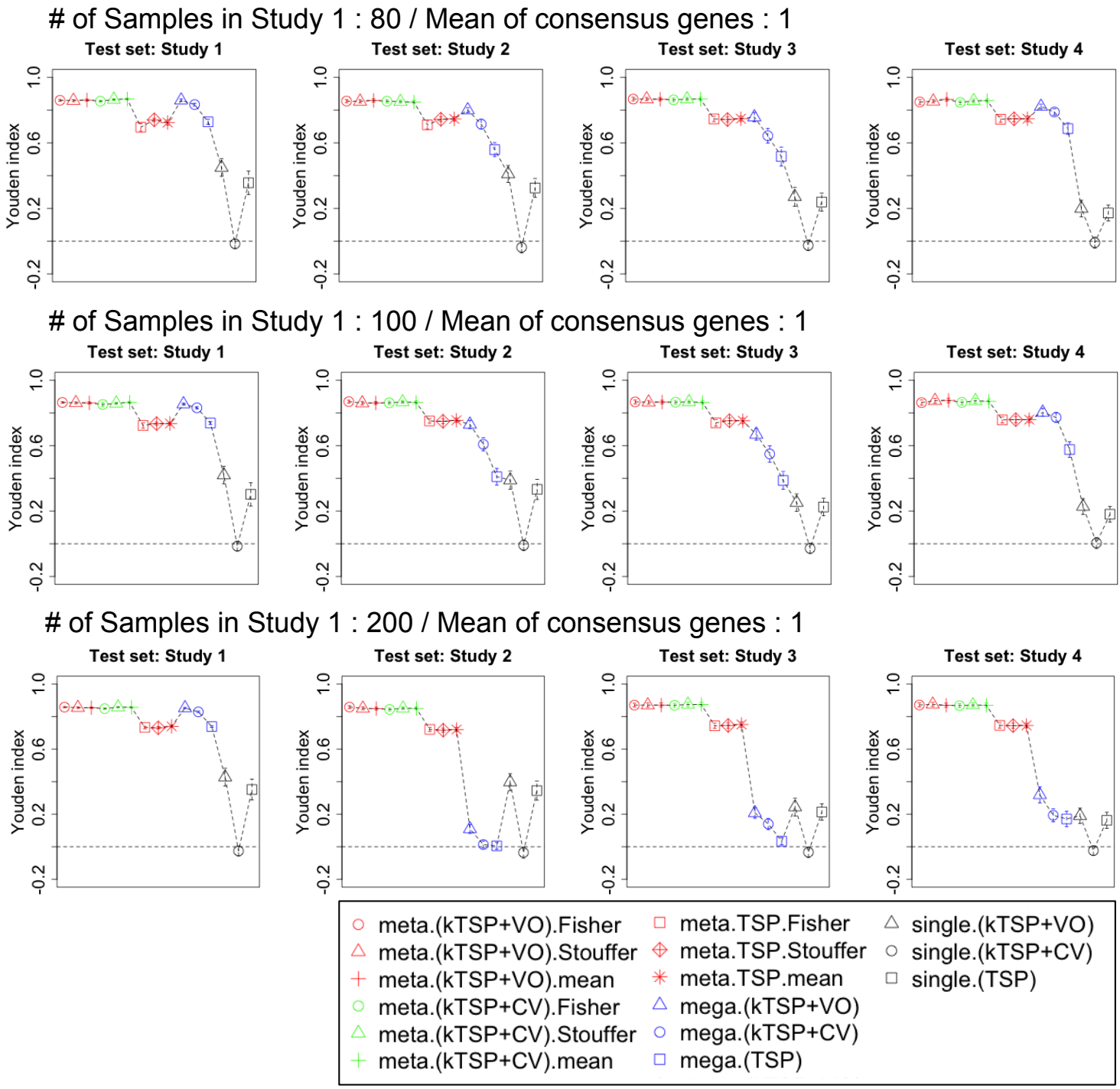

Figure 9: Simulation results of the methods of TSP and MetaTSP family ($\mu_a = 1$).

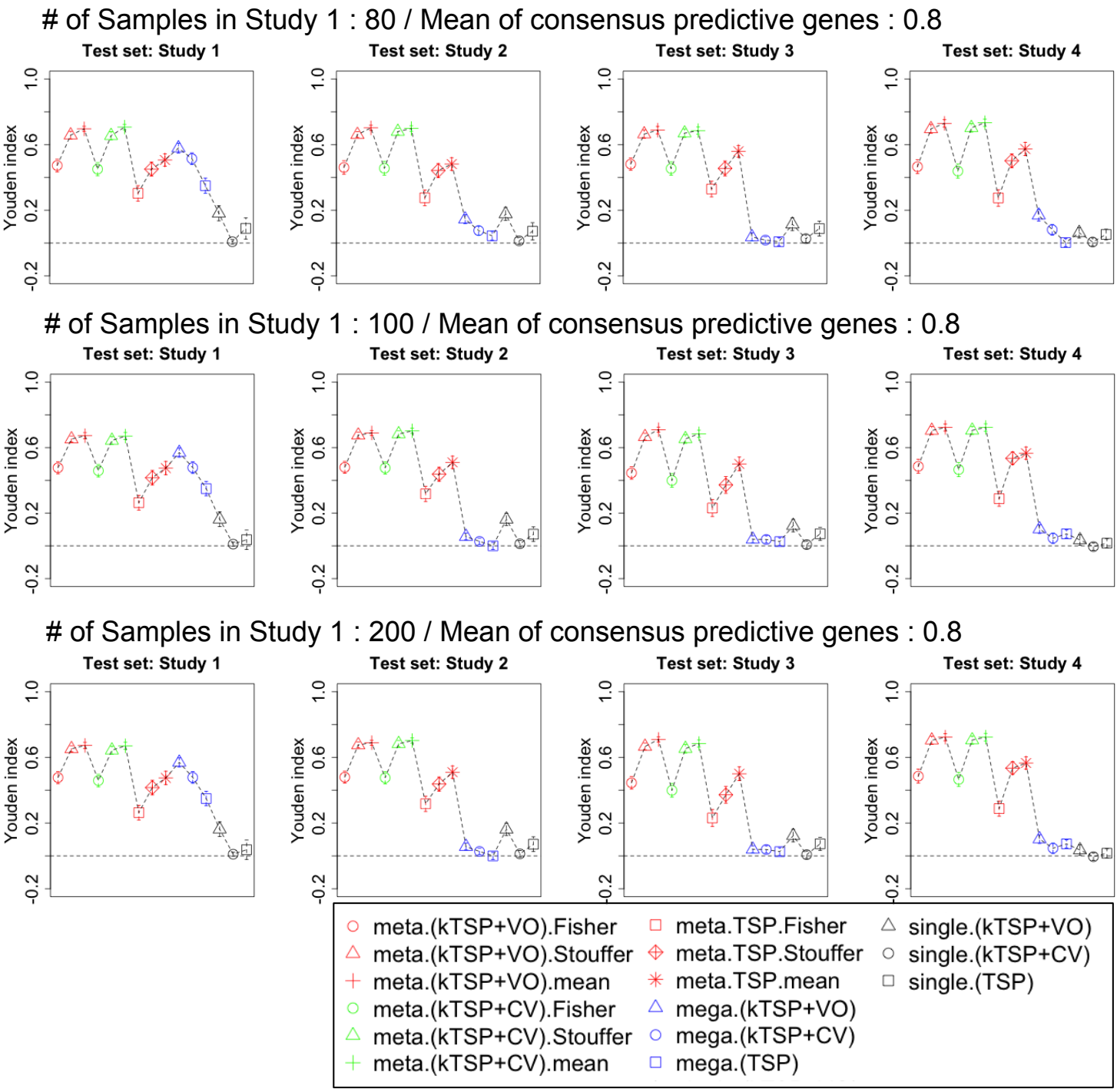

Figure 10: Simulation results of the methods of TSP and MetaTSP family ($\mu_a = 0.8$).

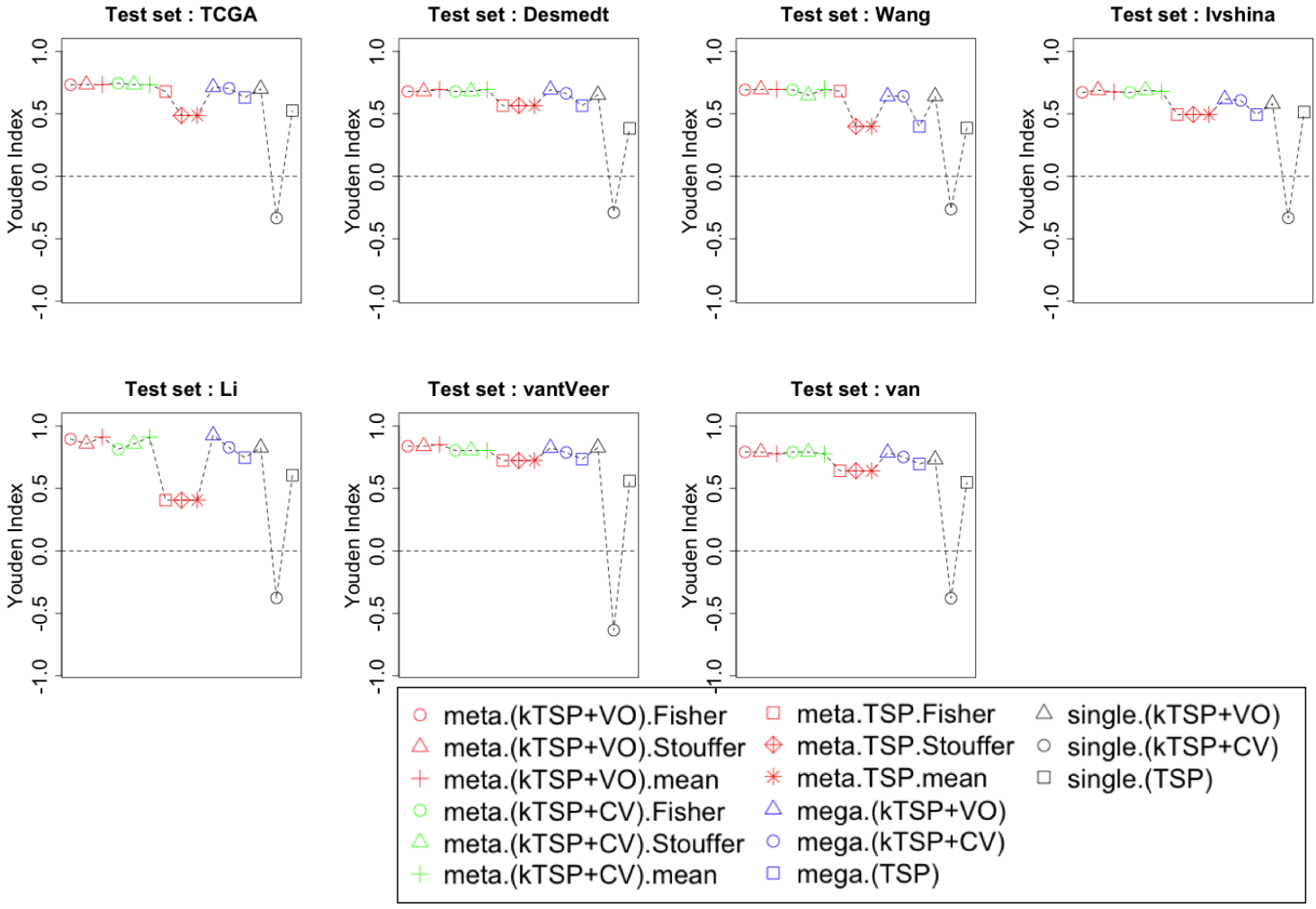

Figure 11: Performance comparisons of the methods of TSP and MetaTSP family using breast cancer mRNA data.

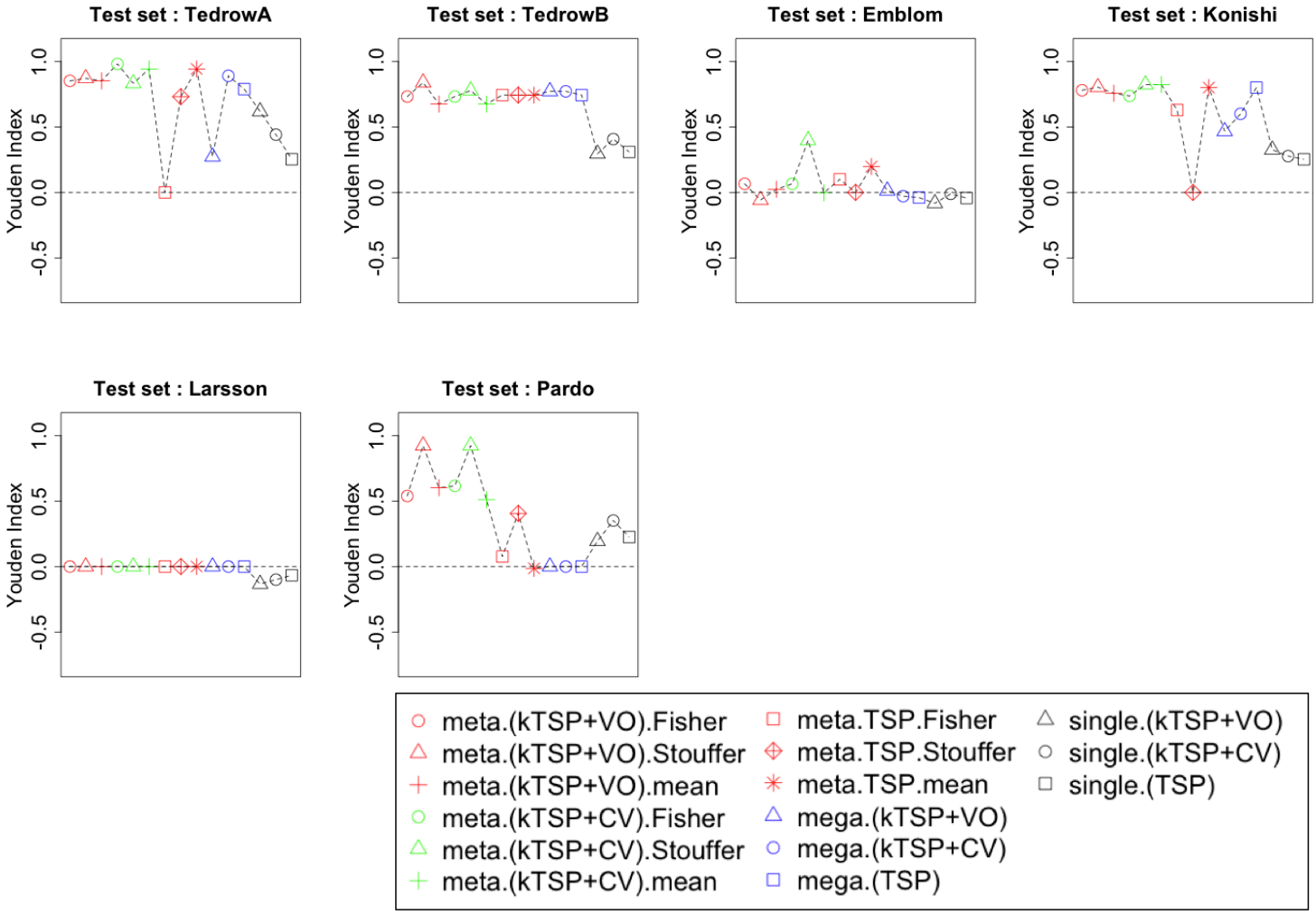

Figure 12: Performance comparisons of the methods of TSP and MetaTSP family using lung disease mRNA data.

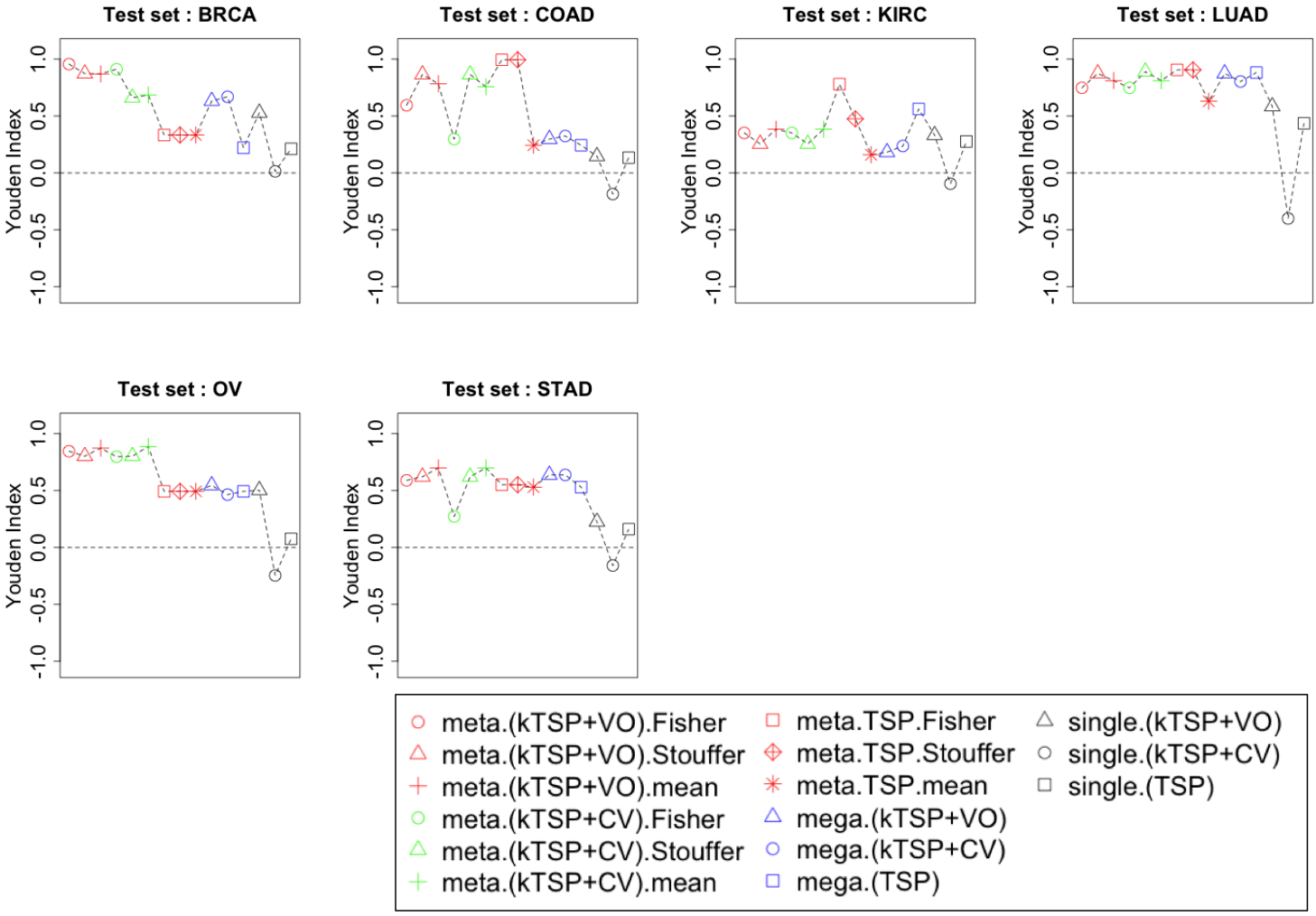

Figure 13: Performance comparisons of the methods of TSP and MetaTSP family using TCGA pan cancer methylation data.

3.0 INTEGRATIVE MULTI-OMICS CLUSTERING FOR DISEASE SUBTYPE DISCOVERY BY SPARSE OVERLAPPING GROUP LASSO AND TIGHT CLUSTERING

3.1 INTRODUCTION

The problem of disease subtype discovery using clustering algorithms has received wide attention in the analysis of microarray data (Pan et al., 2007; Golub et al., 1999; Ghosh et al., 2002). Many classical algorithms such as hierarchical clustering, K-means, self-organizing maps, Gaussian mixture model-based clustering and Bayesian clustering have been developed. The purpose is to cluster samples in an expression profile based on their expression intensity patterns with the goal to identify biologically (e.g. different pathway activation and disease progression mechanisms) and clinically (e.g. different drug response or survival) meaningful disease subtypes. Breast cancer was initially divided into two major subtypes of estrogen receptor (ER)-positive and ER-negative disease. The well-known paper from Perou's lab (Perou et al., 2010) applied hierarchical clustering in their microarray dataset, identified five molecular breast cancer subtypes (Luminal A, Luminal B, Basal, Her2, and Normal-like) and demonstrated their biological and clinical relevance. With additional multi-omics data sources, several methods for disease subtype discovery of vertical omics integration have been proposed. Rey and Roth (2012) introduced a copula mixture model for dependency-seeking clustering of multi-omics data. Lock et al. (2013B) proposed a Bayesian consensus clustering to account for consensus and source-specific information when identifying clusters. Shen et al. (2009, 2013) developed an integrative clustering approach (iCluster) via Gaussian latent regression model. iCluster has several advantages that made it popular in many, particularly cancer, applications. First, the computation of the matrix decompo-

sition is much more efficient than MCMC in other Bayesian-based methods. Second, the consensus and source-specific information from different omics data are automatically shown in the matrix decomposition without further modeling. The method, however, has two major drawbacks. First, the prior knowledge of regulatory mechanisms between different omics data is not considered in the regularization of feature selection. For example, copy number variation (CNV) is known to likely regulate a gene's expression. When there exists an association between a gene's CNV and mRNA expression, we should encourage the selection of both features to improve accuracy and biological interpretation. Secondly, current iCluster assigns all samples into clusters and does not allow any "scattered" samples, a concept commonly seen in microarray data analysis (Tseng and Wong, 2005; Tseng, 2007; Maitra and Ramler, 2009). Considering the biological complication of disease processes, there may exist a portion of "outlier" samples who do not belong to or form any meaningful tight cluster. We propose group structured tight iCluster (GST-iCluster) method by applying a sparse overlapping group lasso technique and an additional regularization on samples in the iCluster modeling to circumvent the two aforementioned problems.

3.2 INTEGRATIVE CLUSTERING (ICLUSTER)

The latent regression model in integrative clustering was first proposed by Shen et al. (2009), and a penalized EM algorithm was developed in Shen et al. (2013). For n subjects, suppose we have M different omics datasets such as mRNA expression, miRNA expression, and DNA methylation. Let $X^{(m)}$ be the m^{th} dataset with p_m features, where each column of $X^{(m)}$ consists of mean-centered features from n subjects ($1 \leq m \leq M$). The combined dataset is $X = \left(X^{(1)^T}, X^{(2)^T}, \ldots, X^{(M)^T}\right)^T$, which is a $\sum_{m=1}^{M} p^{(m)} \times n$ matrix. The joint latent regression model is

$$X^{(m)} = B^{(m)} Z + E^{(m)} \text{ for } 1 \leq m \leq M,$$

where Z is a $\ell \times n$ matrix whose rows are latent variables and columns are samples. The matrix $B^{(m)}$ is used to control the degree of relation between feature intensities and latent

variables. Under the multivariate normal assumption, the conditional distribution is

$$X^{(m)}|Z \sim N(B^{(m)}Z, \psi^{(m)}) \text{ for } 1 \leq m \leq M,$$

where $\psi^{(m)} = \text{diag}(\sigma_1^2, \ldots, \sigma_{p^{(m)}}^2)$. To achieve the sparse estimation of $B^{(m)}$, L_1-lasso penalty (Tibshirani, 1996) is used and the penalized complete log likelihood function becomes

$$\log(L_p(B, \psi)) = -\frac{n}{2}\log(|\psi|) - \frac{1}{2}\text{tr}\left[(X - BZ)^T\psi^{-1}(X - BZ)\right] - \frac{1}{2}\text{tr}(ZZ^T) - P_\lambda(B), \quad (3.1)$$

where $B = \left(B^{(1)^T}, B^{(2)^T}, \ldots, B^{(M)^T}\right)^T$, $P_\lambda(B) = \sum_{m=1}^{M} \lambda^{(m)} \sum_{j=1}^{\ell} \sum_{i=1}^{p^{(m)}} \left|\beta_{ij}^{(m)}\right|$, and ℓ is the number of effective latent variables, which is defined by $K-1$ (Ding and He, 2004), where K is the number of sample clusters. The expectation-maximization (EM) algorithm (Dempster et al., 1977) is then applied to estimate $\hat{B}$ and $\hat{\psi}$.

3.3 GROUP STRUCTURED AND TIGHT INTEGRATIVE CLUSTERING

3.3.1 Sparse overlapping group lasso

We first define feature modules using multiple omics features. Let $G = \{f_i^{(m)}\}$ be the set of all features ($1 \leq i \leq p^{(m)}$ and $1 \leq m \leq M$), and let $g_v \subset G$ ($1 \leq v \leq V$) be the v^{th} feature module. Features in a feature module are associated with the potential genomic feature regulation flow and can come from the same or different omics datasets. The corresponding model parameters are defined as $\beta_{g_v j} = \{\beta_{ij}\} \in R^{|g_v|}$, where $1 \leq i \leq |g_v|$, $1 \leq j \leq \ell$, and $|g_v|$ denotes the number of features in module g_v. For example, we define a feature module based on the gene symbol annotation of gene EST1 in the TCGA data. One mRNA (f_i), one CNV (f_j), and two methylation probes (f_k, f_ℓ) are aligned to gene EST1. A feature module using those four features can be defined as $\{f_i, f_j, f_k, f_\ell\}$. It is possible that multiple modules share the same feature (e.g. two miRNAs may potentially regulate the same gene). If no feature modules share common features, it is a non-overlapping problem (no features are overlapped across pre-defined feature modules), and otherwise the modules are overlapping. These feature modules are the basis for group lasso regularization.

Below we present an overview of how to solve an overlapping group lasso problem. Chen et al. (2012) developed a Smoothing Proximal Gradient Descent (SPG) approach to tackle with intractable nature of the sparse overlapping group lasso (non-separable and non-smooth penalty) under the penalized regression framework. Consider the sparse overlapping group lasso with L_1/L_2 mixed penalties (Obozinski et al., 2008, 2010),

$$\Omega(B) = \lambda \sum_{j=1}^{\ell} \sum_{v=1}^{V} w_{g_v} \|\beta_{g_v j}\|_2. \tag{3.2}$$

Using a dual norm representation (Chen et al., 2012), the penalty function (3.2) becomes

$$\begin{aligned} \Omega(B) &= \lambda \sum_{j=1}^{\ell} \sum_{v=1}^{V} w_{g_v} \max_{\|\alpha_{g_v j}\|_2 \le 1} \alpha_{g_v j}^T \beta_{g_v j} \\ &= \max_{\alpha \in Q} \sum_{j=1}^{\ell} \sum_{v=1}^{V} \lambda w_{g_v} \alpha_{g_v j}^T \beta_{g_v j} = \max_{\alpha \in Q} \alpha^T C B, \end{aligned} \tag{3.3}$$

where $C \in R^{\sum_{v=1}^{V} |g_v| \times \sum_{m=1}^{M} p^{(m)}}$, $\alpha = \left(\alpha_{g_1}^T, \alpha_{g_2}^T, \ldots, \alpha_{g_V}^T\right)^T \in R^{\sum_{v=1}^{V} |g_v| \times \ell}$, $\alpha_{g_v} = \{\alpha_{ij}\} \in R^{|g_v| \times \ell}$, $Q = \{\alpha | \|\alpha_{g_v j}\|_2 \le 1, 1 \le j \le \ell, 1 \le v \le V\}$, and λ and w_{g_v} are constants used to adjust sparseness for β_{g_v}. Then, $C_{(k,g_v),i} = \lambda w_{g_v}$, if $k = i$ or 0 otherwise (For details, see the selection of C in Section 3.3.4).

To circumvent the non-differentiable property of $\Omega(B)$ at 0, a smooth approximation of $\Omega(B)$,

$$f_\mu(B) = \max_{\alpha \in Q} \{\alpha^T C B - \mu d(\alpha)\}, \tag{3.4}$$

is used as shown in Nesterov (2005). Here, μ is a positive constant of smoothing parameter and $d(\alpha) = \frac{1}{2}\|\alpha\|_2^2$ is a smoothing function. The maximum difference of $f_\mu(B)$ and $f_0(B)$ is $\max_{\alpha \in Q}\{\mu d(\alpha)\}$. $f_\mu(B)$ is smooth on B by Theorem 1 in Nesterov (2005), and convex and continuously differentiable on B. The gradient of $f_\mu(B)$ is $\nabla f_\mu(B) = C\alpha^*$, where α^* is the solution to obtain $f_\mu(B)$ (Proposition 1 in Chen et al. 2012). Note that $\nabla f_\mu(B)$ is Lipschitz continuous with Lipschitz constant $L_\mu = \frac{1}{\mu}\|C\|_S^2$, where $\|C\|_S$ is the matrix spectral norm defined by $\max_{\|v\|_2 \le 1} \|Cv\|_2$. As a result, we can replace $\Omega(B)$ with $f_\mu(B)$ and solve the optimization with the parameter μ that determines the degree of smoothness.

3.3.2 Group structured integrative clustering (GS-iCluster)

Regulation mechanisms of complex diseases such as cancer are very complicated. The signals from each omics data are often weak and dependent. This motivates us to develop a group structured framework to deal with multiple omics data by applying a sparse overlapping group lasso penalty. By adding approximation of group lasso penalties (3.4), the penalized log likelihood (3.1) becomes

$$\log(L(B,\psi)) = -\frac{n}{2}\log(|\psi|) - \frac{1}{2}\mathrm{tr}\big[(X-BZ)^T\psi^{-1}(X-BZ)\big] - \frac{1}{2}\mathrm{tr}(ZZ^T) - P(B), \quad (3.5)$$

where the aggregated penalty $P(B)$ is $f_\mu(B) + \lambda\|B\|_1$ and μ is a pre-defined smoothing parameter.

In order to estimate the parameters of the latent variable matrix Z, we apply the penalized EM-algorithm. In E-step of iteration t, we take the expected value of complete-data log-likelihood (3.5) with respect to $f(Z|X, B^{(t)})$,

$$\begin{aligned} Q(B|B^{(t)}, X) &= E_{Z|X,B^{(t)}}\Big(\log L(B; X, Z)\Big) \\ &= \int \log L(B; X, Z) f(Z|X, B^{(t)}) dZ. \end{aligned}$$

Note that given X and $B^{(t)}$ an estimated solution for B at iteration t, $E(Q)$ solely depends on $E(Z)$ and $E(ZZ^T)$. Using the property of multivariate normal distribution, we have

$$\begin{aligned} E_{Z|X,B^{(t)}}(Z|X) &= {B^{(t)}}^T\Sigma^{-1}X, \\ E_{Z|X,B^{(t)}}(ZZ^T|X) &= I - {B^{(t)}}^T\Sigma^{-1}B^{(t)} + E(Z|X)E(Z|X)^T, \end{aligned}$$

where $\Sigma = B^{(t)}{B^{(t)}}^T + \psi$.

In M-step, $B^{(t+1)}$ is obtained by maximizing the function Q given $B^{(t)}$, an estimate at a previous iteration:

$$\begin{aligned} B^{(t+1)} &= \mathrm{argmax}_B Q(B|B^{(t)}, X) \\ &= \mathrm{argmax}_B \left\{ -\frac{1}{2}\mathrm{tr}[(X-BZ)^T\psi^{-1}(X-BZ)] - \frac{1}{2}\mathrm{tr}(ZZ^T) - P(B) \right\}. \end{aligned} \quad (3.6)$$

We iterate the E-step and M-step until convergence (i.e. $\max(|B^{(t)} - B^{(t-1)}|) < 10^{-4}$). To satisfy the Gaussian assumption of latent variables in Z (i.e. standard normal distribution)

(Shen et al., 2009), the latent variables are centered and scaled at each iteration. Once we obtain the solution of B and Z, Partition Around Medoids (PAM; Reynolds et al. 2006) is applied over Z to estimate integrative clustering labels.

When estimating the sparse solution of B, the Smoothing Proximal Gradient Descent (SPG, Chen et al. 2012) algorithm can be applied to obtain $B^{(t+1)}$ in M-step, which improves computation efficiency. The smoothing part of the objective function in equation (3.6) is

$$h(B) = \frac{1}{2}\text{tr}[(X - BZ)^T\psi^{-1}(X - BZ)] + f_\mu(B). \tag{3.7}$$

The gradient and Hessian matrix of $h(B)$ are given by

$$\nabla h(B) = \psi^{-1}(BZ - X)Z^T + C^T\alpha^* = \psi^{-1}B(ZZ^T) - \psi^{-1}XZ^T + C^T\alpha^*,$$
$$\nabla^2 h(B) = H_{\{(i,j),(i,j)\}} = \text{diag}(h), \text{ where } h = \Big\{h_{i,j}|h_{i,j} = z_i^T z_i \frac{1}{\psi_{jj}}, \forall i, j\Big\},$$

where α^* is the optimal solution to 3.7. $\nabla h(B)$ is Lipschitz-continuous (Theorem 1 in Chen et al. 2012) with Lipschitz constant

$$L = \lambda_{\max}\big(\nabla^2 h(B)\big) + \frac{\|C\|_S^2}{\mu}.$$

Since the likelihood function of exponential family is log concave (Bickel and Doksum, 2001), $h(B)$ remain convex and continuous. Therefore, based on Theorem 1 in Chen et al. (2012), we can reformulate objective function in (3.6) using the proximal operator as

$$-\frac{1}{2}\Big\|B - \big(w^t - \frac{1}{L}\nabla h(w^t)\big)\Big\|_2^2 - \frac{\lambda}{L}\|B\|_1.$$

A closed form solution for B can be obtained using soft-thresholding operation (Friedman et al., 2010). The accelerated gradient descent algorithm can be applied until convergence. The summary of the entire algorithm is presented in Table 3.

Table 3: Smoothing proximal gradient descent algorithm for structured likelihood function.

Input: X, Z, C, B^0, Lipchitz constant L, and pre-defined stopping criterion ε.

For $t = 1, 2, \ldots$ until $B^{(t)}$ converges

(1) Minimize proximal operator along with the L_1 lasso term

$$B^{(t+1)} = \text{argmin}_B \Big[\tfrac{1}{2} tr[(X - BZ)^T \psi^{-1} (X - BZ)] + f_\mu(B) + \lambda \|B\|_1 \Big]$$
$$= \text{argmin}_B \tfrac{1}{2} \Big\| B - \big(w^t - \tfrac{1}{L} \nabla h(w^t) \big) \Big\|_2^2 + \tfrac{\lambda}{L} \|B\|_1,$$

where $\nabla h(B) = \psi^{-1}(BZ - X)Z^T + C^T \alpha^* = \psi^{-1} B(ZZ^T) - \psi^{-1} XZ^T + C^T \alpha^*$

(2) Set $\theta_{t+1} = \frac{2}{t+3}$

(3) Set $w^{t+1} = B^{(t+1)} + \frac{1-\theta_t}{\theta_t} \theta_{t+1} \big(B^{(t+1)} - B^{(t)} \big)$

Output $\hat{B} = B^{(t+1)}$

3.3.3 Group structured tight integrative clustering (GST-iCluster)

Motivated by the methods to pursue tight clustering (Tseng and Wong, 2005; Tseng, 2007), we propose GST-iCluster to find stable and coherent clusters by adding an additional L_1 penalty to control the sparseness of samples in Z. In the latent regression model given estimated $\hat{B}$ from penalized EM, Z can be considered as coefficients (i.e. $X = \hat{B}Z + \varepsilon$, where $\varepsilon \sim N(0, \Phi)$, $\Phi = \text{diag}(\tau_1^2, \ldots, \tau_n^2)$). The objective function to be minimized with respect to Z is

$$\|X - \hat{B}Z\|^2 + \lambda_1 \|Z\|_1.$$

Again, a soft-thresholding and PAM is applied to obtain sparse solutions for Z and to produce clustering labels.

3.3.4 Selection of penalization constant for GS-iCluster

In equation 3.4, C determines the degree of sparseness. The selection of elements in C must be appropriate in order to obtain similar overall penalties for the coefficient matrix B. For instance, some features in a group may be highly overlapped (e.g. a miRNA feature links to multiple target gene expression features). If we directly apply the sparse overlapping group lasso (Chen et al., 2012), highly overlapped features are more likely to be penalized. For example, suppose that we have four features f_i $(1 \leq i \leq 4)$ involved in two pre-defined feature modules, $g_1 = \{f_1, f_2, f_3\}$ and $g_2 = \{f_2, f_3, f_4\}$. We can specify components of $C_{(k,g_v),i}$ as illustrated in Figure 14. We note that the sum of columns in $C_{(k,g_v),i}$ associates with overall penalty for f_i, and should be adjusted to assign an equal overall penalty for all f_i $(1 \leq i \leq 4)$. To address this, we force the column elements of C subject to

$$\sum_{k=1}^{\sum_{v=1}^{V}|g_v|} C_{(k,g_v),i} = \lambda \quad \text{for all } i. \tag{3.8}$$

In this particular example, the constraint (3.8) results in

$$\lambda = \lambda\omega_{11} = \frac{\lambda\omega_{12} + \lambda\omega_{22}}{2} = \frac{\lambda\omega_{13} + \lambda\omega_{23}}{2} = \lambda\omega_{24},$$

and so $\omega_{11} = \omega_{24} = 1$, and $\omega_{12} = \omega_{22} = \omega_{13} = \omega_{23} = 0.5$.

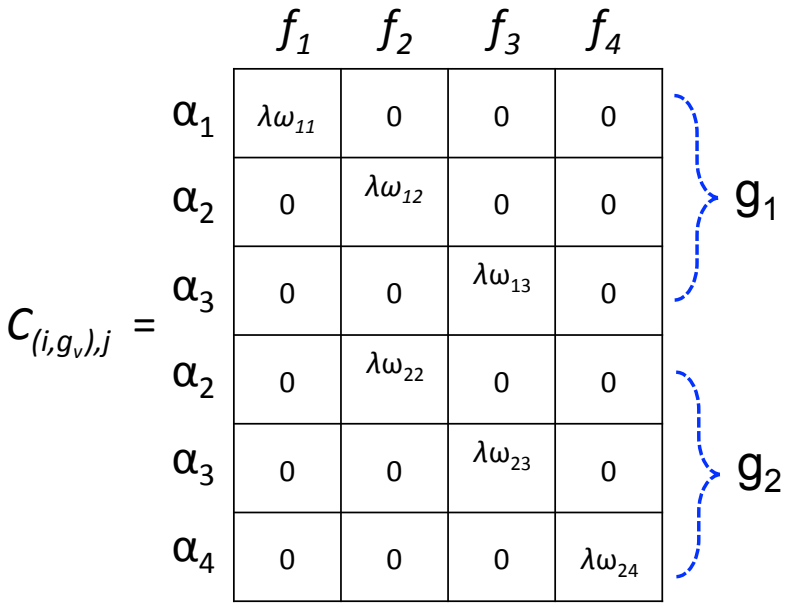

Figure 14: An example of penalization constant C implemented in sparse overlapping group lasso technique.

Two turning parameters, λ of features and K (# of clusters), are involved in GS-iCluster, and an additional λ_1 for sample tight clustering is chosen for GST-iCluster. To select the optimal turning parameter set, we propose sequential searching of all combinations of tuning parameters via S-fold cross validation ($S = 5$ is used). We seek the optimum tuning parameters, λ and K using GS-iCluster, and then given λ and K, we seek λ_1 in GST-iCluster. S-fold cross-validation is applied in finding optimal λ and K. When we apply GS-iCluster to the training data set, class labels of samples in training dataset are estimated and those estimated class labels are treated as true labels. The class labels of testing data set are determined by estimated class labels from training dataset using the shortest Euclidean distance. GS-iCluster is iteratively applied to all training datasets from S-fold validation, and all inferred class labels can be obtained from testing datasets. As an optimization criterion, we use Adjusted Rand Index (ARI, Vinh et al. 2009). ARI is calculated between the class labels estimated from original dataset and the class labels via S-fold cross-validation in the test data sets. Finally, the optimal tuning parameters are selected by searching parameters that maximize ARI values.

3.4 APPLICATIONS

We obtained the mRNA expression, CNV, methylation, and miRNA expression data of breast cancer from TCGA Portal (https://tcga-data.nci.nih.gov/tcga/). The regulation impact of CNV and methylation on mRNA expression is usually stronger than miRNA. Furthermore, probes measuring CNV and methylation features match to gene regions, while one miRNA can potentially affect many genes and many different miRNAs may also impact expression levels of many genes. As a result, we first integrated CNV, methylation and mRNA in section 3.4.1 as an example of non-overlapping group lasso. In section 3.4.2, we consider miRNA and mRNA datasets using the sparse overlapping group lasso method.

3.4.1 Integration of mRNA, methylation and CNV using TCGA breast cancer data

The data obtained from TCGA Data Portal contain CNV for 23,235 genes, methylation levels of 22,529 probes and mRNA expression levels for 17,814 genes with 306 breast cancer samples (ER-positive: 234 samples, ER-negative: 66 samples, Not performed or Intermediate: 6 samples; freeze date: 04/01/2013). Each of CNV and mRNA probe is matched to a gene symbol. We first filtered out genes with low-expressed (mean < 0.9) or non-informative (standard deviation < 0.85) features in the mRNA expression data. 828 genes from mRNA expressions are left for further analysis. We also obtained 1,345 methylation probes and 828 CNV genes by matching 828 mRNA gene symbols. Note that multiple methylation probes may match to one mRNA gene. The features from three different omics datasets that share the same cis-regulatory annotation (same gene symbol) are grouped together to form 828 feature modules. In this case, each module has one mRNA gene expression, one CNV gene and one or more methylation probes. Each module contains multi-omics regulatory information because CNV and methylation may regulate mRNA expression. GS-iCluster and the original iCluster were applied to the data using 828 feature modules information, although iCluster does not incorporate the module information.

Figure 15 shows the results from iCluster and GS-iCluster. The cross-validation analysis was applied and determined the optimal number of clusters at 4 among the choices of 3, 4, 5 and 6. We selected the tuning parameters so that both methods can find similar number of nonzero effect features (1,110 for GS-iCluster and 1,100 for iCluster). Among the selected features from GS-iCluster, 119 features belong to modules covering all three types of omics entries (Category I), 339 features belong to modules covering selected mRNA and methylation (Category II), and 150 features belong to modules covering selected mRNA and CNV (Category III). In total, 608 out of 1,110 (54.8%) features share modules one another and hence may have better biological interpretation, iCluster selected 1,100 features and only 405 (36.8%) share modules. Scatter plots of the top 12 modules of each category I - III selected by the largest absolute values of correlations are shown in Figure 16. Standardized mRNA expression, methylation levels and CNV values are shown on the same plot (y-axis) with

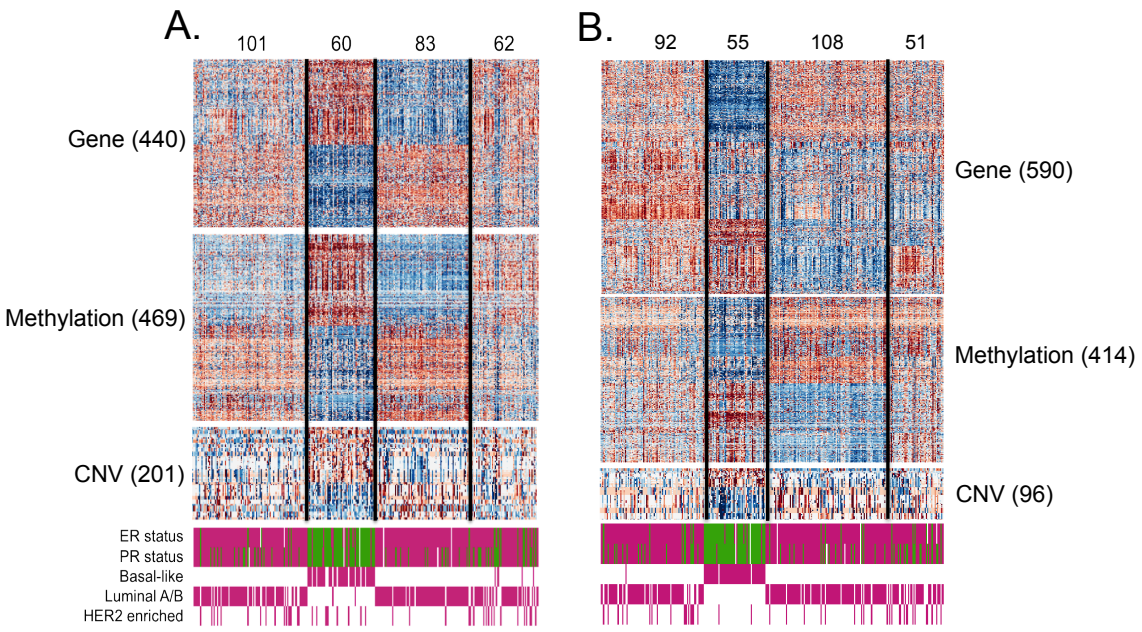

Figure 15: Heatmap of three omics (Gene, Methylation, and CNV) features selected via (A: Group structured / B: iCluster) integrative clustering. For ER and PR status, the pink and green colors represent ER-positive and ER-negative, respectively. For the rest, the pink color refers to Basal-like, Luminal A/B, and HER2 enriched, respectively.

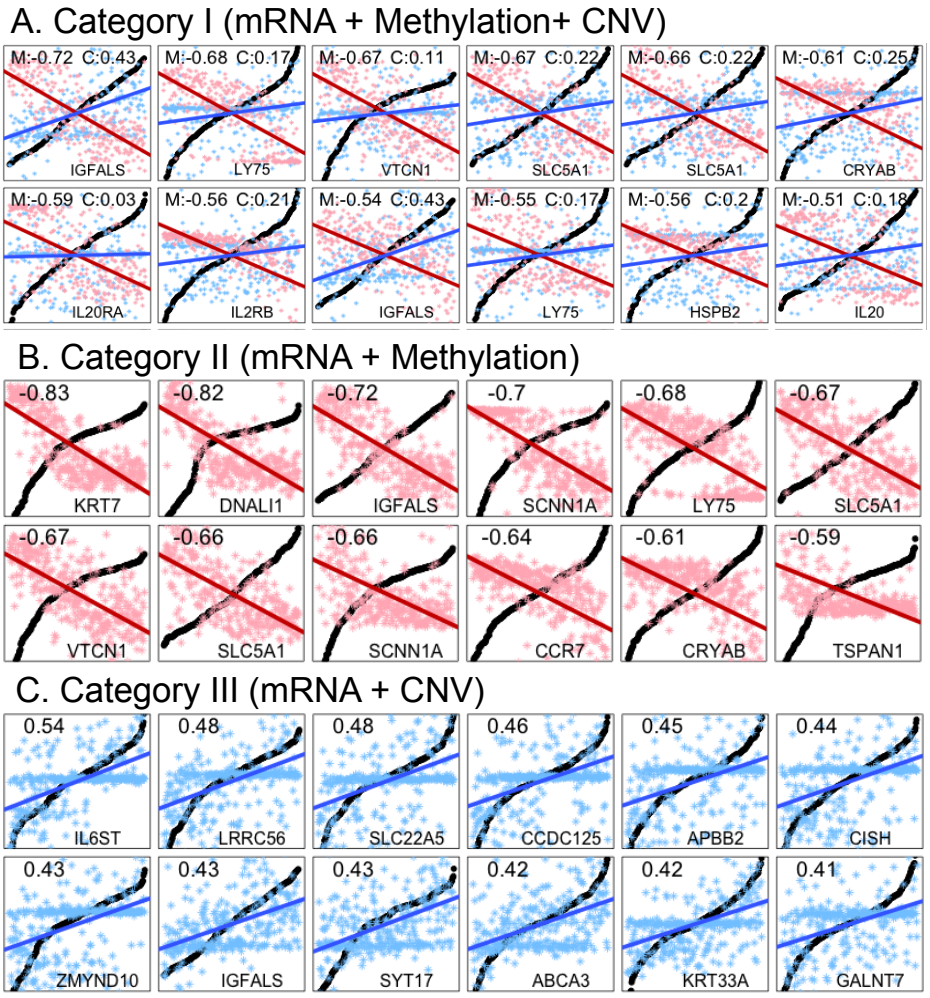

Figure 16: Scatter plots of the top 12 feature modules that are negatively or positively mapped to the ordered mRNA features. Red, Blue, and Black colors represent Methylation, CNV, and mRNA feature intensities, respectively. The values at the corner are correlations between two involving features, and each solid line represents a simple linear regression model of Methylation (Red) and CNV (Blue). Y-axis refers to expression levels, and X-axis samples ordered by mRNA expression.

black, red and blue respectively. The samples are sorted on the x-axis by mRNA expression. As expected, Category II modules with identified mRNA and methylation features showed significant negative Pearson correlation (-0.83 to -0.59), while Category III modules with mRNA and CNV features show positive correlations but with smaller magnitude (0.41 to 0.54). This is consistent with the prior biological knowledge that methylation usually suppressed mRNA expression (negative correlation) and amplification or deletion of CNV usually has positive causal relationship (up-regulation or down-regulation) with mRNA expression. The result provides a good internal validation since the module information integrated in the group lasso does not include the directions of association. The reason that CNV shows a smaller correlation with mRNA expression compared to methylation is probably because most tumors in the data do not have CNV aberrations. The true association of the CNV and gene expression may be much stronger than what we observed based on Pearson correlation.

We further investigated the performances of GS-iCluster and iCluster on identifying biological functions related to breast cancer using pathway enrichment analysis (using Fisher's exact test) with gene symbols of the selected features. Three pathway databases, BioCarta (217 pathways), KEGG (186 pathways) and integrated breast cancer pathway (http://wikipathways.org), were used. Figure 17 shows the Manhattan plot of the pathway enrichment analysis and three significant pathways were identified using GS-iCluster while none were identified using iCluster at the 5% significant level of false discovery rate.

Table 4 lists the top 3 significant pathways related to breast cancer, which were identified by GS-iCluster but not by iCluster: cytokine receptor interaction, Chemokine signaling, and JAK-STAT signaling. Compared to iCluster, results from GS-iCluster identified many more category I modules (mRNA + Methylation + CNV), which indicates high genomic instability and regulatory complexity in these oncogene pathways. Notably, these pathways were already shown to biologically associate with breast cancer (Huan et al., 2014; Palacios et al., 2014; Hernandez-Vargas et al., 2011).

The results from GST-iCluster are shown in Figure 18A. It excluded 88 samples from clustering (inside the dotted green line). These samples appear to have little discriminant subtype patterns. While the number of samples used for clustering is reduced, the patterns looks more coherent to the assigned clusters. In real data analyses, we often notice that

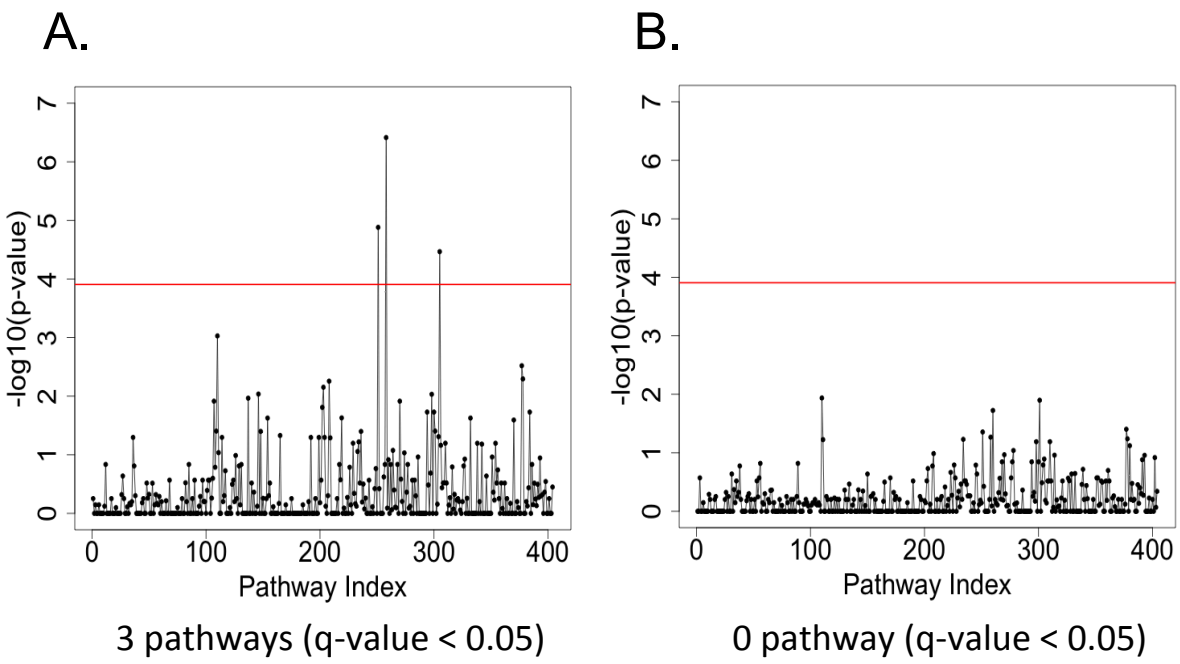

Figure 17: Manhattan plots of pathway enrichment analysis (A: Result from GS-iCluster / B: Result from iCluster).

Table 4: Analysis of three pathways over selected genes from both GS-iCluster and iCluster

.

# of features in the pathway	# of selected features & pathway features (Total # of selected features)	P-value	Q-value	Category I	Category II	Category III
KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION						
(GS-iCluster) 162	91 (1110)	0.0000004	0.00015	18	12	9
(iCluster) 162	71 (1100)	0.054	1	4	21	3
KEGG CHEMOKINE SIGNALING PATHWAY						
(GS-iCluster) 70	44 (1110)	0.000013	0.0053	7	5	5
(iCluster) 70	34 (1100)	0.044	1	2	9	1
KEGG JAK STAT SIGNALING PATHWAY						
(GS-iCluster) 68	42 (1110)	0.000034	0.014	10	3	9
(iCluster) 68	31 (1100)	0.1278	1	2	7	3

Category I : mRNA+methyl+CNV
Category II : mRNA+methyl
Category III : mRNA+CNV

samples used for cancer subtype identification tend to be loosely associated with the subtypes because of heterogeneity and the complicated biological mechanisms of cancer. GST-iCluster may provide deeper insights and better accuracy by identifying accurate subtypes of disease only for patients with clear genomic and epigenetic patterns.

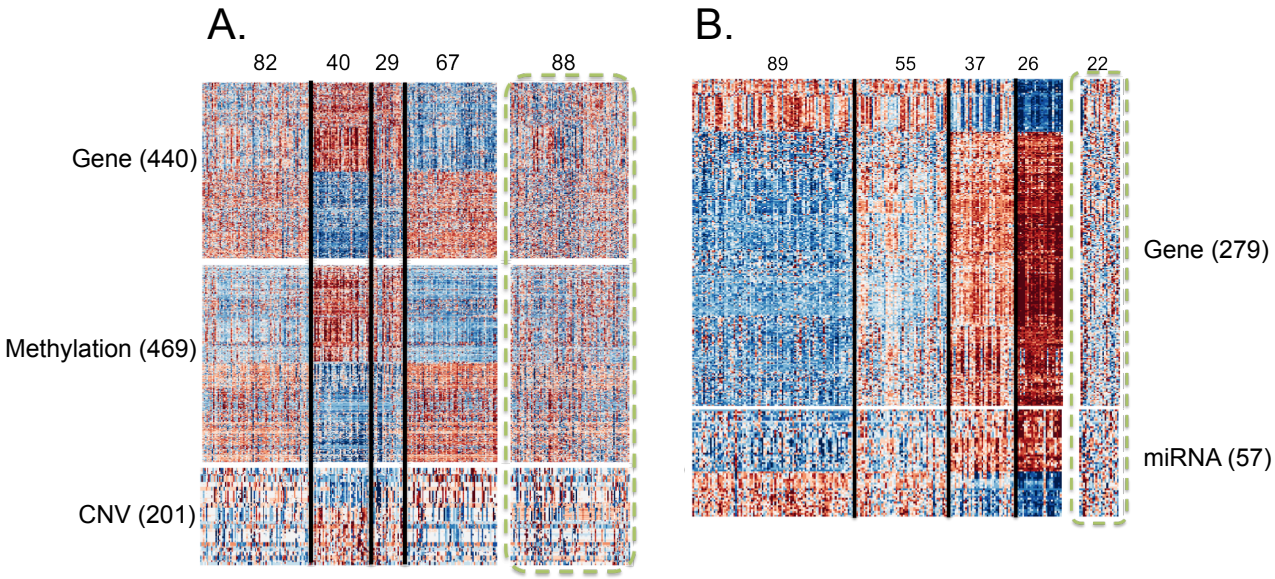

Figure 18: Heatmap of two omics (A:mRNA, Methylation and CNV / B:mRNA and miRNA) features selected via GST-iCluster.

3.4.2 Integration of mRNA and miRNA using TCGA breast data

We collected two microarray datasets that contain mRNA and miRNA expressions of 229 breast cancer samples from TCGA (ER-positive: 174 samples, ER-negative: 51 samples, Not performed or Intermediate: 4 samples; freeze date: 04/01/2013). We selected top 1,500 most variable gene expression probes from 17,814 original probes, and 650 miRNAs (mean < 0.015, SD < 0.13) from 1,046 probes to conduct further analysis. Based on the miRNA database of target gene predictions (miRanda; John et al. (2004)), candidate target mRNA features and the miRNAs were grouped to define feature modules in GS-iCluster.

GS-iCluster identified 4 sample clusters among the choices of 3, 4, 5 and 6 as optimal number of disease subtypes. Again to make the results from both iCluster and GS-iCluster

Table 5: The number of selected features in modules with two or more features.

	Category I (mRNA + Methyl + CNV)	Category II (mRNA + Methyl)	Category III (mRNA + CNV)	Category IV (Methyl + CNV)
GS-iCluster	119	339	150	159
iCluster	43	294	68	57

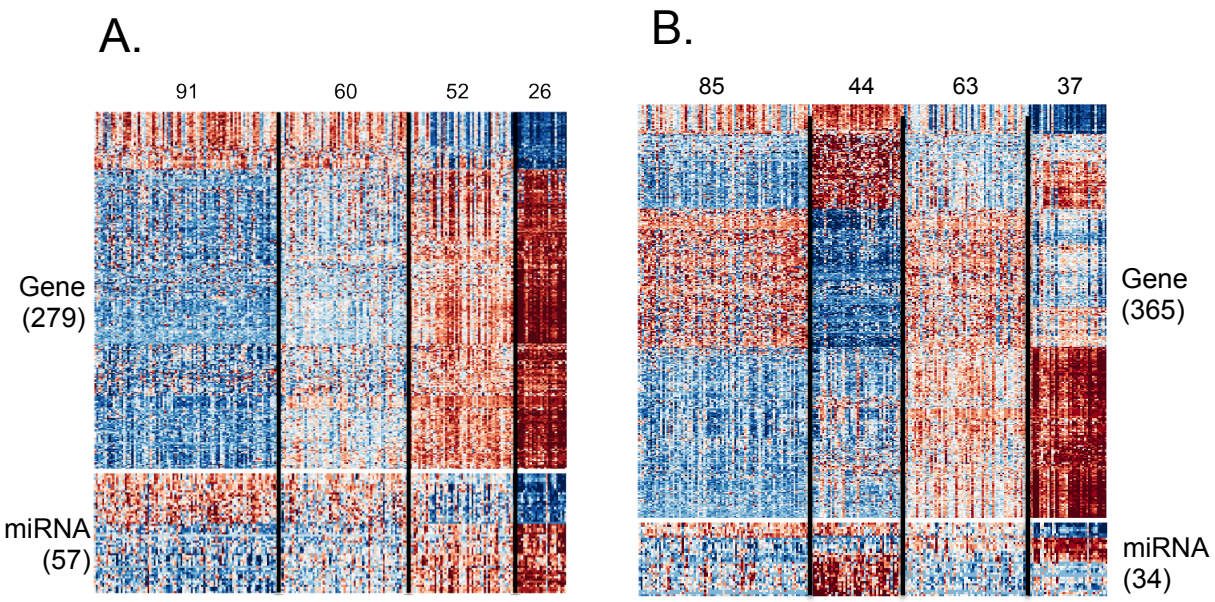

Figure 19: A: Heatmap of two omics (mRNA and miRNA) features selected via Group structured integrative clustering, B:Heatmap of two omics (mRNA and miRNA) features selected via iCluster

comparable, tuning parameters in both methods are selected to have similar number of selected features (552 in iCluster and 537 in GS-iCluster). iCluster identified 33 miRNAs compared to 40 miRNAs from GS-iCluster (Figure 19A). The miRCancer database of miRNAs and cancer associations (Xie et al., 2013) was obtained to evaluate biological relevance of the findings. Out of total 399 miRNAs in the database, 85 miRNA features are related to breast cancer. Among these 85 miRNAs, 56 miRNAs from GS-iCluster overlap with 17 miRNA features, while miRNAs from iCluster overlap with only 9. Based on Fisher's exact test, GS-iCluster are more likely to detect breast cancer related miRNAs (OR = 2.68, $p = 0.0142$) than iCluster (OR= 1.60, $p = 0.3381$)(Table 6). We further applied GST-iCluster in Figure 18B and excluded 15 scattered samples from the clustering. Similar to the previous example in Section 4.1, we obtained much tighter subtype patterns.

3.5 SIMULATION

In this simulation study, we evaluate the performance of GS-iCluster compared to iCluster for module structure identification and clustering accuracy by comparing with true underlying clusters. Our simulation scheme is as follows:

1. Generate two omics datasets ($s = 1, 2$), one for mRNA expression and the other for methylation. Each dataset contains three true sample clusters ($k = 1, 2, 3$) that are characterized by five gene clusters ($h = 1, 2, \ldots, 5$). Simulated data and corresponding parameters are generated sequentially:
   a. Size of sample cluster k: $N_k \sim \text{Poisson}(15)$;
   b. Mean expression levels of gene clusters: $\mu_{kh} \sim Unif(4, 10)$, with restriction $\max_{p,q} |\mu_{ph} - \mu_{qh}| \geq 0.5$ to ensure that there is enough information to infer the sample clusters;
   c. Template cluster patterns: $X'_{skhi} \sim N(\mu_{kh}, 4), s = 1, 2, k = 1, 2, 3, h = 1, 2, \ldots, 5, i = 1, \ldots, N_k$;
   d. Correlation structure of genes in the same gene cluster: $\sum_{skh}$ is obtained by standardizing $\sum'_{skh}$ with the diagonal elements equal to 1, where $\Sigma'_{skh} \sim W^{-1}(\phi, 100)$.

Table 6: miRNAs set enrichment analysis of miRCancer database

GS-iCluster (p = 0.0142)			
	in DB	Not in DB	Sum
Selected miRNAs	17	21	38
Non-selected miRNAs	39	129	128
Sum	56	150	206

iCluster (p = 0.3381)			
	in DB	Not in DB	Sum
Selected miRNAs	9	16	25
Non-selected miRNAs	47	134	181
Sum	56	150	206

W^{-1} denotes inverse Wishart distribution and $\phi = 0.5 \cdot I + 0.5 \cdot J$, where I and J are identity and all-ones matrices respectively.

e. 10 correlated genes in each gene cluster of mRNA expression dataset and 30 correlated probes in each cluster of the methylation dataset: $\left(X_{skhi1}, \ldots, X_{skhiJ}\right)^T \sim MVN(X'_{skhi}, \sum_{skh})$, where $J = 10$ if $s = 1$ and $J = 30$ if $s = 2$. This generates a total of 50 predictive features in the mRNA dataset and 150 predictive features in the methylation dataset.

f. 450 and 1350 noise genes from $N(0, 4)$ for data set 1 and 2 respectively. The final simulated mRNA expression data matrix contains 500 genes and methylation data matrix has 1,500 features for $N_1 + N_2 + N_3$ samples.

2. Generate m feature modules to reflect cross omics regulatory information. The parameters used to generate module information are similar to those estimates from the TCGA breast cancer data. Each module contains one mRNA expression from a specific gene cluster and n_m methylation features from the same gene cluster with the probability p, where n_m sampled from zero-truncated $Poisson(3)$ and $p \in \{0.8, 1\}$. When $p < 1$, each gene module adds noise features with the probability $1 - p$. We set $m = 5$ for each gene cluster in this simulation (i.e. a total of 25 feature modules), and repeat the simulations 100 times.

Figure 20 shows the clustering performance for GS-iCluster and iCluster. In Figure 20A, estimated sample clusters are compared to the underlying truth using the adjusted rand index (ARI, Vinh et al. 2009). ARI compares similarity of two clustering results (the larger the better). The results from 100 simulations are presented using Loess (Cleveland, 1979) with standard error bars. In this simulation, regularization parameters are tuned to generate different number of selected features on the x-axis. GS-iCluster shows improved clustering accuracy, particularly when small number of features are used for clustering (Figure 20A), due to the incorporation of prior module knowledge.

Modules are biologically interpretable and so potentially more biologically relevant. We then compared the number of identified modules from GS-iCluster and iCluster. To make the results comparable, we compared the number of true modules identified (y-axis) when the number of features used to clustering are similar (x-axis). We found that GS-iCluster

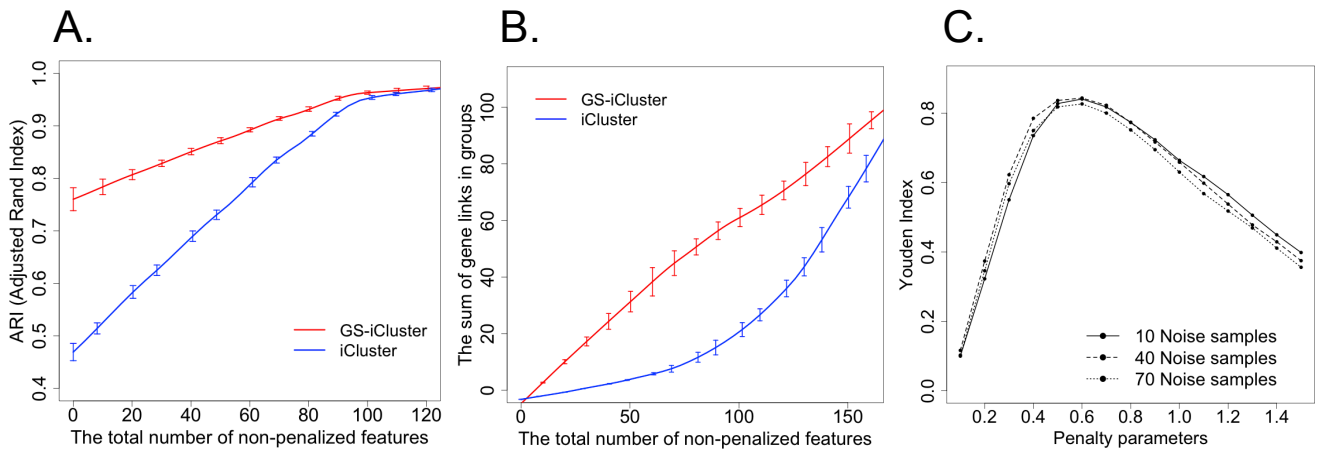

Figure 20: Performance comparisons between Group-structured integrative clustering and standard iCluster.

can find more module genes than iCluster (Figure 20B). To evaluate the robustness of our method, we also compared our methods with iCluster when there exist noise features in the modules (i.e. error knowledge in prior module information). Note that when $p = 1$, there is no noise feature and when $p = 0.8$ there are about 20% noise features. Although adding noise features in the modules reduce the performance of GS-iCluster as expected, GS-iCluster is robust and is still superior to iCluster that does not incorporate module information (Figure 20A-B).

To evaluate the performance of GST-iCluster on tight clustering, N_4 noise samples are added into both miRNA and methylation datasets. The noise features are generated from $N(0, 0.01)$ and the size of N_4 are 10, 40 and 70 respectively. The Youden Index is used as a benchmark to assess the performance of noise sample detection. Large values of the Youden Index (up to 1) represent the better performance. We calculated the Youden index using different penalty parameters $\lambda_1 \in \{0.1, 0.2, \cdots, 1.4\}$. The performance of GST-iCluster mostly depends on the penalty parameters as compared to effects of the number of noise samples. This suggests that we can ignore the proportions of noise samples but focus on the selection of optimal λ_1 (Figure 20C).

3.6 DISCUSSION

The advances of high-throughput experimental technologies and affordable cost result in rapid accumulation of multi-source and multi-platform omics data sets. This trend is expected to continue in the foreseeable future. The complexity of multi-omics datasets brings not only new statistical challenges but also great opportunities for the data integration. Clustering samples for coherent omics signature is an important biological objective, which can lead to findings of novel disease subtypes and deliver meaningful information for tailored treatment and precision medicine. A comprehensive inference of potential inter-omics regulatory mechanisms provides a deeper under- standing of the underlying disease mechanism. We improved the iCluster framework with an sparse overlapping group lasso technique that incorporates prior knowledge of the regulatory information flow (e.g. methylation usually suppresses nearby gene expression) in the modeling. We further adopted a tight clustering concept to allow scattered samples to be left out of meaningful clusters. We demonstrated the benefit of the new framework using simulation and real data with benchmarks of (1) frequency of module identification that reflects consistency of prior regulatory knowledge, (2) clustering accuracy in simulation and (3) pathway enrichment analysis in real data analysis. Results in the two TCGA breast cancer examples shed lights on the disease mechanisms of the discovered subtypes.

4.0 META-ANALYTIC FRAMEWORKS FOR PRINCIPAL COMPONENT ANALYSIS

4.1 INTRODUCTION

Dimension reduction for high-dimension data plays an important role in pattern recognition, classification (Chu et al., 2005), clustering (Bartenhagen et al., 2010) and so on. In particular, principal component analysis (PCA) is the most popular dimension reduction tool to explore high dimensional data through low dimensional space. In addition, principal component space proves to be minimizing the sum of squares of projection errors so that the first r leading eigenvectors and eigenvalues give the best rank-r approximation to an original matrix (Eckart et al., 1936). PCA has been jointly incorporated into many statistical analyses (e.g. regression analysis or multivariate analysis, Jolliffe et al. (2003); Hotelling (1957)) to circumvent the multicollinearity problem, and to alleviate the curse of dimensionality (Bishop et al., 2006). Nevertheless, PCA potentially suffers many practical limitations in high dimension data analysis. For example, noise features of large-scale microarray data often interrupt effective dimension reduction because each principal component involves a linear combination of every variable. Moreover, a number of small but nonzero loadings become huge obstacles for clear interpretations. For a few decades, several alternatives to PCA have been proposed to improve its low interpretability and variable selection, such as non-negative matrix factorization (NMF) (Lee and Seung, 1999) and sparse PCA (Jolliffe et al., 2003). Especially, sparse PCA aims to deal with variable selection problem by regularizing over the eigenvector components. The method is commonly called "sparse PCA" in the literature, and various sparse PCA have been proposed: (Hoyle et al., 2004; Journ ee et al., 2010; Witten et al., 2009; Zou et al., 2006). Zou et al. (2006) proposed sparse PCA (eNet)

based on a regression-type optimization problem via the elastic net penalty. Similar to eNet, Witten et al. (2009) developed sparse PCA that exploits the penalized matrix decomposition (PMD).

Here we introduce three mRNA data sets of mouse metabolism (See the section 4.6). In Figure 21A, PC projections by individual study's eigenvector (row: training study) to each study (column: testing study) clearly loses its discriminant patterns. For example, Figure 21A (training: Liver) shows that the class labels of WT (circle), VLCAD (star) and LCAD (triangle) are mingled together. Lee et al. (2010) reported that PCA often causes distorted PC projections, especially when applied to an independent testing study. To circumvent this problem, we introduce two analytic frameworks to generate common PC space for dimension reduction of multiple homogeneous data. To our best knowledge, an analytic framework for producing common PCA has not been proposed yet.

The Meta-PCA frameworks aim to identify "meta" principal component (Meta-PC) space by using (1) decomposition of the sum of variances matrix (SV) motivated by Flury (1984) and (2) minimization of the sum of squared cosines (SSC) inspired by Krzanowski (1979). With graphical illustrations, we demonstrate how Meta-PCA (SSC) geometrically searches optimum Meta-PC space, and determines the best Meta-PC space dimension. Meta-PCA forms common PC space that simultaneously accounts for variations of multiple omics data sets. Similarly, Lock et al. (2013A) developed an integrative dimension reduction algorithm, the joint and individual analysis of explained variance (JIVE) algorithm. Interestingly JIVE is originally designed for vertical data integration of different omics data but can be applicable to multiple homogeneous data integration (horizontal data integration) by transposing input data, and thus compared with JIVE, we are interested in how effectively Meta-PCA performs visualization and classification among class labels. Extensive simulated data experiments show that Meta-PCAs precisely detects true principal component space and is robust to effects of noise features and outlier samples. With applications to various microarray experiments (Mouse metabolism, Yeast cell cycle, Prostate cancer and TCGA Pan-cancers), we assess whether Meta-PCAs efficiently find separated PC projections onto common eigenvector space. We also propose several Meta-sparsePCA frameworks that penalize components of

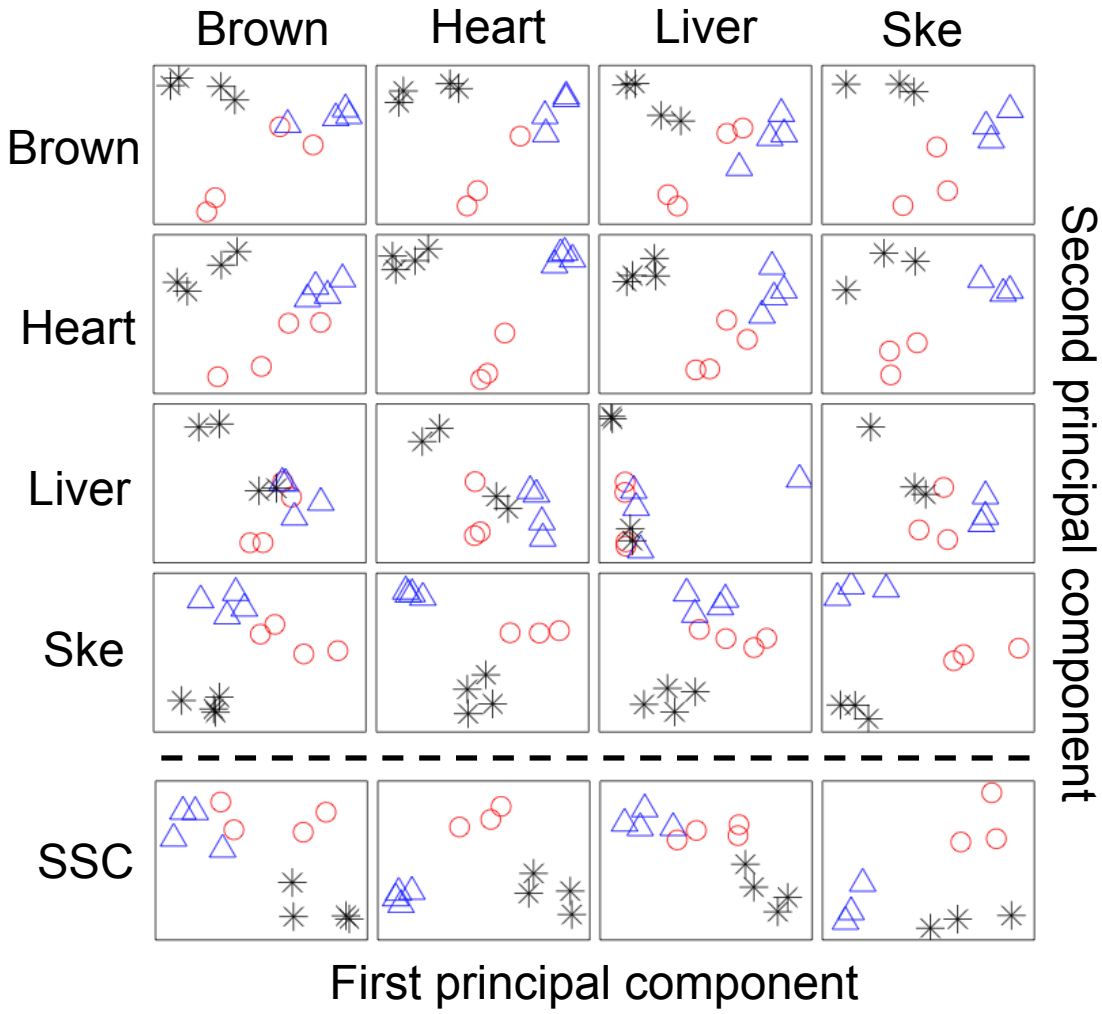

Figure 21: Examples of dimension reduction via PCA and Meta-PCA (SSC) over the four mouse metabolism omics data. The x-axis and y-axis refer to the first and second principal component projection. Red (WT), black (VLCAD), and blue (LCAD) colors represent wild-type, very longchain acyl-coenzyme A dehydrogenase (VLCAD), and longchain acyl-coenzyme A dehydrogenase (LCAD) deficiencies, respectively. Each figure (star, square, circle, and triangle) represents each study label.

common eigenvectors, and test whether Meta-sparsePCA improves Meta-PCA in estimating true eigenvector space by using simulated data and real genomic data.

4.2 METHODS

4.2.1 Meta-PCA via sum of variance decomposition (SV)

Let $X^{(m)}$ be an observed data set of sample size $n^{(m)}$ and p features for study m $(1 \leq m \leq M)$. Denote by $S^{(m)}$ the maximum likelihood (ML) estimate of covariance matrix $\Omega^{(m)}$. To test whether $\Omega^{(m)}$ $(1 \leq m \leq M)$ can produce a common eigenvector space, we consider a null hypothesis, H_0: $L^T\Omega^{(m)}L = \Lambda^{(m)}$ $(1 \leq m \leq M)$, where L is $p \times p$ common eigenvector matrix, and $\Lambda^{(m)}$ is diagonal eigenvalue matrix of study m (Flury, 1984). To circumvent high computational cost, Krzanowski (1979) suggested the simple estimation of common eigenvector matrix L:

$$L^T\Big(\sum_{m=1}^{M} S^{(m)}\Big)L = \Lambda, \tag{4.1}$$

where L and Λ are the eigenvector and eigenvalue matrix of $T = \sum_{m=1}^{M} S^{(m)}$. However, a covariance matrix is subject to a measurement scale and hence a few covariance matrices can excessively dominate components of T. To handle this problem, we propose the weighted covariance matrix by multiplying an inverse of the first leading eigenvalue:

$$T^{SV} = \sum_{m=1}^{M} w^{(m)} S^{(m)}, \tag{4.2}$$

where $w^{(m)}$ is an inverse of the largest eigenvalue of $S^{(m)}$. The largest eigenvalue accounts for a great deal of the variance of principal components (typically more than half of total variance), and hence the inverse of largest eigenvalue adjusts the sum of covariance matrix to be balanced.

By applying (4.2), we propose a meta analytic framework for principal component analysis (Meta-PCA). Define a common eigenvector matrix B subject to:

$$\Big(\sum_{m=1}^{M} w^{(m)} S^{(m)}\Big)B = T^{SV}B = \Lambda^* B, \tag{4.3}$$

where $T^{SV} = \sum_{m=1}^{M} w^{(m)} S^{(m)}$, Λ^* is a diagonal matrix of eigenvalues of T^{SV}, and β_i is an i^{th} column vector of B. To determine the degree of dimension reduction, we adopt by *scree plot* (Cattell et al., 1966) k-column dimensional eigenvector space such that explained variances significantly turn away at the k^{th} principal component. Due to the formulation structure of (4.2), we call this method "Sum of variance decomposition (SV)". Table 7 outlines the framework of Meta-PCA (SV).

Table 7: The algorithm of Meta-PCA via sum of variance decomposition (SV)

(1) Let $X^{(m)}$ be observed data of $n^{(m)}$ samples and p features, by which we estimate the ML estimator of covariance matrix $S^{(m)}$ $(1 \leq m \leq M)$.

(2) Calculate $w^{(m)}$, an inverse of the largest eigenvalue of $S^{(m)}$.

(3) Perform the eigen decomposition of T^{SV} to obtain B, an eigenvector matrix of T^{SV},

$$\Big(\sum_{m=1}^{M} w^{(m)} S^{(m)}\Big) B = T^{SV} B = \Lambda^* B,$$

where Λ^* is a diagonal matrix of eigenvalues of T^{SV}.

(4) Choose the optimal k dimension of eigenvector matrix by *scree plot*, and thereby obtain the meta eigenvector matrix $B^{SV} = (\beta_1, \ldots, \beta_k)$.

4.2.2 Meta-PCA via Sum of squared cosine (SSC) maximization

Suppose that we derive a $j^{(m)}$ column dimensional eigenvector matrix $V^{(m)} = (v_1^{(m)}, \ldots, v_{j^{(m)}}^{(m)})$ of study $X^{(m)}$, where $v_i^{(m)}$ is i^{th} leading eigenvector for study m. Let $\delta^{(m)}$ be an angle between an arbitrary vector $g \in \mathbb{R}^p$ and a vector most nearly parallel to g in the space generated by $j^{(m)}$ principal components of study m. The vector g that maximizes the sum of squared cosine (i.e., $\sum_{m=1}^{M} cos^2 \delta^{(m)}$) is given by an eigenvector of $\sum_{m=1}^{M} V^{(m)} V^{(m)^T}$ corresponding to the largest eigenvalues λ_1 (Krzanowski, 1979):

$$\Big(\sum_{m=1}^{M} V^{(m)} V^{(m)^T}\Big) g = \lambda_1 g. \tag{4.4}$$

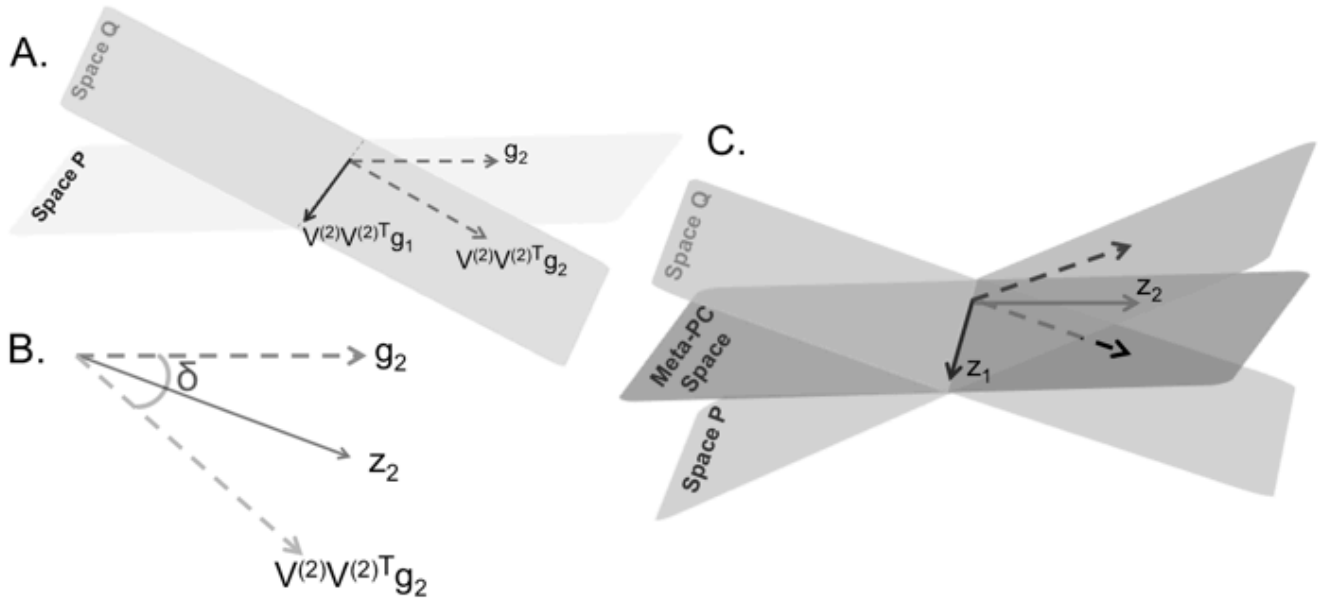

Figure 22: Geometrical illustrations for common principal component space (SSC).

The following example demonstrates an approach exploits topological comparison of multiple principal component (PC) projection to identify common PC projections. Consider two studies $X^{(1)}$ and $X^{(2)}$: Let P and Q be arbitrary subspace of two PC projections generated by three column dimensional eigenvector matrix of $V^{(1)}$ ($\in \mathbb{R}^{p\times 3}$) and $V^{(2)}V^{(2)^T}V^{(1)}$ ($\in \mathbb{R}^{p\times 3}$), respectively (i.e., $P, Q \in \mathbb{R}^3$ and $j^{(1)} = j^{(2)} = 3$) (Figure 22A). Denote by g_1 and g_2 two arbitrary orthogonal vectors in subspace P. In Theorem 1 (Krzanowski (1979)), the vectors $V^{(2)}V^{(2)^T}g_1$ and $V^{(2)}V^{(2)^T}g_2$ laid on space Q are most parallel to g_1 and g_2 on space P, respectively. Here, we define a common PC projection– i.e., bisector z_1 passing by between g_1 and $V^{(2)}V^{(2)^T}g_1$, where $z_1 = \frac{1}{\sqrt{(1+3\lambda_1)}}\left(I + V^{(2)}V^{(2)^T}\right)g_1$, λ_1 is the largest eigenvalue of $\left(I + V^{(2)}V^{(2)^T}\right)$, and I is an identity matrix. The second bisector z_2 perpendicular to z_1 is given by $\frac{1}{\sqrt{(1+3\lambda_2)}}\left(I + V^{(2)}V^{(2)^T}\right)g_2$, where λ_2 is the second largest eigenvalue of $\left(I + V^{(2)}V^{(2)^T}\right)$. In Figure 22B, the second bisector z_2 (solid) topologically best represents both two vectors g_2 (dotted) and $\left(I + V^{(2)}V^{(2)^T}\right)g_2$ (dotted). We also define the space

spanned by k bisectors as k-dimensional meta principal component (Meta-PC) space. For example, Figure 22C presents Meta-PC space (in the middle) spanned by z_1 and z_2, which can topologically best explains both space P and Q. By extension, in case of M (≥ 3) studies we can derive a common eigenvector matrix by eigenvectors of $\sum_{m=1}^{M} V^{(m)}V^{(m)^T}$, which best accounts for all M individual eigenvector spaces (Theorem 3, Krzanowski (1979)).

Motivated by (4.4), we introduce the second Meta-PCA framework. We first estimate a meta eigenvector matrix B^{SSC} by applying the eigen decomposition to $T^{SSC} = \sum_{m=1}^{M} V^{(m)*}V^{(m)*^T}$:

$$\Big(\sum_{m=1}^{M} V^{(m)*}V^{(m)*^T}\Big)B^{SSC} = \Lambda^* B^{SSC}, \tag{4.5}$$

where $V^{(m)*}$ is a matrix consisting of $j^{(m)}$ leading eigenvectors, Λ^* is a diagonal eigenvalue matrix, and $B^{SSC} = (\beta_1^{SSC}, \ldots, \beta_k^{SSC})$. To select dimension $j^{(m)}$ of $V^{(m)}$, we suggest a choice of $j^{(m)}$ such that PC projection explains more than 80% of total variance for each data set, since over 80% is commonly accepted as sufficient in genomic data analysis. To determine the optimal dimension k, we utilize the *scree plot* method (Cattell et al., 1966). By definition, the meta-eigenvector β^{SSC} maximizes $\sum_{m=1}^{M} cos^2\delta^{(m)}$ since $\delta^{(m)}$ is the minimized angle between β^{SSC} and k dimensional PC projection of study m. Considering the formulation property, we name this approach "sum of squared cosine maximization (SSC)". Table 8 outlines the framework of Meta-PCA (SSC).

4.2.3 Variable selection of Meta-PCAs (Meta-sparsePCA)

It is commonplace that large-scale microarray data contains a vast amount of noise features that often undermine effective dimension reduction. Such noise features often interrupts effective dimension reduction because each principal component involves a linear combination of whole variables. Moreover, a number of small but nonzero loadings often causes low interpretability of principle components. In this section, we introduce regularized Meta-PCA frameworks (Meta-SparsePCA) for variable selection for which we consider two sparse PCA methods: (1) regression-type sparse PCA together with elastic net penalty (eNet) (Zou et al., 2006) (2) sparse PCA based on penalized matrix decomposition (PMD) (Witten et al.,

Table 8: The algorithm of Meta-PCA (Sum of squared cosine (SSC) maximization)

(1) Let $X^{(m)}$ be data matrix of $n^{(m)}$ samples and p features, and $V^{(m)}$ be an eigenvector matrix of $X^{(m)}$ $(1 \leq m \leq M)$.

(2) Choose $j^{(m)}$ such that V_m^* consists of $j^{(m)}$ leading eigenvectors.

(3) Perform eigen decomposition of $T^{SSC} = \sum_{m=1}^{M} V^{(m)*}V^{(m)*^T}$:

$$\Big(\sum_{m=1}^{M} V^{(m)*}V^{(m)*^T}\Big)B^{SSC} = \Lambda^* B^{SSC}.$$

(4) Choose the optimal k dimensional by scree plot, and derive $B^{SSC} = (\beta_1^{SSC}, \ldots, \beta_k^{SSC})$.

2009). Here we propose four frameworks of Meta-SparsePCA by applying the two sparse PCA methods. More precisely, Table 9 describes the four methods of Meta-SparsePCAs, where the two sparse PCA methods (eNet and PMD) are applied to two Meta-PCA's objective formulations (T^{SV} and T^{SSC}). Each sparse PCA algorithm is characterized with distinct advantages depending on experimental scenarios, and thus it is worthwhile to evaluate all candidate sparse Meta-PCAs to find the best sparse PCA. The best choice of penalization constant is another problem, for which we develop an *scree plot* tool on the basis of explained variances. (See the section 4.6).

4.3 SIMULATION STUDY

4.3.1 True eigenvector detection of Meta-PCA

In this section, we evaluate the two proposed Meta-PCA frameworks (SV and SSC) compared with the standard PCA and JIVE. We first define a benchmark ω to assess the similarity

Table 9: The four proposed methods of Meta-SparsePCAs for variable selection

Method 1: SSC + PMD
Estimate Meta-sparse eigenvectors (SSC) by the sparse PCA method (PMD),
$(U^*, B^*) = \text{argmax}_{U,B} U^T T^{SSC} B$ subject to $\|B\|_2^2 \leq 1, \|B\|_1 \leq \lambda$ and $\|U\|_2^2 \leq 1$,
where $B^* = (\beta_1^*, \cdots, \beta_k^*) \in R^{p \times k}$
Method 2: SSC + eNet
Estimate Meta-sparse eigenvectors (SSC) by the sparse PCA method (eNet),
$(A^*, B^*) = \text{argmax}_{A,B} \sum_{i=1}^{p} \|t_i - AB^T t_i\|^2 + \lambda \sum_{j=1}^{k} \|\beta_j\|^2 + \sum_{j=1}^{k} \lambda_{1,j} \|\beta_j\|_1$ subject to $A^T A = I$,
where $B^* = (\beta_1^*, \cdots, \beta_k^*) \in R^{p \times k}$ and t_i is the i^{th} column vector of $T^{SSC} (1 \leq i \leq p)$.
Method 3: SV + PMD
Estimate Meta-sparse eigenvectors (SV) by the sparse PCA method (PMD)
$(U^*, B^*) = \text{argmax}_{U,B} U^T T^{SV} B$ subject to $\|B\|_2^2 \leq 1, \|B\|_1 \leq \lambda$ and $\|U\|_2^2 \leq 1$,
where $B^* = (\beta_1^*, \cdots, \beta_k^*) \in R^{p \times k}$.
Method 4: SV + eNet
Estimate Meta-sparse eigenvectors (SV) by the sparse PCA method (eNet),
$(A^*, B^*) = \text{argmax}_{A,B} \sum_{i=1}^{p} \|t_i - AB^T t_i\|^2 + \lambda \sum_{j=1}^{k} \|\beta_j\|^2 + \sum_{j=1}^{k} \lambda_{1,j} \|\beta_j\|_1$ subject to $A^T A = I$,
where $B^* = (\beta_1^*, \cdots, \beta_k^*) \in R^{p \times k}$ and t_i is the i^{th} column vector of T^{SV} $(1 \leq i \leq p)$.

between two principal component space. Consider two eigenvector matrices $V^{(1)}$ and $V^{(2)}$, where $V^{(1)} = (v_1^{(1)}, \ldots, v_{j_1}^{(1)}) \in \mathbb{R}^{p \times j_1}$ and $V^{(2)} = (v_1^{(2)}, \ldots, v_{j_2}^{(2)}) \in \mathbb{R}^{p \times j_2}$. The evaluation measure ω is given by:

$$\begin{aligned} \omega(V^{(1)}, V^{(2)}) &= \sum_{i=1}^{j_1} \lambda_i \\ &= \text{tr}\big(V^{(1)^T} V^{(2)} V^{(2)^T} V^{(1)}\big), \end{aligned}$$

where λ_i is the i^{th} largest eigenvalue of $V^{(1)^T}V^{(2)}V^{(2)^T}V^{(1)}$. Krzanowski (1979) proved $\sum_{i=1}^{j_1}\lambda_i$ is associated with angles between two eigenvectors of $V^{(1)}$ and $V^{(2)}$, and hence $\omega(V^{(1)}, V^{(2)})$ gauges the geometrical similarity between two matrices $V^{(1)}$ and $V^{(2)}$. Using ω, we test whether the Meta-PCA frameworks effectively estimate true eigenvectors. Below we demonstrate details of our simulation scenarios:

Step 1 (Define true eigenvectors): Let $E = (e_1, e_2)$ and $\lambda = (\lambda_1, \lambda_2)$ be a true eigenvector matrix of two feature clusters (c_1 and c_2) and its corresponding true eigenvalues, where $e_1 = (\underbrace{1, 1, \ldots, 1}_{10}, 0, \ldots, 0)/\sqrt{10} \in \mathrm{R}^{200}$ and $e_2 = (\underbrace{0, 0, \ldots, 0}_{10}, \underbrace{1, 1, \ldots, 1}_{10} 0, \ldots, 0)/\sqrt{10} \in \mathbb{R}^{200}$, $\lambda_1 = 1000, \lambda_2 = 800$, $c_1 \in \{1, 2, \ldots, 10\}$ and $c_2 \in \{11, 12, \ldots, 20\}$.

Step 2 (Generate random data sets): We define a true covariance matrix $\Sigma^{(m)}$, where $\Sigma^{(m)} = \Sigma + E^{*(m)}$ for $1 \leq m \leq M$, where $E^{*(m)} = \epsilon^{(m)} \cdot \epsilon^{(m)^T}$, $\epsilon^{(m)} \sim MVN_p(\mathbf{0}, W)$, $W = I \times C$, $C \in \{0.1, 0.5, 1\}$ and $I_{p\times p}$ is an identity matrix. Generate M simulated data sets of 20 samples and 200 features, $X^{(m)} = (x_1^{(m)}, \cdots, x_{200}^{(m)}) \sim MVN_{200}(\mathbf{0}, \Sigma^{(m)})$ for $1 \leq m \leq M$ and $1 \leq M \leq 10$. Note that ten features that belong to each feature cluster (c_1 and c_2) are highly correlated respectively, since e_1 and e_2 are orthogonal and the eigenvalue values (λ_1 and λ_2) are considerably large. By multiplying the true eigenvectors and eigenvalues, we create a symmetric matrix Σ, where $\Sigma = e^T\lambda e$.

We derive an eigenvector matrix $V^{(m)} = (v_1^{(m)}, v_2^{(m)})$ from simulated data sets $X^{(m)}$ ($1 \leq m \leq M$). By utilizing M data sets $\left(\text{i.e., } X^{(1)}, X^{(2)} \ldots, X^{(M)}\right)$, the Meta-PCA frameworks (SV, SSC) produce two dimensional meta-eigenvector matrices, $B^{SV} = (\beta_1^{SV}, \beta_2^{SV}) \in \mathbb{R}^{200\times 2}$ and $B^{SSC} = (\beta_1^{SSC}, \beta_2^{SSC}) \in \mathbb{R}^{200\times 2}$. To evaluate the similarity between the derived eigenvector matrix (B^{SV}, B^{SSC}, V^{JIVE} or $V^{(m)}$) and the true eigenvector matrix ($= E$), we calculate $\omega(E, B^{SV})$, $\omega(E, B^{SSC})$, $\omega(E, V^{JIVE})$, and $\omega(E, V^{(m)})$ such that by definition ω ranges from 0 to 2 due to two column dimensional space of E. The simulations are repeated 50 times and average values are presented.

Figure 23A implicates the meta eigenvectors (B^{SV} and B^{SSC}) more precisely estimate the true eigenvectors (E) than JIVE and standard PCA. When the variance parameter (C) changes (i.e., $C \in \{0.1, 0.5, 1\}$ in Figure 3A, 3B and 3C, respectively) evaluation measures consistently drop down (i.e., the highest values (SSC) of each scenario are 1.95, 1.75 and 1.48), yet the two Meta-PCAs consistently show superior performance against JIVE and

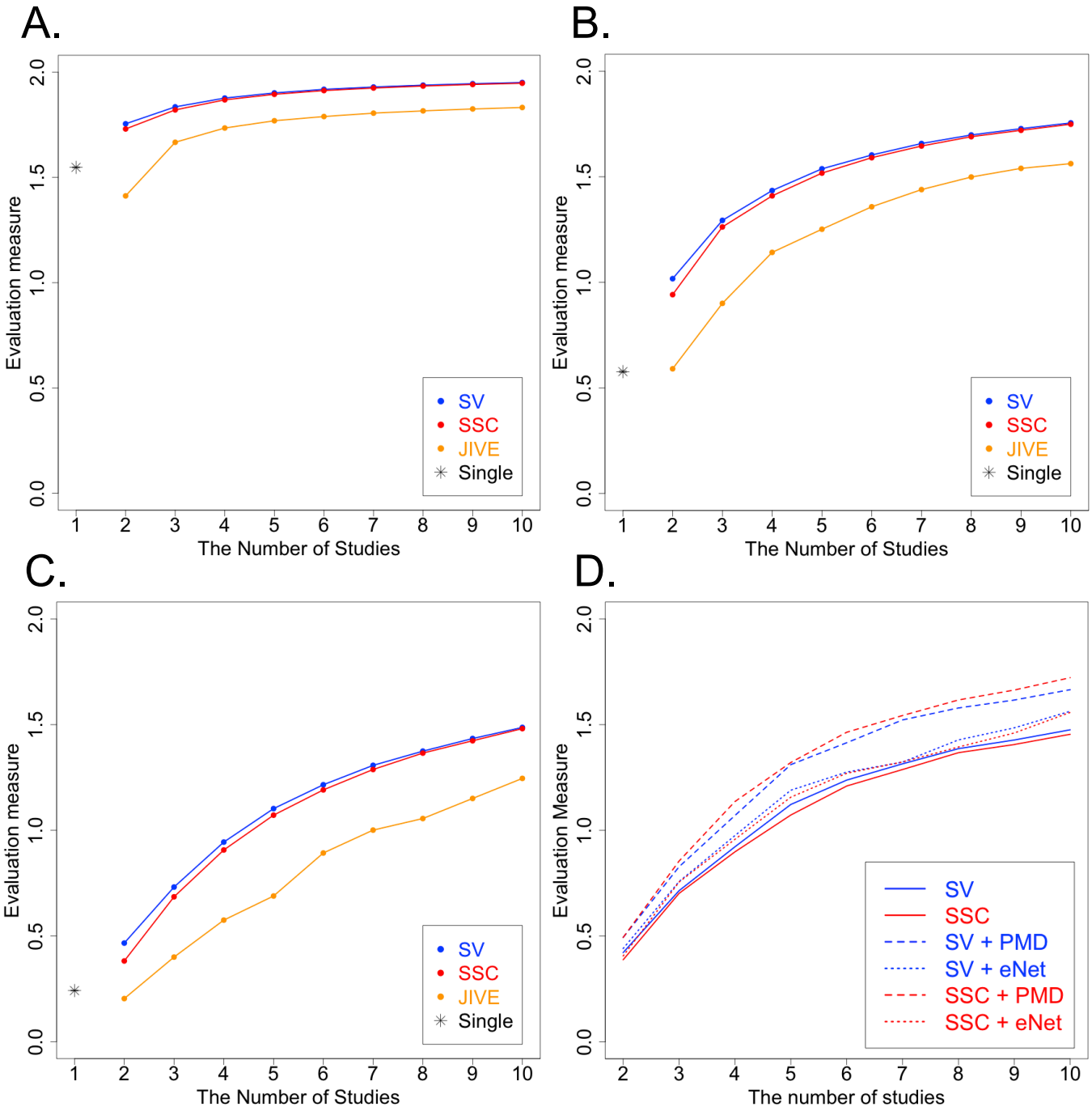

Figure 23: Performance comparisons (Meta-PCAs, PCA and JIVE) of the effects on the number of studies for estimating true eigenvector. "SV", "SSC" refer to Meta-PCA (SV) and Meta-PCA (SSC). "Single" represents standard PCA of each individual study (A: $C = 0.1$, B: $C = 0.5$, C and D: $C = 1$).

standard PCA. The two Meta-PCAs clearly improves in identifying the true eigenvector space. While Meta-PCA (SV) method slightly better performs than Meta-PCA (SSC), the difference does not seem noticeably significant.

Figure 23D compares the four Meta-sparsePCAs and two Meta-PCAs. We notice all the four Meta-sparsePCAs outperform the two Meta-PCAs in identifying the true eigenvectors (the highest values are 1.65 (SSC+PMD), 1.61 (SV+PMD), 1.49 (SV+eNet) and 1.48 (SSC+eNet), respectively). These consequences analytically make sense, in the sense that the sparse Meta-PCA regularizes non-influencing elements, and thereby it promotes to leave only influential eigenvector components. Besides we found the method (SSC + PMD) consistently performs the best among the four Meta-sparsePCAs as the number of studies increases. We therefore recommend to utilize the Meta-sparsePCA (SSC + PMD) for future analysis.

4.3.2 Robustness of Meta-PCA

High-throughput microarray data, for the most part, contain quite a few noise features. Unless we filter noise features, PCA likely fails to perform effective dimension reduction. In this section, we test whether Meta-PCAs are robust to effects of noise features and outlier samples. To mimic real data, we adopt the simulation scenario introduced by Qiu et al. (2006) that generates simulated data sets with an adjustment of cluster separation levels, noise features and outlier samples.

Here we describe details of the simulation scenarios. We generate a full data set of samples that separate one of three clusters, where each cluster includes 100 samples and 100 features (the method to define clusters, see Qiu et al. (2006)). We then add noise features (20, 60 and 100) and outlier samples (0, 10 and 30). We then randomly split the full data set into four subsets, where each subset holds an equal cluster size (i.e., 20 samples from each cluster). Finally we impose equal outlier samples to each subset. Denote by $X^{(m)}$ $(1 \leq m \leq 4)$ four data sets. To generate data, we utilize "clusterGeneration" package Qiu et al. (2006) in R (http://www.r-project.org/). To benchmark each method, we exploit the

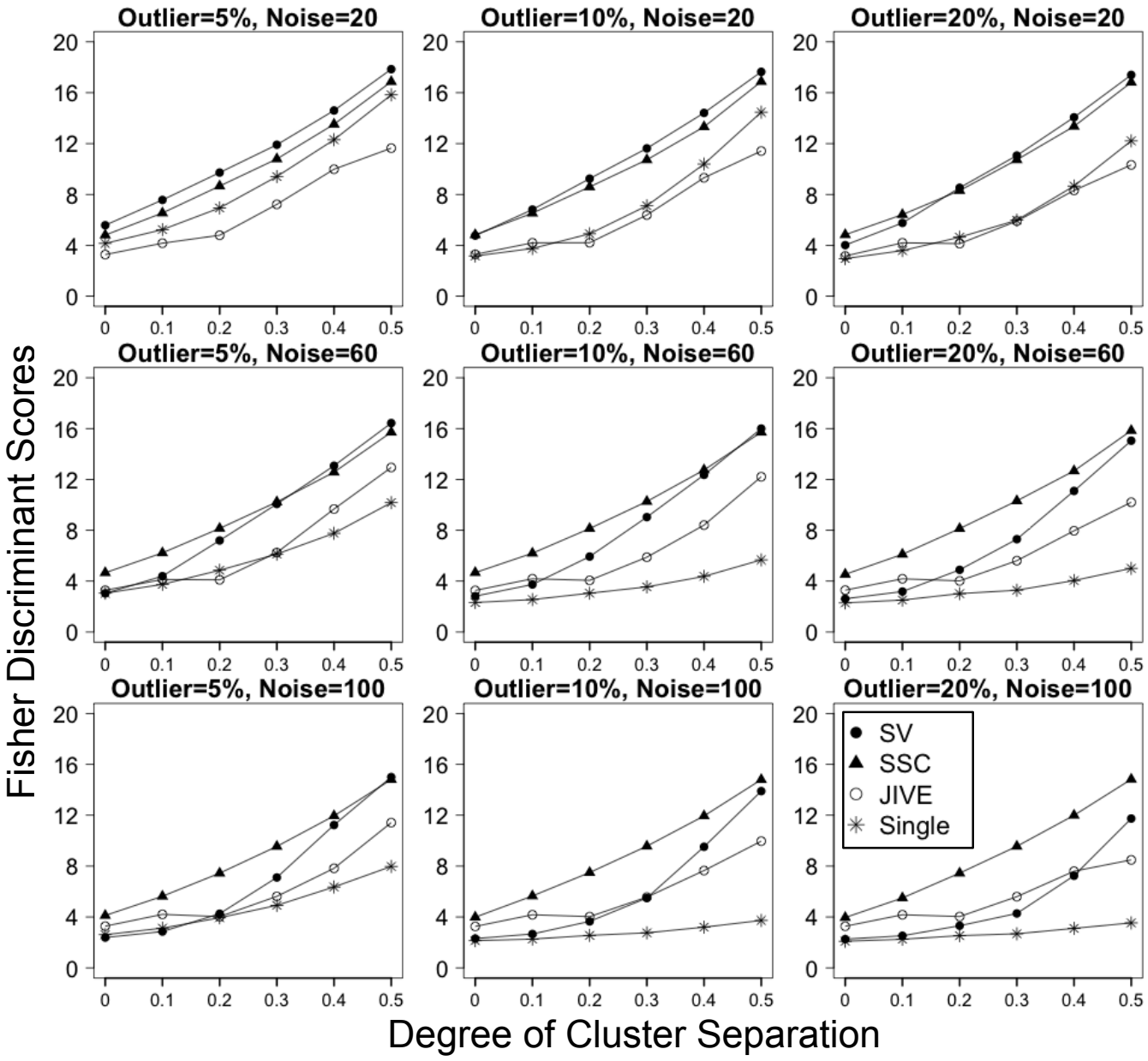

Figure 24: Robustness comparisons of Meta-PCA, JIVE and PCA to outliers and noises. The y-axis represents the averages of Fisher discriminant scores, and the x-axis the magnitude of cluster separation. The figure presents the two MetaPCA methods SV (dot), SSC (triangle), JIVE (circle) and standard PCA (Single, star) applied to each individual study.

Fisher discriminant scores defined by

$$\frac{v^T S_B v}{v^T S_W v} \tag{4.6}$$

where S_B and S_W are the between-group covariance matrix and the within-group covariance matrix, and v is a line direction unit vector that maximizes (4.6) (Duda, 2014). The simulations are repeated 100 times and average values are presented.

Figure 24 reveals that the two Meta-PCAs (SV, SSC) more effectively distinguish the three clusters, considering the two Meta-PCAs (SV, SSC) present higher Fisher discriminant scores than they appear in standard PCA (Single) and JIVE. This result implies the two Meta-PCA frameworks are robust to effects of outliers and noises, compared with JIVE and standard PCA. More interestingly, the magnitude of outliers and noises almost proportionately affects the performance of Meta-PCA (SSC) (i.e., linearly increase of Fisher discriminant scores), whereas Meta-PCA (SV) is more subject to noise and outlier effects and even become worse than Meta-PCA (SSC) in some scenarios (e.g., the panel of Outlier = 20%, Noise = 100). When the separation degree is small, Meta-PCA (SV) tend to poorly separate clusters (e.g., the separation degree from 0 to 0.2 in all scenarios with "Noise" being more than 60"). Taken together, we recommend exploiting Meta-PCA (SSC) for future analysis.

4.4 APPLICATION TO REAL DATA SETS

In this section, we apply Meta-PCA (SSC) and Meta-SparsePCA (SSC + PMD) to various high-throughput microarray data sets. We obtained mRNA expression and methylation expression data of various diseases from GEO (http://www.ncbi.nlm.nih.gov/geo/) and TCGA Portal (https://tcga-data.nci.nih.gov/tcga/). We examine whether the proposed Meta-PCA (SSC) and Meta-SparsePCA (SSC + PMD) effectively visualizes Meta-PC projections for multiple studies.

4.4.1 Spellman's cell cycle data

Spellman's yeast cell cycle data (Spellman et al., 1998) include time-dependent gene expression profiles to monitor transcriptomic variation during yeast cell cycles. Yeast cells were arrested to the same cell cycle stage using four different synchronizing methods: α arrest (alpha), arrest of cdc15 or cdc28 temperature-sensitive mutant, and elutriation (elu). A total of 18, 24, 17 and 14 time points were considered for each synchronization. Since the diverse synchronization methods can potentially lead to heterogeneity, we divided the samples into four data sets depending on synchronizing methods (alpha, cdc15, cdc28, and elu). Due to nature of iterative cell cycle, the expression profiles are well-characterized in cyclic patterns (Spellman et al., 1998). We matched up features across all the four studies and filtered out features using standard deviation (i.e. SD ≤ 0.45, non-informative features with smaller variation) that left 1,025 features. We imputed missing values via R package "impute" (Hastie et al., 2014). We applied the frameworks of Meta-PCA (SSC) and Meta-SparsePCA (SSC + PMD) to assess whether they effectively reveal cyclic patterns of gene expression profiles compared with JIVE and PCA of single study.

In Figure 25, each row of panels refers to training study to estimate the leading top two eigenvectors. The column refers to testing study that produces PC projections onto the trained eigenvector space. The numbers on the lines indicate time points for two cell cycles. PC projections at the panel of first row and first column ("alpha") clearly delineates the cyclic pattern of cell cycles, whereas PC projections of the second row and second column panel ("cdc15") oscillates as time points increase. Interestingly, this non-cyclic effect was already reported in Li et al. (2002). Note that Meta-PCA (SSC) consistently captures the cyclic pattern of PC projections across all the four studies. In particular, Meta-PC (SSC) projections of the study ("cdc15") remarkably recover its cyclic pattern. This result evidences that Meta-PCA (SSC) borrows and combines information derived from underlying true common eigenvectors of other studies, and hence Meta-PCA (SSC) facilitate to estimate true common eigenvector space. In Figure 25, it is obvious that Meta-PCA (SSC) and Meta-SparsePCA (SSC + PMD) presents more noticeable cyclic patterns than they appear in JIVE (e.g., non-clear cyclic patterns of the study ("cdc28")) and standard PCA.

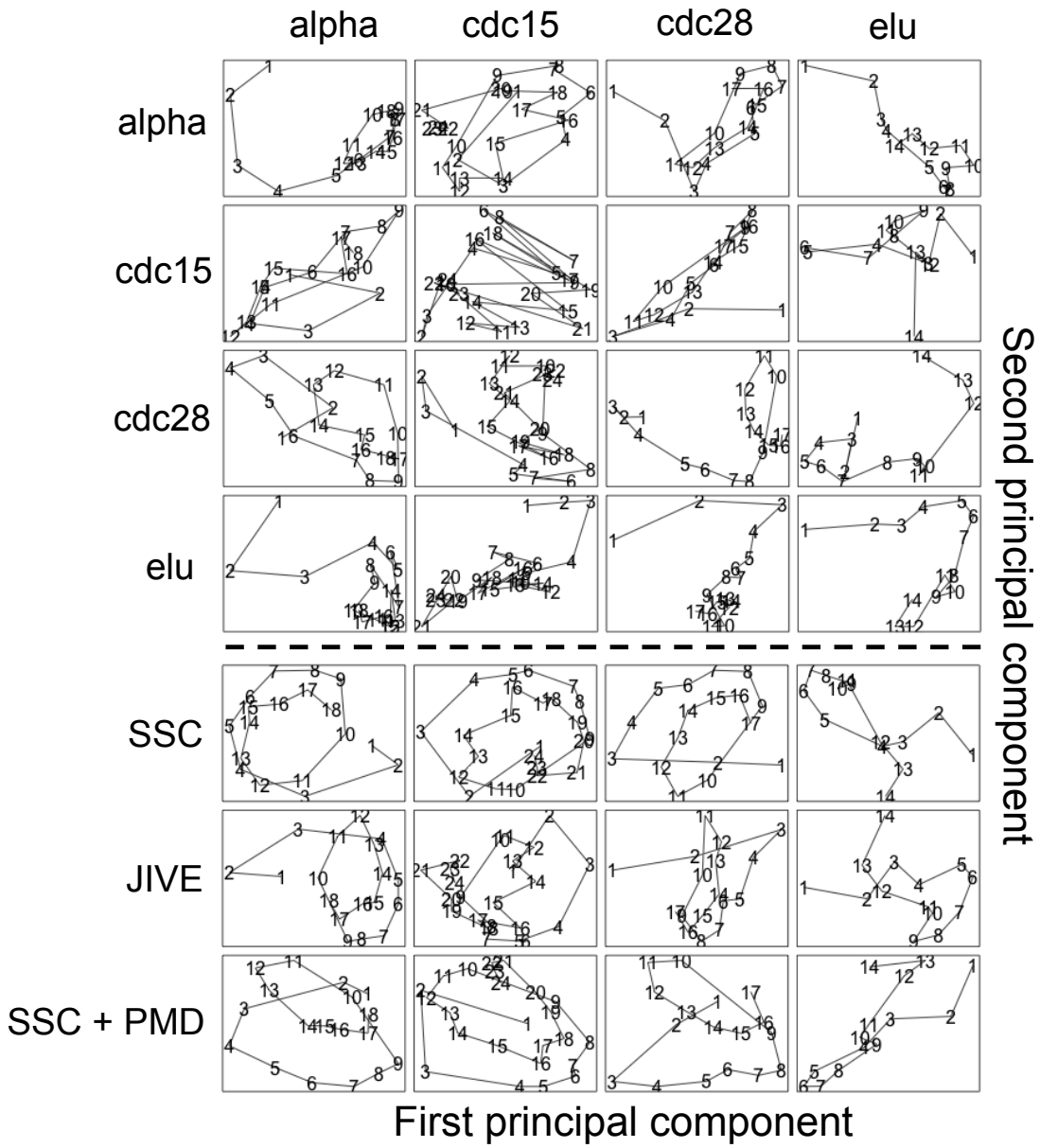

Figure 25: Two dimensional PC projections of PCA, Meta-PCAs (SV, SSC), JIVE using four mRNA expression data sets of Spellman's yeast cellcycle experiment. The numbers on the lines indicate time point during the two cell cycles. The first and second PC projection are on the x-axis and y-axis of each panel, respectively.

4.4.2 Prostate cancer data

In this section, we analyze data sets of four microarray experiments (Lapointe et al. (2004); Tomlins et al. (2006); Varambally et al. (2005); Yu et al. (2004)), where each patient belong to one of three class labels (normal, primary, and metastasis; See Table 10). We matched up features across the four studies and filtered non-informative features by the rank sum of feature mean and standard deviation (mean < 0.1, SD < 0.1; Wang et al. (2012)), and imputed missing values that left 3,056 features.

Table 11 presents the result of Fisher discriminant scores and their average values over the four studies. On average the standard PCA of "Lapointe" and "Tomlins" poorly distinguish the class labels of samples (12.94 and 12.10 for each) compared with "Yu" and "Varambally" (16.50 and 18.81 for each), while Meta-PCA (SSC) results in moderately high discriminant effects (16.56). In Figure 26, Meta-PC projections (SSC) reveal the transition pattern from normal (star) to primary tumor (square) and to metastasis tissues (circle). Note that the first leading Meta-PC (x-axis) projection accounts for larger variances across the class labels than the second leading Meta-PC (y-axis). We also observe the class separation via JIVE appears not as obvious as both Meta-PCA (SSC) and Meta-sparsePCA (SSC+PMD) (i.e., SSC (16.56), SSC+PMD (18.93), JIVE (10.36)).

Table 10: The summary of four prostate cancer data.

Author	Year	Platform	Sample Size	Source
Lapointe et al.	2004	cDNA	103	GSE3933
Tomlins et al.	2006	cDNA	57	GSE6099
Varambally et al.	2005	HG-U133 Plus 2	13	GSE3325
Yu et al.	2004	HG-U95Av2	146	GSE6919

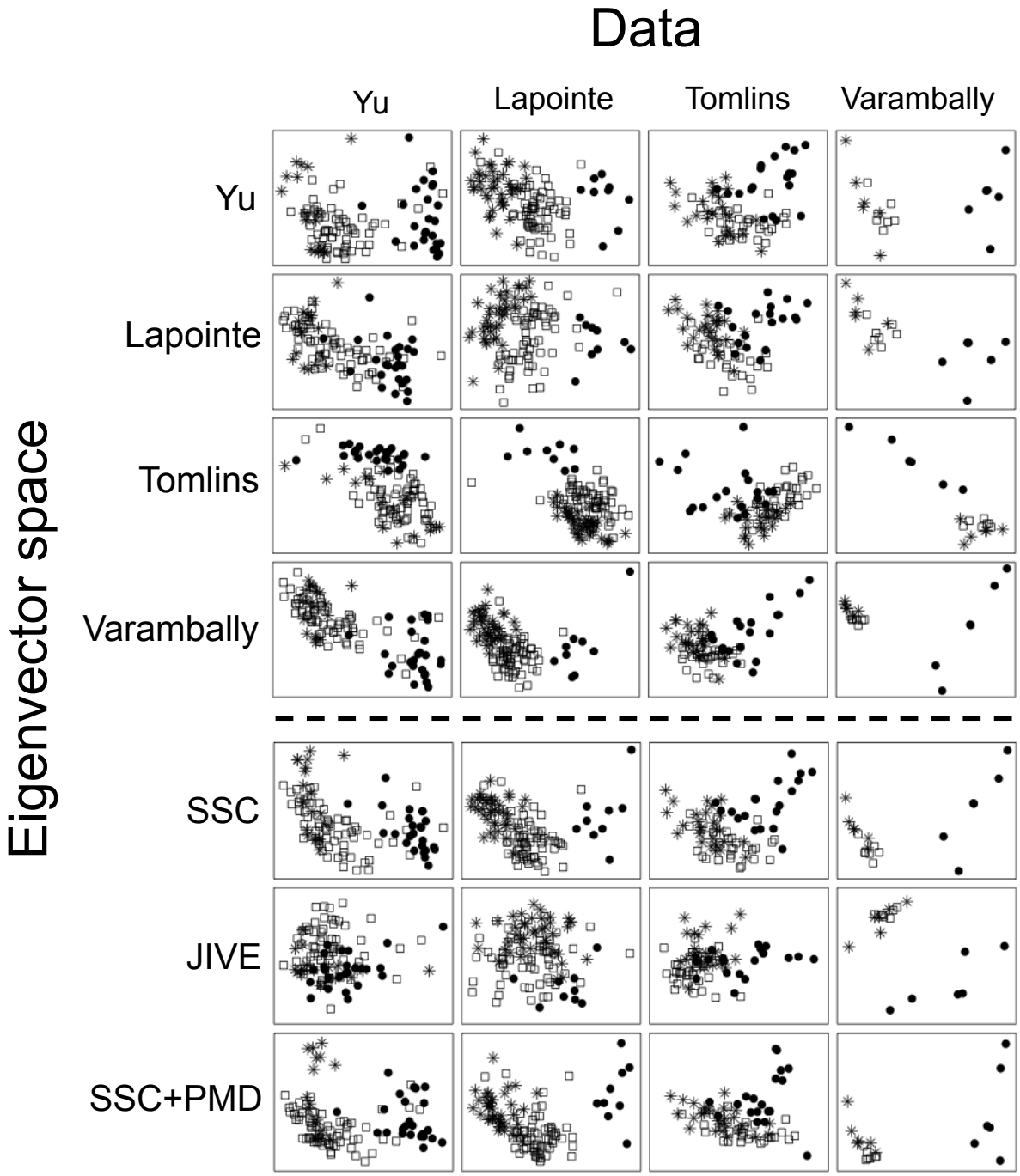

Figure 26: Two dimensional PC projections using four prostate cancer mRNA expression data sets; star (normal), square (primary tumor) and circle (metastasis tissues). The first and second PC projections are on the x-axis and y-axis, respectively.

Table 11: Fisher discriminant scores of PC projections (prostate cancer data).

	Yu	Lapointe	Tomlins	Varambally	Average
Yu	15.374	24.01	10.58	16.04	16.50
Lapointe	9	21.20	11.14	10.41	12.94
Tomlins	9.822	19.86	10.67	8.04	12.10
Varambally	11.955	26.41	10.69	26.17	18.81
SSC	14.712	26.45	11.37	13.69	16.56
JIVE	5.716	11.01	9.07	15.65	10.36
SSC+PMD	15.341	29.80	9.40	21.17	18.93

4.4.3 TCGA cancer data

In this section, we apply Meta-PCA (SSC) to TCGA cancers data sets (Level 3 DNA methylation of beta values targeting on methylated and the unmethylated probes; https://tcga-data.nci.nih.gov/tcga/). We retrieve the six cancer types (Breast carcinoma (BRCA), Colon carcinoma (COAD), Kidney renal clear cell carcinoma (KIRC), Lung adenocarcinoma (LUAD), rectum adenocarcinoma (READ), and Stomach Adenocarcinoma (STAD)) to explore common PC projection patterns between tumor and normal, including but not limited to, between male and female. We matched up features across all studies and filtered out probes by the rank sum of feature mean and standard deviation (mean < 0.7, SD < 0.7; Wang et al. (2012)), and thereby we selected 910 probes. Table 12 describes details of TCGA data.

In Figure 27, Meta-sparsePCA (SSC+PMD) layouts distinguishable PC projections over the six cancer data sets across the four class labels (i.e., Tumor, Normal, Male and Female). The normal samples (dot) are mostly distributed on the right side of first PC (x-axis), whereas the tumor samples (square) are at the left side of the first PC. For the most part, female samples (grey) are projected to the upper side of panels, while male samples (black) are on the bottom side of panels. The projections via JIVE, by contrast, do not assemble to-

gether to several focal points but scatter (see Table 13). Since Meta-sparsePCA (SSC+PMD) gives an clear cut of the class labels (especially between genders), this exploratory analysis implicates there exist common factors of cancer development revealed by the leading meta sparse eigenvectors. Consequently, we may address a potential biological hypothesis associated with common oncogenic factors that lie over the six different cancer types as post-hoc analysis.

Table 12: The summary of six TCGA methylation data.

Type	Platform	# of genes	Sample Size	Source
BRCA	HumanMethylation27	13,311	350	The Cancer Genome Atlas
COAD	HumanMethylation27	13,169	215	The Cancer Genome Atlas
KIRC	HumanMethylation27	12,606	427	The Cancer Genome Atlas
LUAD	HumanMethylation27	12,709	157	The Cancer Genome Atlas
READ	HumanMethylation27	13,295	84	The Cancer Genome Atlas
STAD	HumanMethylation27	13,196	114	The Cancer Genome Atlas

4.4.4 Mouse Metabolism Data

It is commonly known that an energy metabolism disorder in children is relevant to very longchain acyl-coenzyme A dehydrogenase (VLCAD) deficiencies. LCAD-deficient mice have impaired fatty acid oxidation, and suffer from disorders of mitochondrial fatty acid oxidation. We consider microarray experiments of mouse metabolism which were introduced and analyzed in (Li et al., 2011). The data sets include mice profiles of three genotypes: wild-type (WT), LCAD knock-out (LCAD) and VLCAD knock-out (VLCAD). We collected four micro array data distinguished by types of tissues (brown fat, skeletal, liver and heart). We filtered out low-expressed and low-variable features (mean$<$0.7, SD $<$0.7), and matched up features across the four data sets, which left 1,304 features. (See Table 15).

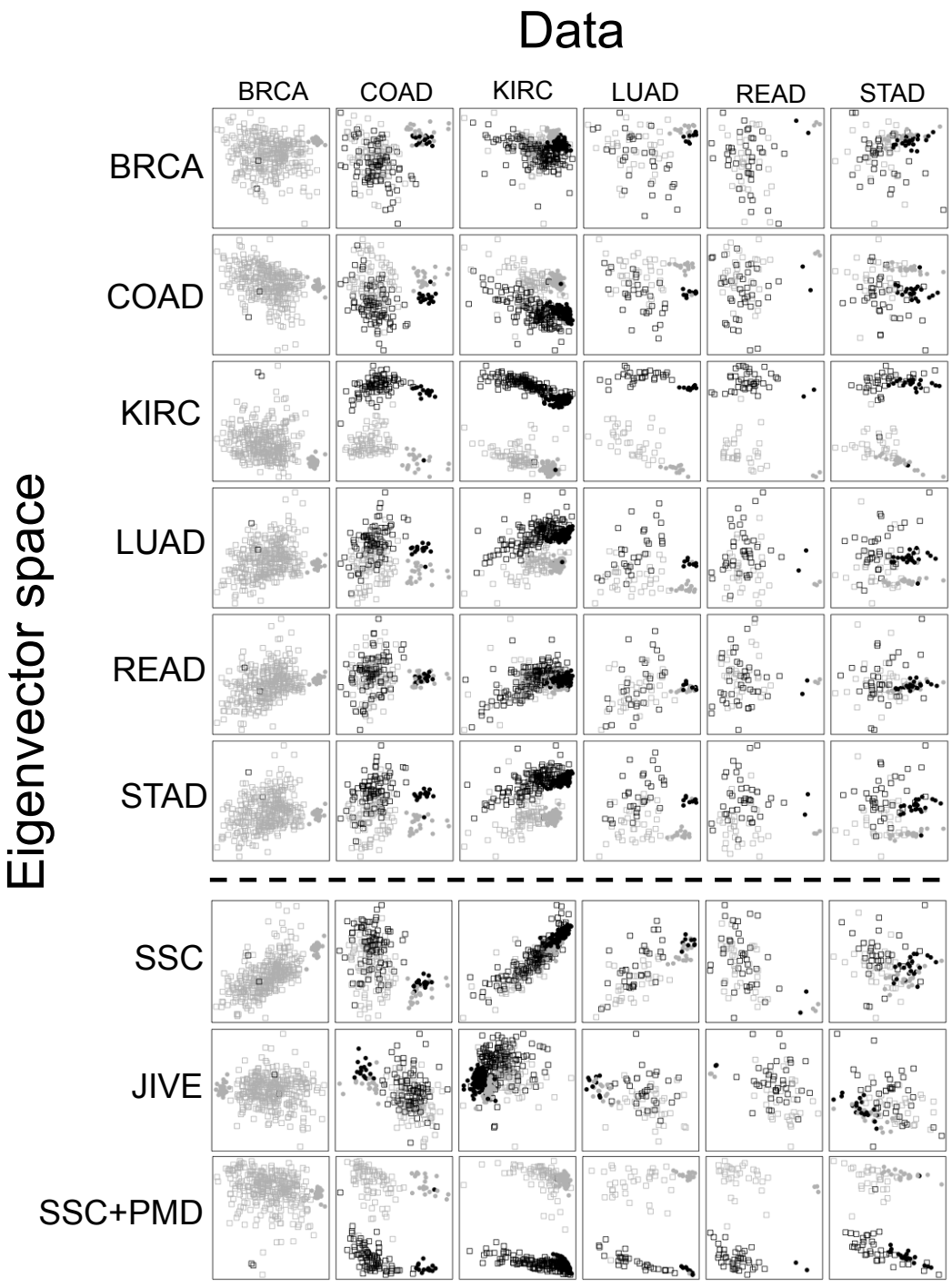

Figure 27: Two dimensional PC projections using methylation expressions of six different cancers (TCGA) data; Tumor (square), Normal (dot), Male (black) and Female (grey).

In Figure 28, PC projections of the standard PCA (e.g., "Liver") does not effectively classifies samples into the three class labels (i.e., WT (square), LCAD (dot) and VLCAD (star). Table 14 shows that the standard PCA of "Liver" data set (training set) produces consistently low Fisher discriminant scores, while Meta-PCA (SSC) efficiently distinguishes the samples into the three labels and appear even better than JIVE (i.e., SSC=10.17, JIVE=4.13). Importantly, Meta-sparsePCA (SSC + PMD) further improves Fisher discriminant scores of Meta-PCA (SSC) (i.e., SSC=10.17, SSC+PMD = 23.47).

4.5 DISCUSSION

We introduce the Meta-PCA and Meta-sparsePCA frameworks to estimate and analyze multiple high-dimensional data through common principal components. The simulation studies demonstrate Meta-PCA (SSC) achieves robustness to outliers and noise features, and more precisely detect true underlying PC space (e.g., common cyclic and transitional patterns), as compared with JIVE and standard PCA. More importantly, Meta-PCA has a potential advantage of computational cost, while JIVE, due to nature of permutation, may not be converged when the size of data is small, yet even if so, computing time can be huge. Our real data examples are shown as homogeneous, which likely lead to effective PC projections. In case of heterogeneous data, we observed that Meta-PCA does not necessarily perform effective dimension reduction. Therefore it is worth collecting homogeneous data so as to obtain reliable and well-separable PC projections. In particular, many high-throughput data sets of prostate cancer have been reported to be more heterogeneous than other cancers (Sboner et al., 2010). To tackle the problem, Kang *et al.* (2012) developed Meta-QC that provides several benchmarks regarding study inclusion in the context of meta-analysis. For future research directions, we may try to enhance the robustness of Meta-PCA. Hubert et al. (2005) introduced "Robust PCA" known to be robust to effects of noise features. This direction is promising since PCA is highly sensitive to noise features or samples and thus often fail to properly project onto robust PC space. Meta-PCA equipped with robustness is expected to enhance the interpretation and visualization of Meta-PC.

4.6 SUPPLEMENTARY MATERIALS

4.6.1 Best choice of Meta-PC dimension

In order to choose the optimal dimension k (=dimension of meta eigenvector matrix), we adopt the *scree plot* technique. We generate simulated data exploiting the same simulation scenario in the section 4.3.1 with different column dimensional true eigenvector matrix. Let $E = (e_1, e_2, e_3, e_4, e_5) \in \mathrm{R}^{200\times 5}$ be the true eigenvector matrix of 200 dimensions, where $e_1 = (1, 0, \ldots, 0)$, $e_2 = (0, 1, 0, \ldots, 0)$, $e_3 = (0, 0, 1, 0, \ldots, 0)$, $e_4 = (0, 0, 0, 1, 0, \ldots, 0)$, $e_5 = (0, 0, 0, 0, 1, 0, \ldots, 0)$. Denote true eigenvalues by $\lambda_1 = 500, \lambda_2 = 300, \lambda_3 = 200, \lambda_4 = 100$, and $\lambda_5 = 50$, and create a diagonal matrix $\lambda = diag(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$. Define $C = 5$ and $j^{(m)} = 5$ $(1 \leq m \leq 10)$. The simulations are repeated 100 times and average values are presented. On the y-axis, we present the values of differences in explained variances of two neighboring values. Figure 29 indicates the elbow point at the 4^{th} index of difference, suggesting to select five as the best.

4.6.2 Penalization constant for Meta-sparsePCA

In order to choose the optimal penalization constant λ, we adopt the *scree plot* technique. We generate simulated data sets using the same scenario in the section 4.3.1. We utilize a proportion of increased explained variance $G(a, b)$ as a benchmark to automatically choose the best λ, where $G(a, b) = \frac{f(b)-f(a)}{f(b)}$, $f(z)$ is explained variance of PC when the z number of non-zero features of eigenvector matrix are applied. We choose λ subject to $G(a, b) < \Delta$, where λ produces b non-zero features of eigenvector and $\Delta = 0.1$. The simulations are repeated 100 times and average values are presented. Figure 30 shows that the stopping rule chooses 20 nonzero features of true eigenvector matrix, suggesting the selection of the penalization constant such that λ leaves 20 non-zero features.

Table 13: Fisher discriminant scores of PC projections (TCGA pan-cancer data; Class lables: Tumor, Normal, Male and Female).

	BRCA	COAD	KIRC	LUAD	READ	STAD	Average
BRCA	18.16	22.22	20.73	12.17	15.04	8.17	16.08
COAD	20.50	25.50	28.23	13.87	17.50	10.70	19.38
KIRC	22.59	29.13	32.70	16.25	20.33	13.78	22.46
LUAD	21.81	25.30	27.64	14.47	17.03	11.09	19.55
READ	20.27	21.29	18.43	11.06	15.35	7.02	15.57
STAD	21.84	26.17	29.34	14.89	17.40	11.98	20.27
SSC	24.93	21.02	16.88	12.52	13.12	7.94	16.07
JIVE	19.69	20.15	18.50	10.68	12.77	8.90	15.11
SSC+PMD	16.96	29.66	27.12	14.72	20.34	13.98	20.46

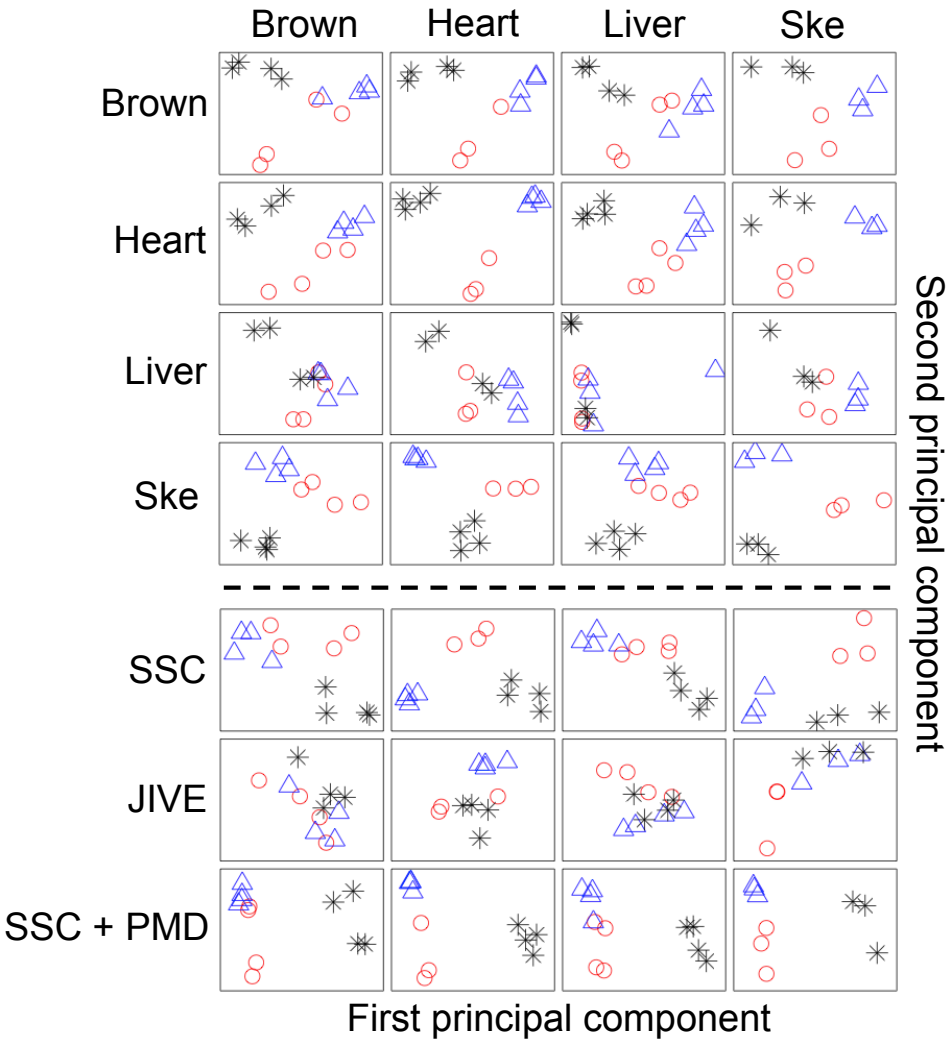

Figure 28: Two dimensional PC projections using mRNA expressions of four mouse metaboloism data; WT (square), LCAD (dot) and VLCAD (star).

Table 14: Fisher discriminant scores of PC projections (mouse metabolism data)

	Brown	Heart	Liver	Ske	Average
Brown	8.64	12.60	7.75	8.15	9.28
Heart	16.65	24.43	15.28	10.91	16.82
Liver	3.83	5.48	2.19	5.23	4.18
Ske	15.51	16.91	12.93	20.93	16.57
SSC	8.28	15.05	8.40	8.93	10.17
JIVE	3.59	5.83	3.75	3.35	4.13
SSC+PMD	19.11	29.17	22.90	22.68	23.47

Table 15: The summary of four mouse metabolism microarray datasets.

Tissue	Type	# of genes	Sample Size	Source
Brown fat	Gene expression	14,495	12	Gerard Vockley, Li et al. (2011)
Liver	Gene expression	14,495	12	Gerard Vockley, Li et al. (2011)
Heart	Gene expression	14,495	11	Gerard Vockley, Li et al. (2011)
Skeletal	Gene expression	14,495	9	Gerard Vockley, Li et al. (2011)

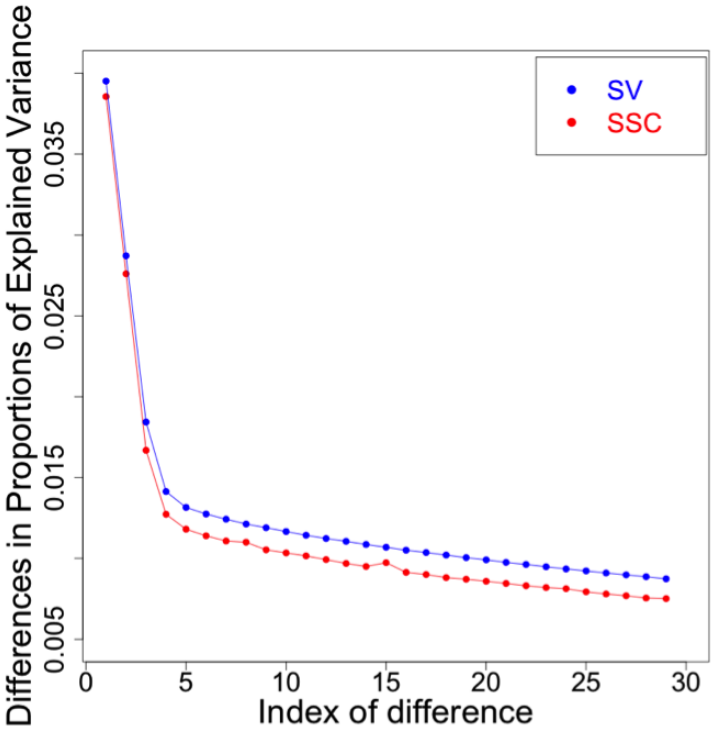

Figure 29: The example of scree plot to determine the optimal dimension reduction of Meta-PCA.

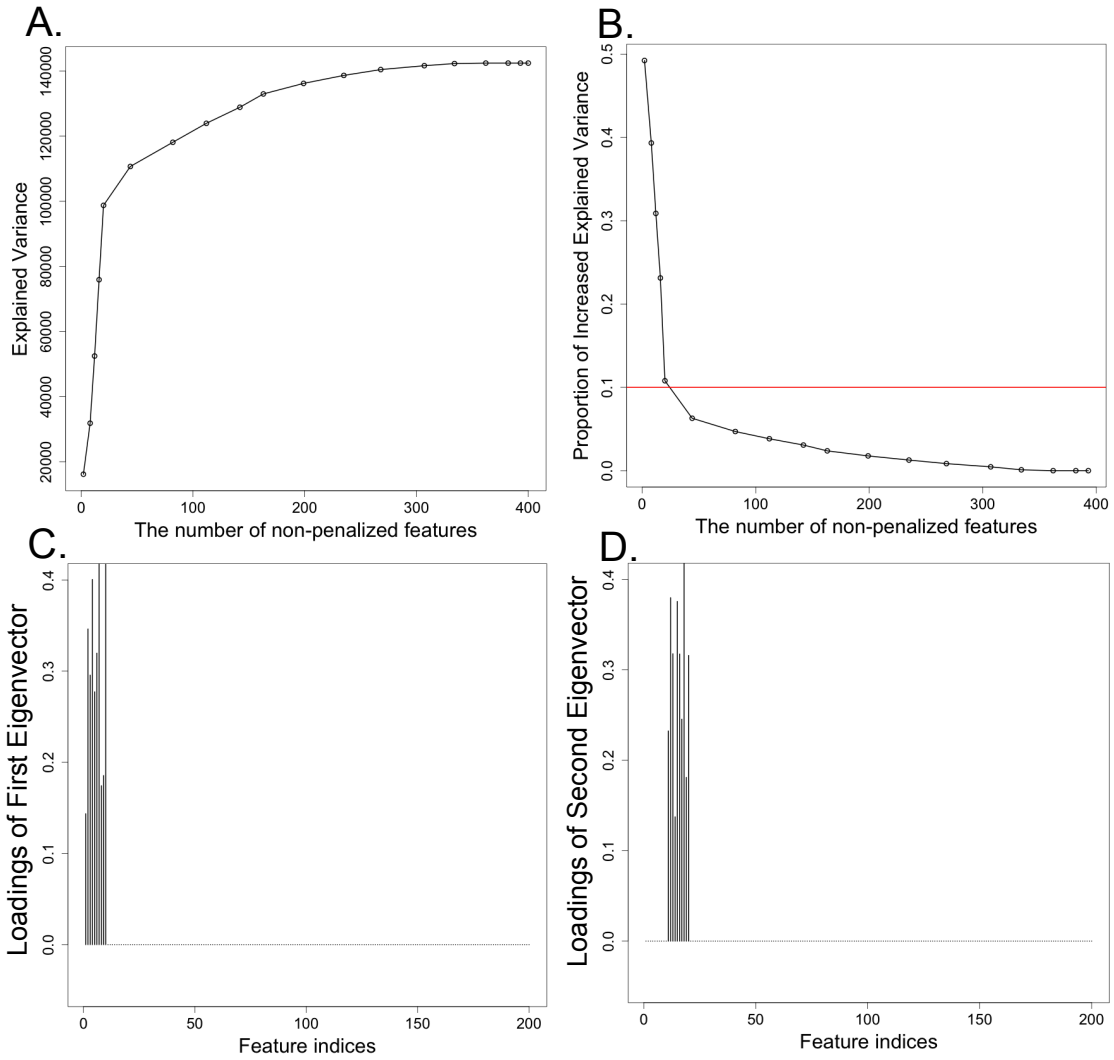

Figure 30: The example of scree plot to determine the penalization constant for MetasparsePCA.

5.0 FUTURE WORKS AND CONCLUSION

5.1 META-KTSP EXTENDED TO MULTI-OMICS AND MULTI-CLASS PROBLEMS

There are a few future directions to consider. The current framework can be extended to a multi-omics or/and multi-platform integration. For instance, a TSP framework that combines high-throughput microarray miRNA and RNA-seq. Due to the fact that TSP relies on a rank based prediction block and RNA-seq data typically contain a multitude of zero counts, RNA-seq might not be suitable to apply. However, it is possible to integrate RNA-seq and microarray data using a copula mixture model (Qunhua et al., 2015), by which we can propose a fine-tuned TSP framework designed to incorporate multi-platform data. Moreover, a TSP model integrating multi-omics data (e.g., mRNA, methylation, CNV and so on) can provide an potential insight into a molecular and cellular mechanism of the diseases. Second, our method and evaluation focus on binary case-control classification. The method could be extended to a multi-class classification scenario. Third, biological knowledge such as pathways or known disease relevant genes can be incorporated to enhance the TSP discovery accuracy. For example, Oncotype DX started with 250 breast cancer related genes to identify the 21 predictive genes in their panel. Although this runs the risk of missing understudied but significant biomarkers, this approach can potentially improve cross-study validation in well-studied diseases. Finally, the current TSP approaches may be extended towards module-based prediction scheme where top scoring pairs of gene modules are sought to provide extra redundancy and robustness (Mi et al., 2010).

5.2 GS-ICLUSTER REFLECTING FEATURE REGULATORY DIRECTIONS

GS-iCluster can accommodate the prior knowledge of regulatory structure, where the model embeds predefined feature modules. However, the presence or absence of exact directional edges between features (e.g. methylation $\rightarrow$ mRNA, CNV $\rightarrow$ mRNA but not CNV $\rightarrow$ methylation) are not explicitly modeled. Instead, the directional information was only used for post-hoc evaluation in applications. Incorporating the directional network prior knowledge is a promising future direction. Zhu et al. (2015) recently proposed a new group lasso to estimate coefficients in a group that are encouraged to keep concordant directions. Motivated by this idea, we may assign an unbalanced penalization weight. More precisely, if the estimated directions of mRNA and CNV are the same, we may impose a relatively small penalization as the identical direction coincides with biologically known evidence.

The proposed GS-iCluster framework contains high computational complexity, mainly from the iterative EM algorithm for the latent variable model and optimization via smoothing proximal gradient (SPG) in the sparse overlapping group lasso. Since the sparse overlapping group lasso applied here is for both multivariate inputs (latent covariates) and multivariate outputs (omics measurements), the optimization is complex and heavy in nature. The modeling is also complex but necessary. Our current package GS-iCluster is written in R and the routines can be further optimized using C programming and parallel computing packages in the future.

5.3 CONCLUSION

As high-throughput experimental data become prevalent, integrative methods to fully capture information of multi-lab, -platform and/or -omics data sets have become popular and critical. Meta-analysis currently encounters new statistical and computational challenges. Regarding data integration problems, this dissertation includes a wide variety of statistical learning methods to combine multiple high-throughput data sets to significantly enhance

the understanding of disease mechanisms and to generate novel biological hypotheses. First, we sought to improve a top scoring pair (TSP) method that is a non-parametric, accurate and easily interpretable model. The method is designed to facilitate cross-study validation for clinical applications. The improved cross-study prediction suggests that the detected biomarkers are robust to apply to any incoming high throughput data. Second, we improved the existing iCluster framework with a sparse overlapping group lasso technique to accommodate prior knowledge of the regulatory information flow in the model. A comprehensive inference of potential inter-omics regulatory mechanisms provides a deeper understanding of disease development. We also proposed a tight clustering to exclude outlier samples out of meaningful clusters. Clustering samples for coherent omics signature is a crucial biological objective. Novel findings of disease subtypes lay a foundation to fulfill tailored treatments down the road. Third, we propose the Meta-PCA frameworks to estimate common PC space for effective visualization and to discover common principal component patterns that multiple studies commonly share. Applications to various examples of high-throughput data sets reveals the superiority of Meta-PCA to distinguish samples into original class labels. In conclusion, we believe that our data integration methods will ultimately promote applications of integrative analysis, and will benefit to translate novel findings towards prediction medicine and/or disease management programs.

BIBLIOGRAPHY

Afsari, B., Neto, U. B. , Geman, D. (2014) Rank Discriminants for Predicting Phenotypes from RNA Expression. *Annals of Applied Statistics*.

Akavia, U., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, HC., Pochanard, P., Mozes, E., Garraway, LA., Pe'er, D. (2010). An integrated approach to uncover drivers of cancer. *Cell* **143**, 100517

Alter, O., Brown, P., and Botstein, D. (2000). Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling. *Proceedings of the National Academy of Sciences* **97**, 10101-10106.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal* **97**, 1382-1408.

Banfield, J.D. and Raftery, A.E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803-821.

Bernau, C., Riester, M., Boulesteix, A., Parmigiani, G., Huttenhower, C. (2014) Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, **30**, i105–12.

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D. (2004) Adjustment of systematic microarray data biases. *Bioinformatics*, **20**, i105–14.

Bhattacharya, S., Mariani, TJ. (2009). Array of hope: expression profiling identifies disease biomarkers and mechanism. *Biochemical Society Transactions* **37**, 855–862

Bickel, P. J. and Doksum, K. A. (2001). Mathematical statistics, volume i.

Bishop, C. M. (2006). Pattern recognition and machine learning. *Springer New York*.

Baek, S., Tsai, C. A., and Chen, J. J. (2009). Development of biomarker classifiers from high- dimensional data. *Brief Bioinform* **10**, 537–546.

Burges, C. J. C. (2011) A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121–167.

Cancer Genome Atlas Research Network. (2012). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **7492**, 315–22 .

Cardoso, F., Piccart-Gebhart, M., Van't Veer, L., Rutgers, E. (2007). "The MINDACT trial: the first prospective clinical validation of a genomic tool" *Mol Oncol* **3**, 246–51.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research* **1**, 245-276.

Chen, X., Lin, Q., Kim, S., Carbonell, J. G., Xing, E. P., et al. (2012). Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics* **6,** 719–752.

Chen, X., Lin, Q., Kim, S., Carbonell, J. G., Xing, E. P., et al. (2012). Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics* **6,** 719–752.

Chen, X., Lin, Q., Kim, S., Carbonell, J. G., Xing, E. P., et al. (2012). Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics* **6,** 719–752.

Cheng, C., Shen, K., Song, C., Luo, J., Tseng, G. *et al.* (2009) Ratio adjustment and calibration scheme for gene-wise normalization to enhance microarray inter-study prediction. *Bioinformatics*, **25**, 1655–1661.

Christoph Bartenhagen, Hans-Ulrich Klein, Christian Ruckert, Xiaoyi Jiang and Martin Dugas (2010). Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data *BMC Bioinformatics* **11**, 567

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association* **74,** 829–836.

Correa ,G., Reis-Filho, J. (2009) Microarray-based gene expression profiling as a clinical tool for breast cancer management: are we there yet? *int J Surg Pathol*, **17**, 285–302.

Cronin, M., Sangli, C., Liu, M., Pho, M., Dutta, D. *et al.* (2007) Analytical validation of the Oncotype DX genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor-positive breast cancer. *Clin Chem.*, **56**, 1084–91.

dAspremont, A., El Ghaoui, L., Jordan, M. I. and Lanckriet, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* **49**, 434-448.

Daniel D. Lee and H. Sebastian Seung (1999). Learning the parts of objects by non-negative matrix factorization *Nature* **401**, 788-791

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F. *et al.* (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res.*, **13**, 3207–14.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 1–38.

De La Torre, F. and Black, M. J. (2003). A framework for robust subspace learning. *International Journal of Computer Vision* **54**, 117-142.

Defays, D. (1977). An efficient algorithm for a complete link method *The Computer Journal* **20,** 364–366.

Ding, C. and He, X. (2004). K-means clustering via principal component analysis. *ACM International Conference Proceeding Series, ACM, Banff, Alberta, Canada.* **69**.

Dobbin, K., Zhao, Y., Simon R., *et al.* (2008) How large a training set is needed to develop a classifier for microarray data? *Clin Cancer Res*, **14**, 108-114.

Domany, Eytan (2014). Using HighThroughput Transcriptomic Data for Prognosis: A Critical Overview and Perspectives. *Cancer research* **74,** 4612-4621.

Duda, R. O., Hart, P. E., Stork, D. H. (2000). Pattern Classification (2nd ed.). *Wiley Interscience* **74,** ISBN 0-471-05669-3.

Dvorkin-Gheva, A., Hassell, J. (2011) Hormone Receptor and ERBB2 Status in Gene Expression Profiles of Human Breast Tumor Samples *Plos one*, **6**, e26023.

Eckart, C, Young, G. (1936). The approximation of one matrix by another of low rank. *Psychometrika* **1,** 211.

Emblom-Callahan,M.C., Chhina,M.K., Shlobin,O.A., Ahmad,S., Reese,E.S. *et al.* (2010) Genomic phenotype of non-cultured pulmonary fibroblasts in idiopathic pulmonary fibrosis. *Genomics*, **96**, 134145.

Feng Chu and Lipo Wang (2005). Applications of support vector machines to cancer classification with microarray data *Int. J. Neur. Syst* **15**, 475

Fisher, R. (1948) Questions and answers #14. *The American Statistician*, **2**, 30-31.

Flury, B. N. (1984). Common principal components in k groups. *Journal of the American Statistical Association* **79**, 892-898

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736* .

Garrett-Mayer, E., Parmigiani, G., Zhong, X., Cope, L., Gabrielson, E. (2008) Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics*, **9**, 333–54

Geman, D., dAvignon, C., Naiman, D., Winslow, R.. (2004) Expression Profiles from Pairwise mRNA Comparisons. *Statistical Applications in Genetics and Molecular Biology*, **3**.

Ghosh, D., and Chinnaiyan, A. M. (2002). Mixture modelling of gene expression data from microarray experiments. *Bioinformatics* **18,** 275-286.

Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D., Lander E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537.

Hastie, T., Tibshirani, R., Narasimhan, B. and Chu, G. (2014). impute: Imputation for microarray data *R package version* (http://www.r-project.org/), **1.40.0**.

Hastie, T., Tibshirani, R., Eisen, M., Brown, P., Ross, D. *et al* (2000). gene Shaving as a Method for Identifying Distinct Sets of Genes With Similar Expression Patterns. *Genome Biology* **1**, 1–21.

Han G., Sidhu, D., Duggan, M., Arseneau, J., Cesari, M., *et al.* (2013) Reproducibility of histological cell type in high-grade endometrial carcinoma. *Mod Pathol*, **26**, 1594-1604.

Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics*, 28, 100–108.

Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat Rev Genet* **11**, 476–486.

Hernandez-Vargas, H., Ouzounova, M., Le, C. F., Lambert. M., McKay-Chopin, S., Tavtigian, S., Puisieux, A., Matar, C., Herceg, Z. (2011). Methylome analysis reveals Jak-STAT pathway deregulation in putative breast cancer stem cells. *Epigenetics* **6,** 428–439.

Hotelling, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. *The British Psychological Society* **10**, 69-79.

Hoyle, D. C. and Rattray, M. (2004). Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Phys. Rev. E* **3**, 69.

Hu, Jun and Tzeng, Jung-Ying (2014). Integrative gene set analysis of multi-platform data with sample heterogeneity. *Bioinformatics* **30,** 1501–1507.

Huan, J., Wang, L., Xing, L., Qin, X., Feng, L., Pan, X., Zhu, L. (2014). Insights into significant pathways and gene interaction networks underlying breast cancer cell line MCF-7 treated with 17β-estradiol (E2). *Gene* **533,** 346–355.

Hubert, M., Rousseeuw, P. J., Branden, K. V. (2005). ROBPCA: a new approach to robust principal components analysis. *Technometrics* **47**, 6479.

Hubert, M., Rousseeuw, P. J., Branden, K. V. (2005). ROBPCA: a new approach to robust principal components analysis. *Technometrics* **47**, 6479.

Ivshina, A., George, J., Senko, O., Mow, B. *et al.* (2006) Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res*, **66**, 10292-301.

Jones, S., Zhang, X., Parsons, D., Lin, J., Leary, R. *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Cancer Cell*, **321**, 1801–1806.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004). Human microRNA targets. *PLoS biology* **2,** e363.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics* **12**, 531-547.

Journ ee, M., Nesterov, Y., Richta rik, P. and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.* **11**, 517-553.

Joyce, A.R. and Palsson, B.O. (2006). The model organism as a system: integrating 'omics' data sets. *Nature reviews* **7**, 198–210 .

Kang, D., Sibille, E., Kaminski, N., Tseng, G. (2012) MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic Acids Res.*, **40**, e15.

Kern, S. (2012) Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures. *Cancer Res*, **72**, 6097–101.

Kohonen, T (1982). Self-organised formation of topologically correct feature map. *Biological Cybernetics* **43,** 56–69.

Konishi, K., Gibson, K., Lindell, K., Richards, T., Zhang, Y. *et al.* (2009) Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med*, **180**, 16775.

Kuo, W., Liu, F., Trimarchi, J., Punzo, C., Lombardi, M.*et al.* (2006) A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat Biotechnol*, **24**, 832-40.

Krzanowski, W. (1979). Between-groups comparison of principal components. *Journal of the American Statistical Association* **74**, 703-707.

Lapointe, J., Li, C., Higgins, J. P., Van De Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U. et al. (2004). Gene expression profiling identifies

clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America* **10**, 811–816

Laua, J. W. and Greena, P. J. (2007) Bayesian Model-Based Clustering Procedures *Journal of Computational and Graphical Statistics* **16**, 526-558.

Larsson, O., Diebold, D., Fan, D., Peterson, M. *et al.* (2008) Fibrotic myofibroblasts manifest genome-wide derangements of translational control. *PLoS One*, **3**, e3220.

Lee, S., Zou, F. and Wright, F. A. (2010) Convergence and prediction of principal component scores in high-dimensional settings *The Annals of Statistics* **38**, 3605–3629

Li, J. and Tseng, G. C. (2011). An adaptively weighted statistics for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics* **5**, 994-1019.

Li, Y., Zou, L., Li, Q., Haibe-Kains, B. *et al.* (2010) Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. *Nat Med*, **16**, 214-8.

Li, W., Zhang, S., Liu, C. and Zhou, X. J. (2012). Identifying Multi-Layer Gene Regulatory Modules from Multi-Dimensional Genomic Data. *Bioinformatics* **28,** 2458-2466.

Li, K., Yan, M., and Yuan S. (2002). A simple statistical model for depicting the CDC15-synchronized yeast cell cycle regulated gene expression data. *Statistica Sinica* **12**, 141–158.

Lock, E. F., and Dunson, D. B. (2013B). Bayesian consensus clustering. *Bioinformatics* **29,** 2610-2616.

Lock, E. F., Hoadley, K. A., Marron, J. S. and Nobel, A. B. (2013A). Joint and individual variation explained (JIVE) for integrated analysis of mutiple data types. *The Annals of Applied Statistics* 523-542.

Ma, X., Wang, Z., Ryan, P., Isakoff, S., Barmettler, A. *et al.* (2004) A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell*, **5**, 607–616.

Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium *et al.* (2013) A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry*, **18**, 497–511.

Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 133–141.

Marisa, L., de Reynis, A., Duval, A., Selves, J., Gaub, M., Vescovo, L., Etienne-Grimaldi, M. *et al.* (2013) Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *Plos Medicine*, **10**, e1001453.

Maitra R. and Ramler I. P. (2009). Clustering in the Presence of Scatter. *Biometrics* **65,** 341-352.

Marshall, E. (2011) Human genome 10th anniversary. Waiting for the revolution. *Science*, **331**, 526-529.

MAQC Consortium, Shi, L., Reid, L., Jones, W., Shippy, R., Warrington, J. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, **24**, 1151–61.

Mi, Z., Shen, K., Song, N., Cheng, C., Song, C., Kaminski, N., and Tseng, G.C. (2010). Module-based prediction approach for robust inter-study predictions in microarray data. *Bioinformatics* **26,** 2586–2593.

Mitchell, S., Brown, K., Henry, M., Mintz, M., Catchpoole, D., LaFleur, B. *et al.* (2004) Inter-platform comparability of microarrays in acute lymphoblastic leukemia. *BMC Genomics*, **5**, 71.

Monti, S., Tamayo, P., Mesirov, J., Golub, T. (2003) Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning* **52**, 91118.

Morris, M., Ricketts, C., Gentle, D., McRonald, F., Carli, N. *et al.* (2011) Genome-wide methylation analysis identifies epigenetically inactivated candidate tumour suppressor genes in renal cell carcinoma. *Oncogene.*, **30**, 1390–401.

McShane, L., Polley, M. Development of omics-based clinical tests for prognosis and therapy selection: the challenge of achieving statistical robustness and clinical utility. *Clin Trials*, **10**, 653–65.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical programming* **103,** 127–152.

Obozinski, G., Taskar, B., and Jordan, M. I. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing* **20,** 231–252.

Obozinski, G. R., Wainwright, M. J., and Jordan, M. I. (2008). High-dimensional support union recovery in multivariate regression. In *Advances in Neural Information Processing Systems*, pages 1217–1224.

Opitz, D., Maclin. R. (1999) Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, **11**, 169–198.

Owen, A. B. (2009) Karl Pearson's Meta-Analysis Revisited. *Annals of Statistics*, **37**, 3867–3892.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J. *et al.* (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*, **351**, 2817–26.

Palacios-Arreola, M. I., Nava-Castro, K. E., Castro, J. I., Garca-Zepeda, E., Carrero, J. C., and Morales-Montor, J. (2014). The Role of Chemokines in Breast Cancer Pathology and Its Possible Use as Therapeutic Targets. *Journal of Immunology Research* **2014,** ID 849720.

Pan, Wei and Shen, Xiaotong (2007). Penalized Model-Based Clustering with Application to Variable Selection *Journal of Machine Learning Research* **8,** 1145-1164.

Pardo, A., Gibson, K., Cisneros, J., Richards, T. *et al.* (2005) Up-regulation and profibrotic role of osteopontin in human idiopathic pulmonary fibrosis. *PLoS Med*, **2**, e251.

Parker, J., Mullins, M., Cheang, M., Leung, S., Voduc, D. *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*, **27**, 1160–7.

Perou, C. M. , Srlie, T., Eisen, M. B., Rijn, Matt., Jeffrey, S. S.(2000). Molecular portraits of human breast tumours. *Nature* **406,** 747-752.

Powe, D., Dhondalay, G., Lemetre, C., Allen, T., Habashy, H. *et al.* (2014) DACH1: its role as a classifier of long term good prognosis in luminal breast cancer. *PLoS One*, **9**, e84428.

Price, N., Trent, J., El-Naggar, A., Cogdell, D., Taylor, E. *et al.* (2007) Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas. *Proc Natl Acad Sci*, **104**, 3414–9.

Qiu, W. and Harry J. (2006). Generation of Random Clusters with Specified Degree of Separation. *Journal of Classification* **23**, 315-334

Quackenbush, J. (2001). Computational analysis of microarray data. *Nat Rev Genet* **2**, 418–427.

Qunhua Li and Yafei Lyv (2015) Integration of RNA-seq and Microarray Data *Statistical and Computational Challenges in Omics Data Integration (SCC-ODI) Workshop*

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci* , **26**, 15149–54.

Ramasamy, A., Mondry, A., Holmes, C. C. and Altman, D. G. (2008). Key issues in conducting a metaanalysis of gene expression microarray datasets. *PLoS medicine* **5,** e184.

Raponi, M., Lancet, J., Fan, H., Dossey, L., Lee, G. *et al.* (2008) A 2-gene classifier for predicting response to the farnesyltransferase inhibitor tipifarnib in acute myeloid leukemia. *Blood*, **5**, 2589-96.

Reynolds, A. P., Richards, G., de la Iglesia, B., and Rayward-Smith, V. J. (2006). Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms* **5,** 475–504.

Reid, J., Lusa, L., De, C., Coradini, D., Veneroni, S. *et al.* (2005) Limits of predictive models using microarray data for breast cancer clinical treatment outcome. *J Natl Cancer Inst*, **97**, 927–30.

Rey, M. and Roth, V. (2012). Copula mixture model for dependency-seeking clustering. *arXiv preprint arXiv:1206.6433* .

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D. and Chinnaiyan, A. M. (2002). Meta-Analysis of Microarrays. *Cancer research* **62**, 4427

Richard, B. and Rand Corporation (1957). Dynamic programming. *Princeton University Pressl*

Rousseeuw, Peter J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics* **20**, 53-65

Ruoqing Zhu, Qingzhao Zhang, Runze Li, Hongyu Zhao, Shuangge Ma (2015) Promoting Sign Concordance in Group Penalized Integrative Analysis of Multiple Datasets *https://sites.google.com/site/teazrq/publication*

Sato, F., Tsuchiya, S., Terasawa, K., Tsujimoto, G. *et al.* (2009) Intra-platform repeatability and inter-platform comparability of microRNA microarray technology. *PLoS One*, **4**, e5540

Sboner, A., Demichelis, F., Calza, S., Pawitan, Y., Setlur, S. R., Hoshida, Y., Perner, S., Adami, H. O., Fall, K., Mucci, L. A. et al. (2010) Molecular sampling of prostate cancer: a dilemma for predicting disease progression. *BMC Medical Genomics* **3**, 8.

Seoane, J. A., Day, I. N., Gaunt, T. R. and Campbell, C. (2014). A pathway-based data integration framework for prediction of disease progression. *Bioinformatics* **30,** 838–845.

Shabalin, A., Tjelmeland, H., Fan, C., Perou, C., Nobel, A. *et al.* (2008) Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, **24**, 1154–60

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25,** 2906–2912.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.* **99**, 1015-1034.

Shen, R., Wang, S., and Mo, Q. (2013). Sparse integrative clustering of multiple omics data sets. *The annals of applied statistics* **7,** 269.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25,** 2906–2912.

Simon, Richard (2005). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* **97,** 866–867.

Simon, Richard, Radmacher, Michael D., Dobbin, Kevin and McShane, Lisa M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* **95,** 14–18.

Slawski, M., Daumer, M., Boulesteix, A. (2008) CMA: a comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics*, **9**, 439.

Smith, D., Saetrom, P., Snve, O., Lundberg, C., Rivas, G. *et al.* (2008) Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation. *BMC Bioinformatics*, **28**, 9:63.

Stouffer, S. (1949) The American Soldier: Adjustment during Army Life. *Princeton University Press*, **1**.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell* **9**, 3273.

Symmans, W., Hatzis, C., Sotiriou, C., Andre, F., Peintinger, F. *et al.* (2010) Genomic index of sensitivity to endocrine therapy for breast cancer. *J Clin Oncol.*, **28**, 4111–9.

Tan, A., Naiman, D., Xu, L., Winslow, R., Geman, D. *et al.* (2005) Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, **21**, 3896–904.

Thakkar, A., Raj, H., Chakrabarti, D., Ravishankar, Saravanan, N. *et al.* (2010) Identification of gene expression signature in estrogen receptor positive breast carcinoma. *Biomark Cancer*, **2**, 1–15.

Tibshirani, R., T. Hastie, *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci*, **99**, 6567–6572.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 267–288.

Teng, S., Zhou, J., Huang, H. (2007) A statistical framework to infer functional gene associations from multiple biologically interrelated microarray experiments.*J. Am. Stat. Assoc*, **104**, 465–473.

Tseng, G., Ghosh, D., Feingold, E. (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res*, **40**, 3785–99.

Tseng, G. C. (2007). Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data. *Bioinformatics* **23,** 2247–2255.

Tseng, G. C. and Wong, W. H. (2005). Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. *Biometrics* **61,** 10–16.

Tomlins, S. A., Mehra, R., Rhodes, D. R., Cao, X., Wang, L., Dhanasekaran, S. M., Kalyana-Sundaram, S., Wei, J. T., Rubin, M. A., Pienta, K. J. et al. (2006). Integrative molecular concept modeling of prostate cancer progression. *Nature genetics* **39**, 4151.

Ulfarsson,M.O.andSolo,V. (2008). Sparse variable PCA using geodesic steepest descent. *IEEE Trans. Signal Process.* **56**, 5823-5832.

Usary, J., Llaca, V., Karaca, G., Presswala, S., Karaca, M. *et al.* (2004) Mutation of GATA3 in human breast tumors. *Oncogene*, **23**, 7669–78.

van 't Veer, L., Dai, H., van de Vijver, M., He, Y., Hart, A. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–6.

van de Vijver, M., He, Y., van't Veer, L., Dai, H., Hart, A. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, **347**, 1999-2009.

Varambally, S., Yu, J., Laxman, B., Rhodes, D. R., Mehra, R., Tomlins, S. A., Shah, R. B., Chandran, U., Monzon, F. A., Becich, M. J. et al. (2005). Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer cell* **8**, 393-406.

Vinh, N. X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1073–1080. ACM.

Wang, Y., Klijn, J., Zhang, Y., Sieuwerts, A. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–9.

Wang, X., Lin, Y., Song, C., Sibille, E. and Tseng, G. C. (2012) Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: With application to major depressive disorder. *BMC Bioinformatics* **13**, 52.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G. and Do, K. (2012). iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* **29,** 149–159.

Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63.

Winslow, R., Trayanova, N., Geman, D., Miller, MI. *et al.* (2012) Computational medicine: translating models to clinical care. *Sci Transl Med*, **4**.

Witten, D. M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534.

Xie, B., Ding, Q., Han, H., Wu, D. (2013) miRCancer: a microRNA-cancer association database constructed by text mining on literature *Bioinformatics* **29,** 638–644.

Xu, L., Tan, A., Winslow, R., Geman, D. *et al.* (2008) Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics*, **9**, 125.

Xu, L., Tan, A., Naiman, D., Geman, D. *et al.* (2005) Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*, **20**, 3905-11.

Yu, J., Koujak, S., Nagase, S., Li, C., Su, T. *et al.* (2008) PCDH8, the human homolog of PAPC, is a candidate tumor suppressor of breast cancer. *Oncogene.*, **27**, 4657–65.

Yu, Y. P., Landsittel, D., Jing, L., Nelson, J., Ren, B., Liu, L., McDonald, C., Thomas, R., Dhir, R., Finkelstein, S. et al. (2004). Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *Journal of Clinical Oncology* **22**, 2790–2799.

Zhang, D., Zhao, W., Liao, X., Bi, T., Li, H., Che, X. *et al.* (2012) Frequent silencing of protocadherin 8 by promoter methylation, a candidate tumor suppressor for human gastric cancer. *Oncol Rep.*, **28**, 1785–91.

Zhang, S., Li, Q., Liu, J. and Zhou, X. J. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory *Bioinformatics* **27,** i401-i409.

Zhang, Y., Schnabel, C., Schroeder, B., Jerevall, P., Jankowitz, R. *et al.* (2013) Breast cancer index identifies early-stage estrogen receptor-positive breast cancer patients at risk for early- and late-distant recurrence. *Clin Cancer Res*, **19**, 4196–205.

Zografos, G., Liakakos, T., and Roukos, D. H. (2013). Deep sequencing and integrative genome analysis: approaching a new class of biomarkers and therapeutic targets for breast cancer. *Pharmacogenomics* **14**, 5–8.

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis. *Journal of computational and graphical statistics* **15**, 265-286.
bibliography