TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Proteomik und Bioanalytik

Application of multivariate methods to the integrative analysis of high-throughput omics data

Chen Meng

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr H.-W. Mewes
Prüfer der Dissertation:	1. Univ.-Prof. Dr. B. Küster
	2. Univ.-Prof. Dr. D. Frischmann

Die Dissertation wurde am 30.11.2015 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 12.01.2016 angenommen

Contents

Abstract

The aim of system-wide molecular profiling studies is to explore the genes or biological molecules of an organism on a global scale. These so-called omics data are mainly collected from high-throughput technologies, which have swept through the whole area of biological studies over the last decades, such as genomics, transcriptomics, proteomics, and metabolomics. Multiple omics datasets that characterize the same set of biological samples are increasingly available. The different types of information the underlying molecules carry require the integrative analysis of these data to understand how different factors interact with one another in diseases. To this end, more computational efforts are required to derive useful biological knowledge from these noisy data.

Multivariate methods represent an old family of statistical methods. These methods account for multiple dependent or independent variables at a time and often involve dimension reduction. Nowadays, multivariate statistics shows great potential for analyzing noisy high-throughput omics data, but the application of these methods to this area is still in its infancy. This thesis explored several applications of multivariate methods for the integration and data analysis of multiple omics datasets. The thesis, first, describes the application of multiple co-inertia analysis (MCIA) for the integrative exploratory of multiple omics data. MCIA simultaneously projects several datasets into the same dimensional space; consequently, concordance and discrepancy can be easily highlighted, and the most variant markers from each dataset can be extracted. Discovering joint patterns among multiple omics data, such as defining subtype models, is another commonly faced task in omics studies. Existing methods for this purpose are often compromised by some problems, notably the problem of nondeterministic solutions. Therefore, a new algorithm was developed, termed moCluster. The moCluster algorithm first employs consensus principal component analysis (CPCA) approaches to define a set of latent variables that represent the joint pattern of multiple dataset; next the defined latent variables are further subjected to an ordinary clustering algorithm to discover joint clusters. The algorithm is computationally efficient (operates 100× to 1000× faster than the competitors); hence, it is particularly suitable for analyzing large-scale omics data. Last, but not least, a novel algorithm, moGSA, was introduced. It is a multivariate-based gene set analysis method with particular focus on the integration of multiple omics data. The method facilitates the detection of concordance and discrepancy between datasets on the gene set level.

All the methods have been evaluated using simulated data. More importantly, the biological application of these multivariate methods illustrated that multiple omics data can complement one another; therefore, integrative analysis may potentially increase the coverage and power of pathway or gene set analysis. Furthermore, in some of the specific applications such as clustering analysis, may also discover knowledge that cannot be generated from a single data set. All the results underscore the importance of analyzing multiple omics data sets in an integrative scheme.

Zusammenfassung

Gene und andere biologische (Makro-)Moleküle können heutzutage System-weit studiert werden. Diese sogenannten 'Omics'-Daten werden mithilfe von Hochdurchsatz-Technologien gewonnen, die Tausende von Features gleichzeitig messen. Getrieben durch den kontinuierlichen technologischen Fortschritt in den vergangenen Jahrzehnten haben diese Technologien ganze Bereiche biologischer Forschung erfasst.

Die enorme Komplexität der Biologie führt zwangsweise dazu, dass die Analyse einer einzelnen Ebene biologischer Information, wie beispielsweise Gene, Transkripte oder Proteine, das biologische System nur unvollständig abbildet. Daher gibt es heutzutage eine zunehmende Menge an Studien und Datensätzen, die mehrere Omics-Ebenen derselben biologischen Proben abbilden. Es sind gerade diese Datensätze, die ungeahnte Möglichkeiten bieten, um die Interaktionen unterschiedlicher Faktoren zu studieren, die für das Funktionieren eines Organismus oder die Entstehung von Krankheiten verantwortlich sind. Allerings sind dafür enorme Rechnerkapazitäten notwendig, um nützliches biologisches Wissen aus diesen verrauschten Datenzu gewinnen.

Multivariate statistische Methoden können gleichzeitig abhängige und unabhängige Variablen analysieren und nutzen häufig Techniken der Dimensionsreduktion. Multivariate Methoden bieten großes Potential für die Analyse von verrauschten Hochdurchsatz-Omics-Datensätzen, aber die Anwendung diese Methoden steckt noch in den Kinderschuhen.

Diese Arbeit hat verschiedene Anwendungen multivariater Methoden für die Integration und Analyse von Multi-Omics-Datensätzen untersucht. Zunächsten beschreibt diese Arbeit die Anwendung der sogenannten *multiple co-inertia analysis* (MCIA) für die integrative Untersuchung von Multi-Omics-Datensätzen. MCIA projiziert simultan mehrere Datensätze in den selben mehrdimensionalen Raum, wodurch Konkordanz und Diskrepanz gut hervorgehoben und die am stärksten variierenden Marker jedes Datensatzes extrahiert werden können.

Ein weiteres, häufig beobachtetes Problem betrifft die Herausforderung gemeinsame Muster in Multi-Omics-Daten zu finden, beispielsweise Tumor-Subtypen. Existierende Lösungen dafür sind oft eingeschränkt, unter anderem durch das Problem nicht-deterministischer Lösungen. Daher wurde im Rahmen dieser Arbeit ein Algorithmus entwickelt, moCluster genannt, der zunächst eine Konsenus-Hauptkomponentenanalyse (principle component analysis, PCA) nutzt und im Anschluss einen Satz an latenten Variablen definiert, die gemeinsame Muster multipler Datensätze definiert. Diese latenten Variablen werden schliesslich mit einem gewöhnlichen Cluster-Algorithmus geclustert, um gemeinsame Muster zu entdecken. Der Algorithmus ist effizient (500x bis 1000x schneller als ähnliche Algorithmen) und daher besonders geeignete um große Omics-Datensätze zu analysieren.

Zu guter Letzt wurde ein neuer Algorithmus eingeführt, moGSA genannt – eine multivariate Gen-Set-Analyse-Methode mit einem besonderen Focus auf multiple Omics-Datensätze. Die Methode vereinfacht die Entdecung von Konkordanz und Diskrepanz zwischen Datensätzen auf der Ebene von Gen-Sets.

Alle drei Methoden wurden mithilfe von simulierten Daten evaluiert. Wichtiger noch, biologische Anwendungen dieser multivariaten Methoden haben gezeigt, dass mehrere Omics-Datensätze einander kompensieren und die integrative Analyse potentiell die Abdeckung und Power von Pathway- und Gen-Set-Analysen verbessern können. Zudem bieten diese Methoden das Potential, biologische Information zu gewinnen, die nicht aus einem einzelnen Datensatz gewonnen werden kann. Alle diese Ergebnisse unterstreichen die Wichtigkeit, die Ebenen biologischer Information in einer integrierten Art und Weise zu analysieren.

CHAPTER I

General Introduction

1.1 Omics data and cancer research

Development of cancer is a dynamic and evolutionary process, which involves multiple genetic and epigenetic changes. Although some oncogenes (promoting cell growth and reproduction) and tumor suppressing genes (inhibiting cell division and survival) were discovered [1], rarely (or even not) a certain type of cancer could be completely explained by the mutation of these genes. Researchers recognized that multiple tumor sub-clones coexist with a primary clone, and these are driven by tumor genetics, epigenetics and the tumor microenvironment [2]. Therefore, a comprehensive understanding of cancer biology is required for the personalized treatment of this heterogenic disease.

Nowadays, the omics studies greatly facilitate cancer studies. Organisms are built upon different molecules including DNA, RNA, protein and others. The interplay within and between these molecules formulates biological functions. In consequence, the study of a single layer is not enough to understand the mechanisms of biological processes. The suffix "-ome" refers to "all constituents considered collectively" [3]. With the emergence of omics technologies, biological systems are being investigated at an unprecedented heterogeneous and comprehensive fashion. Advances in RNA-sequencing and mass spectrometry (MS) based proteomics have dramatically improved coverage and quality of genomic, transcriptomic and proteomic profiling [1-4]. Recent advances of MS-based proteomics provide a complementary approach to genomics and transcriptomic technologies [4, 9] and systematic analyses can identify and quantify the majority of proteins expressed in human cells [4-6]. Large-scale consortia projects shed light on important biological knowledge that would not be revealed by small-scale analysis. For example, The Cancer Genome Atlas (TCGA) project profiles multiple layers of omics data over hundreds of patients with tumor from different tissues of origin. This project identified cancer or subtype-related mutational drivers from the genomic data. Furthermore, potentially related pathways were identified through the integration of mRNA or protein data from the same patients [7, 8]. Another project using the same datasets, pan-cancer analysis, revealed the subtype similarities between different cancers and a potentially new classification scheme of cancers using multiple omics data [9, 10]. These data yielded unprecedented view of molecular building blocks and the machinery of cells. However, interpreting these large-scale datasets

and deriving fundamental and applicable information about biological system still represent a considerable challenge.

1.2 Omics technologies

The field of omics studies is strongly influenced by high-throughput technologies such as sequencing, microarray or MS-based proteomics. A comprehensive introduction is beyond the scope of this thesis. Hence, this section introduces only the basics of the transcriptional profiling using microarray, RNA sequencing and MS-based proteomics, with particular emphasis on data processing and quantification. The data generated by these methods were intensively used within this project.

1.2.1 Microarray based transcriptomics profiling

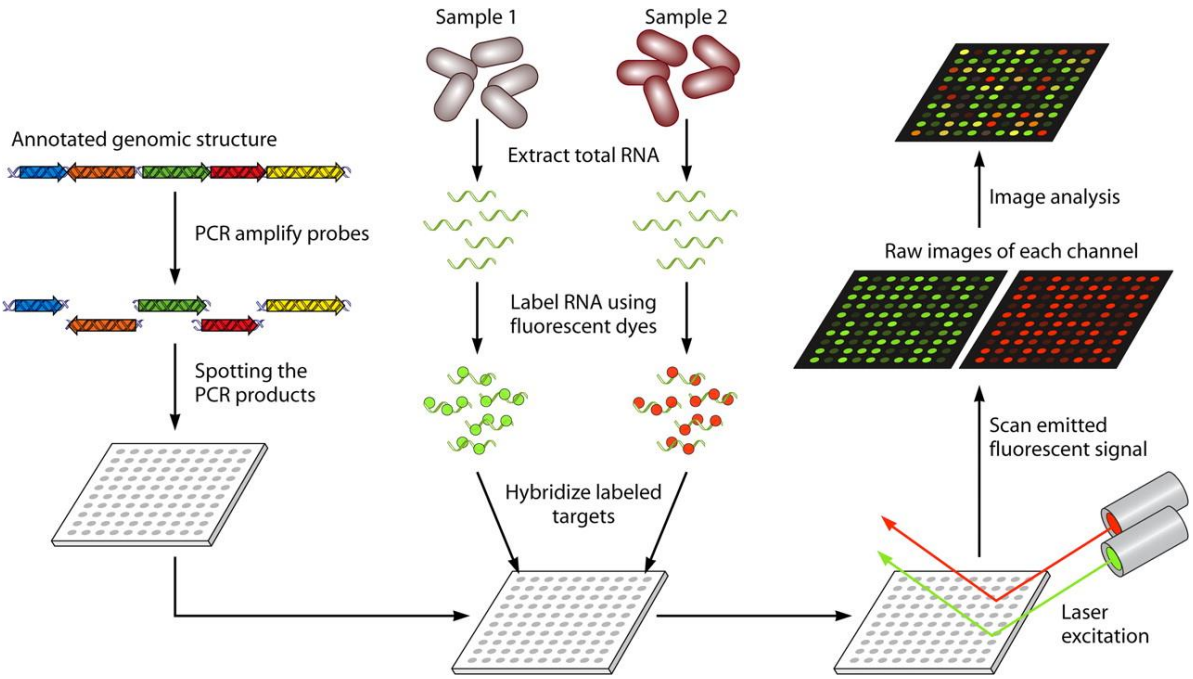

Figure 1.1 *A schematic view of a typical printed two-color microarray experiment. (Copied from [11])*

Starting from late 1990s, gene expression microarray is the first widely used high throughput technology [12]. Regardless of the different variants of microarrays, the basic concept is rather similar. A typical experimental workflow of complementary DNA (cDNA) array is illustrated in Figure 1.1. A microarray experiment consists of preparation of cDNA library and probes and hybridization of the two molecules. First, total RNAs are extracted from the samples of interest, followed by a poly-A enrichment of mRNAs. Then the isolated mRNAs are reverse transcribed to cDNAs using fluorescently labeled nucleotides. A probe consists of a large number of identical oligonucleotide sequences that perfectly match a target gene sequence (perfect-matched probes). In addition, some probes that are not perfectly matched to any target transcript (mismatched probes) are also used. The signal detected from these mismatched probes can be used to control the unspecific binding. Probes can be derived from different sources such as cDNA or pre-synthesized high-density oligonucleotides. The probes are then attached to a solid surface (e.g., silicon wafers used by Affymetrix, glass slide used by Agilent Technologies or silica beads used by illumina beadarray). During the hybridization step, probes will

preferentially bind to the cDNA molecules that are perfectly complementary, whereas the unbound or non-perfectly complement molecules will be washed off from the array. Then a laser scanner will scan the microarray and record the intensity of the fluorescent dye.

The microarray data are noisy and therefore require intensive data preprocessing and normalization, including filtering background noise, adjustments for cross hybridization (nonspecific binding of cDNA and probe), estimation of expression values across appropriate ranges, and removal of unwanted variance within or between microarray experiments. Different normalization methods for microarray data were proposed; the widely used ones include MicroArray Suite 5 (MAS5) and Robust Multiarray Average (RMA) normalization. MAS5 normalizes each individual microarray data and makes use of both the perfect-matched and mismatched probes. RMA normalization does not subtract the mismatched signals, instead it assumes that most of the genes are not differentially expressed between conditions. It normalizes all the arrays simultaneously. Recently, GCRMA, a modification of RMA, was proposed. This modification is based on the observation that the binding affinity between probe and target is dependent on the sequence of different types of nucleotides (T, C, G, A). The inclusion of position-specific effects increased the proportion of explained variance by over 10% in comparison with other methods [13].

1.2.2 RNA sequencing

An alternative method of mRNA expression profiling is RNA sequencing (RNA-seq). Similar to the microarray experiment, an RNA-seq experiment starts with total RNA extraction and an mRNA enrichment (often using the poly-A tail; Figure 1.2). Next, the isolated mRNAs are reverse transcribed to cDNA, which are subsequently fragmented to short nucleotide sequences (or first fragment mRNA followed by reverse transcription). Then, short DNA adaptors are ligated to the termini of the DNA fragments for indexing and deposition on the sequencing platform. The outputs of RNA sequencing are millions of short reads with quality scores. Dependent on the technologies, the lengths of reads range from several tens (Solexa system) to several hundred (454 Sequencing) [15]. The sequencing could be performed from a single end of the fragments or both ends (paired-end sequencing). The paired-end sequencing provides high quality sequences for both the sides of a fragment; therefore, it is particularly useful for the detection of structural variants such as genomic rearrangement, gene fusion or repetitive regions.

In data analysis, the sequencing reads need to be trimmed to remove the adaptors, poly-A tail and low quality regions [16]. Then the reads are used to reconstruct transcript using either *de novo* method or reference genome-guided method. *De novo* assembly does not require prior knowledge of a genome. The genome-guided assembly becomes overwhelming because of its computational efficiency and ever-increasing coverage and accuracy of reference genomes. However, the challenge of read mapping is from both technological (e.g. short reads and error detection) and biological

aspects (e.g., exon-exon junction and mutation). Among these, one of the major challenges is that some reads are mapped to multiple genes or isoforms with high quality scores. Frequently, people will consider the "law of parsimony" (Ocam's razor), so that only a minimal set of compatible isoforms would be reported, such as Cufflinks [17]; on the contrary, some approaches report all isoforms supported by the read data, such as Scripture [18]. Other commonly used methods include bowtie, ELAND, MapSplice, and so on [19].

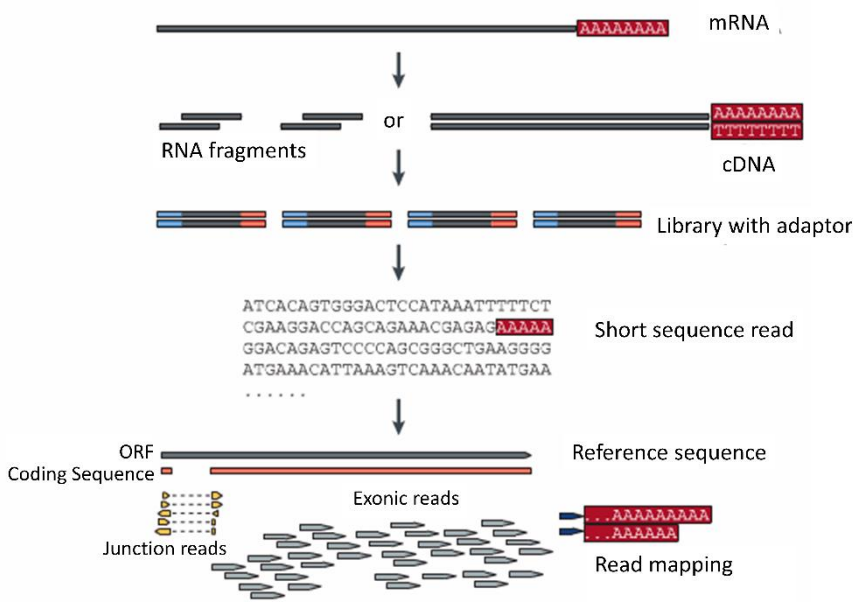

Figure 1.2 *A schematic overview of RNA sequencing experiment. (Adapted from [14])*

For the transcript quantification, a general idea is that the abundance of a transcript in a cell correlates with the number of reads measured. However, longer transcripts will contribute to more reads in comparison with the shorter ones that have the same abundance. In addition, the total number of measured reads may vary among experiments because of artefact reasons. Therefore, the number of reads needs to be normalized according to the length of transcripts and total number of reads in an experiment. This quantitative approximation of transcripts is called reads per kilobase per million (RPKM) [20], which is defined by

$$RPKM = \frac{Number\ of\ mapped\ reads\ to\ a\ transcript}{The\ transcirpt\ length\ in\ kilobase} \times \frac{10^6}{Total\ number\ of\ mapped\ reads\ in\ base}$$

While dealing with paired-end data, one fragment of cDNA could correspond to either one or two reads. Therefore, the number of reads in the above formula can be replaced by the number of fragments, which results in the fragments per kilobase per million (FPKM). However, the problem of read mapping uncertainty also impacts the quantification of transcripts. The simplest way to deal with this problem is to discard reads mapped to multiple regions in the genome. However, the fraction of non-uniquely mapped reads ranges from 17% (mouse) to 52% (maize) in different organisms [19],

hence, simply removing these reads may lose significant amount of information. A more sophisticated method is to heuristically allocate the non-uniquely mapped reads in proportion to uniquely mapped reads originated from the gene. Recently, RNA-seq by expectation maximization (RSEM) was proposed to solve uncertainty in read mapping. It employs a statistical model derived from the sequencing process, allowing modeling of the non-uniform read distributions. It has been shown that this method has more accurate gene expression estimates in comparison to other methods [19, 21]. The RSEM method is also used by TCGA (the RNA-seq data processed by this method is referred as RNASeqV2) and this thesis.

1.2.3 Comparison of microarray and RNA-seq data

Both microarray and RNA-seq are indispensable methods for transcriptomic profiling. Each method has its pros and cons. For the microarray platform, the high background noise (from cross hybridization) may conceal the true signal from low abundant mRNAs, whereas hybridization between highly abundant mRNA species and their probes could be saturated. Therefore, the microarray data have a limited dynamic range and yield a relative quantification across conditions, which result in extra complexity for cross-experiment and meta-analysis.

RNA-seq data are less suffering from signal saturation and therefore better represent the dynamic range of mRNAs in the samples. The accuracy of RNA-seq has been validated by quantitative polymerase chain reaction and spike-in studies [14]. Another advantage of RNA-seq is that it has the potential to detect the genomic events that are not known in prior, such as unknown alternative splicing, single nucleotide polymorphism (SNP) or gene fusion.

1.2.4 Mass spectrometry based proteomics and protein quantification

Proteins are the direct executors of biological functions and the targets of most cancer drugs. The abundance of mRNA only has a moderate correlation with the abundance of their translated proteins [22]. Therefore, the analysis of proteome is indispensable even transcriptomic analysis has been widely used in cancer studies. The development of MS and other instruments such as high-performance liquid chromatograph (HPLC) allows fast and accurate measurement of complex proteomes (such as human proteome) with reasonable costs. Nowadays, MS is the standard method for large-scale proteomic studies [23].

General workflow of MS-based proteomics

MS-based proteomic studies could be generally classified into bottom-up and top-down approaches. The top-down method requires extensive protein purification and is not suitable for large-scale proteome profiling. On the contrary, the bottom-up approach (also referred to as shotgun proteomics) does not require extensive protein separation but provides higher throughput and more sensitive

workflows. This thesis describes only the bottom-up approach, which was also exclusively applied within this project. Figure 1.3 is a schematic illustration of a typical shotgun proteomics analysis procedure.

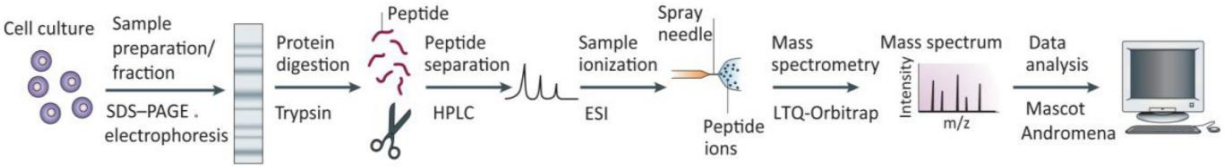

Figure 1.3 *A schematic view of MS-based proteomics. (Copied from [24])*

MS-based proteomics can analyze proteins from different sources such as cell lines, tissues, or body fluids. Proteins in these samples are first extracted and undergo a series of protein purification and separation steps to decrease the sample complexity. For this purpose, sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) is one of the most commonly used techniques. SDS bears negative charge and can denature proteins (destroy the structure of proteins). The binding of SDS and a protein is proportional to the relative molecular mass of the protein. Therefore, the treatment of SDS creates proteins with uniformed mass-to-charge (*m*/*z*) ratio, which is further separated by PAGE.

Then the gel-lanes are sliced and the proteins inside are digested by proteases, most often by trypsin. Trypsin specifically cleaves the carboxyl-terminal side of arginine and lysine residues. Because of the distribution of the two amino acids in proteins, the trypsin-digested peptides are suitable for the MS analysis. Other proteases with high specificity of their cleavage sites can also be used in protein digestion, including Lys-C, Asp-N, and Glu-C.

Because of the complexity of the digested peptide mixture, peptide separation is required for the in-depth identification and quantification of peptides by MS. In a liquid chromatography tandem mass spectrometry (LC–MS/MS) experiment, the peptide mixture is further separated by LC, which separates the peptides by passing them through fine capillaries or columns. The peptides with different properties (such as hydrophilicity) will travel through the columns in different paces. Therefore, the composition of samples eluted out at a time (retention time) is less complex. The peptides with specific modifications could be enriched before this step as well, for instance, by the use of hydrophilic interaction through LC for enriching glycosylated peptides [25] or titanium dioxide columns for phosphopeptides [26, 27].

An MS generally consists of three components: the ionization source (e.g., matrix-assisted laser desorption and ionization (MALDI) and electrospray ionization (ESI)), mass analyzer (e.g., time-of-flight (TOF) or ion trap (IT)), and detector. The peptides eluting out at a time are ionized and transferred into the gas phase by an ionization source. The ionized peptides are then separated according to their *m/z* ratio in a mass analyzer. Finally, the detector in MS results in a full scan of mass spectra (MS1

spectra) at a specific retention time. In a tandem MS experiment, a list of peptides (referred as "precursors" or "parent" ions) would be selected from the MS1 spectra for further fragmentation, most frequently using the collision-induced dissociation (CID), the high energy C-trap dissociation (HCD), or the electron transfer dissociation (ETD) [28]. In the conventional data-dependent acquisition (DDA) approach, a dynamic inclusion list or exclusion list could be used to ensure that the peptides of interest or previously undetected peptides could be preferentially selected. The detection of fragmented ions generates the tandem spectra (MS2 or MS/MS spectra; Figure 1.4), which often contains the amino acid sequence information and could be used for peptide sequencing and identification.

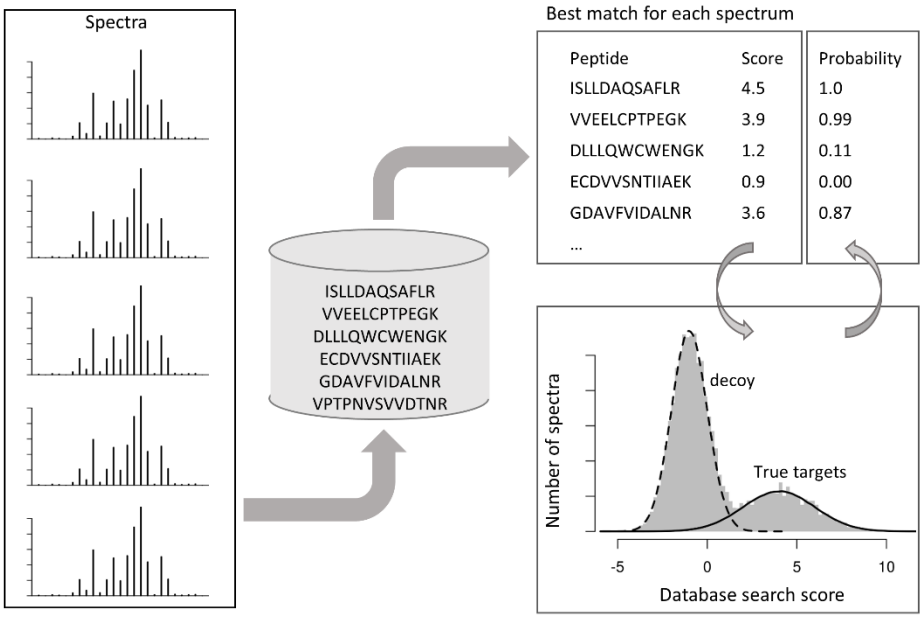

Figure 1.4 *A schematic illustration of target-decoy database search approach for peptide identification. A list of spectra are compared with target and decoy sequences and assigned with scores, which generates distributions of the correct and incorrect scores. The FDR could be estimated from the comparison of the two distributions (see text). (Adapted from [28])*

Peptide identification – the target decoy approach

A typical LC-MS/MS experiment produces a large number of mass spectra. These spectra undergo preprocessing, including smoothing, noise filtering, deisotoping, and peak picking and in the end are converted to peak lists, which are used for peptide sequencing and identification. Peptide sequences could be extracted from spectra with or without referring to a protein sequence database, referred as *de novo* approach and database search approach, respectively.

The database search approach requires a protein database, which includes the peptide sequences of all proteins identified in the sample of interest. The frequently used ones include Uniprot, Swissprot, or Refseq. For peptide identification, the experimental acquired MS/MS spectra are compared with

mass spectra reconstructed from *in silicon* digestion of all protein sequences in the database in use (theoretical spectra), resulting in a list of peptide-spectrum matches (PSMs). The similarities of the theoretical and experimental mass spectra are measured with scores, which are often calculated on the basis of the cross correlation between these two types of spectra. The database search approach is often in conjugation with a target-decoy strategy for assessing the false discovery rate (FDR) of PSMs: in addition to the *in silico* digestion of true peptide sequences in a database (target database), a decoy database is constructed through the *in silico* digestion of the randomly shuffled or reversed peptide sequences. Theoretical spectra in the decoy database are subsequently compared with the MS/MS spectra from the experiment. The comparison of MS/MS spectra with target and decoy spectra generates two score distributions representing the true matches and random matches (Figure 1.4). An empirical score cutoff could be determined from the decoy and target score distribution to control the FDR of PSMs, which indicates how likely (the probability of) the identifications are potentially incorrect. The FDR for a peptide with score *s* is calculated as

$$FDR(s) = \frac{Number\ of\ decoy\ hits\ with\ score\ higher\ than\ s}{Number\ of\ target\ hits\ with\ score\ higher\ than\ s}$$

Protein identification

A confident identification of protein using peptide sequences from MS is challenging. The difficulties are from at least two aspects. First is the nonrandom grouping effect of peptides (Figure 1.5A) [29]. In practice, the proteins presented in the sample of interest generate multiple different peptides, many of which could be identified in an MS experiment (True discoveries). Therefore, although a large number of peptides is correctly identified, the number of identified proteins could be rather small. However, the false discovered peptides match randomly with the protein sequences in the database, hence, each of them are likely uniquely mismatched to a random protein. As a result, the nonrandom grouping of peptides lead to an inflation of FDR at the protein level. It is particularly suspect for the proteins that are only identified by one peptide. The second problem, similar to the read mapping uncertainty problem in RNA-seq, is that many identified peptides correspond to multiple proteins. In this case, a minimum of proteins that account for all the observed peptides can be reported (the Occam's razor) or, in contrast, report all the proteins that match to the set of detected peptides (anti-Occam's razor; Figure 1.5B).

Protein quantification

Considering the difficulty of protein identification, the quantification of proteomic data is even more challenging. The protein quantification can use either reporter ion-based approaches or label-free approaches (Figure 1.6). Within the reporter ion-based approaches, stable isotope labeling by amino acids in cell culture (SILAC) [31] achieves the best accuracy. SILAC requires the cell lines of interest to

be cultured with medium containing isotope labeled amino acid. The cell lines will take up the labeled amino acid from the growth medium so that all proteins are labeled finally. In an MS measurement, samples from different conditions are pooled together and the MS distinguishes proteins from different samples or conditions by their different masses. However, the *in vivo* labeling requires long cultivation or feeding times when applying it to multicellular organisms [32], particularly, patient tissues cannot be analyzed using this method. Two most popular alternatives of SILAC labeling are labeling with tandem mass tags (TMTs) and isobaric tags for absolute and relative quantification (iTRAQ). Both the methods use isobaric labeling. In these approaches, peptides from different experimental conditions (such as the disease and healthy tissue) are labeled with tags having the same mass, but the fragmentation of the tags will produce different reporter ions in MS2, which are used for distinguishing samples and quantification. The MS2-based quantification gives the benefit of multiplexing capacity without increasing MS1 complexity [33]. Therefore, it is suitable for more complicated experimental designs. In addition, it can be used in proteome analysis of samples where *in vivo* labeling cannot be applied, such as patient tissues. Nevertheless, drawbacks are that it is less accurate because of contamination of isobaric ions and that it has a limited dynamic range [34].

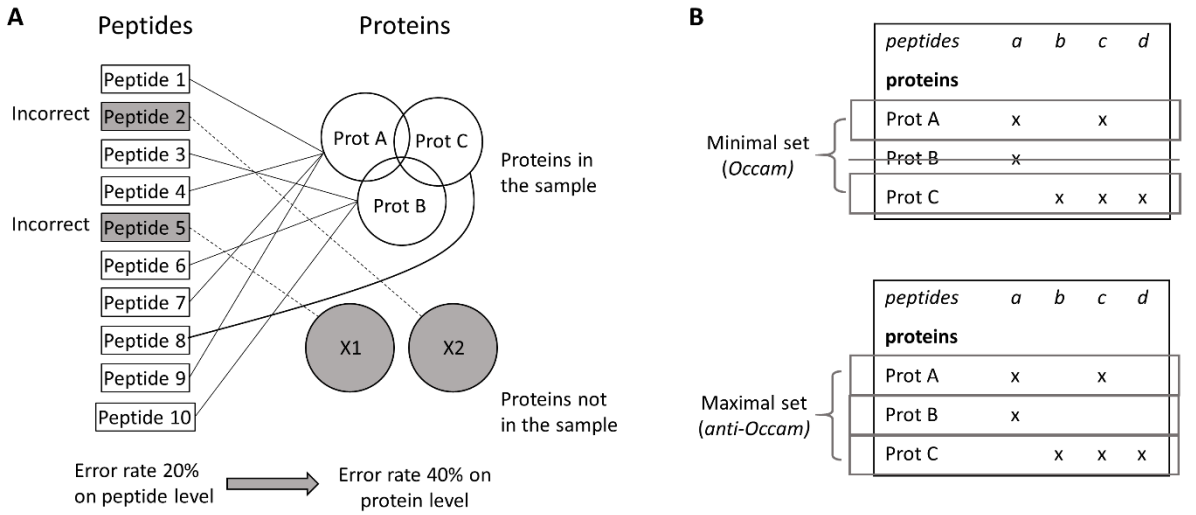

Figure 1.5 *An illustration of problems in protein identification using peptides identified in LC-MS/MS. (A) A toy example showing the problem of non-random grouping of peptides. Left side shows 10 identified peptides, two of which are incorrect PSMs (false discovery, shown as gray boxes). The eight correct peptides correspond to three proteins in the sample (A, B, C), whereas the two incorrect peptides are randomly matched to two other proteins, which are not presented in the sample (X1 and X2). Consequently, the FDR is increased from 20% at peptide level to 40% at the protein level. (Adapted from [29]) (B) An illustration of different conviction of protein identification using peptides matched to multiple proteins. (Adapted from [30])*

Different from the reporter ion-based approaches, label-free quantification (LFQ) measures samples sequentially (not pooled). For the protein quantification, the abundance of a protein could be

indicated by the number of MS/MS spectra from that protein (spectra count method) [34, 35]. However, several issues are associated with this method. First, the quantification relies on the number of MS/MS spectra, which are often acquired in a data-dependent manner. This may affect the robustness of quantification results. Second, this method is biased toward the highly abundant proteins and suffers the saturation problem [34].

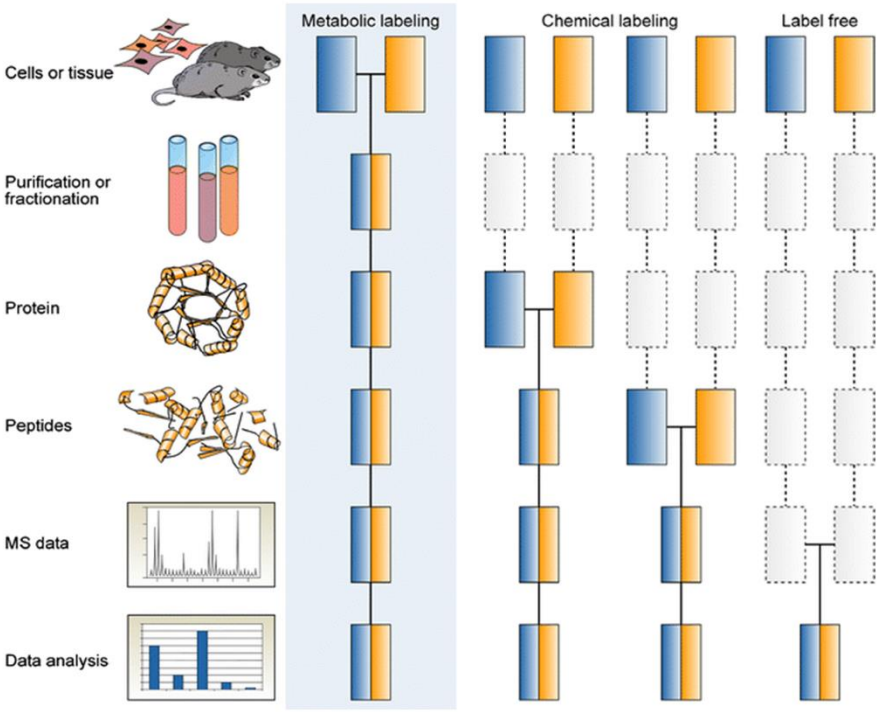

Figure 1.6 *A schematic view of common quantitative MS workflows. Blue and yellow boxes represent two experimental conditions. Horizontal lines indicate when the samples are combined. Dashed lines indicate points at which experimental variation and thus quantification errors can occur. (Copied from [34])*

Instead of using MS/MS spectra, another class of LFQ methods uses the precursor ion intensity, which is represented by the integrated peak area of extracted ion chromatograms (XIC) of a precursor in MS1 scan. For example, the Top 3 Protein Quantification (T3PQ) method approximates the abundance of a protein using the average intensity of the three most abundant peptides originated from the protein [36]. To ensure different experiments having a comparable scale, the average intensities are further divided by the sum of top 3 intensity values in an entire experiment. Similar to RNA-seq where long transcripts produce more reads, long proteins generate more digested peptides in comparison with equal amounts of shorter ones. The intensity-based absolute quantification (iBAQ) solves this problem by dividing the protein intensity by the number of digested peptides in the length of 6 and 30 amino acids [37]. The protein intensity is estimated by the sum of all (unique) peptides originated from it (which is used by MaxQuant [38]). Another LFQ method, which is implemented in MaxQuant (maxLFQ), is based on the assumption that the abundance of majority of the proteins is constant across samples.

Therefore, maxLFQ normalizes the protein intensities using a global optimization procedure that minimizes the proteome variation over all samples [39]. The accuracy of LFQ methods is the lowest in comparison with reporter ion-based methods, but the data generated by LFQ are increasingly available because of the simple sample preparation. The maxLFQ and iBAQ data were frequently used in this project.

1.3 A Multivariate dimension reduction approaches

1.3.1 Dimension reduction and principal component analysis

Dimension reduction methods arose in the early 20th century [40, 41] and have continued to evolve independently in multiple fields, giving rise to a myriad of associated terminologies. Each of these approaches used in the thesis are dimension reduction techniques. Therefore, this thesis starts by introducing the central concepts of dimension reduction.

Boldface uppercase letters are used to denote matrices. In this chapter, the rows of a matrix contain the cases or samples while the columns hold the variables. In an omics study, the variables (also referred as features) generally measure biological molecules including abundance of mRNAs, proteins, or metabolites. Vectors are denoted with boldface lowercase letters. All vectors are column vectors. Scalars are denoted by italic letters. Given an omics dataset, **X**, which is an $n \times p$ matrix of n observations and p variables, it can be represented by:

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p) \quad (1)$$

where **x**s are vectors of length n, as it is the mRNA or other biological variable measurements for n samples. In a typical omics study, p ranges from several hundred to millions. Therefore, samples are represented in a very large dimensional spaces $\mathbb{R}^p$. The goal of dimension reduction is to identify a (set of) new variable(s) using linear combination of the original variables, such that the number of new variables is much smaller than p. An example of such a linear combination is shown in equation (2).

$$\mathbf{f} = q_1\mathbf{x}_1 + q_2\mathbf{x}_2 + ... + q_p\mathbf{x}_p \quad (2)$$

or expressed in a matrix form

$$\mathbf{f} = \mathbf{X}\mathbf{q} \quad (3)$$

In equations (2) and (3), **f** is a new variable, which is often called a latent variable or a component. $\mathbf{q} = (q_1, q_2, ..., q_p)^T$ is a p-length vector storing the coefficients of the linear combination of original variables. These coefficients are also called "loadings." In dimension reduction analysis, additional

constraints are introduced to obtain a meaningful solution. Different optimization and constraints criteria underscore the difference between different dimension reduction methods.

Principal component analysis (PCA) is one of the most widely used dimension reduction methods [42]. Given a column centered (zero mean) and scaled (unit variance) matrix **X**, PCA finds a set of new variables $\mathbf{f}^i=\mathbf{X}\mathbf{q}^i$ (where i is the ith component and $\mathbf{q}^i$ is the variable loading for the ith principal component (PC) (superscript denotes the component or dimension) so that the variance of $\mathbf{f}^i$ is maximized, that is:

$$\arg \max_{q^i} \text{var}(\mathbf{X}\mathbf{q}^i) \tag{4}$$

with the constraints that $||\mathbf{q}^i||=1$ and each pair of components ($\mathbf{f}^i$, $\mathbf{f}^j$) are orthogonal to each other (or uncorrelated, i.e., $\mathbf{f}^{iT}\mathbf{f}^j=0$ for $j\neq i$).

PCA can be computed using different algorithms including eigenanalysis, singular value decomposition (SVD) [43] or linear regression [44]. Among them, SVD is the most widely used approach. Given **X**, an $n\times p$ matrix with rank r, $r \leqslant \min(n, p)$, SVD decomposes **X** into three matrices:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{Q}^T \text{ subject to the constraint that } \mathbf{U}^T\mathbf{U} = \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \tag{5}$$

where **U** is an $n\times r$ matrix and **Q** is a $p\times r$ matrix, and the columns of **U** and **Q** are the orthonormal left and right singular vectors, respectively. **S** is an $r\times r$ diagonal matrix of singular values, which are proportional to the standard deviations associated with r singular vectors. The singular vectors are ordered such that their associated variances are monotonically decreasing. In a PCA of **X**, the PCs comprise an $n\times r$ matrix, **F**, which is defined as:

$$\mathbf{F} = \mathbf{U}\mathbf{S} = \mathbf{U}\mathbf{S}\mathbf{Q}^T\mathbf{Q} = \mathbf{X}\mathbf{Q} \tag{6}$$

where the columns of matrix **F** are the PCs and the columns of matrix **Q,** which is the loading matrix, contains the coefficients of the variables for each PC (**q** in equation (3)). The above formula also emphasizes that **Q** is a matrix that projects the observations in **X** onto the PCs.

PCA can also be calculated using the NIPALS algorithm (algorithm 1.1). It computes PCs in a sequential manner (in contrast SVD calculates all PC simultaneously) – through regression procedure for each dimension (steps 1–3). For a higher order solution, the same procedure is applied on the residual matrix calculated from the deflation step (step 5). The residual matrix can be viewed as a result of regressing $\mathbf{X}^{i-1}$ onto $\mathbf{f}^i$ or removing the variance explained by $\mathbf{f}^i$ from $\mathbf{X}^{i-1}$. This algorithm is particularly useful in the analysis of omics data because it requires less computational time when only a few PCs are of interest, particularly for a large matrix **X**. In addition, another benefit of calculating PCs through

a regression approach is that it can handle a small number of missing values, which are often present in omics data. Of particular interest in multiple omics data analysis, this algorithm may be generalized to discover the correlated structure in more than one dataset (see sections on the analysis of multiple omics datasets).

Algorithm 1.1 – Nonlinear Iterative Partial Least square (NIPALS) algorithm for principal component analysis*.*

Initialize $\mathbf{X}^0 = \mathbf{X}$

for i = 1, ..., r

initialize $\mathbf{f}^i$ as the first column in $\mathbf{X}^{i-1}$

1. $\mathbf{q}^i = \mathbf{X}^{i-1}\mathbf{f}^i / (\mathbf{f}^{iT}\mathbf{f}^i)$
2. $\mathbf{q}^i = \mathbf{q}^i / (\mathbf{q}^{iT}\mathbf{q}^i)^{1/2}$
3. $\mathbf{f}^i = \mathbf{X}^{i-1}\mathbf{q}^i$
4. Check convergence of $\mathbf{f}^i$ and $\mathbf{q}^i$, if not, go back to step 1; otherwise step 5.
5. $\mathbf{X}^i = \mathbf{X}^{i-1} - \mathbf{f}^i\mathbf{q}^{iT}$ # deflation step

The results of PCA can be easily interpreted by visualizing the samples and variables on the same component space using a biplot. For example, PCA was used to analyze mRNA expression data of different cell lines from the NCI-60 cell lines panel (Figure 1.7). To simplify this case study, only 20 genes and cell lines originating from melanoma (ME), leukemia (LE), and central nervous system (CNS) tumors were included. Figure 1.7A is a biplot showing the first and second PCs, where samples are points and genes are arrows from the plot origin. The relative positions of samples are important, since variable profiles that are similar will be projected closer to one another. The lengths of vectors are proportional to the squared correlations between the components and variables. Additionally, variables with a relatively high expression in a sample will be positioned in the same direction with that of the sample, and the greater the distance from the origin, the stronger the association.

Most analyses plot and examine the first few PCs since they explain the most variant trends in the dataset. Generally, the selection of components is subjective and depends on the study purpose. An informal elbow test could help to determine the number of PCs to retain [45, 46]. For example, Figure 1.7B shows the elbow point is the fourth PC because the decrease in PC variance becomes relatively moderate from this point. Another approach that is widely used is to include (or retain) PCs that cumulatively capture a certain proportion of variance (e.g., 70% of variance is modeled with three PCs). If a parsimony model is preferred, the variance proportion cutoff can be as low as 50% [45].

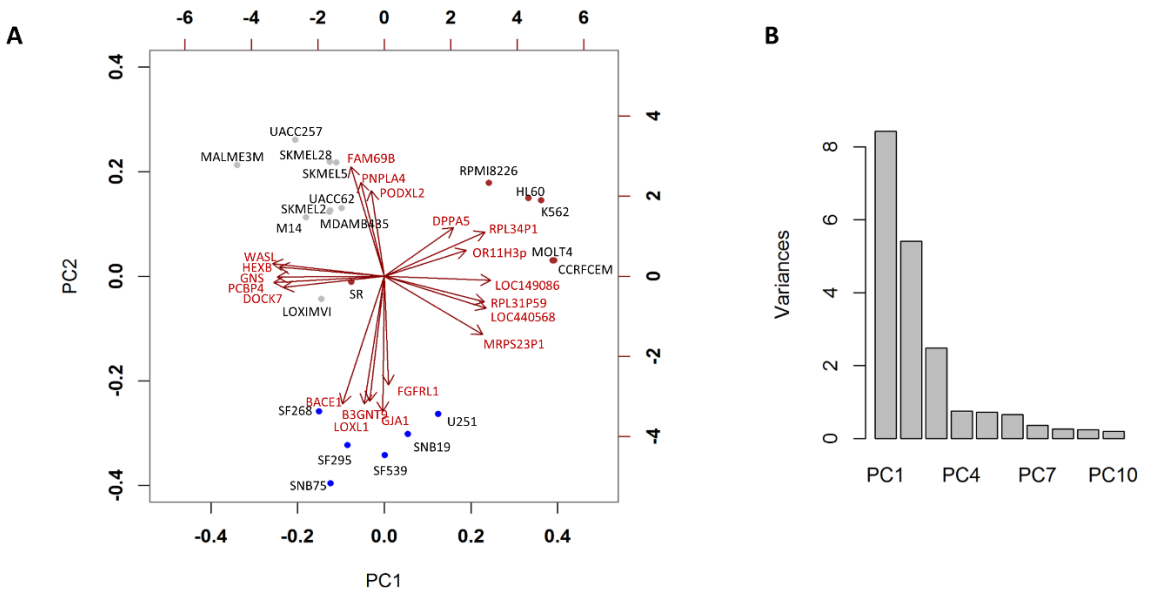

Figure 1.7 *Output of PCA (A) A biplot of the first two PCs, where genes are arrows and dots are cell lines. (B) The scree shows the variance of PCs.*

There are many dimension reduction approaches related to PCA, such as principal coordinate analysis (PCoA), correspondence analysis (CA), and nonsymmetrical correspondence analysis (NSCA). Each of these may be computed by SVD but differ in how the data are transformed [43, 47, 48]. PCoA (also known as classical multidimensional scaling) is an SVD of a distance matrix. CA and NSCA decompose a chi-squared matrix [48, 49]. Although designed for contingency tables of nonnegative count data, CA and NSCA have been successfully applied to continuous data including gene expression and protein profiles [51, 52]. As described previously, gene and protein expression can be seen as an approximation of the number of corresponding molecules present in the cell during a certain measured condition [52]. Additionally, Greenacre [48] emphasized that the descriptive nature of CA and NSCA allows their application on data tables in general, not only on count data. These two arguments support the suitability of CA and NSCA as analysis methods for omics data. While CA investigates symmetric associations between two variables, NSCA captures asymmetric relations between variables.

1.3.2 Integrative analysis of two datasets

Some extensions of one-table dimension reduction methods have been described, allowing the simultaneous decomposition and integrative analysis of paired data matrices into the same space. These include generalized SVD (gSVD) [53], co-inertia analysis (CIA) [54, 55], and sparse or penalized extensions of partial least squares (PLS) and canonical correspondence analysis (CCA) and Canonical Correlation Analysis (CCA) [56-59]. Given two omics datasets **X** (dimension $n \times p_x$) and **Y** (dimension $n \times p_y$), these methods can be expressed as the following latent component decomposition problem:

$$\mathbf{X} = \mathbf{F}_x \mathbf{Q}_x^{\mathrm{T}} + \mathbf{E}_x$$
$$\mathbf{Y} = \mathbf{F}_y \mathbf{Q}_y^{\mathrm{T}} + \mathbf{E}_y \quad (10)$$

where $\mathbf{F}_x$ and $\mathbf{F}_y$ are $n \times r$ matrices, the columns of which are components representing the co-structure between **X** and **Y**. The columns of $\mathbf{Q}_x$ and $\mathbf{Q}_y$ are loading vectors for variables in **X** and **Y**, respectively.

CCA searches for a linear combinations of original variables from a pair of matrices so that their correlation is maximized, that is,

for the *i*th component:

$$\arg\max_{\mathbf{q}_x^i \mathbf{q}_y^i} \mathrm{cor}(\mathbf{X}\mathbf{q}_x^i, \mathbf{Y}\mathbf{q}_y^i) \quad (11)$$

In CCA, the components $\mathbf{X}\mathbf{q}_x^i$ and $\mathbf{Y}\mathbf{q}_y^i$ are called canonical variates and their correlations are the canonical correlations. One of the main limitations of applying CCA to omics data is that it requires an inversion of the covariance matrix [60-62], which cannot be calculated when the number of variables exceeds the sample size [58]. Given the high dimensionality of omics data where *p>>n*, applying CCA to omics data requires a regularization step, which may be accomplished by adding a RIDGE penalty, that is adding a multiple of the identity matrix to the covariance matrix.

PLS is an efficient dimension reduction method and has been used in the analysis of high dimensional omics data. Depending on the algorithm, different objective functions with different constraints are optimized [see review 63]. In summary, PLS maximizes the covariance rather than the correlation between components, which is used in CCA. From the algorithm point of view, PLS components can be calculated via kernel-PLS, an iterative local regression algorithm such as PLS2 or the statistically inspired modification of PLS (SIMPLS) [63], which do not require the inversion of a correlation matrix. For example, the PLS2 algorithm is an extension of the NIPALS algorithm onto two datasets scheme. This algorithm calculates the components and loading vectors using the iterative local regression procedure. Therefore, it does not suffer from the *p>>n* problem as in CCA does. Even though PLS can be applied to data with *p>>n* without penalty, a sparse solution is desired. For example, a sparse PLS method was proposed for the feature selection purpose by introducing a Lasso-penalization to the loading vectors [64]. In a recent comparison, sPLS performed comparably to sparse CCA [57]. For the classification purpose, sPLS was combined with discriminant analysis (sPLS-DA) by coding the response matrix Y with dummy variables. This method has been applied to classification and variable selection of microarray and SNP data [64].

CIA is a descriptive non-constrained approach for coupling pairs of data matrices. It was originally proposed to link two ecological tables [65, 66], but has been successfully applied in omics data analysis [51, 54]. CIA is implemented under the duality diagram framework. In this scheme, CIA analyzes two statistical triplets (**X**,**L**,**D**) and (**Y**,**R**,**D**). In physics, the inertia of a set of points relative to one point is defined by the weighted sum of squared distances between each considered point and the reference point. Correspondingly, the inertia of a centered matrix (zero mean) is simply the sum of the squared

matrix elements. The inertia of the matrix **X** defined by the metrics **L** and **D** is the weighted sum of its squared values, that is:

$$\text{trace}(\mathbf{XLX}^{\mathrm{T}}\mathbf{D}). \tag{12}$$

The inertia equals the total variance of **X** when **X** is centered, **L** is the Euclidean metric and **D** is a diagonal matrix with $l_i = 1/n$. However, the concept of inertia is more flexible since different metrics may be used to account for different types of data. For example, if **L** and **T** are defined as in CA the inertia of **X** is proportional to the χ^2 statistics. When coupling a pair of datasets, the co-inertia between two matrices, **X** and **Y**, is calculated as

$$\text{trace}(\mathbf{XLX}^{\mathrm{T}}\mathbf{DYRY}^{\mathrm{T}}\mathbf{D}). \tag{13}$$

CIA decomposes the co-inertia criteria into a set of orthogonal axes. To benefit from the flexible weighting, CIA is performed in two steps: 1) application of a dimension reduction technique such as PCA, CA or NSCA to the initial datasets depending on the type of data (binary, categorical, discrete counts or continuous data) and 2) constraining the projections of the orthogonal axes such that they are maximally covariant [55, 65]. CIA does not require an inversion step of the correlation or covariance matrix thus can be applied to datasets of genomics data (*p>>n*) without regularization or penalization.

CIA is closely related to CCA [60], but it maximizes the squared covariance between the linear combination of the preprocessed matrix, that is,

for the *i*th dimension:

$$\arg\max_{\mathbf{q}_x^i \mathbf{q}_y^i} \operatorname{cov}^2(\mathbf{X}\mathbf{q}_x^i, \mathbf{Y}\mathbf{q}_y^i). \tag{14}$$

Equation (14) can be decomposed as:

$$\operatorname{cov}^2(\mathbf{Xq}_x^i, \mathbf{Yq}_y^i) = \operatorname{cor}^2(\mathbf{Xq}_x^i, \mathbf{Yq}_y^i)\cdot\operatorname{var}(\mathbf{Xq}_x^i)\cdot\operatorname{var}(\mathbf{Xq}_y^i) \tag{15}$$

The relationship between CIA, Procrustes analysis [66] and CCA [60] have been well described. A comparison between sCCA (with elastic net normalization), sPLS and CIA is provided in [57]. CIA and sPLS both maximize the covariance between components and efficiently identify both joint and individual patterns in the paired datasets. In contrast, CCA maximizes the correlation between components and will mainly discover the concordance structure presented in both datasets, but may fail to discover strong individual effects [57]. In addition, both sCCA and sPLS are sparse methods and variables selected by these methods are similar, whereas, CIA does not integrate a feature selection step. Hence, in terms of feature selection, the result of CIA is more redundant in comparison to the sparse methods [57].

1.3.3 Correlation structure across multiple matrices

Simultaneous decomposition and integration of multiple matrices is the focus of this thesis. It is more complex than the analysis of single or paired data, because each dataset may have different numbers of variables, different scales, different internal structure or different variance. This might produce

components that are dominated by one or a few datasets. Therefore, it is crucial to preprocess the datasets before decomposition. The preprocessing is often performed on two levels. On the variable level, variables are often normalized as PCA, CA or NSC. This procedure enables all the variables to have comparable contribution to the total inertia of a dataset. However, the number of variables may vary between datasets. Therefore, a dataset level normalization is required. Frequently, the datasets are divided by the square root of their total inertia (sum of squares of all elements) or by the square root of the numbers of columns of each dataset [67].

The multiple matrices analysis methods can be generally expressed as the following model:

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{F}\mathbf{Q}_1^{\mathrm{T}} + \mathbf{E}_1 \\ &\vdots \\ \mathbf{X}_k &= \mathbf{F}\mathbf{Q}_k^{\mathrm{T}} + \mathbf{E}_k \\ &\vdots \\ \mathbf{X}_K &= \mathbf{F}\mathbf{Q}_K^{\mathrm{T}} + \mathbf{E}_K \end{aligned} \tag{16}$$

The matrices $\mathbf{X_k}$ (*k* ranges from 1 to *K,* representing *K* omics datasets). For the convenience in expression, **F** is the "component" matrix that integrates information from all datasets, i.e. the common pattern defined by all datasets. The matrices $\mathbf{Q_i}$, with i ranging from 1 to *K* are the loadings or coefficient matrices. A high positive value indicates a strong positive contribution of the corresponding variable to the "Component". The multivariate analysis methods used in this thesis, including multiple co-inertia analysis, multiple factorial analysis and consensus principal component analysis, belong to this family.

1.4 Objective and outline of this thesis

Multivariate analysis has a long history but its application to omics data analysis is rising in recent years. The objective of this thesis is to explore the application of multivariate methods to omics data analysis, particularly for the integrative analysis of multi-omics data. The outline of this thesis is given below.

Chapter II describes the application of multiple co-inertia analysis (MCIA), in combination with non-symmetric correspondence analysis (NSC), to the analysis of transcriptomic and proteomic data. MCIA can be used for the integrative exploratory of multiple omics data. It simultaneously projects several datasets into the same dimensional space, transforming features onto the same scale, to extract the most variant from each dataset and facilitate biological interpretation and pathway analysis.

Integrative exploratory analysis is often used for quality control in early stage of data analysis, such as detecting outliers or batch effect. In the downstream analysis of cancer omics data, clustering methods are often required for the stratification of patients. Traditional clustering methods are applicable to single omics data. Chapter III introduces a multiple omics clustering method, moCluster, which find the joint pattern across omics data. At the same time, it identified the biomarkers associated with each cluster via a sparse operator (L1-penalty). The merits of this method, including stable solution and better computational efficiency, have been established through the comparison with existing methods, including iCluster and iCluster+.

Gene-set annotation is widely used in omics data analysis because it provides functional insights for a large list of biological molecules. Similar to the clustering analysis, gene-set analysis methods are largely confined to the analysis of single datasets. Chapter IV describes an integrative gene-set analysis method − moGSA. The application of moGSA to simulated data demonstrates that the integration of multiple omics data potentially increases the power to detect permuted gene-sets.

Chapter IV gives a general discussion about how the multivariate methods are interlinked and can be included in a more general framework, Regularized Generalized Canonical Correlation Analysis (RGCCA). I will also discuss the challenges of integrative analysis and the merits of multivariate method to solve these challenges. Last, the chapter emphasizes emerging requirements of multivariate analysis that are particularly important in omics data analysis, such as dealing with missing values and more interactive visualization.

ABBREVIATIONS

CA	Correspondence Analysis
CCA	Canonical Correlation Analysis or Canonical Correspondence Analysis *
cDNA	complementary DNA
CIA	Co-Inertia Analysis
CID	collision induced dissociation
DDA	Data dependent acquisition
EM	Expectation Maximization
ENCODE	Encyclopedia of DNA Elements
ESI	electrospray ionization
ETD	electron transfer dissociation
FPKM	Fragments Per Kilobase Per Million
gSVD	generalized Singular Value Decomposition
HCD	high energy C-trap dissociation
HPLC	High-Performance Liquid Chromatography
iBAQ	Intensity Base Absolute Quantification
IT	Ion-trap
LC	Liquid Chromatography
LDA	Linear Discriminant Analysis
LFQ	Label Free Quantification
MALDI	matrix-assisted laser desorption and ionization
MAS5	MicroArray Suite 5
MCIA	Multiple Co-Inertia Analysis
MS	Mass Spectrometry
NSCA	Non-Symmetric Correspondence Analysis
PC	Principal Component
PCA	Principal Component analysis
PCoA	Principal Co-ordinate Analysis
PLS	Partial Least Square

PSM	peptide spectrum match
RGCCA	Regularized Generalized Canonical Correlation Analysis
RMA	Robust Multiarray Average
RPKM	Reads Per Kilobase Per Million
RSEM	RNA-Seq by Expectation Maximization
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SILAC	Stable Isotope Labeling by Amino Acids in Cell Culture
SNP	single nucleotide polymorphism
SVD	Singular Value Decomposition
T3PQ	Top 3 Protein Quantification
TCGA	The cancer genome atlas
TOF	time of flight
XIC	Extracted Ion Chromatograms

* Both Canonical Correspondence analysis and Canonical Correlation Analysis are referred to by the acronym CCA. Canonical Correspondence analysis is widely used in ecologically statistics, and is a constrained form of CA, however it has not been adopted by the genomics community in the analysis of pairs of omics data. By contrast several groups have applied extensions of Canonical Correlation Analysis to omics data integration. Therefore, this thesis uses CCA to describe Canonical Correlation Analysis.

REFERENCES

1. Croce CM: **Oncogenes and cancer**. *N Engl J Med* 2008, **358**(5):502-511.
2. Alizadeh AA, Aranda V, Bardelli A, Blanpain C, Bock C, Borowski C, Caldas C, Califano A, Doherty M, Elsner M *et al*: **Toward understanding and exploiting tumor heterogeneity**. *Nat Med* 2015, **21**(8):846-853.
3. Prohaska S, Stadler P: **The Use and Abuse of -Omes**. In: *Bioinformatics for Omics Data Methods and Protocols.* Edited by Mayer B: Humana Press; 2011.
4. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Paabo S, Mann M: **Deep proteome and transcriptome mapping of a human cancer cell line**. *Mol Syst Biol* 2011, **7**:548.
5. Moghaddas Gholami A, Hahne H, Wu Z, Auer FJ, Meng C, Wilhelm M, Kuster B: **Global proteome analysis of the NCI-60 cell line panel**. *Cell Rep* 2013, **4**(3):609-620.
6. Geiger T, Wehner A, Schaab C, Cox J, Mann M: **Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins**. *Mol Cell Proteomics* 2012, **11**(3):M111 014050.
7. Cancer Genome Atlas Research N: **Comprehensive molecular profiling of lung adenocarcinoma**. *Nature* 2014, **511**(7511):543-550.
8. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA *et al*: **Mutational landscape and significance across 12 major cancer types**. *Nature* 2013, **502**(7471):333-339.
9. Cancer Genome Atlas Research N: **Comprehensive molecular characterization of urothelial bladder carcinoma**. *Nature* 2014, **507**(7492):315-322.
10. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V *et al*: **Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin**. *Cell* 2014, **158**(4):929-944.
11. Miller MB, Tang YW: **Basic concepts of microarrays and potential applications in clinical microbiology**. *Clin Microbiol Rev* 2009, **22**(4):611-633.
12. Bumgarner R: **Overview of DNA microarrays: types, applications, and their future**. *Curr Protoc Mol Biol* 2013, **Chapter 22**:Unit 22 21.
13. Gharaibeh RZ, Fodor AA, Gibas CJ: **Background correction using dinucleotide affinities improves the performance of GCRMA**. *BMC Bioinformatics* 2008, **9**:452.
14. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**(1):57-63.
15. Metzker ML: **Sequencing technologies - the next generation**. *Nat Rev Genet* 2010, **11**(1):31-46.
16. Garber M, Grabherr MG, Guttman M, Trapnell C: **Computational methods for transcriptome annotation and quantification using RNA-seq**. *Nat Methods* 2011, **8**(6):469-477.
17. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, **28**(5):511-515.
18. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C *et al*: **Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs**. *Nat Biotechnol* 2010, **28**(5):503-510.
19. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA: **Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)**. *Bioinformatics* 2011, **27**(18):2518-2528.
20. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat Methods* 2008, **5**(7):621-628.

21. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC Bioinformatics* 2011, **12**:323.
22. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H *et al*: **Mass-spectrometry-based draft of the human proteome**. *Nature* 2014, **509**(7502):582-587.
23. Aebersold R, Mann M: **Mass spectrometry-based proteomics**. *Nature* 2003, **422**(6928):198-207.
24. Steen H, Mann M: **The ABC's (and XYZ's) of peptide sequencing**. *Nat Rev Mol Cell Biol* 2004, **5**(9):699-711.
25. Hagglund P, Bunkenborg J, Elortza F, Jensen ON, Roepstorff P: **A new strategy for identification of N-glycosylated proteins and unambiguous assignment of their glycosylation sites using HILIC enrichment and partial deglycosylation**. *J Proteome Res* 2004, **3**(3):556-566.
26. Pinkse MW, Uitto PM, Hilhorst MJ, Ooms B, Heck AJ: **Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns**. *Anal Chem* 2004, **76**(14):3935-3943.
27. Larsen MR, Thingholm TE, Jensen ON, Roepstorff P, Jorgensen TJ: **Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns**. *Mol Cell Proteomics* 2005, **4**(7):873-886.
28. Chalkley R: **Instrumentation for LC-MS/MS in Proteomics**. In: *LC-MS/MS in Proteomics Methods and Applications.* Edited by Cutillas PR, Timms JF: Humana Press; 2010.
29. Nesvizhskii A: **Protein Identification by Tandem Mass Spectrometry and Sequence Database Searching**. In: *Mass Spectrometry Data Analysis in Proteomics.* Edited by Matthiesen R: Humana press; 2007.
30. Martens L, Hermjakob H: **Proteomics data validation: why all must provide data**. *Mol Biosyst* 2007, **3**(8):518-522.
31. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics**. *Mol Cell Proteomics* 2002, **1**(5):376-386.
32. Kirchner M, Selbach M: **In vivo quantitative proteome profiling: planning and evaluation of SILAC experiments**. *Methods Mol Biol* 2012, **893**:175-199.
33. McAlister GC, Huttlin EL, Haas W, Ting L, Jedrychowski MP, Rogers JC, Kuhn K, Pike I, Grothe RA, Blethrow JD *et al*: **Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses**. *Anal Chem* 2012, **84**(17):7469-7478.
34. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B: **Quantitative mass spectrometry in proteomics: a critical review**. *Anal Bioanal Chem* 2007, **389**(4):1017-1031.
35. Bantscheff M, Lemeer S, Savitski MM, Kuster B: **Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present**. *Anal Bioanal Chem* 2012, **404**(4):939-965.
36. Grossmann J, Roschitzki B, Panse C, Fortes C, Barkow-Oesterreicher S, Rutishauser D, Schlapbach R: **Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods**. *J Proteomics* 2010, **73**(9):1740-1746.
37. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: **Global quantification of mammalian gene expression control**. *Nature* 2011, **473**(7347):337-342.
38. Cox J, Mann M: **MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification**. *Nat Biotechnol* 2008, **26**(12):1367-1372.
39. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M: **Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ**. *Mol Cell Proteomics* 2014, **13**(9):2513-2526.
40. Pearson K: **LIII. On lines and planes of closest fit to systems of points in space**. *Philosophical Magazine Series 6* 1901, **2**(11):559-572.

41. Hotelling H: **Analysis of a complex of statistical variables into principal components**. *Journal of educational psychology* 1933, **24**(6):417.
42. Ringner M: **What is principal component analysis?** *Nature biotechnology* 2008, **26**(3):303-304.
43. Wall ME, Rechtsteiner A, Rocha LM: **Singular value decomposition and principal component analysis**. *A practical approach to microarray …* 2003.
44. Legendre P, Legendre LFJ: **Numerical ecology**. *Numerical ecology* 2012.
45. Abdi H, Williams LJ: **Principal component analysis**. *Wiley Interdisciplinary Reviews* 2010.
46. Dray S: **On the number of principal components: A test of dimensionality based on measurements of similarity between matrices**. *Computational Statistics & Data Analysis* 2008, **52**(4):2228-2237.
47. Wouters L, Gohlmann HW, Bijnens L, Kass SU, Molenberghs G, Lewi PJ: **Graphical exploration of gene expression data: a comparative study of three multivariate methods**. *Biometrics* 2003, **59**(4):1131-1139.
48. Greenacre M: **Correspondence analysis in practice**. *Correspondence analysis in practice* 2007.
49. Beh EJ, Lombardo R: **Correspondence Analysis: Theory, Practice and New Strategies**. *Correspondence Analysis: Theory, Practice and New Strategies* 2014.
50. Greenacre MJ: **Theory and applications of correspondence analysis**. *Theory and applications of correspondence analysis* 1984.
51. Fagan A, Culhane AC, Higgins DG: **A multivariate analysis approach to the integration of proteomic and gene expression data**. *Proteomics* 2007, **7**(13):2162-2171.
52. Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M: **Correspondence analysis applied to microarray data**. *Proc Natl Acad Sci U S A* 2001, **98**(19):10781-10786.
53. Alter O, Brown PO, Botstein D: **Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms**. *Proc Natl Acad Sci U S A* 2003, **100**(6):3351-3356.
54. Culhane AC, Perriere G, Higgins DG: **Cross-platform comparison and visualisation of gene expression data using co-inertia analysis**. *BMC Bioinformatics* 2003, **4**:59.
55. Dray S: **Analysing a Pair of Tables: Coinertia Analysis and Duality Diagrams**. *Visualization and Verbalization of Data* 2014.
56. Witten DM, Tibshirani R, Hastie T: **A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis**. *Biostatistics* 2009, **10**(3):515-534.
57. Le Cao KA, Martin PG, Robert-Granie C, Besse P: **Sparse canonical methods for biological data integration: application to a cross-platform study**. *BMC Bioinformatics* 2009, **10**:34.
58. Braak CJF: **Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis**. *Ecology* 1986.
59. Hotelling H: **Relations between two sets of variates**. *Biometrika* 1936.
60. Thioulouse J: **Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods**. *The Annals of Applied Statistics* 2011.
61. Hong S, Chen X, Jin L, Xiong M: **Canonical correlation analysis for RNA-seq co-expression networks**. *Nucleic Acids Res* 2013, **41**(8):e95.
62. Lee S, Epstein MP, Duncan R, Lin X: **Sparse principal component analysis for identifying ancestry-informative markers in genome-wide association studies**. *Genet Epidemiol* 2012, **36**(4):293-302.
63. Boulesteix AL, Strimmer K: **Partial least squares: a versatile tool for the analysis of high-dimensional genomic data**. *Brief Bioinform* 2007, **8**(1):32-44.
64. Le Cao KA, Boitard S, Besse P: **Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems**. *BMC Bioinformatics* 2011, **12**:253.
65. Dolédec S, Chessel D: **Co‐inertia analysis: an alternative method for studying species–environment relationships**. *Freshwater biology* 1994.

66. Dray S, Chessel D, Thioulouse J: **Co-inertia analysis and the linking of ecological data tables**. *Ecology* 2003.
67. Abdi H, Williams LJ, Valentin D: **Statis and distatis: optimum multitable principal component analysis and three way metric multidimensional scaling**. *Wiley Interdisciplinary*2012.

CHAPTER II

Multiple Co-Inertia Analysis: A Multivariate Approach to the Integration of Multiple Omics Datasets

SUMMARY

To leverage the potential of multi-omics studies, exploratory data analysis methods that provide systematic integration and comparison of multiple layers of omics information are required. In this chapter, an exploratory data analysis method, multiple co-inertia analysis (MCIA), is introduced. It identifies co-relationships between multiple high dimensional datasets. Based on a covariance optimization criterion, MCIA simultaneously projects several datasets into the same dimensional space, transforming features onto the same scale, to extract the most variant from each dataset and facilitate biological interpretation and pathway analysis.

MCIA was applied to two typical scenarios of integrative analysis of omics data. The integration of transcriptome and proteome profiles of cells in the NCI-60 cancer cell line panel revealed distinct, complementary features, which together increased the coverage and power of pathway analysis. The analysis highlighted the importance of the leukemia extravasation signaling pathway in leukemia that was not ranked highly when datasets are analyzed individually. Secondly, I compared transcriptome profiles of high grade serous ovarian tumors that were obtained on two different microarray platforms and next generation RNA-sequencing. The results suggested that the variance of RNA-seq data processed using RPKM had greater variance than that with MapSlice and RSEM. In addition, MCIA detected novel markers highly associated to tumor molecular subtype combined from four data platforms.

2.1 BACKGROUND

Possibly the most straightforward way of integrating multiple omics data is to map identifiers from each platform to a common set and analyze the common subset of genes [1, 2]. However, this overlooks several fundamental platform and biological biases. Platforms are not universal and target different molecules. Filtering genes to their intersection considerably reduces data coverage and because correlations between different platforms are probably lower than expected [3], it may not provide large gains in data quality or study power. Such filtering may also introduce bias because platform discrepancy in molecule quantification could reflect biological variation. Poor correlation between a transcript and its translated protein may be a result of biological processes, such as post-transcriptional repression of genes by miRNA [4, 5]. Correlations between proteins and metabolites of pathways can be divergent if proteins are expressed in an inactive form, in which case the abundance of a protein does not represent activity. In addition to technological and biological gene variance, the many-to-many mapping of gene identifiers from multiple platforms complicates direct comparison of genes across multiple levels.

Ordination methods, such as principal component analysis (PCA) and correspondence analysis (COA), are exploratory data analysis approaches that have been applied to analyze omics data including transcriptome and proteome studies [6-9]. Graphical representation of measurements (samples) and variables (genes, proteins) onto a lower dimensional space facilitates interpretation of global variance structure and identification of the most informative (or variant) features across datasets. These methods permit visualization of data that have considerable levels of noise and data where the number of variables exceeds the number of measurements, which is typical in omics studies. However, these approaches do not solve the problem of comparing many datasets simultaneously. Some efforts have been done to couple two datasets together. One such approach is co-inertia analysis [CIA; 10]. CIA was originally applied to study of ecological and environmental tables, where it was employed to link environmental variables with species characteristics [11]. Culhane and colleagues introduced CIA in genomics, when they compared data from two microarray platforms [12]. An advantage of this method is that it does not require the mapping or filtering of genes to a common set. In addition, CIA couples with several dimension reduction approaches, including PCA or correspondence analysis, such that it can accommodate both discrete count data (e.g. somatic mutation) and continuous data. However, CIA is limited to two datasets, which confines its application in modern multi-omics studies.

Here, multiple co-inertia analysis (MCIA), an extension to CIA, is described for the analysis of more than two omics datasets. The application of MCIA is illustrated by two different examples, and show that integrated analysis is more insightful than analysis of the individual datasets. First, I demonstrated the power of MCIA via applying it to the integration and comparison of multi-omics data independent of data annotation. MCIA is able to identify common relationships among multiple genes and protein expression data of the NCI-60 cancer cell line panel of the National Cancer Institute [14-16]. The integrated analysis revealed that cell lines are clustered according to anatomical tissue source and showed a significant degree of correlation between transcript and protein expression. Second, MCIA was employed to assess the concordance in gene expression data obtained from microarray and next

generation RNA-sequencing of 266 samples of high grade serous ovarian cancer. Despite the fact that the majority of genes exhibit considerable variation between different platforms, MCIA revealed that, across all platforms, several biomarkers consistently relevant to ovarian cancer subtypes.

2.2 METHODS

2.2.1 Mathematical basis of MCIA

A typical omic dataset is a matrix where the number of features exceeds the number of measurements (row and columns of the matrix, respectively). Prerequisite for MCIA is a set of such matrices where the observations are matched. In this study, equal weights were applied to all the observations. MCIA is performed in two steps to represent features and observations as points along several axes. In the first step, a one matrix ordination method, such as PCA, COA or non-symmetric correspondence analysis [NSC; 17] is applied on each dataset separately, which transforms data into the same scale and comparable space.

In the analysis, given an omics data matrix **M** = [m_{ij}] with 1 ≤ i ≤ n and 1 ≤ j ≤ q, where **M** is a (*n* x *q)* matrix, *i* indicates row index and *j* for column index. Here, the row and column sums of **M** are denoted as m_{i+} and m_{+j} respectively, and m_{++} as the grand total. The relative contribution or weight of row *i* to the total variation in the dataset is denoted r_i and calculated as $r_i = m_{i+}/m_{++}$. While the relative contribution of column *j* is denoted as $c_j = m_{+j}/m_{++}$. Similarly, the contribution of each individual element of **M** to the total variation p_{ij} can be calculated as $p_{ij} = m_{ij}/m_{++}$. Then, a new matrix **X** with the values defined above could be calculated as

$$x_{ij} = \frac{p_{ij}}{r_i} - c_j \quad (1)$$

where x_{ij} is the centered row profile, i.e. the relative abundance of selected variable to the measurement's weight.

The second step is MCIA, which is the analysis of a set of statistical triplet ($\mathbf{X_k}$, $\mathbf{Q_k}$, **D**) where $k = 1, \ldots, k, \ldots, K$ and $\mathbf{X_k}$ are set of transformed matrices. $\mathbf{Q_k}$ is a $q_k \times q_k$ matrix with r_{ij} in diagonal elements, indicating the hyperspace of features metrics. Let **D** be an $n \times n$ matrix with 1 in the diagonal indicating equal weight across all columns in all matrices. The matrix $\mathbf{X_k}$ can be weighted and concatenated to one matrix as $\mathbf{X} = [\omega_1\mathbf{X_1} \,|\ldots|\, \omega_K\mathbf{X_K}]$, where ω is the weight of each matrix (reverse of sum of eigenvalue for each matrix).

MCIA maximizes the correlation of each individual matrices with a consensus reference structure through synthetic analysis, which finds a set of reference axes sequentially. In order to find the solution for the first dimension, MCIA determines a single reference axis (referred to as common component $\mathbf{v_1}$) and a set of auxiliary axes for each matrix ($\mathbf{u_{11}}$ .. $\mathbf{u_{1K}}$) so as to maximize the sum of the co-variances between each of the auxiliary axes $\mathbf{u_{1k}}$ and the $\mathbf{v_1}$. The first order solutions of $\mathbf{u_{11}}$ to $\mathbf{u_{1K}}$ and $\mathbf{v_1}$ are given by the first principal component of the concatenated weighted matrix **X**. The subsequent solutions are found with residual matrices from the calculation of the first order solution

with the constraint that the rest order axes are orthogonal with the previous sets. These steps are repeated so that the desired number of axes (principal components, dimensions) are generated. As a result, MCIA provides a simultaneous ordination of columns (measurements) and rows (features) of multiple matrices within the same hyperspace, with features or measurements sharing similar trends will be closely projected. The detailed description of MCIA and the proof that these axes are maximally co-variant are given in Chessel and Hanafi [18].

2.2.2 Datasets

This work analyzes two publicly available datasets: (i) transcriptomic [14, 16, 19] and proteomic [15] datasets of the NCI-60 cancer cell lines and (ii) an ovarian cancer dataset generated as part of the TCGA project [20]. Both comprise multiple datasets of the same samples.

NCI-60 data

The NCI-60 panel is a collection of 59 cancer cell lines originating from leukemia, lymphomas, melanomas and carcinomas of ovarian, renal, breast, prostate, colon, lung and CNS origin. The transcriptome data were downloaded from Cellminer [21] representing four different platforms: Affymetrix HG-U133 plus 2.0, HG-U133, HG-U95 and Agilent GE 4x44K [22]. Affymetrix data were normalized using GC robust multichip averaging [GCRMA; 23] and Agilent data were log transformed as obtained from the Cellminer. Microarray data were filtered to exclude probes that do not map to an official HUGO gene symbol. When multiple probes are mapped to the same gene, the probe with highest average value was retained. This filtering produced datasets of 11,051, 8,803, 9,044 and 10,382 genes on Agilent, HG-U95, HG-U133 and HG-U133 plus 2.0 platforms, respectively. The lung cancer cell line NCI-H23 was excluded since its expression profile was not available on the HG-U133 platform.

The proteome profiles of cell lines were produced from a conventional GeLC-MS/MS approach and label-free quantification, as described in [15]. The international protein index (IPI) identifiers were mapped to official gene symbol to facilitate subsequent pathway interpretation. Data were log transformed (base 10) and no filtering or additional normalization were performed. This dataset represents 7,150 protein expressions across 58 cell line in NCI-60 panel.

Ovarian cancer datasets

Gene expression of ovarian cancer tissues were profiled using two microarray platforms (Affymetrix customized platform G4502A and HG U133 plus 2.0) and RNA-sequencing (Illumina HiSeq platform). Data were downloaded from the NCI-TCGA data portal [07/08/2013; 20]. The Agilent and Affymetrix data were normalized and summarized by lowess and multichip averaging respectively [RMA; 24] respectively. Two different pre-processing pipelines were applied to the Illumina RNA-sequencing data to determine the transcript expression levels, denoted as RNASeq and RNASeqV2. RNASeq used the RPKM method [25] for normalization and quantification. RNAseqV2 employed MapSplice for alignment and RSEM for gene expression quantification [26, 27]. In this analysis, missing values were replaced with a small positive value (10e-15 in RNASeq and 10e-10 in RNASeqV2) and then the

expression values were log10 transformed. 266 out of 489 patient samples were present across all four datasets and hence included in the analysis. Only genes mapped to an official HUGO gene symbol were retained and duplicated genes were excluded. In the RNA sequencing datasets, 20,657 and 20,135 genes were detected and those with more than 15 missing values were removed, yielded 17,814; 12,042; 16,769 and 15,840 gene expressions in Agilent, Affymetrix, RNASeq and RNASeqV2 respectively.

2.3 RESULTS AND DISCUSSION

2.3.1 Integrated analysis of transcriptome and proteome of the NCI-60 cell lines

The NCI-60 panel, a collection of 59 cancer cell lines derived from nine different tissues (brain, blood and bone marrow, breast, colon, kidney, lung, ovary, prostate and skin) have been extensively used in *in vitro* high throughput drug screen assays. Their genetic composition on different levels have been extensively profiled, including comparative genomic hybridization array [28], karyotype analysis [29], DNA mutational analysis [30, 31], transcripts expression array [14, 32], microarrays for microRNA expression [16] and protein expression [15]. MCIA was applied as an exploratory analysis of four transcriptomic studies (Agilent n=11,051; HGU95 n=8,803; HGU133 n=9,044 and HGU133 plus 2.0 n=10,382) and one proteomic study (GeLC-MS/MS; n=7,150) of the 58 cell lines. Figure 2.1A shows the projection of cell lines on the first two components. The MCIA plot of the first two principal components (which accounted for 17.4% and 14.2% of the variance) shows similar trends within transcriptome as well as between transcriptome and proteome profiles, indicating that the most variant sources of biological information were similar in both transcriptome and proteome data. Generally, the cell lines originating from the same or closely related anatomical source converged into groups. The coordinates of cell lines from each dataset are connected by lines, the lengths of which indicate the divergence (the shorter the line, the higher the level of concordance between the mRNAs and proteins for a particular cell line). MCIA revealed colon, leukemia, melanoma, CNS, renal and ovarian cell lines segregated largely according to their tissue of origin. There was greater divergence in the cell lines from tumors with more intrinsic molecular heterogeneity (e.g. breast and NSCLC cell lines). The transcriptome and proteome profiles of these cell lines shared a high degree of consensus, providing further evidence that the observed spread reflected the tumor cell lines heterogeneity, as opposed to technical or other stochastic variance between gene or protein expression profiles. For instance, it was observed that the estrogen receptor positive breast cancer cell line MCF7 displays an epithelial phenotype and was clustered to colon cancer lines. In contrast, the cell line negative for the estrogen receptor, HS578T, clustered with the stromal/mesenchymal cluster of glioblastoma and renal tumor cell lines. This suggests that HS578T exhibits more invasive mesenchymal features comparing to T47D and MCF7. Seven out of eight melanoma lines clustered together, and the remaining one, LOX-IMVI, has been reported to lack melanin production [LOX-IMVI; 33]. These results are in consistent with hierarchical clustering analysis (Figure 2.2).

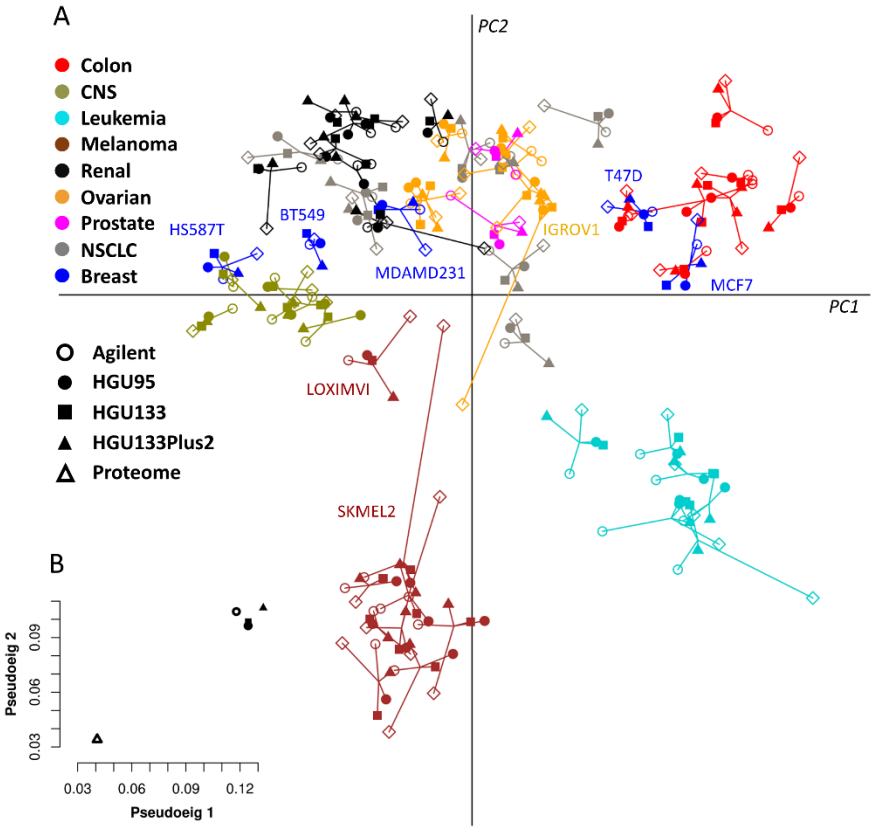

Figure 2.1 *MCIA projection plot. (A) The first two principal components (PCs) of MCIA representing transcriptomic and proteomic datasets of the NCI-60 panel. Different shapes represent the respective platforms and are connected by lines where the length of the line is proportional to the divergence between the same cell lines from each dataset. Lines are joined by a common point, representing the reference structure which maximizes covariance of multiple datasets. Reference coordinates emphasis on the similarity by showing which cell lines are driving this variation and which are responsible for the deformation of the common variation. Colors represent the nine NCI-60 cell lines from different tissues. The epithelial and mesenchymal features are separated along the first axis (PC1, horizontal). Melanoma and leukemia cell lines were projected on the on the negative side of second axis (PC2, vertical). (B) Summarizing the concordance between platforms by representing pseudo-eigenvalue space of NCI-60 datasets. The pseudo-eigenvalue space represents overall co-structure between datasets and shows which platform contributes more to the total variance.*

2.3.2 Overall co-structure comparison using MCIA

Each principal component (PC) has an associated eigenvalue, which represents its variability. In this analysis, the first three PCs of the MCIA accounted for 17.4%, 14.2% and 9.7% of variance respectively. Therefore the first two PCs (shown in Figure 2.1A) captured less than a third of the structure in the datasets, reflecting the complexity inherent in cell lines of 58 tumors from nine different organs. When characterizing the trends or co-relationships between more than two high-dimensional datasets, it is useful to probe the contribution of each dataset to overall structure, that is, to what extent each dataset deviates or agrees with what the majority of datasets support. This information is shown by the MCIA pseudo-eigenvalues. Figure 2.1B shows the pseudo-eigenvalues associated with the first two principal components of each dataset. Comparison within microarray data revealed that Affymetrix HGU133 Plus 2.0 accounts for the highest variance on axis 1 and 2, possibly because this platform contains informative features that are poorly detected or absent on other platforms. As expected, the

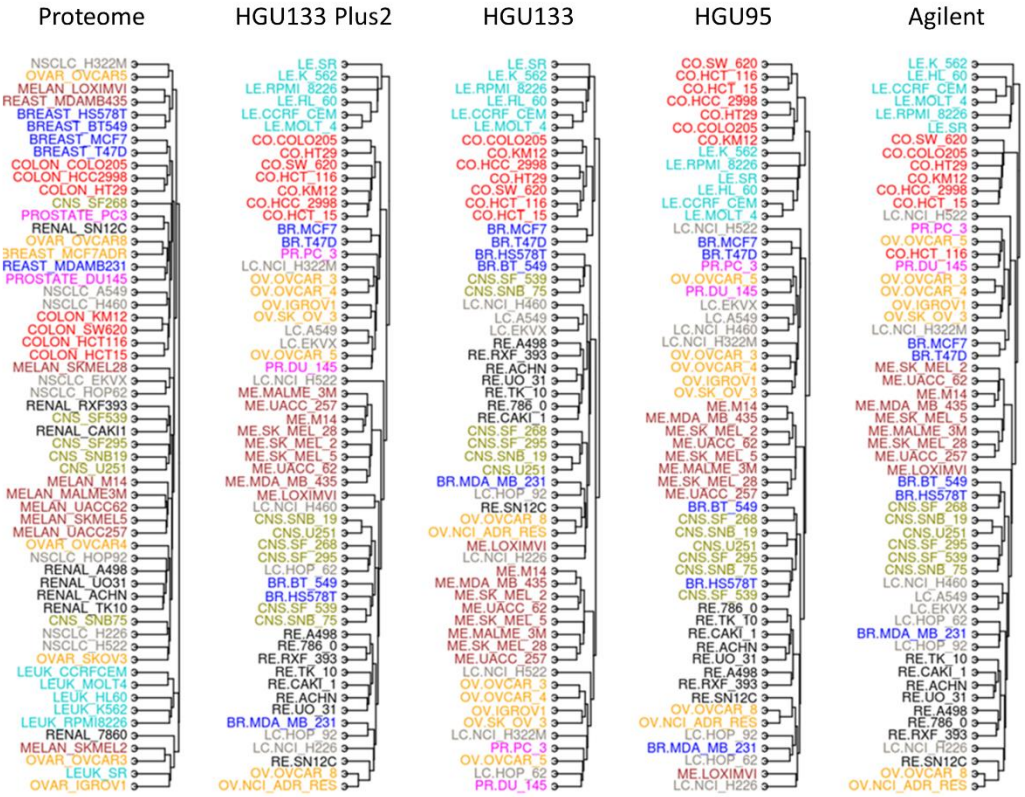

Figure 2.2 *Hierarchical clustering of mRNA and protein expression profiles of NCI-60 dataset. Dendrograms showing the average linkage hierarchical clustering using Euclidean distance. The cell lines are colored as in Figure 2.1 (Abbreviations for cancer types: LE: Leukemia; NSCLC/LC: Non-small cell lung cancer; CO: Colon cancer; BR: Breast cancer; PR: Prostate cancer; OV: Ovarian cancer; ME: Melanoma; CNS: central nervous system (CNS) lymphoma; RE: Renal Cancer)*

results suggested that the similarity within transcriptome datasets is greater than the similarity between transcriptome and proteome datasets, which is also consistent with the results shown by Figure 2.1A. For instance, there is, in particular, a large variation between the protein and transcript expression patterns of melanoma SKMEL2 and ovarian IGROV1. The coordinate of SKMEL2 in the proteome was closer to the origin comparing with in the transcriptomic dataset. According to the genes and proteins expressed in the corresponding direction, the divergence of SKMEL2 cell line could result from the lack of expression of melanin related genes on protein level. Similarly, the divergence of the ovarian cell line IGROV1 in the proteome dataset may be due to expression of less epithelial markers that projected on the positive direction of axis 2. This is an attractive feature of MCIA that can be used to highlight lack or presence of co-structure between datasets, thus it allows selection of the strongest features from each datasets for subsequent analysis.

The overall correlations between pairs of such high dimensional datasets were quantified using RV-coefficient, a multivariate generalization of the squared Pearson correlation coefficient [34]. For each pair of datasets, the RV-coefficient is calculated as the total co-inertia (i.e. sum of eigenvalues of the product of two cross product matrices) divided by the square root of the product of the squared total inertia from the individual analysis. It has a range 0 to 1 where a high RV-coefficient indicates a high degree of co-structure. The overall similarity in the structure of all microarray datasets were very high

and resulted in RV coefficient of >0.8 whereas the RV coefficients were on average 0.76 across microarrays and proteomics data (Figure 2.3). To examine the effect of missing values to the overall co-structure between transcriptome and proteome datasets, MCIA was applied to analyze transcriptome data and 524 proteins that were quantified across all NCI-60 cell lines (core proteome). Interestingly, when filtering out missing values, the similarity between proteome and transcriptome data significantly increased (Figure 2.4). Thus, a better correlation between transcriptome and proteome could be expected when the coverage of the proteomic data is increased in the future (i.e. less missing value presented in the proteomic data).

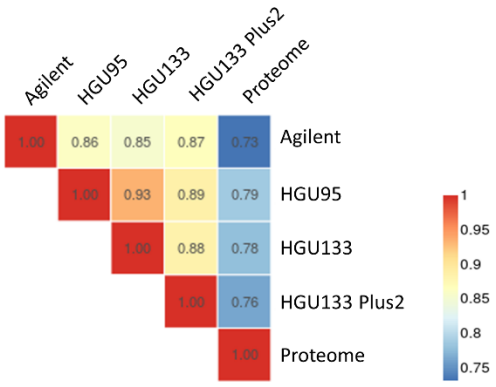

Figure 2.3 *The heatmap shows the RV coefficients between each pair of normalized datasets. The RV coefficients between different microarray datastes range from 0.85 to 0.93. The relative high RV coefficients indicates a high degree of overall similarity in the structure among datasets. The coefficients between proteomic data and microarray is from 0.73 to 0.79, which indicate relative high correlation between transcriptomics and proteomic datasets.*

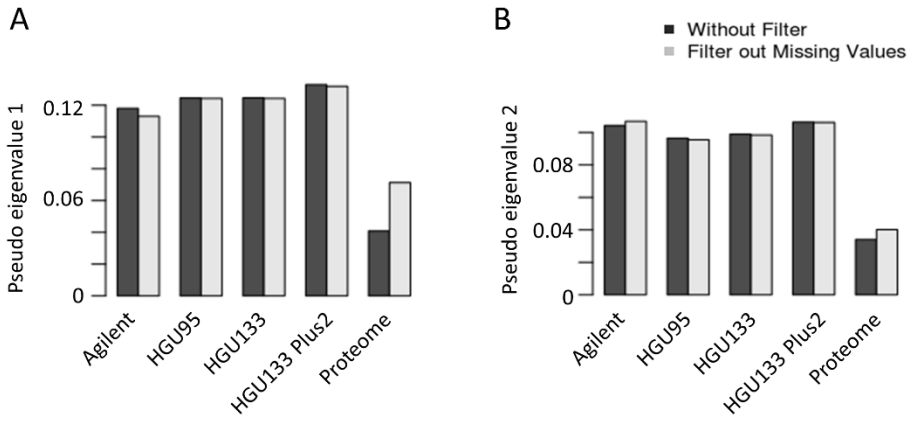

Figure 2.4 *The barplots show the pseudo-eigenvalue space of the NCI60 data, representing the contribution of each dataset to the first principal component (left) and second principal component (right). In both cases filtering out missing values in the proteome data increases the correlated variance between transcriptome and proteome.*

2.3.3 MCIA axes describe biological properties

In contrast to traditional clustering methods (based on the Euclidian distance or other dissimilarly measurement), MCIA projects original data onto a lower dimension space, maximizing the covariance of each dataset with a reference structure. In MCIA plots a gene that is highly expressed in a certain

cell line will be projected in the direction of this cell line and the greater the distance to the center, the stronger the association. In order to identify biomarkers that are highly associated with cancer cell lines of different origins, the feature space, onto where the mRNAs from all microarray platforms are projected, were explored (Figure 2.5). Epithelial-mesenchymal transition (EMT) plays an important role in the malignancy and metastasis of epithelial cells. The first axis (PC1, horizontal axis), which explains the largest variance, separated cells with epithelial or mesenchymal characteristics, suggesting that EMT is an essential mechanism defining different classes of cancers (Figure 2.5A). Epithelial markers, including SLC27A2, CDH1, SPINT1, S100P and EPCAM had high weights on the positive side of PC1 (Figures 2.5B-F). At the opposite end mesenchymal and collagen markers, including GJA1, which is involved in epicardial to mesenchymal cell transition, and TGFβ2 were observed. The second (vertical) axis, PC2, clearly separated melanoma and leukemia from other epithelial cancer types. The strongest determinant of the vertical axis is a set of melanoma-related genes, namely melanoma-associated antigens (MAGEA), melanogenic enzyme (GPNMB) as well as TYR, DCT, TYRP1, MALANA, S100B and BCL2A1. The top 100 genes with greatest weights on PC1 and PC2 were selected from four microarray datasets. Among 1,377 selected genes, 145 presented in three or more datasets (Figure 2.5B-E) as robust markers that could be subjected for further analysis.

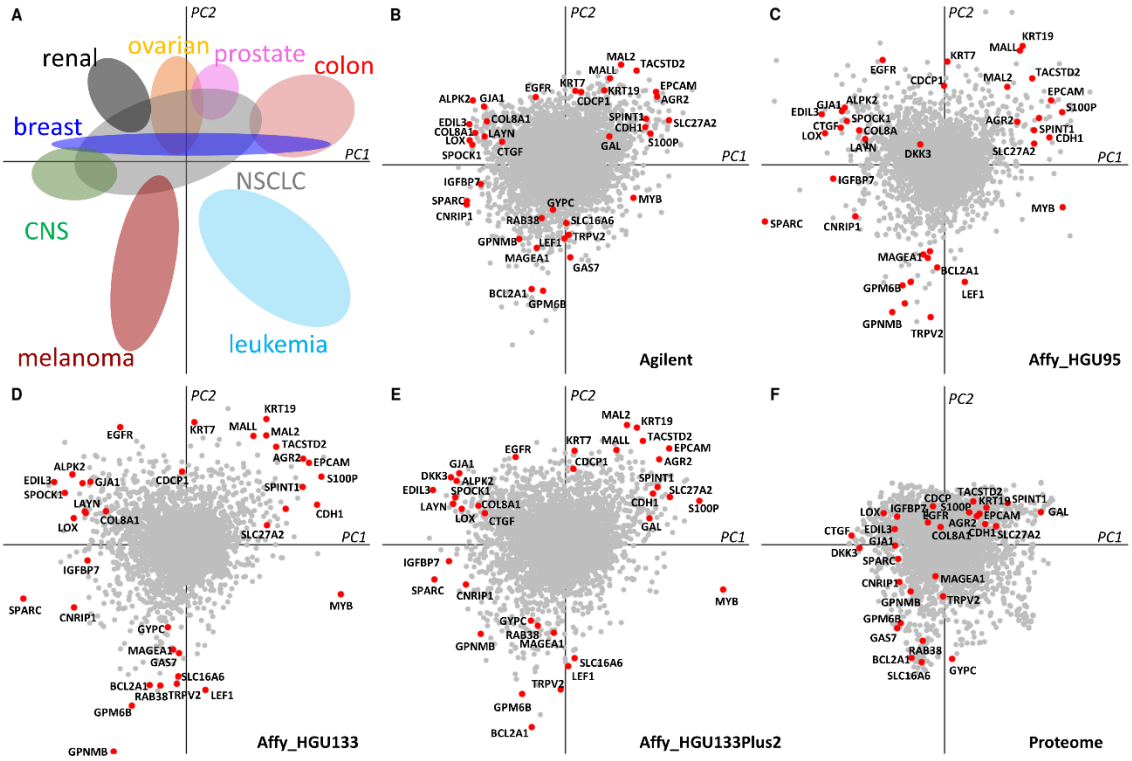

Figure 2.5 *Detecting robust markers defining major trends using MCIA. (A) Shows the projection of the respective cell lines from the NCI-60. Colors represent tissue types as in Figure 2.1. (B-E) represent the coordinates of genes in transcriptomic data. The top 100 genes at the positive or negative ends of MCIA axes (PC1 and PC2) are selected, and those identified in at least three platforms are labeled in red in panel (B-E), indicating the robust markers defining axes. (F) shows proteins from proteomics dataset. The corresponding protein product of genes labeled in (B-E) are in red color and labeled.*

2.3.4 Integration of proteomics and transcriptomics complements the biological information

To further evaluate the biological significance of the features selected by MCIA, the Ingenuity Pathway Analysis (IPA: Ingenuity Pathways Analysis, http://www.ingenuity.com) was used to reveal significant canonical pathways which discriminate different cell lines (Figure 2.6). In MCIA plots, features strongly associated to each tissue type (most extreme) from both transcriptome and proteome datasets can be concatenated and mapped to signaling pathways using IPA. For instance, in contrast to microarray platforms, there is a clear trend of leukemia associated features in the proteomics dataset (Figure 2.5A and F). The most extreme features for leukemia cell lines from all platforms were selected according to their coordinates and subjected to the functional and pathway analysis. While numerous over-expressed genes were co-detected in transcriptome and proteome, a confined number (HCLS1, PECAM and two integrins, ITGAL, ITGB2) was identified exclusively in the proteome dataset, indicating the complementary nature of the information obtained by integrating available data from different platforms. The leukocyte related biological functions including activation of mononuclear leukocyte, mobilization of Ca^{2+} and activation of lymphocyte are strongly associated with these selected features. Enrichment analysis suggested that the most enriched pathways are leukemia extravasation signaling pathway (p = 1.04e-11; Figure 2.6B), which is known to be responsible for leukocyte migration and related to metastasis in leukemia cell lines [35], T cell receptor signaling (p = 5.25e-5) and iCOS-iCOSL signaling in T helper cells (p = 8.32e-5). In order to show the advantage of combining multiple layers of information in pathway analysis, the identical analysis were performed purely based on transcriptome markers. Although leukocyte extravasation signaling was still the most enriched pathway, it did not reach the same level of significance (p = 1.14e-4). In addition, the mRNA-based analysis detected some pathways that are not strongly associated with leukemia (p < 0.01; e.g. hereditary breast cancer signaling and NFAT in Cardiac Hypertrophy). Moreover, pathways that are associated with leukemia and detected in the combined analysis were absent in this case, including NF-kB pathway and PI3K Signaling in B lymphocytes.

The analysis was repeated on melanoma associated features identified by MCIA. The selected features comprised proteins highly expressed in melanoma cell lines and the biological functions or pathways associated with proteins in this group include eumelanin biosynthesis and disorder of pigmentation including the melanocyte development and pigmentation signaling pathway (Figure 2.6C). Melanocytic development and pigmentation is regulated in large part by the bHLH-Lz microphthalmia-associated transcription factor (MITF) and MITF activity is controlled by at least two pathways: MSH and Kit signaling. BCL2A1 is transcriptionally activated by MITF and serves as an anti-apoptosis factor [36]. Interestingly, the upstream regulator of MITF, lEF1, is also consistently identified on the same direction in all transcriptome datasets (Figure 2.5). It is of note that, although all five datasets contributed to the coverage of this pathway, MITF was solely detected in the Affymetrix data. This propensity of MCIA to increase coverage and, hence, power of pathway (and other annotation) analyses result form that it does not require mapping or pre-filtering of features to those present in all datasets. Thus, it allows easy integration of multiple omics levels to identify classes that are relevant in the given biological context.

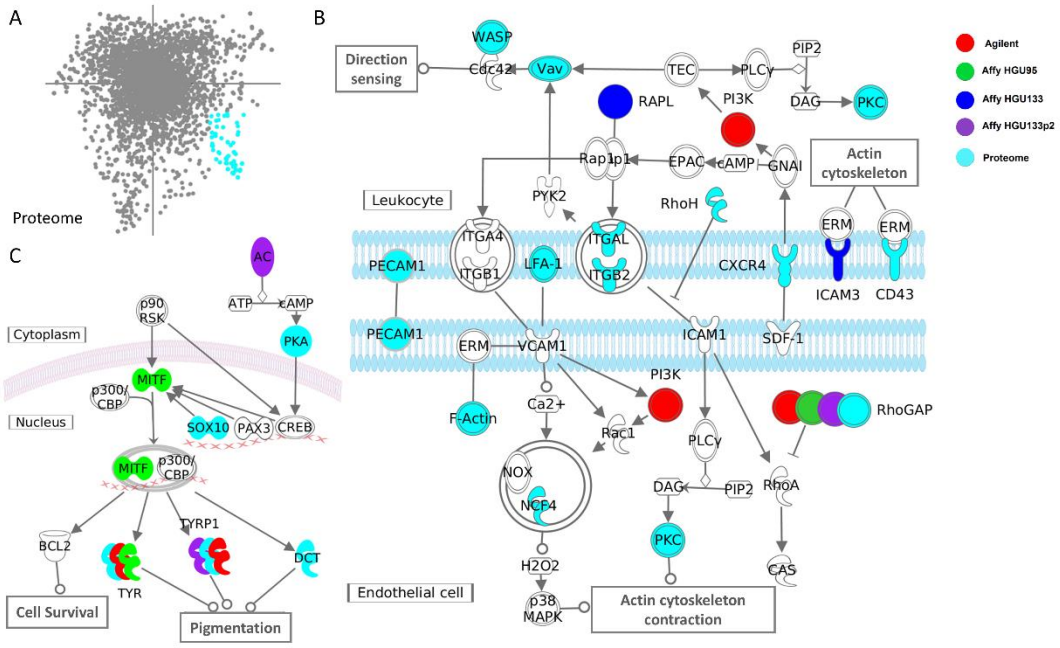

Figure 2.6 *Integrative pathway analysis. (A) Shows the protein coordinates from the proteomics dataset, where proteins from the leukemia tail are highlighted. In contrast to the microarray data, the leukemia genes are clearly represented in the proteome dataset. (B) Represents the leukocyte extravasation signaling pathway significantly enriched by the integration of leukemia features from all platforms. Colors indicate the presence of features in different datasets. (C) Melanocyte development and pigmentation signaling enriched by the melanoma genes. Genes identified in different platforms are in different colors.*

2.3.5 Integrated analysis of microarray and RNA-sequencing ovarian cancer datasets

In the ovarian cancer datasets, MCIA was carried out on the dataset containing 17,814; 12,042; 16,769 and 15,840 genes denoted as Agilent, Affymetrix, RNASeq, RNASeqV2 datasets respectively. In the MCIA space, the first PC (horizontal axis) accounted for 19.6% of the total variance and the second PC (vertical axis) accounted for 10.6% of variance. After filtering genes with more than 15 missing values in RNA-seq data, the four datasets comparably contributed to the total variance (Figures 2.7). RNASeq and RNASeqV2 represented two different pre-processing approaches applied to the same Illumina RNA-sequencing data. MCIA result indicated that normalization and quantification with the RPKM method (RNASeq) outperforms MapSplice and RSEM (RNASeqV2). The variance in RNA sequencing data tended to be sensitive to pre-processing and filtering algorithms which is expected given that methods for processing these data are still emerging. In addition, Affymetrix profiles were generally more variant than Agilent ones, which is shown by the greater pseudo-eigenvalues on PC1-4. Comparison of microarray to RNASeq data, MCIA detected several outlier genes that were highly variant on PC1 and PC2 on RNASeq that are absent on the microarray platforms. These include CDHR4 and HESRG which were suggested highly expressed by the differentiated subtype (Figure 2.7) [37].

2.3.6 MCIA identified ovarian subtypes

In order to compare consistency and discrepancy of ovarian samples obtained from RNA-sequencing and microarray data, the samples were projected onto the MCIA space (Figure 2.7A). The results

revealed that the overall similarity in the structure of four platforms were high. Recent studies have reported four subtypes of ovarian cancer (proliferative, immunoreactive, mesenchymal and differentiated) based on microarray data [20, 38]. The MCIA resulted suggested that the previously reported subtypes of HGS-OvCa can be clearly distinguished by MCIA along the first two axes. Generally, the first axis separated samples with immunoreactive versus proliferative characteristics, whereas the second axis distinguished mesenchymal and differentiated ovarian cancer samples, suggesting that proliferative and immunoreactive subtypes were mutually exclusive, much like the mesenchymal and differentiated subtypes. Moreover, it has been shown that different subtypes have distinct survival characteristics [39]. Mesenchymal features have been implicated in invasiveness and metastasis for numerous carcinomas. As expected, in contrast to the differentiated subtype, the mesenchymal subtype shows a short survival time [39]. Consistent with other studies, the MCIA also identified a large overlap between the four subtypes, indicating that most samples exhibit features of multiple subtype signatures [39]. In order to find whether this classification correlates with clinical factors, I compared the first two PCs with clinical records provided in TCGA data portal and [39]. This comparison revealed that the percentage of stromal cells is positively correlated with PC2 (Pearson correlation test p = 1.79e-3), which is in consensus with the mesenchymal subtype having greater percentage of stromal cells. Age at diagnosis is significantly negatively correlated with PC1 and positively correlated PC2 (Pearson correlation test p = 1.29e-3 and p = 3.56e-4 respectively), with differentiated and immunoreative patients tending to present at younger age. Other clinical factors, such as somatic mutation, drug treatment and tumor stages did not significantly correlate with neither axis.

2.3.7 MCIA suggests robust subtype biomarkers

Both microarray and RNA sequencing data resulted in a similar ordination of samples in the MCIA space. In order to identify which genes contribute significantly to the divergence of samples, the gene expression variables were superimposed onto the same space. Figures 2.7B-E show the projection of transcripts into the feature space. The top 100 genes from the end of each axis were selected. Of the total 9,755 genes common between these four datasets, 27 were projected within the top 100 genes at the ends of the first two axes in all datasets and 82 genes present in at least three platforms, which are identified as robust markers. Of these robust markers, some have been suggested previously [20, 39]. These include many T-cell activation and trafficking genes, such as CXCL9, CD2 and CD3D that were projected onto the positive end of the first axes, indicating that they are highly associated with the immunoreactive subtype. In addition, MCIA suggested new markers associated to the immunoreactive subtype, most of them related to the immune system, such as SH2D1A, RHOH, SAA1, SAA2 and GNLY. This is further corroborated by numerous GO terms significantly associated with genes on this end of the axis [DAVID functional annotation; 40]. For instance, the significantly enriched gene set includes, glycoprotein, chemotaxis, defense and immune response (BH corrected $p < 0.01$). The genes at the opposite end of the MCIA axes included transcriptional factors SOX11, HMGA2, along with several cell cycle promoters, such as BEX1, MAPK4 as well as nerve system development regulators (TBX1, TUBB2B), which characterize the proliferative subtype. Genes that are expressed on

the positive end of axis 2, such as POSTN, CXCL14 and HOXA5, define the mesenchymal cluster. Other potential mesenchymal subtype markers for ovarian cancer include ASPN, homeobox alpha genes as well as collagens. ASPN is a critical regulator of TGF-beta pathway that induces the epithelial mesenchymal transition. Gene set analysis revealed that mesenchymal genes are enriched in GO terms including cell adhesion, skeletal system development, collagen and ECM receptor interaction pathway.

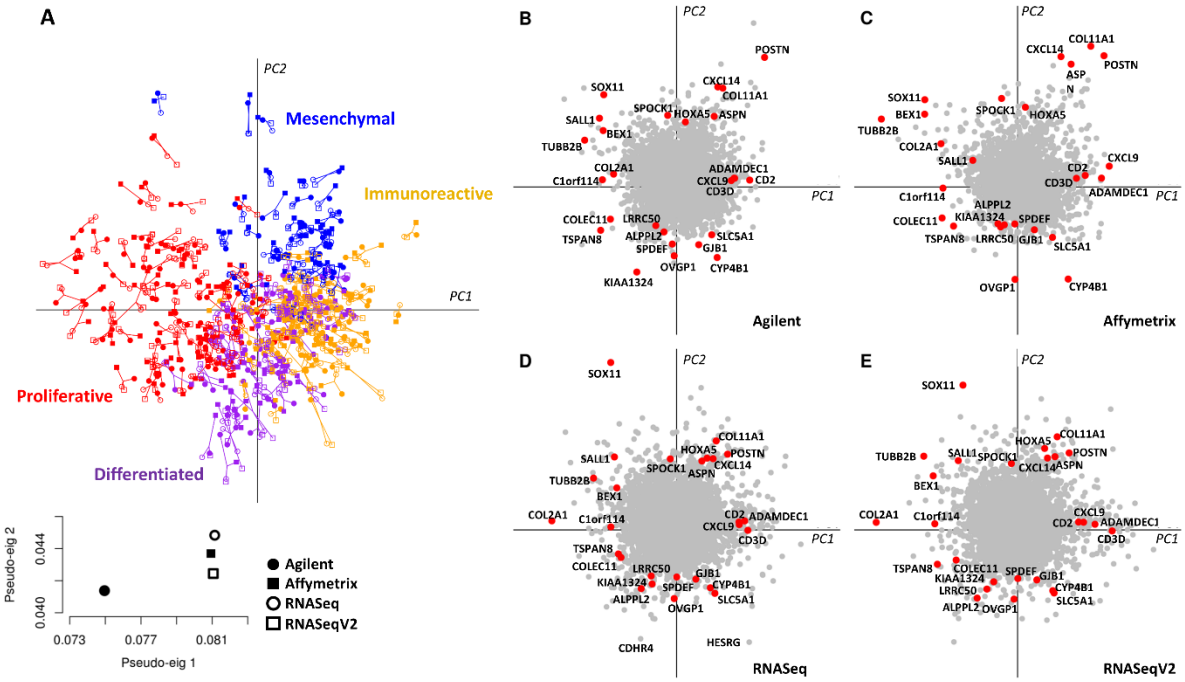

Figure 2.7 *Cross-platform comparison of transcriptional expression profiles of ovarian cancer using MCIA. (A) Visualization of the concordance of patients from multiple platforms. The samples are colored according to four subtypes of patients [53]. PC 1 clearly separates the proliferative and immunoreactive subtypes, whereas the mesenchymal and differentiated subtypes are separated by PC 2. The inset represents the pseudo-eigenvalues of each dataset on the first two PCs. (B-E) Shows the coordinates of genes from each platform. Top consensus genes of all platforms from the end of each axis are colored and labeled.*

The robust markers at the differentiated end include oviductal glycoprotein 1 (OVGP1/MUC9), SPDEF, KIAA1324, GJB1 and ALPPL2, some of which have already been reported as ovarian biomarkers. For instance, OVGP1 has been suggested as a possible serum marker for the detection of low grade ovarian cancer [41]. According to this analysis, this gene was highly expressed in differentiated subtype. Human SPDEF protein plays a significant role in tumorigenesis in multiple cancers, including ovarian cancer and has been reported suppress prostate tumor metastasis. A recent study based on prostate cancer demonstrated that SPDEF suppresses cancer metastasis through down-regulation of matrix metalloproteinase 9 and 13 (MMP9, MMP13), which are required for the invasive phenotype of cells [42]. This analysis implied that SPDEF and matrix metalloproteinase plays a similar role in the development of ovarian cancer. In addition, it has been shown that, in a mouse model, POSTN down-regulates ALPP mRNA [43]. POSTN and ALPPL2 were projected onto the diametral ends of axes 2, which implies that the same mechanism of regulation exists in ovarian cancer and can be exploited to

distinguish subtypes of ovarian. Interestingly, the DAVID gene set analysis of markers for the differentiated phenotype did not reveal as strong GO term enrichments as described for the other subtypes (lowest BH corrected p = 0.0022 vs. 10^{-47} to 10^{-9}) indicating that this subtype exhibits a considerably higher degree of heterogeneity.

2.4 CONCLUSIONS

In the present study, I described multiple co-inertia analysis (MCIA), an exploratory data analysis method that can identify co-relationships between multiple high dimensional datasets. MCIA projects multiple none or only partially overlapping features onto the same dimensional space and provides a simple graphical representation for the efficient identification of concordance between datasets. By transforming multiple sources of data onto the same scale, the most variant features are scaled and concatenated, which significantly facilitates biological interpretation. The integrative analysis of the NCI-60 cell line panel indicated that, although both transcriptome and proteome clustered cell lines according to their lineage, integration of these two different layers of data provide complementary insights suggesting their combined use yields more information than a single one alone. MCIA analysis highly ranked the leukemia extravasation signaling pathway that was overrepresented by features predominantly identified in the proteomics dataset with the most enriched biological functions of activation of mononuclear leukocyte and lymphocyte. The MCIA analysis of high grade serous ovarian cancer revealed four previously described subtypes of ovarian cancer as well as provided novel markers highly associated to different tumor subtypes. An advantage of this method is that MCIA couples multiple matrices at the data level rather than the annotation level, thus it does not require prior mapping or filtering of genes/proteins to common matched set. Therefore, it is not limited by the immaturity of annotations, and there is no subsequent loss of information, considerably increasing data coverage and quality.

ABBREVIATIONS

BH corrected *p*	Benjamini-Hochberg Procedure corrected *p* value
CIA	Co-Inertia Analysis
CNS	Central Nervous System
COA	Correspondence Analysis
CPCA	Consensus Principal Component Analysis
EMT	Epithelial-Mesenchymal Transition
ENCODE	The Encyclopedia of DNA Elements
GCCA	Generalized Canonical Correlation Analysis
gcRMA	GC Robust Multichip Averaging
GeLC-MS/MS	In-Gel Digestion and Liquid Chromatography Tandem Mass Spectrometry
HGS-OvCa	High Grade Serous Ovarian Cancer
IPA	Ingenuity Pathways Analysis
IPI	International Protein Index
MFA	Multiple Factorial Analysis
MCIA	Multiple Co-Inertia Analysis
MS	Mass Spectrometry
NCI	the National Cancer Institute
NSC	Non-Symmetric Correspondence Analysis
NSCLC	Non-Small-Cell Lung Carcinoma
PAGE	Polyacrylamide Gel Electrophoresis
PC	Principal Component
PCA	Principal Component Analysis
RPKM	Reads Per Kilo base per Million
RMA	Robust Multichip Averaging
RSEM	RNA-Seq by Expectation Maximization
TCGA	The Cancer Genome Atlas

REFERENCES

1. Shen K, Tseng G: **Meta-analysis for pathway enrichment analysis when combining multiple genomic studies**. *Bioinformatics (Oxford, England)*, **26**(10):1316-1323.
2. Tyekucheva S, Marchionni L, Karchin R, Parmigiani G: **Integrating diverse genomic data using gene sets**. *Genome biology*, **12**(10).
3. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: **Analysis of matched mRNA measurements from two different microarray technologies**. *Bioinformatics* 2002, **18**(3):405-412.
4. Ebert M, Sharp P: **Roles for microRNAs in conferring robustness to biological processes**. *Cell*, **149**(3):515-524.
5. Fagan As, Culhane An, Higgins D: **A multivariate analysis approach to the integration of proteomic and gene expression data**. *Proteomics* 2007, **7**(13):2162-2171.
6. Raychaudhuri S, Stuart J, Altman R: **Principal components analysis to summarize microarray experiments: application to sporulation time series**. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2000:455-466.
7. Yeung K, Ruzzo W: **Principal component analysis for clustering gene expression data**. *Bioinformatics (Oxford, England)* 2001, **17**(9):763-774.
8. Fellenberg K, Hauser N, Brors B, Neutzner A, Hoheisel J, Vingron M: **Correspondence analysis applied to microarray data**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(19):10781-10786.
9. Fagan A, Culhane AC, Higgins DG: **A multivariate analysis approach to the integration of proteomic and gene expression data**. *Proteomics* 2007, **7**(13):2162-2171.
10. Dray S, Chessel D, Thioulouse J: **co-inertia analysis and the linking of ecological data tables**. 2003, **84**(11):11.
11. Dray S, Chessel D, Thioulouse J: **Co-inertia analysis and the linking of ecological data tables**. *Ecology* 2003.
12. Culhane An, Perriere G, Higgins D: **Cross-platform comparison and visualisation of gene expression data using co-inertia analysis**. *BMC bioinformatics* 2003, **4**:59.
13. Le Cao KA, Martin PG, Robert-Granie C, Besse P: **Sparse canonical methods for biological data integration: application to a cross-platform study**. *BMC bioinformatics* 2009, **10**:34.
14. Shankavaram UT, Reinhold WC, Nishizuka S, Major S, Morita D, Chary KK, Reimers MA, Scherf U, Kahn A, Dolginow D *et al*: **Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study**. *Molecular cancer therapeutics* 2007, **6**(3):820-832.
15. Moghaddas Gholami A, Hahne H, Wu Z, Auer FJ, Meng C, Wilhelm M, Kuster B: **Global proteome analysis of the NCI-60 cell line panel**. *Cell reports* 2013, **4**(3):609-620.
16. Liu H, D'Andrade P, Fulmer-Smentek S, Lorenzi P, Kohn KW, Weinstein JN, Pommier Y, Reinhold WC: **mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities**. *Molecular cancer therapeutics* 2010, **9**(5):1080-1091.
17. Kroonenberg PM, R. L: **Nonsymmetric correspondence analysis: a tool for analysing contingency tables with a dependence structure**. *Multivariate Behavioral Research* 1999, **34**(3):367–396.
18. Chessel D, Hanafi M: **Analysis of the co-inertia of K tables Analyses de la co-inertie de K nuages de points**. *Revue de statistique appliquée* 1996, **44**(2):35-66.
19. Pfister TD, Reinhold WC, Agama K, Gupta S, Khin SA, Kinders RJ, Parchment RE, Tomaszewski JE, Doroshow JH, Pommier Y: **Topoisomerase I levels in the NCI-60 cancer cell line panel determined by validated ELISA and microarray analysis and correlation with indenoisoquinoline sensitivity**. *Molecular cancer therapeutics* 2009, **8**(7):1878-1884.
20. Cancer Genome Atlas Research N: **Integrated genomic analyses of ovarian carcinoma**. *Nature* 2011, **474**(7353):609-615.

21. Shankavaram UT, Varma S, Kane D, Sunshine M, Chary KK, Reinhold WC, Pommier Y, Weinstein JN: **CellMiner: a relational database and query tool for the NCI-60 cancer cell lines**. *BMC genomics* 2009, **10**:277.
22. Shankavaram U, Varma S, Kane D, Sunshine M, Chary K, Reinhold W, Pommier Y, Weinstein J: **CellMiner: a relational database and query tool for the NCI-60 cancer cell lines**. *BMC genomics* 2009, **10**:277.
23. Wu Z, Irizarry RA, Gentleman R, Murillo FM, Spencer F: **A Model Based Background Adjustment for Oligonucleotide Expression Arrays**. *Journal of the American Statistical Association* 2004, **99**.
24. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.
25. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nature methods* 2008, **5**(7):621-628.
26. Li B, Dewey C: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC bioinformatics*, **12**:323.
27. Wang K, Singh D, Zeng Z, Coleman S, Huang Y, Savich G, He X, Mieczkowski P, Grimm S, Perou C *et al*: **MapSplice: accurate mapping of RNA-seq reads for splice junction discovery**. *Nucleic acids research*, **38**(18).
28. Bussey KJ, Chin K, Lababidi S, Reimers M, Reinhold WC, Kuo WL, Gwadry F, Ajay, Kouros-Mehr H, Fridlyand J *et al*: **Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel**. *Molecular cancer therapeutics* 2006, **5**(4):853-867.
29. Roschke AV, Tonon G, Gehlhaus KS, McTyre N, Bussey KJ, Lababidi S, Scudiero DA, Weinstein JN, Kirsch IR: **Karyotypic complexity of the NCI-60 drug-screening panel**. *Cancer research* 2003, **63**(24):8634-8647.
30. Abaan OD, Polley EC, Davis SR, Zhu YJ, Bilke S, Walker RL, Pineda M, Gindin Y, Jiang Y, Reinhold WC *et al*: **The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology**. *Cancer research* 2013, **73**(14):4372-4382.
31. Ikediobi ON, Davies H, Bignell G, Edkins S, Stevens C, O'Meara S, Santarius T, Avis T, Barthorpe S, Brackenbury L *et al*: **Mutation analysis of 24 known cancer genes in the NCI-60 cell line set**. *Molecular cancer therapeutics* 2006, **5**(11):2606-2612.
32. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT *et al*: **A gene expression database for the molecular pharmacology of cancer**. *Nature genetics* 2000, **24**(3):236-244.
33. Stinson SF, Alley MC, Kopp WC, Fiebig HH, Mullendore LA, Pittman AF, Kenney S, Keller J, Boyd MR: **Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen**. *Anticancer research* 1992, **12**(4):1035-1053.
34. Robert P, Escoufier Y: **A unified tool for linear multivariate statistical methods: The RV-coefficient.** *Applied statistics* 1976, **25**(3):8.
35. Springer TA: **Traffic signals on endothelium for lymphocyte recirculation and leukocyte emigration**. *Annual review of physiology* 1995, **57**:827-872.
36. Wu Z, Moghaddas Gholami A, Kuster B: **Systematic identification of the HSP90 candidate regulated proteome**. *Molecular & cellular proteomics : MCP* 2012, **11**(6):M111 016675.
37. Virant-Klun I, Stimpfel M, Cvjeticanin B, Vrtacnik-Bokal E, Skutella T: **Small SSEA-4-positive cells from human ovarian cell cultures: related to embryonic stem cells and germinal lineage?** *Journal of ovarian research* 2013, **6**(1):24.
38. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, Johnson DS, Trivett MK, Etemadmoghadam D, Locandro B *et al*: **Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2008, **14**(16):5198-5208.

39. Verhaak RG, Tamayo P, Yang JY, Hubbard D, Zhang H, Creighton CJ, Fereday S, Lawrence M, Carter SL, Mermel CH *et al*: **Prognostically relevant gene signatures of high-grade serous ovarian carcinoma**. *The Journal of clinical investigation* 2013, **123**(1):517-525.
40. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature protocols* 2009, **4**(1):44-57.
41. Maines-Bandiera S, Woo MM, Borugian M, Molday LL, Hii T, Gilks B, Leung PC, Molday RS, Auersperg N: **Oviductal glycoprotein (OVGP1, MUC9): a differentiation-based mucin present in serum of women with ovarian cancer**. *International journal of gynecological cancer : official journal of the International Gynecological Cancer Society* 2010, **20**(1):16-22.
42. Steffan JJ, Koul S, Meacham RB, Koul HK: **The transcription factor SPDEF suppresses prostate tumor metastasis**. *The Journal of biological chemistry* 2012, **287**(35):29968-29978.
43. Bonnet N, Conway SJ, Ferrari SL: **Regulation of beta catenin signaling and parathyroid hormone anabolic effects in bone by the matricellular protein periostin**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**(37):15048-15053.

CHAPTER III

moCluster: Identifying Joint Patterns across Multiple Omics Datasets

SUMMARY

This chapter describes a novel algorithm, termed moCluster, which discovers joint patterns among multiple omics data. The method first employs a multi-block multivariate analysis to define a set of latent variables representing joint patterns across input datasets, which is further passed to an ordinary clustering algorithm in order to discover joint clusters. Using simulated data, it is shown that moCluster performance is not compromised by issues present in iCluster/iCluster (notably the nondeterministic solution), and operates 500x to 1000x faster. In addition, moCluster was applied to cluster proteomic and transcriptomic data from the NCI-60 cell line panel. The resulting cluster model revealed different phenotypes across cellular subtypes, such as doubling time and drug response. Applying moCluster to methylation, mRNA and protein data from a big study on colorectal cancer identified four molecular subtypes, including one characterized by microsatellite instability and high expression of genes/proteins involved in immunity such as PDL1, a target of multiple drugs currently in development. The other three subtypes have not been discovered before using single datasets, which clearly underscores the complexity of oncogenesis and the importance for holistic, multi-data analysis strategies.

3.1 BACKGROUND

Cancer is a heterogeneous disease. Even when originating from the same tissue, the underlying molecular mechanisms of cancer development can vary dramatically between patients. Therefore, every patient or patient subgroup should motivate a personalized treatment strategy. Broad profiling of genetic mutations, mRNAs, proteins and other biological molecules are valuable sources for the stratification of cancer patients into such molecular subgroups. However, due to the complexity of the underlying biology, a study focusing only on one of the different molecular levels (e.g. genome, transcriptome or proteome) may fail to reveal important factors contributing to oncogenesis. Conversely, an integrated analysis using multiple levels of information may provide a much more detailed picture not readily available from a single dataset [1]. Therefore, an increasing number of projects like the Cancer Genome Atlas (TCGA), the cancer cell line encyclopedia (CCLE) and the Encyclopedia of DNA elements (ENCODE) aim to systematically measure multiple levels of “omics” data from a large number of samples in an attempt to derive a comprehensive understanding of the oncogenesis mechanisms. Recently, researchers used the clustering of cluster algorithm (COCL) and identified 13 integrative subtypes (including copy number variation, DNA methylation, mRNA, miRNA and protein expression data) from thousands of cancer samples originating from 12 different tumor sites [2]. The COCL algorithm is a simple method involving two steps of clustering. First, a clustering algorithm is performed on each individual omics dataset. Next, the clustering results are represented by dummy matrices that contain binary vectors indicating the cluster assignments for each subtype. Last, the dummy matrices for all omics data are concatenated and passed to a clustering method (such as consensus clustering [3]) to identify the joint pattern of multiple omics data. However, this study failed to detect the common pathways or mechanisms shared by cancers from different origins. Hence, there is an increasing need for methods capable of integrative clustering of multiple datasets.

Traditional methods for analyzing two datasets rely on the analysis of a correlation matrix [4]. Shen et al. criticized that the correlation matrix based methods aim for identifying the correlated pattern, which is insufficient for the identification of unique or complementary patterns in each dataset [5]. Therefore, the same authors proposed the "iCluster" algorithm in attempts to address this issue [5]. In the iCluster framework, multiple high-dimensional datasets are represented by a low number of common variables, called "latent variables". The latent variables may account for distinct molecular subtype related biological molecules in each dataset. Therefore, a clustering of the latent variables is able to represent the integrated patterns of multiple datasets [5]. The method was already successfully applied in several studies. For example, an integrative analysis of copy number and gene expression data of 2,000 breast cancer tissues resulted in a novel subtype model with distinct clinical relevance [6]. Recently, an extension of the iCluster algorithm, named iCluster+, was developed to cluster a more diverse range of data types [7]. In iCluster+, different models are used to account for diverse data types, for example, logistic regression for binary variables (mutation data); the multi-logit regression model for multi-category variables (copy number variation) or Poisson regression for count variables (sequencing data) [7]. Despite their widespread application, there are limitations for these methods. For example, they use an iterative expectation-maximization algorithm, which does not

necessarily converge to a deterministic optimal solution. In addition, these algorithms are computational intensive, which is a particular limitation for nondeterministic algorithms since the usual remedy for nondeterministic results is to run the algorithm multiple times and select a consistent output.

In this study, I introduce a new method termed “moCluster”, which is based on a multiple-table multivariate analysis. I evaluated the iCluster/iCluster+ algorithms and compared them to the novel algorithm using simulated datasets. The results demonstrated that moCluster outperforms iCluster/iCluster+ since moCluster defines a subspace that distinguishes subtypes more clearly and always converges to a deterministic solution. At the same time, moCluster is 500x to 1000x faster than the other two algorithms. The method was also applied to the transcriptomic and proteomic dataset of the NCI 60 cell line panel and to methylation, transcriptomics and proteomics data from the TCGA colorectal cancer study. The later analysis resulted in a four-subtype model, one of which was previously known, while the others could only be discovered on the basis of the integrated data. Therefore, this work does not only provide a novel method for the integration of multiple levels of omics data, it also forms the starting point for future research on the newly discovered molecular subtypes.

3.2 METHODS

3.2.1 moCluster algorithm

The first step of the moCluster algorithm is defining Joint Latent Variables (JLV) using the modified consensus PCA (CPCA). CPCA can be calculated using a multiple-block extension of the NIPALS algorithm. I describe the algorithm using notations by [8]. The input of the algorithm is a set of matrices ($\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_k$, ..., $\mathbf{X}_K$). In the algorithm, $\mathbf{t}$ is the joint latent variables; $\mathbf{p}_k$ and $\mathbf{t}_k$ are the feature coefficients vector and the Block Latent Variable (BLV) for matrix *k*, respectively:

Transform, center and scale all matrices
For each latent variable:

1. *Randomly choose start* $\mathbf{t}$
2. *loop until convergence of* $\mathbf{t}$ *:*
   - 2.1 $\boldsymbol{p}_k = \boldsymbol{X}_k^T \cdot \boldsymbol{t} \ / \boldsymbol{t}^T \boldsymbol{t}$
   - 2.2 *normalize* $\mathbf{p}_k$ *to* $\| \mathbf{p}_k \| = 1$
   - 2.3 $\boldsymbol{p}_k = \text{soft}(\boldsymbol{p}_k)$ # soft-thresholding operator, introduce sparsity
   - 2.4 $\mathbf{t}_k = \mathbf{X}_k \cdot \mathbf{p}_k$
   - 2.5 $\mathbf{T} = [\mathbf{t}_1 ... \mathbf{t}_K]$
   - 2.6 $\mathbf{w}_T = \mathbf{T} \cdot \mathbf{t} \ / \mathbf{t}^T \cdot \mathbf{t}$ # $\mathbf{w}_T$ is the linear combination coefficients of BLV to construct the JLV
   - 2.7 *normalize* $\mathbf{w}_T$ *to* $\| \mathbf{w}_T \| = 1$
   - 2.8 $\mathbf{t} = \mathbf{T} \cdot \mathbf{w}_T$

   end
3. $\mathbf{p}_k = \mathbf{X}_k^T \mathbf{t} \ / \mathbf{t}^T \cdot \mathbf{t}$ # final coefficients for features in matrix *k*

4. $\mathbf{X}_k = \mathbf{X}_k - \mathbf{t} \cdot \mathbf{p}_k^T$ # deflation by global score

end

In addition to the centering and scaling of features in each dataset, the integrative analysis of multiple datasets requires a normalization on dataset level because the one with more features (columns) often has more overall variance. In this study, this issue was solved by weighing each data matrix by the reverse of its first eigenvalue, allowing different matrices to contribute comparable variance to the first (few) JLV(s). After proper normalization of datasets, the algorithm calculates JLVs in a sequential manner. First, a random JLV ***t*** is initialized (step 1), which is further iteratively updated until convergence is reached (step 2.1-2.8). In each iteration, the coefficient vector $\boldsymbol{p}_k$ (loading) of each individual matrix $\boldsymbol{X}_k$ is derived by regressing $\boldsymbol{X}_k$ to the initialized ***t*** (step 2.1). Then, the Block Latent Variable (BLV) for dataset *k* is calculated by regressing $\boldsymbol{X}_k$ to the coefficient vector $\boldsymbol{p}_k$ (2.2-2.4). This work modified the CPCA approach for feature selection by introducing a soft-thresholding operator in this process (step 2.3). The soft(·) is defined as "soft(***x****, a*) = sign(***x***)(|***x***|- *a*)$_+$". By tuning parameter *a* one can explicitly control the number of non-zero coefficients in $\boldsymbol{p}_k$. Steps 2.5-2.8 show that JLV ***t*** is updated according to the weights computed from the BLVs $\boldsymbol{t}_k$. Steps 2.1 to 2.8 are iterated until convergence is reached. It has been shown that this algorithm is convergent [9]. The higher order JLVs are based on the residual matrices, which are calculated in step 4. The different ways of calculation of residual matrices underscore the difference between CPCA, GCCA and MCIA: GCCA deflates matrices with respect to BLVs, ($\boldsymbol{t}_k$; i.e. step 4 is $X_k = X_k - \boldsymbol{t}_k \cdot \boldsymbol{p}_k^T$), which do not need to be directly related to the variation of the global variation pattern; MCIA deflates matrices with respect to coefficient vectors ($\boldsymbol{p}_k$; step 4 is $X_k = X_k - X_k \cdot \boldsymbol{p}_k \cdot \boldsymbol{p}_k^T$) [9]. The advantage of deflation matrices with respect to JLV ***t***, as in CPCA, is that this strategy guarantees the orthogonality of JLVs. At the same time, the coefficient vectors $\boldsymbol{p}_k$ can incorporate a sparse operator.

Next, the latent variables can be clustered by conventional clustering algorithms, such as *K*-means or consensus clustering [3]. In this study, I used the hierarchical clustering algorithm because of its simplicity (the Euclidean distance measurement and Ward linkage method [10]). The optimal number of clusters was evaluated by the gap statistic [11].

3.2.2 iCluster

The "iCluster/iCluster+" algorithms were called from the CRAN package "iCluster" (version 2.1.0) and the bioconductor R package "iClusterPlus" (version 1.2.0), respectively. The "iCluster" function in the iCluster package implement the method described in [5, 7].

3.2.3 Data acquisition and processing

NCI-60 data

The NCI-60 drug sensitivity data (DTP NCI60 *Z*-scores) and mRNA data (average *Z*-score from five microarray platforms) were downloaded from CELLMINER (download date: 2014-06-23) [12]. The proteomics data were downloaded from the supplementary table of [13]. The proteome data were

quantified and normalized using the iBAQ method [14]. Because the proteomics data for the melanoma cell line MDAN and the mRNA data for the CNS cell line SNB19 were not available, the two cell lines were excluded from the analysis, resulting in 58 cell lines. The mRNA expression data were filtered based on the standard deviation of a gene across the 58 cell lines − genes with standard deviation greater than 0.8 were retained. In the proteomics data, proteins with total intensity across the panel greater than 10 were retained. Then, the iBAQ values were transformed by $x_i = \log_{10}(iBAQ_i + 1)$.

TCGA Colorectal cancer data

To evaluate the integrative subtypes of colorectal cancers, the mRNA expression (RPKM) matrix and methylation matrix were downloaded from the cBio cancer genomics portal [15]. The spectral count proteomics data and related subtypes, oncogene mutation and clinical information were downloaded from the supplementary information of [16]. Patients with all three types of data (mRNA, methylation, proteomics) were retained in the analysis, resulting in 83 patients. The association between each integrated subtype and other factors was evaluated by two-sided Fisher's exact test.

The methylation level was represented by the beta-value. Beta-values were transformed to M-values as $M_i = \log_2(\frac{Beta_i}{1 - Beta_i})$ [17]. The RPKM values were transformed by $x_i = \log_{10}(RPKM_i + 1)$. Then, all data underwent an unspecific filtering. For each methylation point in the methylation data, the sum of M-values across the 83 patients was calculated. Methylation sites with a sum of M-values lower than -300 were excluded. In addition, due to the distinct methylation status of the X chromosome between genders, all methylation sites located on the X chromosome were also removed. For the mRNA expression data, genes with median absolution deviation (MAD) greater than 0.1 were retained. Due to the presence of missing values in the proteomics data, only proteins with less than 60 missing values across the whole panel were retained. As a result, 11282, 12503 and 5708 sites/genes/proteins were retained in the methylation, mRNA and proteomics datasets, respectively.

3.2.4 Data simulation

In order to simulate data that mimic the true variance level of real data and at the same time have a clearly defined cluster pattern, the following procedure was used:

TCGA bladder cancer transcriptomic and copy number variation (CNV) data were downloaded using TCGA assembler (Date: 26/09/2014). The transcriptome data (RNAseqV2) were quantified by the MapSplice and RSEM approach and were subsequently logarithm transformed. The gene level CNV is estimated by the mean copy number of the genomic region corresponding to a gene (retrieved by TCGA assembler directly [18]). I randomly selected 3000 genes and 240 matched patients in the two datasets, referred to as $\mathbf{X}_{cnv}$ and $\mathbf{X}_{mrna}$, where rows are patients and columns are genes. Then, the two matrices underwent singular value decomposition (SVD)

$$\mathbf{X}_{cnv} = \mathbf{U}_{cnv}\mathbf{D}_{cnv}\mathbf{V}_{cnv}^{T} \text{ and } \mathbf{X}_{mrna} = \mathbf{U}_{mrna}\mathbf{D}_{mrna}\mathbf{V}_{mrna}^{T}$$

U and **V** are orthogonal matrices known as left and right singular matrix, respectively; the columns of them are singular vectors. **D** is the diagonal matrix, where the diagonal elements represent the standard deviation of the corresponding singular vector.

The study planned to clearly define two subtypes in each of the datasets, which should be described by the first left singular vector. In order to reflect its exceptional importance, the diagonal elements of **D** were modified – the first diagonal elements in **D** was kept and the other diagonal elements were replaced by their means, that is, $\hat{d}_i = \begin{cases} d_i & \textbf{if} \quad i = 1 \\ \text{mean}(d_{i \neq 1}) & \textbf{if} \quad i \neq 1 \end{cases}$ for the *i*th diagonal elements d_i.

Therefore, the pattern defined by the first singular vector represents the "true" signal and remaining singular vectors are "noise". The modified diagonal singular value matrices were denoted as $\hat{\mathbf{D}}_{cnv}$ and $\hat{\mathbf{D}}_{mrna}$.

So far, the first singular vectors in $\mathbf{U}_{cnv}$ and $\mathbf{U}_{mrna}$ are exceptional more variant than others but do not have a clear cluster structure. Therefore, $\mathbf{U}_{cnv}$ and $\mathbf{U}_{mrna}$ need to be simulated as well to including clear subtype pattern. To do so, two matrices $\mathbf{X}_{sim1}$ and $\mathbf{X}_{sim2}$ were defined, which have the same dimensions as $\mathbf{X}_{cnv}$ and $\mathbf{X}_{mrna}$, by a simple linear additive model:

$$x_{ij} = \tilde{x}_{ij} + \varepsilon_{ij}$$

where $\varepsilon_{ij} \sim \mathcal{N}(0,1)$ represents random noise; $\tilde{x}_{ij} \sim \mathcal{N}(0, sd_{signal})$ is the pseudo gene expression level, $\tilde{x}_{ij_1} = \tilde{x}_{ij_2}$ if j_1 and j_2 belongs to the same cluster. In $\mathbf{X}_{sim1}$, the two clusters are samples 1-120 and 121-240, whereas in $\mathbf{X}_{sim2}$, the two clusters are samples 1-60/121-180 and 61-120/181-240. Therefore, the two datasets define four subtypes as samples 1-60, 61-120, 121-180 and 181-240. In order to create orthogonal matrices to replace $\mathbf{U}_{cnv}$ and $\mathbf{U}_{mrna}$, I calculated the SVD of $\mathbf{X}_{sim1}$ and $\mathbf{X}_{sim2}$:

$$\mathbf{X}_{sim1} = \mathbf{U}_{sim1}\mathbf{D}_{sim1}\mathbf{V}_{sim1}^{\mathrm{T}} \text{ and } \mathbf{X}_{sim2} = \mathbf{U}_{sim2}\mathbf{D}_{sim2}\mathbf{V}_{sim2}^{\mathrm{T}}$$

Finally, the simulated CNV and mRNA data were generated by

$$\mathbf{X}_{simCNV} = \mathbf{U}_{sim1}\hat{\mathbf{D}}_{cnv}\mathbf{V}_{cnv}^{\mathrm{T}} \text{ and } \mathbf{X}_{simMRNA} = \mathbf{U}_{sim2}\hat{\mathbf{D}}_{mrna}\mathbf{V}_{mrna}^{\mathrm{T}}$$

Therefore, the first left singular vectors in $\mathbf{U}_{sim1}$ and $\mathbf{U}_{sim2}$ capture the two-cluster structure in each dataset; the four-cluster structure defined by $\mathbf{X}_{simCNV}$ and $\mathbf{X}_{simMRNA}$ could be represented by exactly two latent variables. To simulate different signal-to-noise ratios, I defined sd_{signal} = 1, 0.5 and 0.2 for high, medium and low signal-to-noise ratio, respectively.

In addition, to simulate data with sparsity, 1000 genes were randomly selected in the first column of $\mathbf{V}_{cnv}^{\mathrm{T}}$ and $\mathbf{V}_{mrna}^{\mathrm{T}}$ as non-sparse genes. All other values were set to 0. Then, the vectors were rescaled so that the sum of square of all values equals one.

3.3 RESULTS AND DISCUSSION

3.3.1 moCluster approach

The idea behind and the steps taken in the moCluster algorithm can be summarized as follows (Figure 3.1):

1. Using sparse consensus PCA to find latent variables
2. Permutation and elbow test to determine the number of latent variables
3. Clustering of latent variables (using e.g. hierarchical or *K*-means clustering)
4. Selection of the best subtype model

When multiple omics datasets are available for the same set of observations, the data can be represented by a set of matrices ($\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_K$). The columns of the matrices refer to the same set of observations (e.g. samples, cell lines or patients) while the rows refer to different features such as genes, mRNAs or proteins (Figure 3.1A). The core idea of moCluster is similar to iCluster. Both use linear combinations of original features to define a set of joint latent variables, which represent the most important patterns as defined by multiple omics data [5]. This can be expressed as

$$\begin{aligned} \boldsymbol{X}_1 &= \boldsymbol{T}\boldsymbol{P}_1^T + \boldsymbol{E}_1 \\ \boldsymbol{X}_2 &= \boldsymbol{T}\boldsymbol{P}_2^T + \boldsymbol{E}_2 \\ &\vdots \\ \boldsymbol{X}_K &= \boldsymbol{T}\boldsymbol{P}_K^T + \boldsymbol{E}_K \end{aligned} \qquad (1)$$

where $\boldsymbol{T} = |\boldsymbol{t}^1, \ldots, \boldsymbol{t}^s, \ldots, \boldsymbol{t}^S|$ is a matrix that comprises the Joint Latent Variables $\boldsymbol{t}^s$ (JLVs) in columns; ($\boldsymbol{P}_1$, $\boldsymbol{P}_2$, ..., $\boldsymbol{P}_K$) are the matrices of coefficients for features in each datasets (also known as loading matrices). In contrast to iCluster, which uses an expectation-maximization algorithm to optimize a log-likelihood function derived from this model, the moCluster algorithm employs the consensus PCA (CPCA) approach to estimate the latent variables. In this study, the CPCA algorithm was modified to introduce sparsity in feature coefficient vectors (columns of $\boldsymbol{P}_k$; see methods) in order to facilitate the biological interpretation of clustering results. For a single matrix, principal component analysis (PCA) models the high dimensional data with a lower number of variables (principal components; PC). The PCs discovered by PCA are the optimal representation for a single matrix and are widely used in conjunction with clustering or regression methods, such as spectral clustering. CPCA is an extension of PCA for the analysis of multiple matrices. In order to calculate the first latent variable, CPCA finds a set of "suboptimal" latent variables for each individual matrix $\boldsymbol{t}_k^1$ (denoted as Block Latent Variable; $\boldsymbol{t}_1^s, \boldsymbol{t}_2^s, \boldsymbol{t}_3^s$ in Figure 3.1B) by regressing individual matrices to the their respective feature coefficient vectors (see methods). The Block Latent Variables (BLV) are "suboptimal" because they are not necessarily the "best representation" of the matrix itself. Instead, CPCA aims at finding a Joint Latent Variable (JLV; $\boldsymbol{t}^s$ in Figure 3.1B) via linear combination of BLVs so that the summed covariance between BLVs and JLV, i.e. $\sum_{k=1}^{K} \mathrm{cov}^2(\boldsymbol{t}_k^1, \boldsymbol{t}^1)$, is maximized.[9] As a result, the JLVs represent the joint patterns of multiple datasets. To calculate a subsequent (higher order) JLV, CPCA first computes the residual

matrices from each original matrix by removing the variance that accounted for the previously calculated JLV. This procedure is called "deflation". Next, the JLV and feature coefficients are calculated from the residual matrices using the same procedure as before. These processes are repeated until the desired number of JLVs is calculated. Figure 3.1C shows an example of a two dimensional JLV space and the detailed algorithm is described in the methods section.

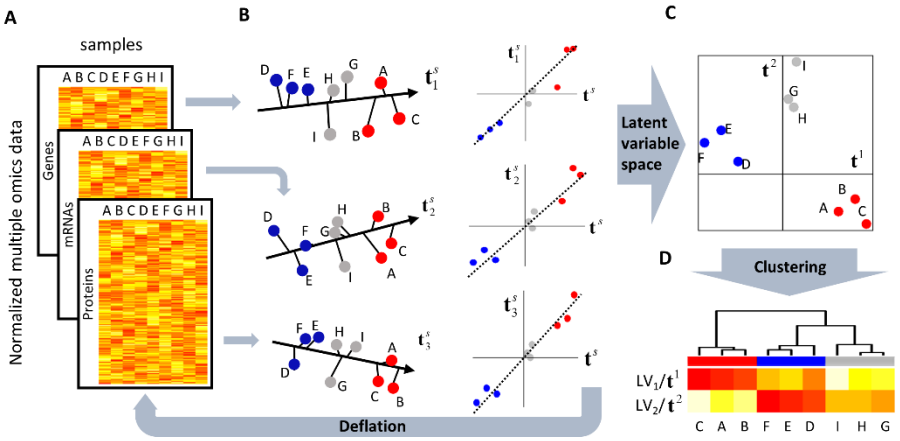

Figure 3.1 *A schematic view of the moCluster algorithm. The input of the CPCA algorithm is a set of matrices that have matched observations, such as genomic, transcriptomics and proteomics data describing the same patient cohort (A). (B) For a latent variable s, CPCA uses a linear combination of original features to define a set of block scores (* $\mathbf{t}_1^s$, $\mathbf{t}_2^s$ *and* $\mathbf{t}_3^s$ *in the figure) and a global score (***t**^{s}*), so as to maximize the summed correlation between each block score and the global score. The global score is termed "joint latent variable" in this study. To derive the higher order solution, the variance accounted for by the global score is removed from the input matrices (deflation step) and the same process reiterated until all the latent variables are defined (C) A scatter plot showing the space defined by two joint latent variables (***t**^{1} *and* **t**^{2}*). (D) A clustering algorithm can be applied on the latent variables to find the joint patterns across datasets.*

Like PCA, the relevance of a JLV can be measured by its pseudo-eigenvalue (explained variance). The pseudo-eigenvalues are monotonically decreasing by definition, but an "elbow" point can be found from a scree plot of pseudo-eigenvalues where the slope of its decreasing goes from steep to flat. I used this method to determine the number of JLVs that should be included in the cluster analysis (see below in the application studies). In addition, a permutation test can be used to evaluate the concordant or divergent structures between datasets. To this end, moCluster permutes samples in each of the datasets and passes the permuted datasets to the CPCA. An empirical confidence interval for each eigenvalue is derived by repeating the permutation for a specified number of times. The permutation analysis provides a further reference for choosing the JLV. Eigenvalues significantly higher than the permutation eigenvalues represent the concordant structures across the datasets, while including extra JLVs enables the detection of divergent structures in multiple datasets. In practice, an elbow test in combination with the permutation test is used to determine the necessary number of JLVs.

With the principle of parsimony in mind, CPCA algorithm was modified by introducing a soft-thresholding operator to ensure sparsity of the features' coefficients ($\mathbf{Q}_k$ in equation 1) for each of the JLVs (see methods). The sparsity of the coefficient matrix leads to the violation of the maximized

covariance criteria (i.e. will explain less variance), but greatly improves the interpretability of the results. In the last step, an ordinary clustering algorithm is applied on the JLVs (Figure 3.1D).

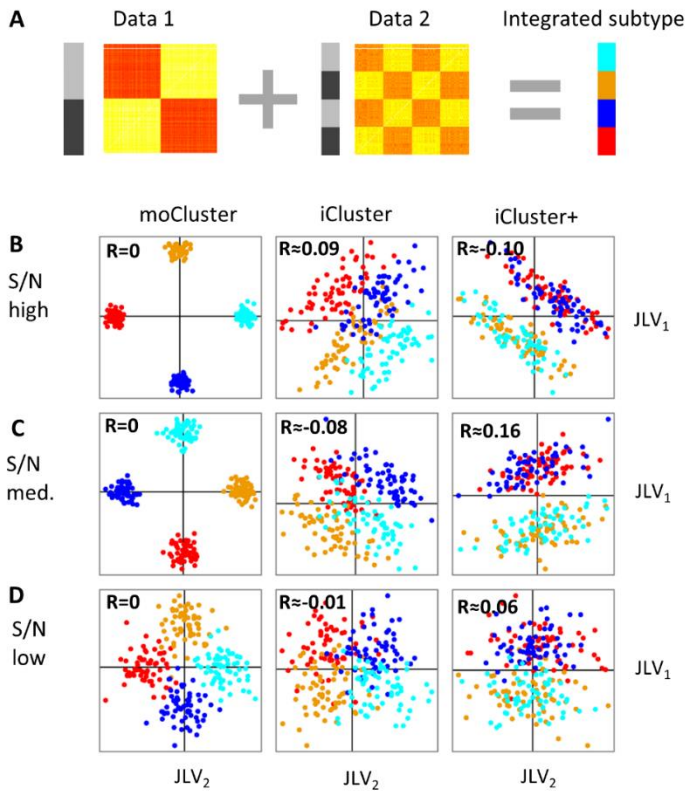

Figure 3.2 *Comparison of moCluster to iCluster and iCluster+ using simulated data. (A) Two datasets were simulated, each of them consisting of two clusters (indicated by the dark and gray bars on the left side of the heatmaps). The combination of them resulted in four clusters as shown in different colors. (B) The space defined by the two latent variables discovered by moCluster, iCluster and iCluster+ using high signal-to-noise ratio data. Different colors indicate samples of different simulated clusters. The horizontal and vertical axes are the JLV1 and JLV2 respectively. (C-D) The same as in (B) but the three algorithms were applied to data with different signal-to-noise ratios (S/N). R indicates the Pearson correlation coefficient, showing that only moCluster returns uncorrelated latent variables.*

A robust estimation of the JLVs is crucial for integrative clustering and this is the essential difference between the moCluster and the iCluster/iCluster+ approaches. Several other generalizations of PCA for multiple-table problems have been proposed and applied to omics data analysis. Two other methods use algorithms that are similar to CPCA. These are the generalized canonical correlation analysis (GCCA) and multiple co-inertia analysis (MCIA) [1, 9]. However, they use different deflation strategies during the calculation of JLVs [5]. The deflation step in the GCCA relies on the block scores. As a result, the JLVs (global scores) derived by GCCA can be un-orthogonal, which is unfavorable since it indicates different JLVs that can be driven by the same or a subset of correlated features (e.g. genes or proteins). In the MCIA approach, the residual matrices are calculated based on the coefficient matrix ($\mathbf{Q}_k$ in equation 1). Subsequently, when sparsity is introduced into the feature coefficients, it dramatically influences the JLVs and may further lead to sparse and correlated JLVs. Therefore, the CPCA approach is particularly suitable for the integrative clustering problem.

3.3.2 Comparison of moCluster to iCluster/iCluster+ using simulated data

In order to have exact control of a putative molecular subtype pattern in molecular profiling data of say cancer patients, I first used simulated data to compare moCluster to iCluster and iCluster+. Two datasets (Figure 3.2A), each consisting of 240 samples, were generated to represent two omics datasets (referred as Data 1 and Data 2). Data 1 was further divided into two sub-clusters: the first sub-cluster contained samples 1-120, while the second cluster consisted of samples 121-240. Data 2 also consists of two sub-clusters: the first cluster is composed of samples 1-60 and 121-180, while the second cluster consists of samples 61-120 and 181-240. Thus a combination of Data 1 and Data 2 results in 4 different integrated subtypes, namely samples 1-60, 61-120, 121-180 and 181-240 (the light blue, orange, blue and red color bars in Figure 3.2A). In order to facilitate comparison and visualization, these datasets were simulated in a way that the four-cluster pattern could be captured within two joint latent variables (JLV; see methods). In order to determine the sensitivity of each method, the data were simulated three times with varying signal-to-noise ratios (see methods). Then the three algorithms (modified CPCA, iCluster and iCluster+) were applied to calculate two JLVs. The corresponding results are shown in figure 3.2B-D, which shows that the moCluster algorithm can always distinguish the four clusters defined by the two datasets. In line with the expectation, as soon as the signal-to-noise ratio becomes lower, the samples are more disperse in their defined clusters. The iCluster algorithm can also roughly separate the four subtypes of simulated samples, however it does not benefit from increasing the signal-to-noise ratio. Interestingly, the iCluster+ algorithm was more likely to discover the two subtypes defined by Data 1. Apart from a better discovery of the joint patterns, the moCluster algorithm also runs 500x to 1000x times faster than the other two algorithms (Table 3.1). The long computation time is particularly problematic for iCluster+. This is because iCluster+ uses a Monte Carlo Newton-Raphson algorithm, which provides nondeterministic results due to the Monte Carlo sampling procedure (Figure 3.3) [7]. To address this problem, the algorithm should be performed multiple times and the best result can then be selected according to a defined criterion. However, the long computation time requirement may impair the application of iCluster+ to the analysis of very large omics datasets.

Table 3.1 - computation time comparison of different algorithms

	time (second)*		
Data Description**	moCluster	iCluster***	iCluster+***
S/N high	0.8	713.4	1042.4
S/N mid	1.0	712.5	1016.6
S/N low	1.0	711.9	992.7

* 1 core; Intel(R) Core(TM) i5 CPU 650 @ 3.20GHz 16 GB RAM.
** Two datasets, each of them have 240 samples and 3000 variables.
*** The number of maximum iterations is 10

In this study, a maximum of 10 iterations was initially allowed for iCluster/iCluster+, but these algorithms do not necessarily converge after 10 iterations. To evaluate the impact of the allowed iteration number, the maximum number of allowed iterations was increased from 10 to 50 for both

methods. The result suggested that iCluster with maximum 50 iterations resulted in highly correlated latent variables; even more problematically, the algorithm converged to a solution which only distinguishes two subtypes (Figure 3.4). In fact, although both methods try to define orthogonal latent variables [5, 7], the results can still correlate with each other (Figure 3.2B). For the iCluster+ algorithm, the increase of the number of iterations did not affect the results at all, while the result from iCluster may be strongly influenced by the number of iterations. In practice, a proper number of iterations may need to be determined so that results represent a good trade-off between processing time and confidence in the results.

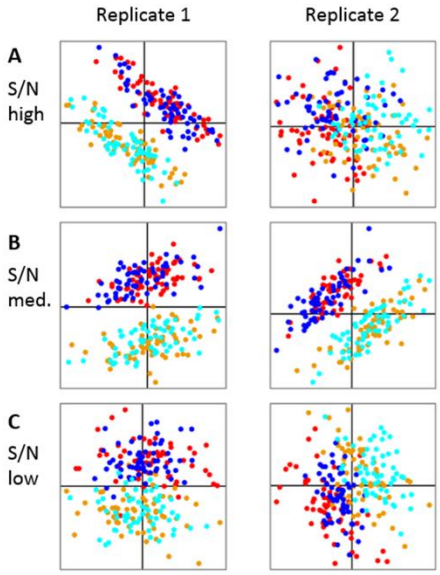

Figure 3.3 *Two replicates of iCluster+ algorithm. Identical datasets and parameters were used but it resulted in different solutions. The datasets are the same with ones used in Figure 3.2. The color code is the same with Figure 3.2. The horizontal and vertical axes are the JLV1 and JLV2 respectively.*

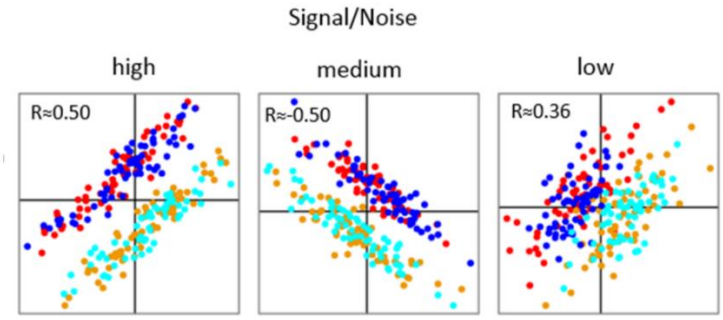

Figure 3.4 *The iCluster with maximum 50 iterations resulted in highly correlated latent variables and only distinguished two clusters. The datasets are the same with ones used in Figure 3.2. The color code is the same with Figure 3.2. The horizontal and vertical axes are the JLV1 and JLV2 respectively. R indicates the Pearson correlation coefficient.*

3.3.3 Application of moCluster to molecular profiling data of NCI-60 cell line panel

Simulated data are valuable for initial benchmark tests of an algorithm, but tend to over-simplify a biological system and may therefore not reflect its true complexity. Hence, I applied moCluster to

cluster the mRNA and proteomics data of the NCI-60 cell line panel [13]. The NCI-60 cell line panel consists of 60 cancer cell lines originating from 9 different tissues (skin, breast, lung, ovary, prostate, blood, central nervous system, colon and kidney) and has been extensively studied, for example in the context of drug sensitivity [12].

The microarray transcriptomic and mass spectrometry-based proteomic data consist of 11,826 and 8,069 genes/proteins, respectively. I had to exclude two of the cell lines because of data availability, resulting in 58 cell lines (see methods). Due to the inherent complexity of cancer, neither the transcriptomic nor the proteomic data necessarily distinguished all the cell lines with respect to their tissue origin. However, the melanoma cell lines and leukemia cell lines were significantly different compared to others due to the expression of genes related to melanogenesis and the immune response [1]. Therefore, the evaluation of moCluster focused on distinguishing these two types of cell lines from the others. As a result, the latent variable space defined by moCluster (using transcriptomic and proteomic data) showed that most of the melanoma and leukemia cell lines clustered according to their tissues of origin and are clearly separated from other cell lines (Figure 3.5). Next, the permutation test and scree plot were used to determine the number of latent variables that should be included in the analysis. The permutation test showed that the top four joint latent variables (JLV) account for significantly correlated structures. The eigenvalue scree plot does not show a clear elbow point and the top four JLVs are significantly higher than the rest (Figure 3.6A). Therefore, this analysis included four JLVs in the subsequent analysis.

In order to derive a model based on sparse feature coefficients, a gradient of non-sparse feature coefficients in each of the datasets were evaluated, i.e. 10%, 20% and 40%. The latent variables with as low as 10% sparse coefficients are well correlated with the ones from non-sparse coefficients. Therefore, I chose 10% non-sparse coefficients (the most parsimonious model) and hypothesized that JLVs with sparse coefficients capture essentially the same biological information like their non-sparse counterparts. Then, hierarchical clustering was employed to cluster the 4 latent variables. In order to determine the optimal number of clusters, the gap statistic was employed [11]. For a given cluster model, the gap statistic calculates the difference (gap) between the within-cluster dispersion of models being evaluated and the model derived from a proper null reference distribution. The within-cluster dispersion is measured as the pooled within-cluster sum of square to the center of each cluster. A high gap statistic indicates that a certain cluster model outperforms a random model generated from the null distribution. In the analysis of the NCI-60 panel, the gap statistic kept increasing until a number of 10 clusters was reached. However, from 7 clusters onwards, the increase of the gap statistic became moderate (Figure 3.6B). Therefore, a 7-cluster model (C1-C7) was selected, resulting in a good compromise between accuracy and parsimony.

The seven resulting subtypes are shown in Figure 3.6C, with most of the leukemia and melanoma cell lines converging to the same cluster. The heatmap of latent variables depicts that leukemia cell lines have high values of the first latent variable whereas melanoma cell lines are showing high values in the second JLV.

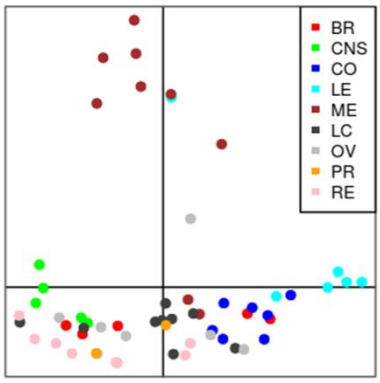

Figure 3.5 *The latent variables space defined by moCluster (NCI60 data; two latent variables). The melanoma cell lines are projected onto the positive end of the second latent variable; whereas the leukemia cell lines are located on the positive end of the first latent variable. (Abbreviations for cancer types: LE: Leukemia; LC: Non-small cell lung cancer; CO: Colon cancer; BR: Breast cancer; PR: Prostate cancer; OV: Ovarian cancer; ME: Melanoma; CNS: central nervous system (CNS) lymphoma; RE: Renal Cancer)*

In order to better understand the underlying biological processes driving the subtypes, genes and proteins with non-zero coefficients were extracted and passed to a gene set over-representation analysis (DAVID functional annotation) [19]. This analysis revealed that genes and proteins positively associated with the first latent variable are related to DNA replication, lymphocyte/T cell activation and DNA repair processes (Table 3.2). The over-representation of lymphocyte/T cell activation is in concordance with leukemia cell lines showing high values of this latent variable. It is noteworthy that the leukemia cell line SR clustered together with the melanoma cell lines due to its protein but not its mRNA expression profile. This might indicate a miss labelling of the cell line in the proteomics data. The result also suggested that other clusters with high values of the first JLV – including C2 (mainly comprising the colon) and C3 (leukemia cell lines) – possess the shortest doubling times (ANOVA p = 0.030; Figure 3.6D). This is in agreement with the over-representation of DNA replication genes and proteins in this latent variable.

Cell adhesion and motions are associated with negative values of latent variables 1 and 2 as well as positive values of JLV 3. These three JLVs mainly drive one big cluster C1 and two small clusters C5 and C6. This is in concordance with the epithelial nature of these subtypes. High values of the second latent variable, a defining characteristic of melanoma cell lines, are strongly associated with genes and proteins related to melanogenesis and pigmentation function. One of the melanoma cell lines, LOXIMVI, known to lack the proteins/genes for melanogenesis was therefore correctly clustered away from the other melanoma cells and was associated with the Non-small cell lung cancer cells, both being epithelial cells.

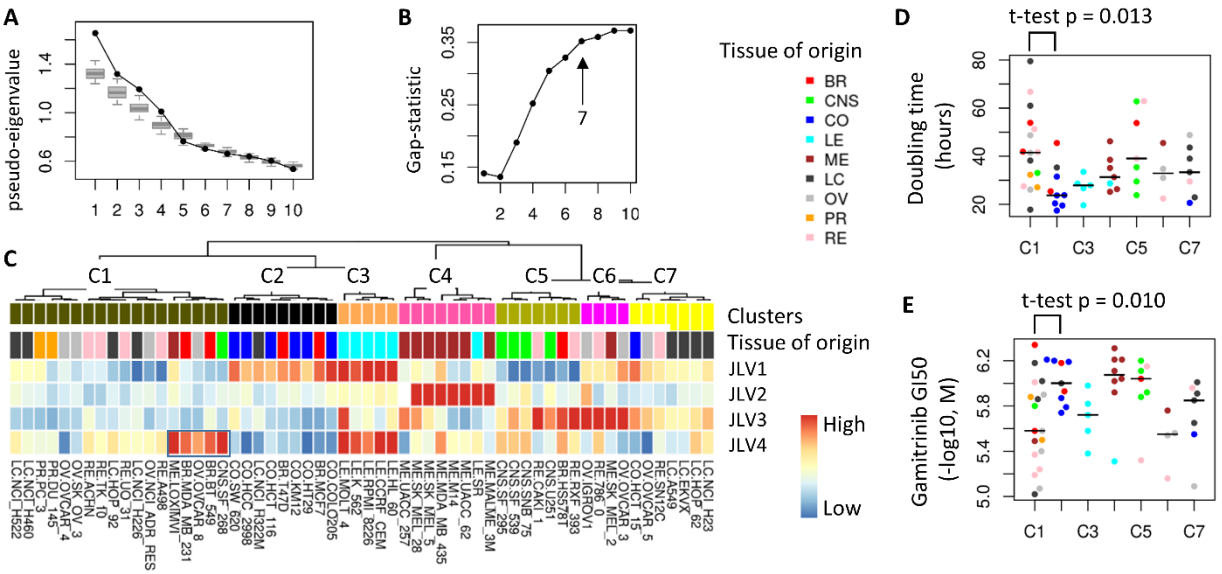

Figure 3.6 *Application of moCluster to NCI-60 mRNA and proteomics data. (A) Permutation test to determine the number of latent variables that should be included in the clustering. The light gray boxplot shows the eigenvalues for 30 permutation test. The first four latent variables represent the concordant structure in the two datasets and are significantly higher than the others. Therefore, this analysis included the first four latent variables. (B) Gap-statistic for the clustering of the top 4 latent variables. From 2 clusters onwards, the gap statistic continuously increased until reaching saturation at approximately 7 clusters. (C) Clustering of the top 4 latent variables using hierarchical clustering. Color-bars indicate the subtypes, tissue of origin and latent variables. The abbreviations for cancer types are the same as in Figure 3.5. (D) The cell doubling time of different clusters. (E) Different sensitivity to the HSP90 inhibitor Gamitrinib across different subtypes of cancer cell lines.*

The mRNAs associated with high values of the fourth JLV (LV4) are also enriched in "T cell activation" and "lymphocyte activation". Apart from the leukemia subtypes, high values of this latent variable additionally defined a subgroup in the C1 subtype. This subgroup includes two Claudin low breast cancer cell lines, BT549 and MDAMB231. It was shown before that Claudin low cancers have a higher expression of genes related to T-cell, B-cell and granulocyte function [20]. Interestingly, three other cell lines, the CNS cell line SF268, the ovarian cell line OVCR 8 and the melanoma cell line LOXIMVI were also found in this this subgroup, which implies that these cancer cell lines may possess characteristics comparable with the Claudin low breast cancer cell lines (JLV4 in Figure 3.6C).

Other functions enriched by selected genes and proteins (gene/proteins with non-zero coefficients) are closely related to GTPase regulation (Table 3.2) and mitochondrial activity, including the electron transport chain, protein oxidative phosphorylation (protein data with negative values of JLV3) and oxidative reduction (proteins on negative end of JLV4). Therefore, this cancer subtype may show a distinct drug sensitivity profile for drugs targeting the mitochondrion [21]. In an attempt to validate this hypothesis, I compared the growth inhibition effect (GI50) of Gamitrinib, a mitochondrial HSP90 ATPase inhibitor, between different subtypes (ANOVA p = 0.010; Figure 3.6E). This analysis suggested that different subtypes indeed show different sensitivity towards this drug, with the subtypes C2, C4 and C5 being more resistant to this compound. In summary, moCluster successfully captures the important joint pattern defined by both proteomic and transcriptomic data. The biological interpretation of the clusters was facilitated by exploring the features associated with latent variables.

Table 3.2 - The enrichment analysis of features positively and negatively associated with each latent variables (NCI-60 data)

	positive			Negative		
latent variable	Term	data*	BH p-value	Term	data	BH p-value
1	DNA replication	R	2.92E-11	regulation of cell motion	R	2.31E-14
	T cell activation	R	5.86E-10	cell adhesion	R	8.42E-13
	DNA replication	P	0.00019	biological adhesion	P	2.46E-17
	lymphocyte activation	P	0.001392	cell adhesion	P	7.95E-17
2	pigmentation	R	9.32E-12	regulation of cell motion	R	4.81E-06
	melanocyte differentiation	R	2.82E-06	cell adhesion	R	1.45E-04
	pigmentation	P	2.99E-07	cell adhesion	P	3.48E-07
	phosphatidylinositol binding	P	6.82E-06	regulation of cell motion	P	2.44E-05
3	skeletal system development	R	2.51E-09	GTPase regulator activity	R	0.000323
	cell adhesion	R	4.70E-08	electron transport chain	P	1.51E-25
	cell adhesion	P	4.31E-06	oxidative phosphorylation	P	2.71E-20
4	T cell activation	R	2.24E-19	GTPase regulator activity	R	5.44E-08
	lymphocyte activation	R	7.62E-18	nucleoside-triphosphatase regulator activity	R	1.01E-07
	cytoskeleton organization	P	3.10E-08	oxidation reduction	P	1.15E-18
	actin cytoskeleton organization	P	5.87E-06	cofactor binding	P	4.93E-10

* P - protein, R - RNA

3.3.4 Application of moCluster to molecular profiling data of colorectal cancer patients

The Cancer Genome Atlas (TCGA) and Clinical Proteomic Tumor Analysis Consortium (CPTAC) recently published multidimensional, genome-scale and proteome-scale analyses on colon and rectum carcinoma [7, 22]. Several different subtype models were proposed by clustering individual omics data from methylation, transcriptomic and proteomic studies: Four subtypes were discovered in the methylation dataset, including two subtypes with elevated methylation, designated "CIMP high" and "CIMP low" and the non-CIMP clusters, Cluster 3 and Cluster 4 [22]; three transcriptomic subtypes were designated “microsatellite instability/CpG island methylater phenotype” (MSI/CIMP), “invasive”, and “chromosomal instability” (CIN) [22]; Zhang et al. reported five proteomic subtypes, designated subtype A – E [16]. However, the proteomics study observed a limited correlation between mRNA and protein levels [16]. Therefore, it is interesting to evaluate whether colorectal cancer subtypes can be better represented with an integrative clustering based on the three types of data. For this purpose,

this study applied the moCluster algorithm to a subset of 83 tumors that had all three types of data available.

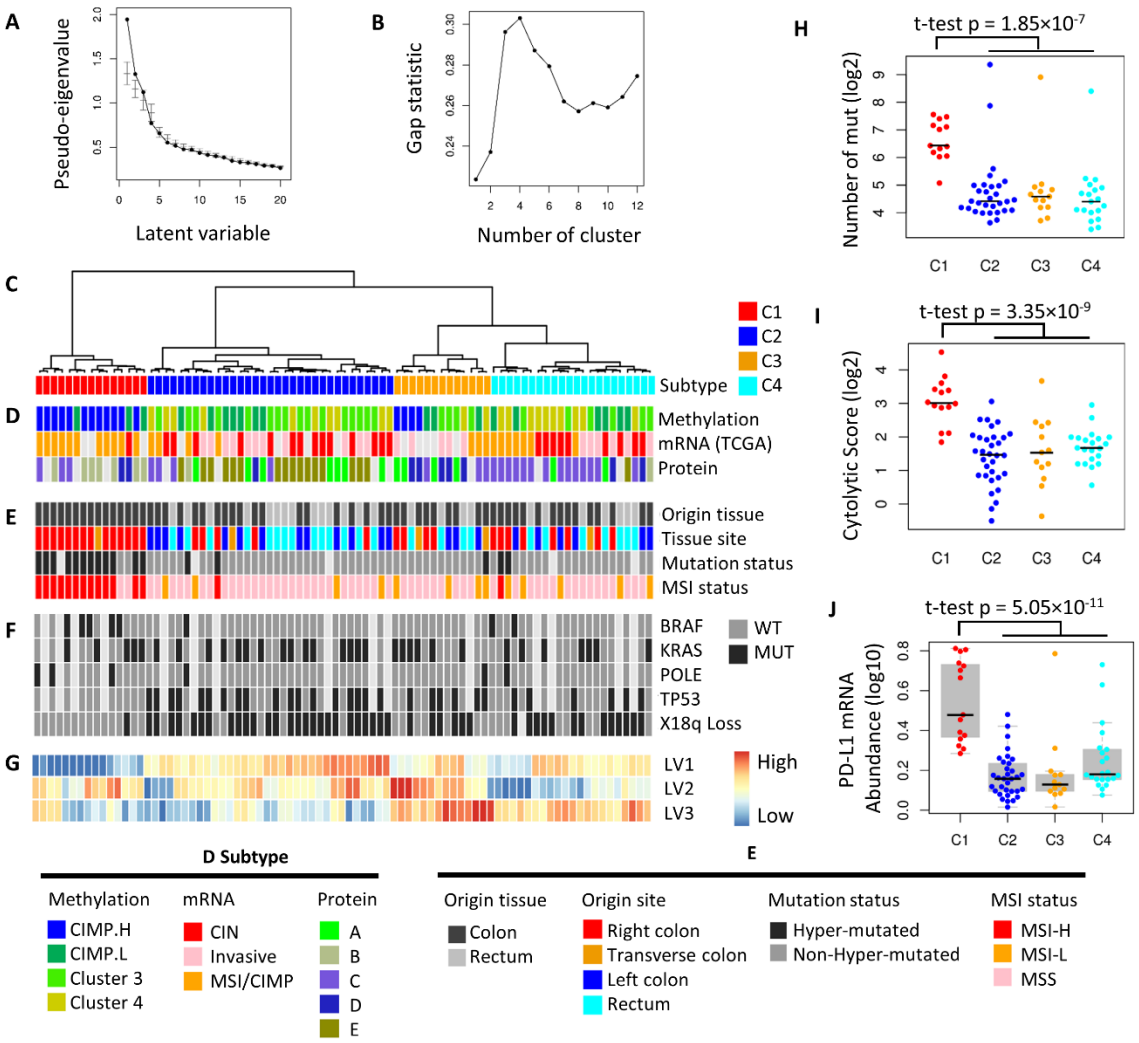

Figure 3.7 *Application of moCluster to the TCGA colorectal cancer dataset. (A) The variance of the top three latent variables are significantly higher than the others. Two of them represent the concordant structure across the three datasets as suggested by the permutation test (the error bars represent the 95% confidence interval). (B) The gap statistic with respect to 1 to 12 subtypes indicates that a 4-subtype model is the optimal choice. (C) Cluster assignment of patients. (D) Comparison of the integrative subtype model with other subtype models derived from individual datasets. (E) Comparison of the integrative subtype model with clinical and genomic information. (F) Mutation patterns of colorectal cancer related genes in different subtypes. (G) Heatmap showing the latent variable expression pattern. (H) The number of mutations of patients in different subtypes. Subtype C1 harbors significantly more mutations than the others. (I) Different distribution of cytolytic scores across subtypes. Subtype C1 is significantly more cytolytic than others. (J) Different PDL1 expression in different subtypes.*

A permutation test suggested that the first two JLVs represent a significant coherent structure among the datasets. In conjunction with the eigenvalue plot, three JLVs were selected for the clustering (Figure 3.7A). 10% non-zero coefficient (the same selection procedure as for the NCI-60 example) in each dataset were retained and there was a good correlation between sparse and non-sparse JLVs. Then, top three latent variables were clustered using hierarchical clustering algorithm (Euclidean distance with Ward method) and four robust integrative subtypes were suggested based on the gap

statistic (Figure 3.7B), denoted as C1 to C4. Each of them consists of 15, 33, 13, 22 cases, respectively (Figure 3.7C).

Table 3.3 - association between subtype and other subtype/clinical factors.

subtype		C1	C2	C3	C4	
		15**	33	13	22	BH P-value*
low grade	FALSE	7	19	3	9	0.61
	TRUE	8	14	10	13	0.70
TP53 mutation	0	11	9	6	12	0.29
	1	0	18	6	6	0.025
BRAF mutation	0	6	26	12	16	0.57
	1	5	1	0	2	0.040
X18Q mutation	0	14	7	7	8	0.049
	1	1	26	6	13	0.032
Proteome subtype	A	0	7	4	2	0.17
	B	7	1	0	1	0.0019
	C	3	4	3	14	0.035
	D	2	1	5	2	0.043
	E	0	16	0	1	6.4E-4
methylation subtype	CIMP.H	12	0	4	0	7.27E-07
	CIMP.L	3	7	3	6	0.98
	Cluster3	0	16	4	6	0.052
	Cluster4	0	10	2	10	0.065
mRNA subtype	CIN	1	14	0	9	0.029
	Invasive	0	11	5	6	0.12
	MSI/CIMP	11	4	3	7	0.037
methylation status	Hyp	12	2	1	2	3.9E-4
	Non-Hyp	2	29	12	18	0.047
MSI status	MSI-H	13	1	0	1	3.95E-06
	MSI-L	0	4	4	5	0.19
	MSS	2	28	9	16	0.072
Site	1 - right colon	14	5	4	8	0.017
	2 - transverse colon	1	1	2	1	0.46
	3 - left colon	0	13	3	6	0.11
	4 - rectum	0	14	4	7	0.087

* Fishers exact test (two sided) was used. Red indicates BH corrected *p* value lower than 0.05.
** Number of patients

Next, this study tested the association between the integrated subtype model with other established subtype models using Fisher's exact test (Figure 3.7D, Table 3.3). Integrated subtype C1 is significantly enriched with microsatellite instable patients and ones from proteomics subtype B; methylation subtype CIMP.H (Figure 3.7E, Table 3.2). In addition, this subtype is significantly associated with the absence of the P53 mutation, absence of chromosome 18q and is moderately enriched in patients with BRAF mutations (Figure 3.7F, Table 3.3). The independent discovery of this subtype in several different studies suggests that this subtype is very different from the other subtypes on several regulatory levels. In this analysis, the dendrogram suggested this subtype is most distinct from the

remaining subtypes (Figure 3.7C), which is specifically characterized by low values in the first JLV (Figure 3.7G).

Noteworthy, the other three integrative subtypes were not discovered in previous studies. Particularly the mRNA subtype "chromosome instability (CIN)", a well-accepted genetic property of colorectal cancer, can be subdivided into the subtypes C2 and C4. This result implies that different mechanisms of oncogenesis may be present in tumors of the mRNA CIN subtype. The C2 subtype also included most of the proteome subtype E, which is characterized by HNF4A amplification and in consequence by higher protein levels of the HNF4alpha protein [16]. Furthermore, there are only weak associations present between the proteomics subtype D and C3, as well as proteome subtype C and the integrative cluster C4 (Table 3.3).

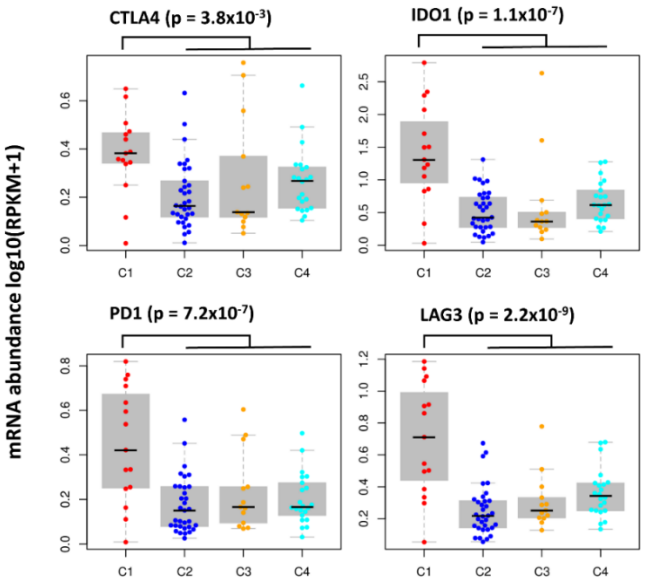

Figure 3.8 *The different mRNA expression levels of four immune checkpoints across subtypes. The four genes are highly expressed in C1 subtype.*

Colorectal cancers are generally divided into two main groups microsatellite unstable (MSI) or microsatellite stable (MSS) but chromosomally unstable tumors [22]. MSI tumors are also characterized as being primarily located to the right colon, harboring the CpG island methylator phenotype (CIMP) and being hyper-mutated [22]. In the integrative model, C1 patients have a higher mutational frequency (t-test $p = 1.85\times10^{-7}$; Figure 3.7H). Therefore, this subtype corresponds to the integrative subtype C1. An enrichment analysis of the associated features (features with negative coefficients in LV1) suggested low values of the first latent variable are associated with immune related genes and proteins (Table 3.4). This implies that this subtype may also represent an immune infiltrated subtype. In order to quantify the degree of immune-activation of samples, I used the cytolytic activation score from Rooney and colleagues [23]. The result confirmed that C1 tumors have the highest degree of immune activation (t-test $p = 3.35\times10^{-9}$; Figure 3.7I). An increasing mutational load is thought to increase the number of cancer associated antigens presented by tumor cells, thus eliciting an enhanced immune response. These cancers are thought to elude immune surveillance and eradication through the expression of PDL1 and in early clinical trials, cancers with high levels of

mutational heterogeneity have responded well to anti-PD-1 therapy [24, 25]. Specifically, Llosa et al. reported that a specific immune microenvironment is associated with MSI colorectal cancer and suggested five immune checkpoint genes, including PD-1, PD-L1, CTLA-4, LAG-3, and IDO as potential drug targets [26]. Of note, all of these checkpoints were significantly (t-test $p = 5.05\times10^{-11}$) elevated in the integrative subtype C1 (MSI and immune activation subtype; Figure 3.7J and3.8). This is particularly interesting since there are several colorectal cancer drugs targeting immune checkpoints in early clinical trials, including anti-PD-1 antibodies Nivolumab (NCT02060188), MEDI0680 (NCT02118337; NCT02013804) and pembrolizumab (NCT01876511); anti-CTLA-4 antibodies ipilimumab (NCT02060188) and tremelimumab (NCT02205333); an anti-CD27 antibody varlilumab (NCT01460134) and a LAG-3 antibody BMS-986016 (NCT01968109). Therefore, the C1 subtype identified by this analysis may represent a subset of colorectal cancers, which might be well susceptible to drugs targeting these immune checkpoint genes. Conversely, the other three subtypes may be less sensitive to this treatment.

Subtype C2-C4 were not described before, hence, the integrative subtype model provides a new basis to study the mechanisms driving these colorectal cancers. In particular, C2 subtype tumors are characterized by negative weight on the third JLV, while subtypes C3 and C4 are distinct on the second JLV. To understand the related functions, enrichment analysis was used to analyze the features associated with the second and third latent variables. The results indicated that, although a limited correlation between mRNA and protein was reported [16], a relatively good correlation was observed between them on the gene set level, namely, the selected non-zero weight mRNAs and proteins seem to be enriched in the same gene sets (Table 3.4). The result suggested that C2 tumors have an elevated ribosome biogenesis activity, the associated gene sets include functions related to "ribosome biogenesis" and "RNA processing". An increased demand for ribosome biogenesis has been associated with tumorigenesis and an increased risk of neoplastic transformation [27]. The C3 subtype has high values of the second JLV; the associated genes include collagens, integrins and cadherins, which are functionally involved in cell adhesion and immune related processes. These molecules play a role in the attachment of malignant cells in their original site, while the downregulation of these genes may support the metastasis of cancer cells to foreign tissues in colon cancer [28]. Therefore, C2 might be a subtype with more advanced neoplastic transformation whereas C3 may represent a subtype with a more epithelial phenotype and less metastatic potential. This analysis provides a hypothesis that may be further tested in the future.

Table 3.4 - the enrichment analysis of variables positively and negatively associated with each latent variables (colorectal data)

	positive			Negative		
latent variable	Term	data*	p-value	Term	data	p-value
1	response to drug	R	1.27E-07	neuron differentiation	M	2.08E-22
	digestion	R	9.54E-05	neuron development	M	8.29E-16
	inorganic anion transport	R	1.15E-04	sequence-specific DNA binding	M	6.88E-13
	translation	P	3.69E-17	immune response	R	2.12E-72
	structural constituent of ribosome	P	5.42E-13	defense response	R	2.96E-44
	ligase activity, forming carbon-carbon bonds	P	4.31E-08	response to wounding	R	1.01E-31
	structural molecule activity	P	2.42E-07	defense response	P	4.31E-36
				immune response	P	7.42E-34
				response to wounding	P	2.90E-31
2	neuron differentiation	M	4.08E-35	cell adhesion	R	1.41E-41
	neuron development	M	8.07E-23	biological adhesion	R	1.70E-41
	neuron projection development	M	1.34E-17	response to wounding	R	1.66E-33
	RNA binding	P	4.36E-09	extracellular matrix structural constituent	P	8.35E-37
	ribosome biogenesis	P	4.40E-08	response to wounding	P	6.75E-30
	ribonucleoprotein complex biogenesis	P	6.47E-08	cell adhesion	P	3.59E-23
3	regionalization	M	2.55E-06	second-messenger-mediated signaling	M	1.47E-06
	pattern specification process	M	9.95E-06	elevation of cytosolic calcium ion concentration	M	2.98E-06
	response to hormone stimulus	R	2.62E-06	negative regulation of transport	R	6.32E-05
	oxidation reduction	R	1.91E-05	regulation of apoptosis	R	1.13E-04
	oxidation reduction	P	1.36E-14	ribosome biogenesis	P	4.17E-13
				ncRNA metabolic process	P	3.06E-12
				RNA processing	P	5.52E-12

* R - mRNA, P - protein, M - methylation.

3.4 CONCLUSIONS

This chapter presents a new method - moCluster - to identify the joint molecular patterns in multiple omics data. The applications of the method showed that the algorithm can generate clustering models that cannot be obtained by single data analysis alone. Harnessing the benefits of multiple-table multivariate analysis, moCluster identifies robust latent variables and runs hundreds of times faster than alternatives such as the iCluster algorithm. At the same time, a sparse operator is incorporated enabling the selection of important features associated with each joint latent variable. This greatly facilitates the biological interpretation of latent variables and therefore aid in the identification of novel biological hypothesis which can subsequently tested by further experiments.

However, moCluster, like other computational methods, should not be used blindly. Our previous work in this area (the MICA approach) showed that the concordance between transcriptomic and proteomic data is increased when filtering out missing values in the proteomics data [1]. Therefore, differences in results between omics data could result from technical artifacts (e.g. missing values), which may lead to further artifacts in joint patterns across datasets. Thus, careful quality control of each individual dataset is required before any integrative analysis. Furthermore, this study mainly considered data that can be modeled using normal distributions (such as log-transformed microarray normalized intensity, RPKM and normalized intensity protein expression data). Appling moCluster to other types of data requires different normalization steps. For example, count data encountered in RNA-Seq (read count) or mass spectrometry (spectrum count) may be converted to a chi-square matrix as in correspondence analysis (CA) or non-symmetric correspondence analysis (NSCA) [1, 29]. It worth noting that because of their descriptive nature, the normalization methods in CA or NSCA could also be used for other data types [30]. Hence, in some cases, such normalization methods may lead to better clustering results [31].

ABBREVIATIONS

CCLE	the Cancer Cell Line Encyclopedia
CNV	Copy Number Variation
COCL	the Clustering of Cluster
ENCODE	the Encyclopedia of DNA elements
iBAQ	intensity Base Absolute Quantification
JLV	Joint Latent Variable
RPKM	Reads Per Kilobase per Million Mapped Reads
TCGA	the Cancer Genome Atlas

REFERENCES

1. Meng C, Kuster B, Culhane AC, Gholami AM: **A multivariate approach to the integration of multi-omics datasets**. *BMC Bioinformatics* 2014, **15**:162.
2. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V *et al*: **Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin**. *Cell* 2014, **158**(4):929-944.
3. Monti S, Tamayo P, Mesirov J, Golub T: **Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data**. *Machine Learning* 2003, **52**(1-2):29.
4. Lee H, Kong SW, Park PJ: **Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes**. *Bioinformatics* 2008, **24**(7):889-896.
5. Shen R, Olshen AB, Ladanyi M: **Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis**. *Bioinformatics* 2009, **25**(22):2906-2912.
6. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y *et al*: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups**. *Nature* 2012, **486**(7403):346-352.
7. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R: **Pattern discovery and cancer gene identification in integrated cancer genomic data**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(11):4245-4250.
8. Westerhuis JA, Kourti T, Macgregor JF: **Analysis of multiblock and hierarchical PCA and PLS moels**. *Journal of Chemometrics* 1998, **12**(5):21.
9. Hassani S, Hanafi M, Qannari EM, Kohler A: **Deflation strategies for multi-block principal component analysis revisited**. *Chemometrics and Intelligent Laboratory Systems* 2013, **120**:15.
10. Murtagh F, Legendre P: **Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?** . *Journal of Classification* 2014, **31**(3):22.
11. Tibshirani R, Walther G, Hastie T: **Estimating the number of data clusters via the Gap statistic**. *Journal of the Royal Statistical Society B* 2001, **63**:13.
12. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, Doroshow J, Pommier Y: **CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set**. *Cancer research* 2012, **72**(14):3499-3511.
13. Moghaddas Gholami A, Hahne H, Wu Z, Auer FJ, Meng C, Wilhelm M, Kuster B: **Global proteome analysis of the NCI-60 cell line panel**. *Cell Rep* 2013, **4**(3):609-620.
14. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: **Global quantification of mammalian gene expression control**. *Nature* 2011, **473**(7347):337-342.
15. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E *et al*: **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data**. *Cancer discovery* 2012, **2**(5):401-404.
16. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S *et al*: **Proteogenomic characterization of human colon and rectal cancer**. *Nature* 2014, **513**(7518):382-387.
17. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM: **Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis**. *BMC bioinformatics* 2010, **11**:587.
18. Zhu Y, Qiu P, Ji Y: **TCGA-assembler: open-source software for retrieving and processing TCGA data**. *Nat Methods* 2014, **11**(6):599-600.
19. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature protocols* 2009, **4**(1):44-57.

20. Sabatier R, Finetti P, Guille A, Adelaide J, Chaffanet M, Viens P, Birnbaum D, Bertucci F: **Claudin-low breast cancers: clinical, pathological, molecular and prognostic characterization**. *Molecular cancer* 2014, **13**:228.
21. Fulda S, Galluzzi L, Kroemer G: **Targeting mitochondria for cancer therapy**. *Nature reviews Drug discovery* 2010, **9**(6):447-464.
22. Cancer Genome Atlas N: **Comprehensive molecular characterization of human colon and rectal cancer**. *Nature* 2012, **487**(7407):330-337.
23. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N: **Molecular and genetic properties of tumors associated with local immune cytolytic activity**. *Cell* 2015, **160**(1-2):48-61.
24. Champiat S, Ferte C, Lebel-Binay S, Eggermont A, Soria JC: **Exomics and immunogenics: Bridging mutational load and immune checkpoints efficacy**. *Oncoimmunology* 2014, **3**(1):e27817.
25. Powles T, Eder JP, Fine GD, Braiteh FS, Loriot Y, Cruz C, Bellmunt J, Burris HA, Petrylak DP, Teng SL *et al*: **MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer**. *Nature* 2014, **515**(7528):558-562.
26. Llosa NJ, Cruise M, Tam A, Wicks EC, Hechenbleikner EM, Taube JM, Blosser RL, Fan H, Wang H, Luber BS *et al*: **The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints**. *Cancer discovery* 2015, **5**(1):43-51.
27. Montanaro L, Trere D, Derenzini M: **Nucleolus, ribosomes, and cancer**. *The American journal of pathology* 2008, **173**(2):301-310.
28. Paschos KA, Canovas D, Bird NC: **The role of cell adhesion molecules in the progression of colorectal cancer and the development of liver metastasis**. *Cell Signal* 2009, **21**(5):665-674.
29. Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M: **Correspondence analysis applied to microarray data**. *Proc Natl Acad Sci U S A* 2001, **98**(19):10781-10786.
30. Greenacre M: **Correspondence analysis in practice**. *Correspondence analysis in practice* 2007.
31. Wouters L, Gohlmann HW, Bijnens L, Kass SU, Molenberghs G, Lewi PJ: **Graphical exploration of gene expression data: a comparative study of three multivariate methods**. *Biometrics* 2003, **59**(4):1131-1139.

CHAPTER IV

moGSA: A Multivariate Approach for Integrative Gene-Set Analysis of Multiple Omics Data

SUMMARY

Gene-set analysis (GSA) greatly facilitated the biological functional interpretation of the omics data. However, there is a lack of GSA methods that annotate multiple-omics data in an integrative manner. This chapter introduces moGSA, a multivariate gene-set analysis based method aims to address this challenge. The application of moGSA to simulated data demonstrated that integrating multiple omics data potentially increases the power to detect subtle changes in a predefined set of genes. Comparing with single sample GSA based methods, integrative analysis by moGSA outperforms existing methods on simulated data. In the analyses of biological data, moGSA was first applied to examine mRNA, protein and phosphorylation profiling of four pluripotent cell lines. The results showed that by using the information in multiple datasets, moGSA not only discovers the gene-set that was identified in single data gene-set analysis, but also identifies the concordance and discrepancy between datasets. Encourage by these results, moGSA was used to analyze larger scale datasets – the copy number variation and mRNA profiling data of 308 bladder cancer from The Cancer Genome Atlas (TCGA). The method, in conjunction with a simple eigenvector based clustering approach, discovered three integrative subtypes of bladder cancer (C1 to C3). The gene-set annotation suggested biological processes associated with subtypes, most of them could be confirmed by other literature. In addition, the gene influential scores provided by moGSA can help selecting potentially interesting markers for further studies.

4.1 INTRODUCTION

Technological innovations have enabled the acquisition of unprecedented amount of multi-scale molecular, genotype and phenotype information. Advances in high-throughput sequencing allow quantification of global DNA variation and RNA expression in tissue or blood samples [1, 2]. Mass spectrometry (MS)-based proteomics has undergone rapid progress in recent years, and systematic MS analyses can now identify and quantify the majority of proteins expressed in a human cell line [3]. Increasingly, studies report comprehensive molecular profiling of the same set of biological samples using multiple different experimental approaches. These data can potentially yield insights into the molecular machinery of biological systems. However, integrating, interpreting and generating biological hypothesis from such complex datasets is a considerable challenge.

Multivariate analysis (MVA) approaches have been used to uncover the latent correlated structure within and between omics datasets [4-6]. In MVA, data are projected onto a lower dimensional space so that trends or relationships between multiple datasets, observations (cases) and features (e.g. genes) can be identified. One attractive feature of these methods is that supplementary information such as gene-set (e.g. Gene Ontology annotations) can be projected onto the MVA to aid interpretation [5, 7, 8]. In addition, MVA methods do not require that gene identifiers in each dataset to be matched to a common intersecting subset of features (genes/proteins), which may lead to information loss particularly when experimental platforms use identifiers that are difficult to be mapped uniquely.

Gene-set analysis (GSA) is widely used in the analysis of genome scale data and is often the first step in the biological interpretation of lists of genes or proteins that are differentially expressed between phenotypically distinct groups [9]. These methods use external biological information to reduce thousands of genes or proteins into short lists of functional related gene-sets (e.g. cellular pathways, subcellular localization, transcription factors or miRNA targets), thus facilitating hypothesis generation. The simplest GSA based methods rely on over-representation analysis and only require a list of genes as input. Hypergeometric tests or Fisher's exact test are often used to identify statistically significant overlap between a shortlist of genes or proteins and a database of gene-sets [10]. Gene-set enrichment analysis (GSEA) and significance analysis of function and expression (SAFE) not only require a list of genes, but also take advantage of quantitative information in omics data [11, 12]. More recently, pathway topology approaches also consider the network structure of biological pathways in over-representation analysis [13]. However, these methods are supervised tests that depend on discriminate predefined groups of experimental clinical, phenotypic or conditional data (e.g. tumor vs. normal cases).

Modern omics studies frequently explore a panel of experimental conditions or tissue samples with multiple phenotypes, illustrated by e.g., The Cancer Genome Atlas (TCGA), ENCyclopedia of DNA Elements (ENCODE) projects [14] and many other studies (see review [15]). Studies frequently aim to discover new molecular subtypes beyond known conditions and thus traditional GSA methods which require known subsets of conditions have limited application in such cases. To address this issue,

several unsupervised, single sample GSA methods have been developed [16-19]. These methods do not require prior availability of phenotypic or clinical data. One of the most popular approaches is the single-sample GSEA (ssGSEA), which ranks genes according to the empirical cumulative distribution function and calculates a single sample-wise gene-set score by comparing the scores of genes that are inside and outside a gene-set [18]. Another recently described method, gene-set variation analysis (GSVA), uses a similar Kolmogorov-Smirnov-like random statistics to assess the enrichment score, but genes are ranked by the statistics calculated through the kernel estimation of a cumulative density function [16]. Each of these unsupervised GSA methods may generate valuable biological information on a single dataset, but none is designed for the analysis of two or more datasets simultaneously.

This chapter presents a novel unsupervised gene-set analysis method for the integrated analysis of multiple datasets of omics information, termed multiple omics GSA (moGSA). Using simulated data, it was suggested that moGSA has greater sensitivity and specificity for detecting altered gene-sets compared to GSA of individual datasets. In addition, moGSA outperforms existing unsupervised GSA methods when applied to simulated data. To further demonstrate the power of moGSA, it was also applied to a small scale and a large scale biological datasets in this study.

4.2 METHODS

4.2.1 moGSA algorithm

Input data and gene-set annotation matrix

The inputs of moGSA are pairs of multiple matrices ($\mathbf{X_K}$, $\mathbf{G_K}$). $\mathbf{X_K}$ is a set of matrices, denoted $\mathbf{X}_1$, … $\mathbf{X}_K$, where K is the total number of quantitative matrices each collected on the same n observation. The matrices $\mathbf{X}_1$, … $\mathbf{X}_K$ will each have a corresponding annotation matrices, $\mathbf{G}_1$, …, $\mathbf{G}_k$. The matrix $\mathbf{X_k}$ is a p_k x n matrix of quantitative omic data which contains p_k rows of features (genes) measured over the same n observations. The gene-set annotation matrix $\mathbf{G_k}$ is a p_k x m binary incidence matrix of gene to gene-set membership associations called the gene-set annotation matrix, where m is the number of gene-sets and p_k is the number of rows of $\mathbf{G_k}$ that match with the original omic data $\mathbf{X_k}$. An element in $\mathbf{G_k}$ has the value 1 if the feature p_i is a member of the gene-set m_j and 0 otherwise. $\mathbf{G_k}$ is constructed according to predefined gene-set information such as the Gene Ontology [20], GeneSigDb [21] or MSigDB [12]. In this paper, “gene” or “feature” is used to indicate rows and “observation” is used to indicate columns without loss of generality.

moGSA step 1 multivariate integration

The first step of the moGSA involves the data integration with a multiple-table multivariate analysis method. In this study, multiple factorial analysis (MFA) [22] was employed because of its simplicity and computational efficiency. MFA can be viewed as a generalization of principal component analysis (PCA) for a multiple-table problem. Here I briefly describe MFA using the nomenclature in [22].

Similar to the integrative clustering problem in Chapter III, one of the most significant issues when applying PCA on a concatenated matrix is that the analysis may be dominated by the matrix or matrices with largest variance or more features. MFA circumvent this problem via dividing all the

matrices by their first eigenvalues, so that each individual matrix contributes almost equal variance to the first principal component. Therefore, the weight of each table is expressed as

$$\alpha_k = \frac{1}{\lambda_k^1} \quad (1)$$

Where λ_k^i is the first eigenvalue of matrix $\mathbf{X}_k^T\mathbf{X}_k$. For convenience, the weights of matrices are stored in a diagonal matrix **A**, whose diagonal elements are

$$\mathrm{diag}\{\ \mathbf{A}\} = [\alpha_1 \mathbf{1}_1^T, \cdots, \alpha_k \mathbf{1}_k^T, \cdots, \alpha_K \mathbf{1}_K^T] \quad (2)$$

The transpose of a matrix is denoted by superscript ^{T}. $\mathbf{1}_k^T$ is a vector of 1 in the length of p_k. Similarly, the weight of each observation is a *n*×*n* diagonal matrix, **M**. In the present study, uniform weighting of samples (i.e. $m_{ii}=1/n$) were used.

A grand matrix **X** (*pxn* where $p = \sum_k p_k$) is constructed through concatenating all individual matrices:

$$\mathbf{X} = [\mathbf{X}_1^T | \ldots | \mathbf{X}_k^T | \ldots | \mathbf{X}_K^T]^T \quad (3)$$

After deriving the matrices' weights (**A**), observation weights (**M**) and the concatenated matrix (**X**), MFA is reduced to a problem of analyzing the triplet (**X**, **A**, **M**). The solution is given by generalized singular value decomposition (GSVD):

$$\mathbf{X}^T = \mathbf{P}\Delta\mathbf{Q}^T \text{ with the constraint that } \mathbf{P}^T\mathbf{M}\mathbf{P} = \mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{I} \quad (4)$$

X is transposed so that the PC or factor scores for observations, **F**, are given by

$$\mathbf{F} = \mathbf{P}\Delta \quad (5)$$

In the PCA framework, the matrix **P** contains the PCs or latent variables, which is also called *sample space* in this chapter. The matrix **Q** is the loading matrix or *gene space*. Because **X** is a concatenation of multiple tables, the gene space matrices $\mathbf{Q}_1$ to $\mathbf{Q}_k$ may also be concatenated or partitioned in the same manner, namely,

$$\mathbf{Q} = [\mathbf{Q}_1^T | \cdots | \mathbf{Q}_k^T | \cdots | \mathbf{Q}_K^T]^T \quad (6)$$

moGSA step 2 project gene-set annotation matrix as supplementary data

Next, moGSA projects the annotation matrices as supplementary tables [23] to generate the gene-set space $\mathbf{W}_k$, which is calculated as a product of the gene annotation matrix $\mathbf{G}_k$ and $\mathbf{Q}_k$

$$\mathbf{W}_k = \mathbf{G}_k^T\mathbf{Q}_k \quad (7)$$

The annotation matrices may be concatenated in the same way as data matrices

$$\mathbf{G} = [\mathbf{G}_1^T | \cdots | \mathbf{G}_k^T | \cdots | \mathbf{G}_K^T]^T \quad (8)$$

Therefore, the overall gene-set space is

$$\mathbf{W} = \mathbf{G}^T\mathbf{Q} = \sum_{k=1}^{K} \mathbf{W}_k \quad (9)$$

moGSA step 3 reconstruction of gene-set-observation matrix

The main output of moGSA is a *gene-set score (GSS)* matrix, denoted by **Y**, whose rows are *m* gene-sets and columns are *n* observations. The GSS matrix is reconstructed with single PCs over single dataset manner so the contribution of single dataset or PC could be evaluated separately. For examples, the GSS matrix for dataset $\mathbf{X}_k$, principal component *t* is calculated as

$$\boldsymbol{Y}_k^t = \boldsymbol{W}_k^t \boldsymbol{F}_k^{t\mathrm{T}} \tag{10}$$

Where $\mathbf{W}_{k,t}$ denotes the gene-set space of *t*th PC in matrix $\mathbf{X}_k$ and $\mathbf{F}_{k,t}$ is the sample space of *t*th PC. The outer product of the two vectors results in a GSS matrix for a specific PC and dataset. Consequently, the overall gene-set score for *t*th PC (i.e. PC-wise decomposed gene scores) is the sum of the gene-set score matrix of the PC across all datasets, that is,

$$\boldsymbol{Y}^t = \sum_k \boldsymbol{Y}_k^t = \sum_{k=1}^{K} \boldsymbol{W}_k^t \boldsymbol{F}_k^{t\mathrm{T}} \tag{11}$$

Similarly, the overall gene-set score matrix by a single dataset (i.e. data-wise decomposed gene scores) is the sum of the matrices by all the axes retained.

$$\boldsymbol{Y}_k = \sum_t \boldsymbol{Y}_k^t = \sum_{t=1}^{T} \boldsymbol{W}_k^t \boldsymbol{F}_k^{t\mathrm{T}} \tag{12}$$

Finally, the complete gene-set score matrix is given by

$$\boldsymbol{Y} = \sum_t \boldsymbol{Y}^t = \sum_k \boldsymbol{Y}_k = \sum_{k=1}^{K} \sum_{t=1}^{T} \boldsymbol{W}_k^t \boldsymbol{F}_k^{t\mathrm{T}} \tag{13}$$

which is the sum of all contributions by individual PC and dataset. In practice, only the PCs with greatest variances (highest eigenvalues) should be retained in the analysis. If all PCs are retained, the result would be similar or the same as naïve matrix multiplication (NMM, see later in the Methods section).

Gene-sets contain different numbers of genes. Therefore in order to compare the dataset contribution across genes, gene-sets should be normalized by the number of their candidate genes. The "normalized gene-sets score" is the gene-set score divided by the number of candidate genes in the gene-set.

Gene influential score

Gene-sets are composed of genes, therefore the contribution of each feature (gene) to the GSS need to be evaluated. Genes with greatest contributions are of interesting from a biological point of view as they could be the "driver" genes for a gene-set. In moGSA, feature contribution, denoted by gene influential score (GIS), is calculated via a leave-one-out procedure. The GSS of gene-set *i*, $\mathbf{Y}_{[i]}$, for all the observations are

$$\mathbf{Y}_{[i]} = \mathbf{G}_{[i]} \mathbf{Q}_{[r]} \boldsymbol{\Delta}_{[r]} \mathbf{P}_{[r]}^T \tag{14}$$

where $\mathbf{G}_{[i]}$ is the gene-set annotation vector for gene-set *i*. $\mathbf{Q}_{[R]}$ and $\mathbf{P}_{[R]}$ are the gene space and observation space with top *r* axes. $\mathbf{\Delta}_{[r]}$ is the diagonal matrix that top *r* singular values. Correspondingly, the gene-set score for *i*th gene-set excluding gene *g* is

$$\hat{\mathbf{Y}}_{[i]}^{-g} = \mathbf{G}_{[i]}^{-g}\mathbf{Q}_{[r]}\mathbf{\Delta}_{[r]}\mathbf{P}_{[r]}^{\mathrm{T}} \tag{15}$$

where $\mathbf{G}_{[i]}^{-g}$ is the gene-set annotation vector for gene-set *i* but without gene *g*. The GIS of gene *g* is measured by

$$E_{[i]}^{g} = -\log_2 \frac{sd(\hat{\mathbf{Y}}_{[i]}^{-g})}{sd(\mathbf{Y}_{[i]})} \tag{16}$$

where $\mathbf{Y}_{[i]}$ is the gene-set scores of gene-set *i* over all observations and $\hat{\mathbf{Y}}_{[i]}^{-g}$ is the gene-set score of gene-set *i* without the gene *g*. $sd(\cdot)$ stands for the function of calculating standard deviation. For convenience, the feature influential score then is rescaled so that gene with maximum influence is 1. Therefore, a positive $E_{[i]}^{g}$ suggests that gene *g* tends to have a positive correlation with gene-set scores of gene-set *i*, whereas a gene with a negative value tends to have a negative correlation.

Statistical inferential aspect

The gene-set score of gene-set *i* observation *j* is

$$\mathbf{Y}_{[i,j]} = \mathbf{G}_{[i]}^{\mathrm{T}}\hat{\mathbf{X}}_{[j]} \tag{17}$$

where $\hat{\mathbf{X}}_{[j]}$ is the *j*th observation in matrix reconstructed with top *r* axes

$$\hat{\mathbf{X}} = \mathbf{P}_{[r]}\mathbf{\Delta}_{[r]}\mathbf{Q}_{[r]}^{\mathrm{T}} \tag{18}$$

Hence, for each observation, a gene-set score could be viewed as the sum of reconstructed expression values of genes in the particular gene-set. If all candidate genes are randomly selected (null hypothesis), the distribution of the means of selected genes is given by central limited theorem (CLT),

$$\bar{x} \sim N(\mu, \sigma_{\bar{x}}) \text{ with } \sigma_{\bar{x}} = c\frac{\sigma}{\sqrt{n}} \tag{20}$$

Where μ is the mean of each observation and $\sigma_{\bar{x}}$ is the sampling standard deviation of means. $c = \sqrt{(N-n)/(N-1)}$ is the finite population correction factor. It is used since each of the genes could be only selected once. Finally, the gene-set score, calculated as the sum of gene in a particular gene-set, follows a normal distribution under null hypothesis, that is,

$$S = \sum_{i=1}^{n} x_i = n\bar{x} \sim N(n\mu, n\sigma_{\bar{x}}) \tag{21}$$

4.2.2 Data simulation

Multiple omics data were simulated as matrix-triplet (K=3), each containing three data matrices of 30 matched observations (n=30). Whilst the number of features in each matrix can be different, in this simulation, each matrix comprised of 1,000 features. The annotation matrix was a binary matrix with dimensions 1,000 features (genes) with 20 non-overlapping "gene-sets" (m=20). Each "gene-set" contains 50 different genes. There were 6 clusters in the 30 observations (n=30). This study hypothesized that observations within the same cluster are driven by a set of differentially expressed (DE) gene-sets. To fulfill this assumption, I and co-workers defined the observations in the same cluster shares the same DE gene-sets. In each of the observations, 5 out of 20 gene-sets were selected as differentially expressed (DE). Within each cluster, the same set of DE gene-sets were randomly selected. Among DE gene-sets, 5, 10 and 25 out of 50 genes were randomly selected as DE genes (DEG) in each of the three simulated data matrix, denoted as DEG_j.

The following linear addictive model adapted from [16] was used, the expression or abundance of gene on *i*th row and *j*th column is simulated as

$$y_{ij} = \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij} \tag{22}$$

where $\alpha_i \sim \mathrm{N}(\mu = 0, \sigma = 1)$ with *i = 1, …, n* is gene specific effect. $\beta_j \sim \mathrm{N}(\mu = 0, \sigma = s)$ is the cluster effect so $\beta_{j_2} = \beta_{j_1}$ if observation j_1 and j_2 belong to the same cluster. The cluster effect factor (categorical variable) is introduced following the hypothesis that observations from the same clusters are driven by some common pathways or "gene-sets" and ensures that observations from the same cluster have a higher within than between cluster correlation. The six correlated clusters in the simulated data are captured by top five PCs. The cluster effect $\beta_j \sim \mathrm{N}(\mu = 0, \sigma = s)$ is sampled from a distribution with a mean of 0 and standard deviation *s.* The standard deviation (*s*) adjusts the correlation between observations in the same cluster, and thus each cluster can have different variance. In this study, *s* equals 0.3, 0.5 and 1.0, which lead to 25%, 30% and 50% of total variance are captured by the top 5 PCs. $\varepsilon \sim \mathrm{N}(\mu = 0, \sigma = 1)$ is the noise factor. γ_{ij} is a factor, if a gene is differentially expressed (DE):

$$\gamma_{ij} \begin{cases} \sim \mathrm{N}(\mu = m, \sigma = 1) & if \quad i \in DEG_j \\ = 0 & otherwise \end{cases} \tag{23}$$

Apart from the retained variance, two other parameters are tuned in the simulation study. First is the number of DEGs in a DE gene-set (5, 10 and 25 out of 50 DEGs). The second parameter is different signal-to-noise ratio, which is tuned through modifying *m* in (23). The candidate *m* are 0.3, 0.5 and 0.8 standing for low, medium and high signal-to-noise ratio. In total, 100 triplets were generated and analyzed by moGSA. NMM, GSVA and ssGSEA, only accept one matrix as input; therefore the three simulated matrices in one triplet-set were concatenated before passing to these methods. The performance was assessed by the area under the ROC curve (AUC) of identifying DE gene-sets.

4.2.3 Preprocessing of Bladder Cancer TCGA data

Normalized gene expression, copy number variation (CNV), microRNA (miRNA) expression data and clinical information of BLCA were downloaded from TCGA (Date: 26/09/2014) using TCGA assembler [24]. Gene expression was profiled with Illumina HiSeq platform. The MapSplice and RSEM algorithm were used for the short read alignment and quantification (Referred as RNASeqV2 in TCGA) [25, 26]. The gene level CNV is estimated by the mean of copy number of genomic region of a gene (retrieved by TCGA assembler directly). Patients that were present in both gene expression and the CNV data were included in the analysis (n=308).

Before applying moGSA, non-specific filtering was performed on both datasets. RNA sequencing data (normalized count + 1) were logarithm transformed (base 10). Genes were filtered to retain those with a total sum greater than 300 and median absolute deviation (MAD) greater than 0.1, which retained 14,692 unique genes (out of 20.531 genes). Then, RNA-seq gene expression data were median centered. For the CNV data, genes with standard deviation greater than the median were retained, resulted in 12,469 unique genes. There were 7,644 genes presented in both datasets.

4.2.4 Gene-set annotation

MSigDB (version 4) [12] categories online pathway database (C2), GO gene-set (C4) and TFT target gene-set (C3) and GO gene-sets (C5, including BP, CC and MF) were used in this study. Gene-set annotations of more than five genes in both of the datasets were retained for the analysis.

4.2.5 Other GSA methods (including NMM)

Single gene-set method, including GSVA and ssGSEA methods were implemented using the R/Bioconductor package GSVA [16]. Default settings were used for these methods.

Naïve gene-set score Y_{naive} was calculated through matrix multiplication (NMM)

$$\mathbf{Y}_{naive} = \mathbf{G}^{\mathrm{T}}\mathbf{X} \tag{24}$$

Therefore, the result of NMM is the same as moGSA when all of the axes are retained.

4.2.6 Clustering latent variable

Consensus clustering was used [27, 28] to cluster the latent variables with Pearson correlation distance and Ward linkage for the inner loop clustering. Eighty percent of patients were used in the re-sampling step of clustering. Average agglomeration clustering was used in the final linkage (linkage for consensus matrix) [27].

4.2.7 Prediction strength to determine the optimal number of subtypes

The “prediction strength” algorithm was used to assess the number of subtypes [29]. In prediction strength method, all samples were assigned a “true” subtype label, based on those predicted by one subtype model. Then, the patients were then divided into “training” and “testing” sets. KNN classifier

(9 nearest neighbors) was used to classify the patients in test set. For each test, the proportion of consensus in assignment between predicted and true labels were computed. The prediction strength was defined by the lowest proportion among all the subtypes and indicates the similarity between the true and predicted labels. The prediction values ranges from 0 to 1, where a value of at least 0.8 suggests a robust subtype classification [29]. Therefore, the prediction with both a prediction strength > 0.8 and the most subtypes can be considered "optimal". In this study, 100 random sampling of training and testing sets were performed and the corresponding prediction strength of each randomized samples was calculated.

4.2.8 Processing of the iPS ES 4-plex data

The transcriptomic (RNA-sequencing), proteomic and phoshphoproteomics data were downloaded from Stem Cell-Omic Repository (Table S1, S2 and S5 from http://scor.chem.wisc.edu/data.php) [30]. The 4-plex data were used in this study, which consists of 17347 genes, 7952 proteins and 10499 sites of phosphorylation. For the transcriptomics data, the expression levels of genes were represented by RPKM values. Three replicates were available and the mean of the three replicates were used. Genes with duplicated symbols and low expression (summed RPKM < 12) were removed. The iTRAQ quantification of protein and phosphorylation sites were performed by TagQuant [31], as describe in [30]. The protein and sites of phosphorylation with low intensity (summed intensity values <20) were removed. In the proteomics data, proteins that are not mapped to an official symbol were removed. Finally, all the data were logarithm transformed ($\log_{10}$(value+1)). After filtering, a few missing values still present and replaced with zero. The enrichment analysis was done on the gene symbol levels, the specific phosphorylation site were not considered.

4.2.9 Subtype calling and comparison with integrative clustering of bladder cancer

This study considered four recently reported bladder cancer subtypes [32-35]. The mRNA markers of each subtype model were selected according to the recent review [36]. Sjödahl et al. firstly defined five major subtypes termed urobasal A (UroA), UroB, genomically unstable (GU), squamous cell carcinoma-like (SCCL) and 'infiltrated'. The Cancer Genome Atlas (TCGA) study defined four expression clusters (I–IV). In other study, the two subtype model, consists of the basal-like and luminal subtypes, were defined by Damrauer et al.; in a study of Choi et al., researcher defined a 'p53-like' luminal subtype apart from basal-like and luminal subtype.

4.2.10 Exam of somatic mutations of patients

GISTIC2.0 [37] data for copy number gains/deletion were downloaded from TCGA firehouse (http://gdac.broadinstitute.org/; download date 2015-03-09). In total, 24776 unique genes were downloaded. The GISTIC code -2, -1, 1 and 2 represent homozygous deletion, heterozygous deletion, low-level gain and high-level amplification respectively. The four types of events were counted for each of the patients. The total number of events were calculated by sum all four types of events.

4.3 RESULTS AND DISCUSSION

4.3.1 Overview of the moGSA algorithm

A typical omics study generates multiple data matrices representing molecular profiles on different levels. In each, the number of features frequently exceeds the number of observations (rows and columns of the matrix, respectively). Data matrices may be RNA sequencing counts of gene expression, measurements of proteins, metabolites, lipids, DNA copy number variations or other biological molecules that can be mapped to gene-sets. moGSA discovers permuted gene-sets that are defined by features from two or more omics data matrices (Figure 4.1). An attractive feature of moGSA is that each omics data matrix can contain different or unmatched features. The method only requires an incidence matrix of gene to gene-set membership associations called "gene-set annotation matrix" for each data matrix. In the gene-set annotation matrix, a value of 1 indicates that a feature (e.g. gene) is a member of a gene-set.

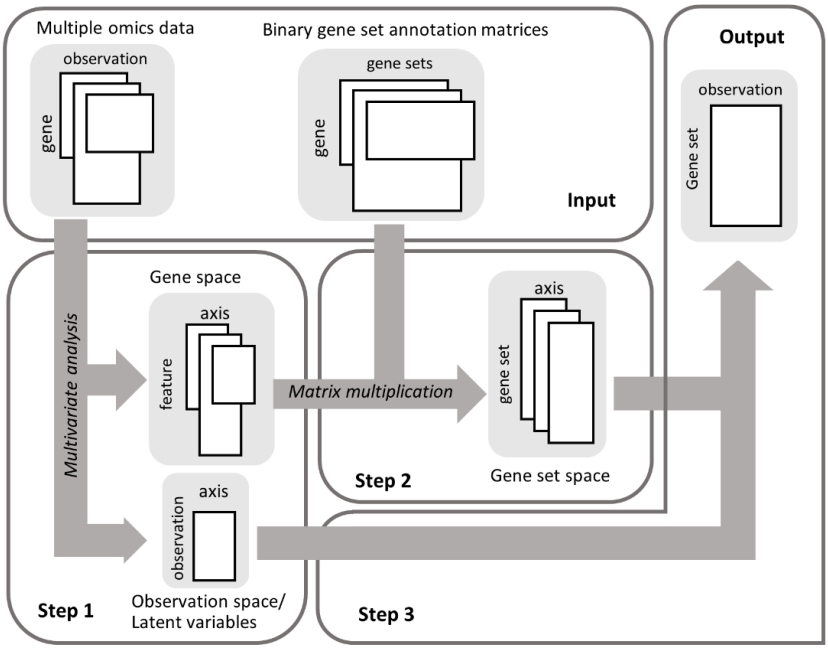

Figure 4.1 *Schematic view of the moGSA algorithm. The algorithm requires two set of matrices as input data, multiple omics data matrices and corresponding gene-set (GS) annotation matrices. In step 1, the multiple matrices are analyzed with a multivariate analysis (MVA) method resulting in the observation space and gene space. Next, the gene-set annotation matrices are projected on the same space, the result is called gene-set space. The last step is to reconstruct gene-set-observation matrix through multiplying the observation and gene-set spaces.*

The algorithm consists of three steps (Figure 4.1). In the first step, input quantitative data matrices are integrated using multiple factor analysis (MFA) [5, 22]. MFA is a multiple-table extensions of principal component analysis (PCA) that discover a small number of latent variables (referred as principal components (PCs) below) capturing the most prominent correlated structure among different datasets [22]. Similar to PCA, the first PC in MFA explains the highest variance for the common structures in multiple data tables. In the next step, the gene-set annotation matrices are projected as additional information into the gene-set space, generating a score for each gene-set in

the same space. The PCs that explain the largest proportion of the variance in the gene-set and observation spaces are selected and, in the final step, moGSA generates a gene-set score (GSS) for each gene-set in every observation. A high absolute value of GSS indicates that the gene-set is up- or down-regulated. Furthermore, the GSS matrix may be decomposed with respect to each dataset or PC so that the contribution of individual dataset or PC to the overall score can be evaluated (see methods).

4.3.2 moGSA outperforms single sample GSA methods

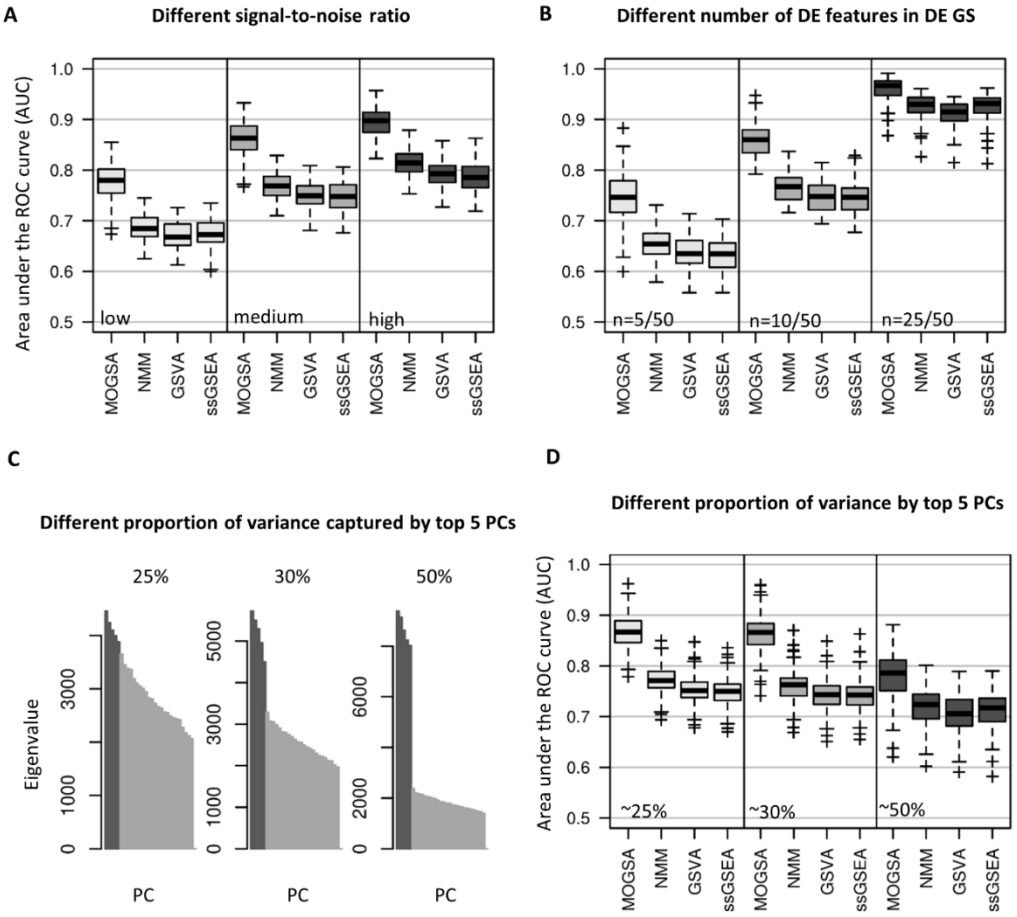

Figure 4.2 *Comparison of moGSA with NMM, GSVA and ssGSEA. The performance of the methods were accessed by their ability to identify differentially expressed gene-sets over 100 simulations (as indicated by the area under the ROC curve; AUC). (A) Comparison of GSA methods using data with different signal-to-noise ratios. (B) Comparison of data with different number of differentially expressed (DE) genes in each of the DE gene-set. From left to right, 5, 10 and 25 of total 50 genes are differentially expressed in each of the three simulated data matrices if a gene-set is defined as DE gene-sets. (C) Scree plots show representative eigenvalues of the datasets used in (D). Different proportions of variance are captured by the top 5 PCs, which are selected and included in the analysis. The darker bars represent the top 5 PCs. (D) AUCs with different proportion of variance are capture by top 5 PCs. From left to right, 25%, 30% and 50% of total variance are captured.*

The moGSA method is compared with naïve matrix multiplication (NMM) and other widely used single sample GSA methods (GSVA and ssGSEA) [16, 18] using simulated datasets. In the data simulation, matrix-triplets (mimics three omics data) contain 1,000 features and 30 observations; their corresponding gene-set annotation matrices (see methods section) consist of 1,000 features with 20

gene-sets. In each of the observations, 5 out of 20 gene-sets were defined as differentially expressed (DE). Within DE gene-sets, 5, 10 and 25 out of 50 genes were randomly selected as DE genes (DEG). Specificity and sensitivity of the methods detecting the DE gene-sets (measured as the area under the receiver operating characteristic curve; AUC) were evaluated. The triplets were analyzed by moGSA directly, however, these matrices were concatenated for NMM, GSVA and ssGSEA because these methods can only accept one matrix as input.

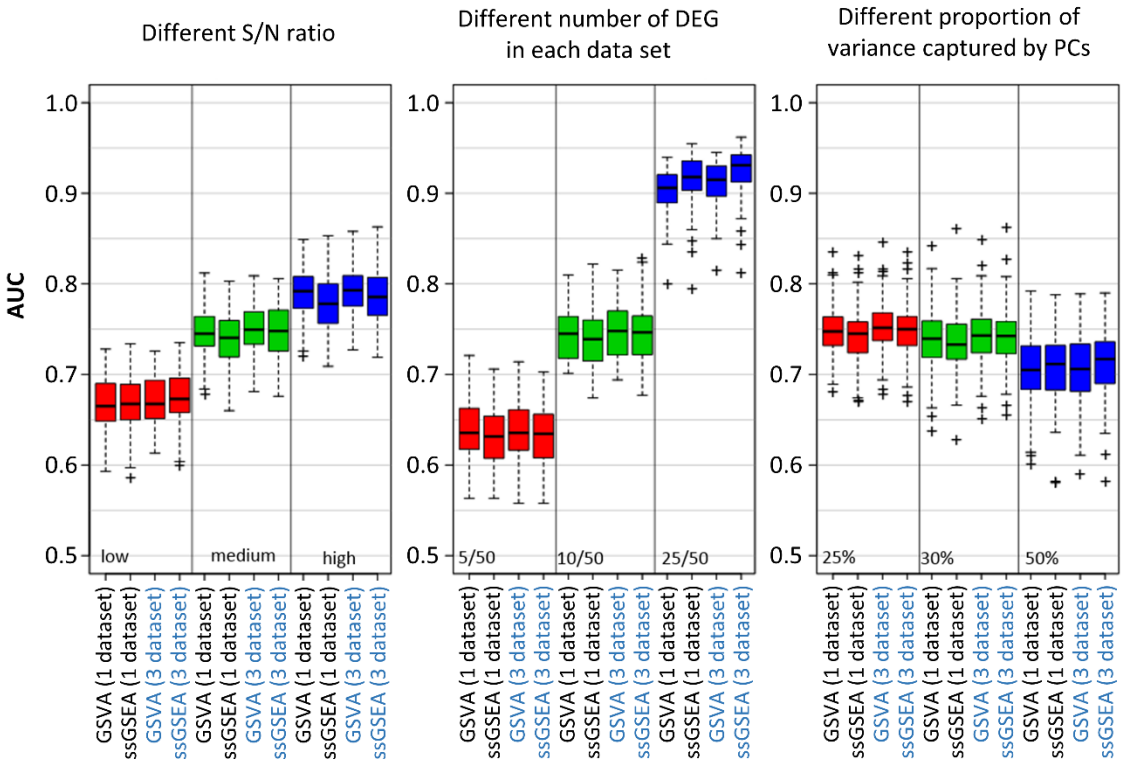

Figure 4.3 *Comparison of the results from analyzing single data (referred as 1 dataset) and concatenated datasets (referred as 3 datasets) by GSVA and ssGSEA. The conditions and evaluation methods are the same with conditions in Figure 4.2. The results show that the simple concatenation of multiple datasets did not improve the performance of GSVA and ssGSEA.*

Simulations were performed to explore 1) the effect of signal-to-noise ratios (low, medium and high), 2) different numbers of DE genes in DE gene-sets (5, 10 and 25 genes) and 3) different levels of variance captured by the axes used in the matrix reconstruction step in moGSA (25%, 30% and 50%, see method section for details). Figure 4.2 shows the performance of each method applied to 100 simulated datasets. As expected, the performance of all methods improved when signal-to-noise ratio or the number of DE genes in DE gene-sets increased (Figure 4.2A-B). moGSA consistently outperformed the other methods and the difference were even more apparent when the signal to noise was low or when there were few DE genes (5 or 10 of 50 genes; Figure 4.2B). Last, I examined the performance of each method when different levels of variance are captured by retained PCs. The datasets were simulated in a way that six clusters could be captured by top five PCs, which were simulated to explain 25%, 30% and 50% of the total variance (Figure 4.2C). Again, moGSA outperformed the other methods and was relatively robust to changes in the variance retained (Figure 4.2D). The performance (AUC) for all the methods decreased when greater variance was captured by

the kept axes. This can be explained by the higher intra-cluster correlation that leads to a lower signal-to-noise ratio (see methods).

In GSVA and ssGSEA, concatenating multiple data matrices did neither improve nor decrease the performance when comparing to the analysis of single datasets, most likely because the signal-to-noise ratio increased proportionally with concatenation (Figure 4.3). Data integration with moGSA may be especially powerful at identifying altered gene-sets in heterogeneous or noisy data because, within moGSA, only a subset of the most informative PCs (n=5) are selected which enables noise-filtering by excluding PCs which may potentially account for noise.

4.3.3 Application of moGSA to IPS and ES cell line data

Next I applied moGSA to a simple dataset consisting of mRNA, protein and phospho-protein profilings of four cell lines − two embryonic stem cell lines (ESC; H1 and H9), one induced pluripotent cell line (iPSC; DF19.7) and one fibroblast cell line (newborn foreskin fibroblast; NFF).

After filtering low quality measurements, the mRNAs, proteins and phosphorylation sites have 10,961; 5,817; and 7,912 unique features (see method). The three datasets were annotated with gene ontology (GO) biological processes (BP) terms. MFA shows that the first PC clearly describes the difference between NFF and other cell lines, which account for most of the variance in the data (Figure 4.4). The second and third PCs represent the differences between iPSC or ESC lines, which may represent important biological information, thus top 3 PCs were retained in the analysis.

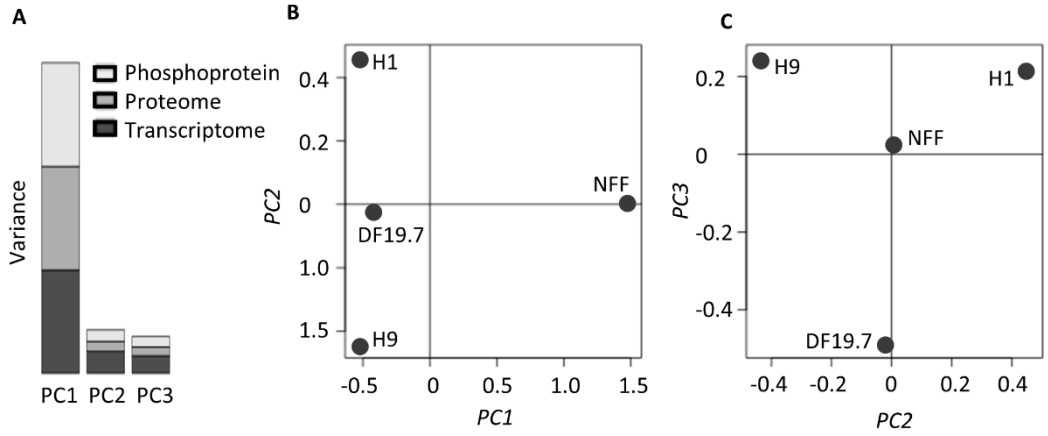

Figure 4.4 *moGSA of the iPS ES 4-plex data. (A) The eigenvalues resulted from MFA. Colors indicates the contribution of each individual dataset. The first axis represent of the variance. (B) The first two PCs from MFA, the first PC accounts for the difference between NFF and pluripotent cell lines. (C) The third PC (The veritical axis)) represents the difference between iPSC and ESC lines.*

The analysis resulted in 228 GO BP terms (out of 825) that are significantly up- or down-regulated in at least one of the cell lines (BH corrected $p < 0.01$). The 228 GO BP terms are generally classified into 19 categories. The gene-set scores of the representative GO BP terms for each category are shown in Figure 4.5A. Integrative analysis suggested that BP terms up-regulated in ES cell lines (H1, H9 and DF19.7) are enriched in functions related to RNA processing, cell cycle related processes and DNA metabolic process, whereas NFF cell line has relative hyper-activation of wounding, cell adhesion and

tissue development processes (Figure 4.5A). These observations agree with previous findings [30]. More importantly, moGSA enabled us to evaluate the relative contribution (either concordant or discrepant) of each dataset to the overall gene-set score. Here, each gene-set score was decomposed with respect to the corresponding dataset (see method). Data-wise decomposition of gene-set scores is shown in Figure 4.3B. The results suggest that the three types of data have concordant contributions to most of the GO terms, including vesicle mediate transport, cell matrix adhesion, cell cycle processes in NFF line; chromosome organization and biogenesis in H9 and NFF cell lines. However, the down-regulation of wound healing in H9 cell line was mainly contributed by the mRNA data, whereas the other two datasets represent a limited role in this function. In addition, the up or down-regulation of "chromosome organization and biogenesis" is mostly contributed by phosphorylation data. Noteworthy, moGSA suggested that "glycoprotein metabolic process" is up-regulated in NFF and mainly resulted in the up-regulation on the protein level. However, the down-regulation of this term in DF19.7 ascribes the low expression of related mRNAs. For the H1 cell lines, the mRNA and proteins suggested strong controversial roles. These results suggest that moGSA in comparison with single data GSA is more sensitive in detecting gene-sets that have subtle but consistent changes in multiple datasets. More importantly, the contribution of individual gene-set can be evaluated by the decomposition of gene-set score with respect to datasets.

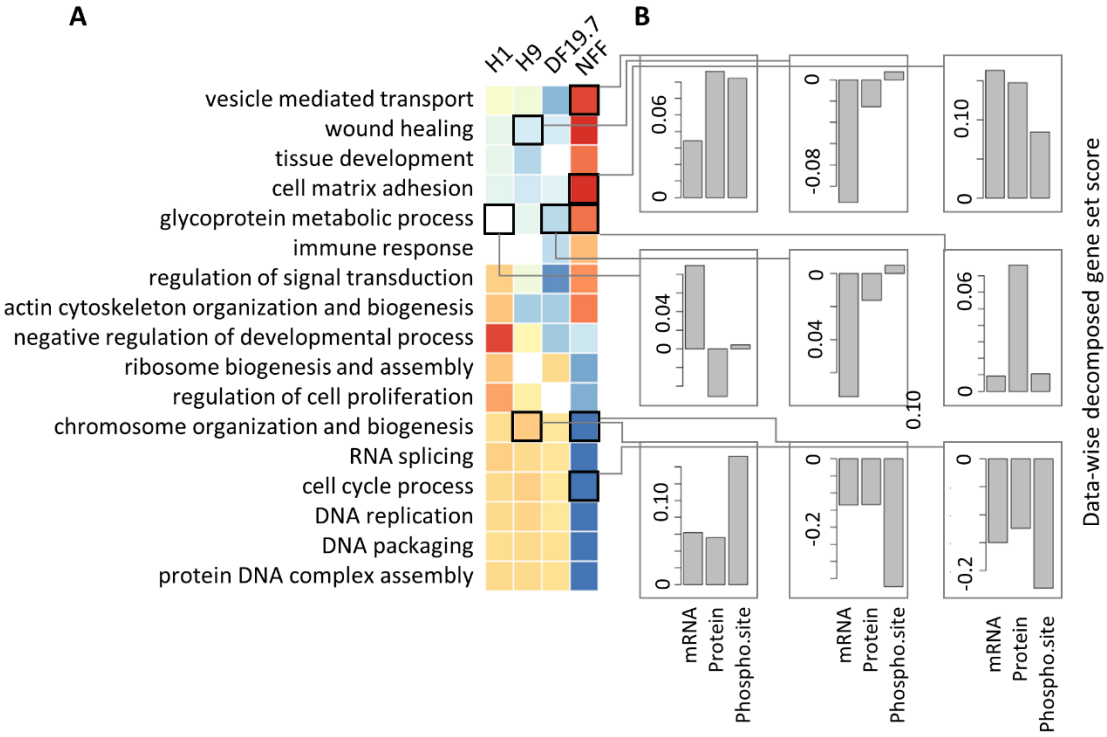

Figure 4.5 *integrative gene-set analysis of iPS ES 4-plex data. (A) A heatmap shows the gene-set score (GSS) for significantly regulated gene-sets in the cell lines, the white colored blocks/cells indicates the change of gene-sets are non-significant. (B) Data-wise decomposition of the GSS for some of the gene-sets. The contribution of each of the data is represent by a bar. The Y-axis is the data-wise decomposed gene-set score.*

4.4.4 Application of moGSA to TCGA Bladder cancer data analysis

Encouraged by these results, moGSA was applied to integrate and annotate copy number variation (CNV) and mRNA data from 308 muscle invasive urothelial bladder cancer (BLCA) patients (obtained as part of the TCGA project). After filtering out low variance genes (see methods), CNV and RNA-seq data contained 12,447 and 14,710 genes respectively, in which 7,644 genes were common to both datasets. Because of the large number of patients, it is hard to analyze the gene-set score of an individual patient. Therefore, the analysis first defined a subtype model of BLCA using a clustering method based on top-k principal components. Figure 4.6A shows the eigenvalues of resulting PCs. The top five PCs captured a quarter of the total variance and were not dominated by either CNV or mRNA (CNV 50.6%, mRNA 49.4%). In addition, these PCs are not correlated with batches (TCGA batch ID), plates, shipping date or tissue source sites. Next, consensus clustering, in conjugation with prediction strength [29] resulted a three-subtype model from the five PCs (Figure 4.6B). The respective three subtypes of bladder cancer consisted of two larger subtypes C1, C2 containing 148 and 103 patients respectively, and a smaller group, C3 of 57 patients. The smaller subtype, C3, was the most robust (highest silhouette width in Figure 4.6C). In general, subtypes identified in previous studies correlated well with the three subtypes identified in the integrative analysis (Table 4.1). Specifically, the integrative subtype C1 harbored most patients of the type III and IV of the TCGA subtypes, the infiltrated and SCCL subtypes of the Sjödahl study and the basal-like subtype identified by Damrauer (BH corrected $p < 0.05$, Table 4.1). Subtypes C2 and C3 were more similar to the Damrauer luminal subtype. In particular, the C3 subtype shows a strong overlap with the UroA subtype of the Sjödahl study and type I of the TCGA subtype model. Finally, subtype C2 contains patients classified into the genomically unstable subtype defined by Sjödahl (Table 4.1). Accordingly, there is a higher mutation rate in the C2 patients (Figure 4.6D).

Table 4.1: The Chi square test of association between integrative subtypes and previously published subtypes.

		Integrative subtypes			
		C1	C2	C3	
		48.1 (148)*	33.4 (103)	18.5 (57)	BH corrected *p* value
Sjodahl subtype	Infiltrated	85.71 (12)	14.29 (2)	0 (0)	$2.3x10^{-2}$
	Genomic Unstable(GU)	25.00 (21)	53.57 (45)	21.43 (18)	$7.8x10^{-4}$
	UroA	23.81 (20)	36.90 (31)	39.29 (33)	$4.3x10^{-5}$
	UroB	54.55 (30)	36.36 (20)	9.09 (5)	$2.2x10^{-1}$
	SCCL	91.55 (65)	7.04 (5)	1.41 (1)	$2.3x10^{-9}$
TCGA subtype	I	11.54 (9)	43.59 (34)	44.87 (35)	$4.7x10^{-9}$
	II	35.40 (40)	46.02 (52)	18.58 (21)	$4.2x10^{-2}$
	III	84.78 (78)	14.13 (13)	1.09 (1)	$4.7x10^{-9}$
	IV	84.00 (21)	16.00 (4)	0 (0)	$2.4x10^{-3}$
Damrauer subtype	Luminal	31.63 (68)	42.79 (92)	25.58 (55)	$1.3x10^{-3}$
	Basal	86.02 (80)	11.83 (11)	2.15 (2)	$2.7x10^{-9}$

* entries are: percentage (number of patients)

To gain further functional insights into the different subtypes of BLCA, moGSA was applied to annotate tumors with Gene Ontology (GO) gene-sets (Figure 4.7A). Gene-sets (n=1,454) were filtered to exclude those with less than 5 genes in a list of the concatenated features of CNV and mRNA data resulting in 1,125 retained gene-sets. The number of significant gene-sets per patient ranged from 183 to 595 and these contained both gene-sets with positive and negative GSS. Only 73 gene-sets had significantly positive or negative GSS in more than 200 patients (out of 308 in total) and these were selected for further analysis. Almost half of the gene-sets with high GSS variance are related to the "immune response" (n=31/73). The remaining 42 gene-sets could broadly be defined by biological processes of "cell cycle" (n=9), "mitochondrion" (n=4), "DNA and chromosome related" g (n=7) and other cancer related gene-sets, including "apoptosis" (n=2), and "G protein coupled receptor" (n=6). Most of these gene-sets have been associated with subtype of bladder cancer. For example, in consistent with findings in this work, the strong up-regulation of immune-associated signatures was observed in basal-like/SCC-like bladder cancer (C1 in the integrative model) [32], and the cell-cycle up-regulation is correlated with "genomically Unstable" (C2 in the integrative model) [38]. The mitochondrial component has been described in bladder cancer and other cancers previously [38, 39], this analysis particularly associates this function with C3 subtype in BLCA.

Next, the importance of an individual gene in a gene-set is determined by gene influential score (GIS), which is calculated using a leave-one-out procedure (see methods). The maximum GIS value for a gene in a gene-set is one, which indicates that gene contributes a high proportion of variance to the overall variance of the gene-set scores. At the same time, a GIS close to one often suggests a high correlation between the gene expression value and gene-set score. The most significant gene-set, "immune response" and "immune system process" have significant positive or negative GSS in 270 and 265 of 308 patients respectively. The scores clearly distinguished subtypes C1-C3 (Figure 4.7A). The median GSS for the gene-set "immune system process" was 0.82, -0.75, -0.61 in C1, C2 and C3 respectively, representing immune related processes are up-regulated in the C1 subtype. Gene influential score (GIS) suggested that the top ranked genes included *ITGB2*, *SPI1*, *DOCK2*, *LILRB2* and *LAT2*. Other highly ranked genes included drug target genes such as *CD4*, *IL6*, the interferon induced proteins *IFITM2* and *IFITM3* and the G protein coupled receptors *GPR183* and *CMKLR1*. Top positive influencers in "regulation of apoptosis" were also related to the immune response, such as *STK17A*, *ANXA5* and *BCL2A1*, *STAT1*, Serpin B, *TGFB* and *ANXA1*. Moreover, several epithelial to mesenchymal transition (EMT) related gene-sets, such as "collagen" (including *COL6A3*, *COL1A1*, *COL5A1* and *COL3A1*), "extracellular matrix proteins" (e.g. glycoproteins *SRGN* and *FBN1*) and mesenchymal gene-sets were elevated in C1.

The C3 subtype tumors had higher gene-set scores in mitochondrial related gene-set and lower expression of genes are related to cell cycle process and DNA replication. The GIS suggested that two families of genes, NADH dehydrogenases (NDUFs) and mitochondrial ribosomal proteins (*ABCC1*/*MRP*) influenced the mitochondrial proteins. The relative over expression of mitochondrial genes in the C3 subtype of tumors may imply a different metabolic status than the C1 or C2 subtypes.

To identify transcription factors (TF) that may regulate gene expression in the three tumor subtypes, transcriptional factor target (TFT) gene-sets were used to annotate the tumors. Similar to the selection of GO terms, this study focused on TFT gene-sets with more than 200 significant gene-set scores (GSSs) across 308 patients. The GSSs of the E2F family target gene-set were significantly different in most of the tumors and are particularly low for the C3 tumors. The rest of the four identified TFs were highly elevated in the C1 subtype, including an *MADS* (*MCM1*, Agamous, Deficiens, and *SRF*) box superfamily member, *SRF* and several TFs associated with transactivation of cytokine and chemokine genes, e.g. *NFkB1*, *ETS1* and *IRF1* (Figure 4.7B). The genes exhibiting the largest GIS in the *IRF1* and *NFkB1* target gene-sets include *ACTN1*, *CXorf21*, *ICAM1*, *MSN*, *TNFSF13B*, *IL12RB1* and *CDK6*. Furthermore, I examined the correlations between gene-set scores and the mRNA expression. All five TFs showed that the TF mRNA and gene-set scores are significantly correlated (Figure 4.7C).

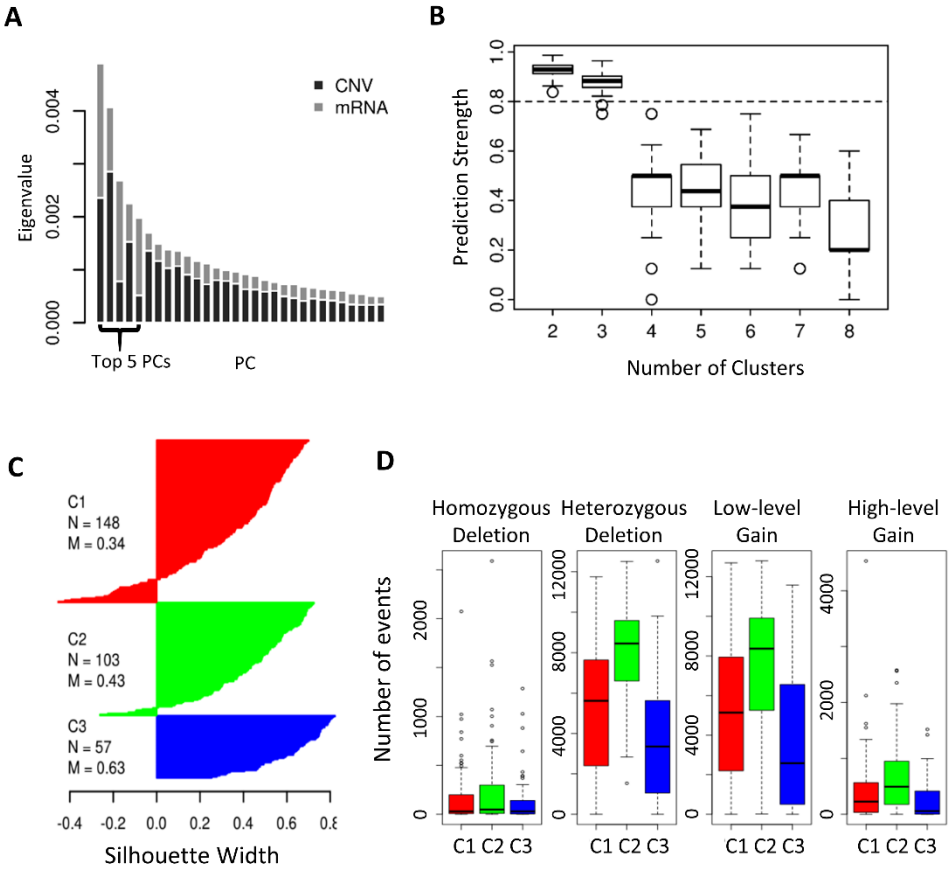

Figure 4.6 *Data integration with moGSA and integrative subtype defined by latent variables. (A) Bar plot showing the eigenvalues of principal components (PCs, also known as latent variable). The top 5 PCs were selected in the analysis. (B) Prediction strength was used to evaluate the robustness of classification into two to eight subtypes. The boxplot shows the prediction strength of 100 randomizations. Two and Three are relative robust subtype models (prediction strength > 0.8). (C) Silhouette plot representing the stability of the identified clusters. The plot suggests there might be unstable patients in C1 and C2, whereas C3 is highly robust. (D) Genomic instability are different in subtypes. C2 subtype have higher number of mutation events.*

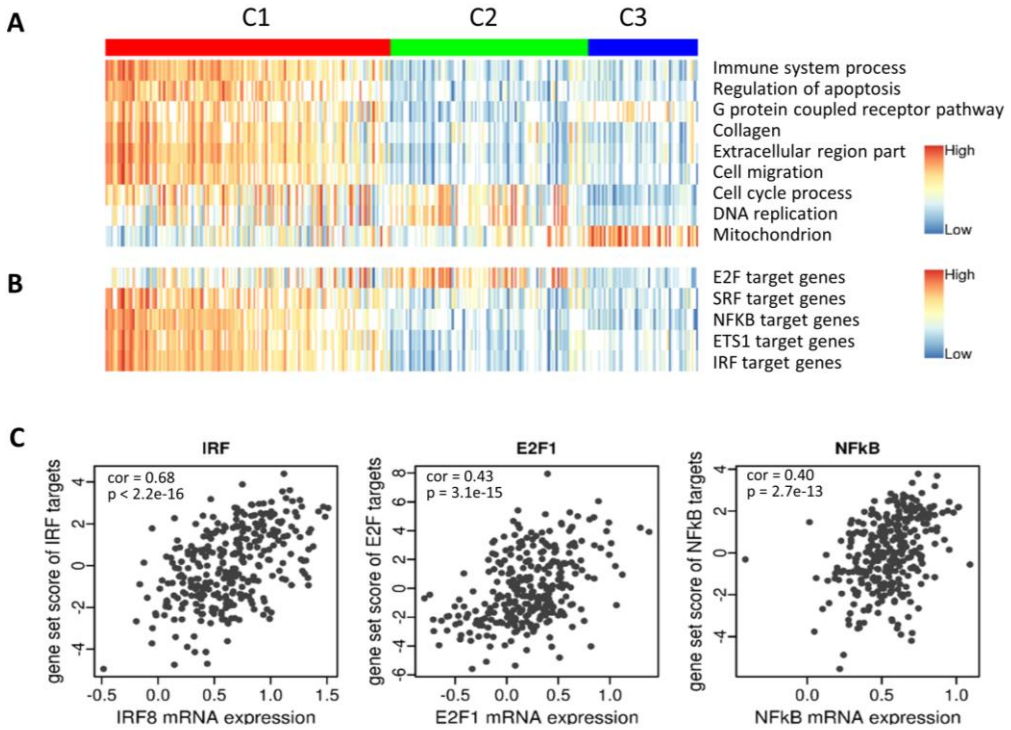

Figure 4.7 *(A) Gene ontology (GO) and transcriptional target (TFT) gene-sets annotation of tumors. Heatmap showing the GSSs for selected gene-sets. The gene-sets "immune-related", "apoptosis", "G protein receptor", "collagen", "extracellular region" and "cell migration" are strongly associated with the C1 (basal-like) subtype, whereas the mitochondrial related gene-sets are over represented in the C3 (luminal A-like) subtype of tumors. (B) The most significant transcriptional factor (TF) target gene-sets. The gene-set scores suggest that 4 out of the 5 TFs are hyperactive in the C1 subtype, except E2F family is active in the C2 subtype of cancer. The white spaces in (A) and (B) denote non-significant GSSs. (C) The scatter plots display the correlation between gene-set scores and the mRNA level of selected TFs. The expression of selected TFs is highly correlated with their gene-set scores.*

In order to identify the contribution of each dataset or PC to the overall gene-set score, a gene-set score could be decomposed with respect to the datasets (referred as data-wise decomposition; see methods). Figure 4.8A shows the normalized means (see methods) of data-wise decomposed GSSs in each subtype for "cell cycle process". The result suggested that mRNA expression strongly influenced the GSS, particularly the low GSS of the C3 subtype patients. The gene influential score (GIS) analysis supports this as the top 30 most influential genes are all from mRNA expression (Figure 4.8B), including *RACGAP1*, *DLGAP5*, *FBXO5*, *AURKA*, *KERA* (*CNA2*) and *CDKN3* (Figure 4.8C). By contrast, both CNV and mRNA data influenced the gene-set "G protein coupled receptor activity" (Figure 4.8D) and the GIS analysis shows that the most influential genes include those from both mRNA and CNV data (Figure 4.8E). However, the CNV and mRNA expression patterns in the C3 subtype shows a clear difference for this gene-set (Figure 4.8F). Top gene influencers of "G protein couple receptor activity" included CNV of *GRM6*, *NMUR2*, *PDGFRB* and adrenergic receptors, the gene expression of *ADGRL4* (*ELTD1*), *CMKLR1* and *PDGFRB* (Figure 4.8F).

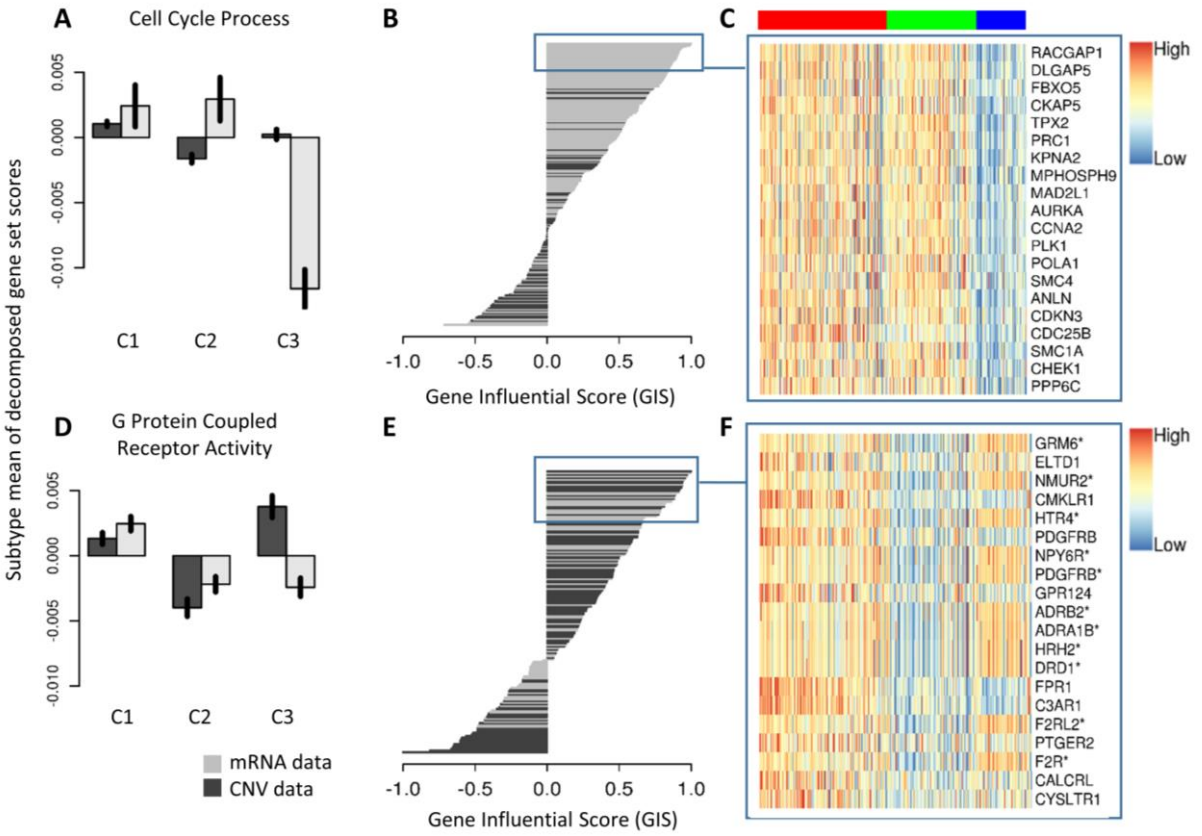

Figure 4.8 *CNV and mRNA data contribute unequally to defining subtype and gene-set scores. (A) Data-wise decomposition of gene-set scores for "cell cycle process". The bar plot shows the normalized mean of data-wise decomposed GSSs in each subtype (the black vertical line on the bars show the 95% confidence interval of the mean). (B) The bar plot shows the gene influential scores (GISs) of genes in the "cell cycle process" gene-sets. The expression of the top 30 most influential genes in the gene-set are shown in (C). (D-F) Same as (A-C) for "G protein couple receptor activity". Gene names in (F) with asterisks indicate genes from CNV data.*

4.4 CONCLUSIONS

This chapter introduced a new multivariate single sample gene-set analysis approach, moGSA that enables discovery of biological pathways with correlated profiles across multiple complex datasets. moGSA uses multivariate latent variable analysis to explore correlated global variance structure across datasets and then extracts the set of gene-sets or pathways with highest variance and most strongly associated with this correlated structure across observations. The combining multiple data types can compensate for missing or unreliable information in any single data type so there is a greater potential to find gene-sets that cannot be detected by single omics data analysis alone [4].

moGSA uses the maximum variance of the concordant structure across of datasets to calculate the gene-set scores for each observation. This is fundamentally different from other gene-set enrichment analysis methods which use a 'within observation summarization' such as the mean or median of gene expression of genes in a gene-set. It has several characteristics that make it attractive for data integration. First moGSA uses MFA, a multiple-table extension of PCA, to reduce the complexity of the original data by transforming high dimensional data to a small number of PCs (latent variables). The PCs with highest eigenvalues (largest variance) capture the most prominent structure among the

different datasets. Noise or artifact variance in the datasets can be filtered by excluding PCs with low variance or those that potentially exhibit noise, which strengthens the signal-to-noise ratio in the data [40, 41]. In moGSA, the entire set of features from each platform is decomposed onto a lower dimension space. The linear combination of feature loadings is used in the calculation of the gene-set scores. Features that contribute low variance contribute little to the score and thus the dimension reduction within moGSA comes with an intrinsic filtering of noise. The advantages of intrinsic variance filtering of features could be clearly seen in the application of moGSA to the simulated data. moGSA outperformed other single sample GSA approaches including ssGSEA and GSVA which do not include a noise-filtering component. Second, data integration of features is achieved at the gene-sets level rather than scoring individual features. This greatly facilitates the biological interpretation of the integrated data. There is no requirement to pre-filter the features in a study or map features from different datasets to a set of common genes. Therefore, it can be used to compare technological platforms that have different or missing features.

There is great potential for applying multiple-table unsupervised GSA approaches for discovery of new subtypes and pathways in integrated data analysis of complex diseases such as cancer. Chapter 3 introduced a latent variable based interactive clustering method. In this study, moGSA was also combined with clustering method. Dimension reduction approaches such as moGSA and MFA are well suited to cluster discovery data because these approaches consider the global variance in the data and as such are complementary to hierarchical or k-means clustering approaches which focus on the pair-wise distance between observations [41-43].

There are a few considerations when applying moGSA for gene-set analysis or cluster discovery. First, the variance of retained PC should not be dominated by one or few of the datasets. This could occur when there is low correlated structure among datasets or a very different number of features presented in each dataset. This problem could be solved by the normalization step before the data integration step in moGSA [22]. Second, issues might arise with clustering latent variables if the PCs with the largest variance do not capture cluster structures [40], for example if technical artifacts (such as batch effects) dominate the top PCs. Therefore, application of the method requires careful batch effect control, especially for the large scale omics study. For example, PCA or clustering methods should be applied on each individual dataset to check effects of potential artifact factors. Batch effects are likely to present if the correlation of PCs with artifact factors is high. A more detailed description of batch effect detection is described in [44]. Similar approaches could be also applied to the PC from MFA. Finally, the method is most efficient in detecting gene-sets resulting from broadest correlation patterns in dataset; and might disregard some gene sets with few genes, particularly when their variances are not captured by the selected PCs. Therefore, the method is more useful as a supportive or descriptive approach in early step data analysis.

ABBREVIATIONS

ANOVA	Analysis Of Variance
AUC	Area Under the ROC Curve
BLCA	Bladder Cancer
BP	Biological Process
CC	Cellular Component
CCA	Canonical Correlation Analysis
CIA	Co-Inertia Analysis
CLT	Central Limited Theorem
DE	Differentially Expressed
DEGS	Differentially Expressed Gene-Set
EMT	Epithelial-to-Mesenchymal Transition
GIS	Gene Influential Score
GO	Gene Ontology
GS	Gene-Set
GSA	Gene-Set Analysis
GSEA	Gene-Set Enrichment Analysis
GSS	Gene-Set Score
KNN	K-Nearest Neighbors
MAD	Median Absolute Deviation
MCIA	Multiple Co-Inertia Analysis
MF	Molecular Function
MFA	Multiple Factorial Analysis
MVA	Multivariate Analysis
NMM	Naïve Matrix Multiplication
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristic
SVD	Singular Value Decomposition
TCGA	The Cancer Genome Atlas
TF	Transcriptional Factor
TFT	Transcriptional Factor Target

REFERENCES

1. Ozsolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities**. *Nat Rev Genet* 2011, **12**(2):87-98.
2. Metzker ML: **Sequencing technologies - the next generation**. *Nat Rev Genet* 2010, **11**(1):31-46.
3. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H *et al*: **Mass-spectrometry-based draft of the human proteome**. *Nature* 2014, **509**(7502):582-587.
4. Meng C, Kuster B, Culhane AC, Gholami AM: **A multivariate approach to the integration of multi-omics datasets**. *BMC Bioinformatics* 2014, **15**:162.
5. de Tayrac M, Le S, Aubry M, Mosser J, Husson F: **Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach**. *BMC Genomics* 2009, **10**:32.
6. Le Cao KA, Martin PG, Robert-Granie C, Besse P: **Sparse canonical methods for biological data integration: application to a cross-platform study**. *BMC Bioinformatics* 2009, **10**:34.
7. Fagan A, Culhane AC, Higgins DG: **A multivariate analysis approach to the integration of proteomic and gene expression data**. *Proteomics* 2007, **7**(13):2162-2171.
8. Busold CH, Winter S, Hauser N, Bauer A, Dippon J, Hoheisel JD, Fellenberg K: **Integration of GO annotations in Correspondence Analysis: facilitating the interpretation of microarray data**. *Bioinformatics* 2005, **21**(10):2424-2429.
9. Khatri P, Sirota M, Butte AJ: **Ten years of pathway analysis: current approaches and outstanding challenges**. *PLoS Comput Biol* 2012, **8**(2):e1002375.
10. Huang da W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Res* 2009, **37**(1):1-13.
11. Barry WT, Nobel AB, Wright FA: **Significance analysis of functional categories in gene expression studies: a structured permutation approach**. *Bioinformatics* 2005, **21**(9):1943-1949.
12. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.
13. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R: **A novel signaling pathway impact analysis**. *Bioinformatics* 2009, **25**(1):75-82.
14. Consortium EP: **The ENCODE (ENCyclopedia Of DNA Elements) Project**. *Science* 2004, **306**(5696):636-640.
15. Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, Ballestar E, Bongcam-Rudloff E, Conesa A, Tegner J: **Data integration in the era of omics: current and future challenges**. *BMC Syst Biol* 2014, **8 Suppl 2**:I1.
16. Hanzelmann S, Castelo R, Guinney J: **GSVA: gene set variation analysis for microarray and RNA-seq data**. *BMC Bioinformatics* 2013, **14**:7.
17. Tomfohr J, Lu J, Kepler TB: **Pathway level analysis of gene expression using singular value decomposition**. *BMC Bioinformatics* 2005, **6**:225.
18. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, Schinzel AC, Sandy P, Meylan E, Scholl C *et al*: **Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1**. *Nature* 2009, **462**(7269):108-112.
19. Lee E, Chuang HY, Kim JW, Ideker T, Lee D: **Inferring pathway activity toward precise disease classification**. *PLoS Comput Biol* 2008, **4**(11):e1000217.
20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

21. Culhane AC, Schroder MS, Sultana R, Picard SC, Martinelli EN, Kelly C, Haibe-Kains B, Kapushesky M, St Pierre AA, Flahive W *et al*: **GeneSigDB: a manually curated database and resource for analysis of gene expression signatures**. *Nucleic Acids Res* 2012, **40**(Database issue):D1060-1066.
22. H A, LJ W, D V: **Multiple factor analysis: principal component analysis for multitable and multiblock data sets**. *Wiley Interdisciplinary Reviews: Computational Statistics* 2013, **5**(2):31.
23. H A, LJ W: **Principal component analysis**. *Wiley Interdisciplinary Reviews: Computational Statistics* 2010, **2**(4):27.
24. Zhu Y, Qiu P, Ji Y: **TCGA-assembler: open-source software for retrieving and processing TCGA data**. *Nat Methods* 2014, **11**(6):599-600.
25. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM *et al*: **MapSplice: accurate mapping of RNA-seq reads for splice junction discovery**. *Nucleic Acids Res* 2010, **38**(18):e178.
26. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression estimation with read mapping uncertainty**. *Bioinformatics* 2010, **26**(4):493-500.
27. S M, P T, J M, T G: **Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data**. *Machine Learning* 2003, **52**(1-2):28.
28. Wilkerson MD, Hayes DN: **ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking**. *Bioinformatics* 2010, **26**(12):1572-1573.
29. R T, G W: **Cluster Validation by Prediction Strength**. *Journal of Computational and Graphical Statistics* 2005, **14**(3):18.
30. Phanstiel DH, Brumbaugh J, Wenger CD, Tian S, Probasco MD, Bailey DJ, Swaney DL, Tervo MA, Bolin JM, Ruotti V *et al*: **Proteomic and phosphoproteomic comparison of human ES and iPS cells**. *Nat Methods* 2011, **8**(10):821-827.
31. Wenger CD, Phanstiel DH, Lee MV, Bailey DJ, Coon JJ: **COMPASS: a suite of pre- and post-search proteomics software tools for OMSSA**. *Proteomics* 2011, **11**(6):1064-1074.
32. Sjodahl G, Lauss M, Lovgren K, Chebil G, Gudjonsson S, Veerla S, Patschan O, Aine M, Ferno M, Ringner M *et al*: **A molecular taxonomy for urothelial carcinoma**. *Clin Cancer Res* 2012, **18**(12):3377-3386.
33. Choi W, Porten S, Kim S, Willis D, Plimack ER, Hoffman-Censits J, Roth B, Cheng T, Tran M, Lee IL *et al*: **Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy**. *Cancer Cell* 2014, **25**(2):152-165.
34. Damrauer JS, Hoadley KA, Chism DD, Fan C, Tiganelli CJ, Wobker SE, Yeh JJ, Milowsky MI, Iyer G, Parker JS *et al*: **Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology**. *Proc Natl Acad Sci U S A* 2014, **111**(8):3110-3115.
35. Cancer Genome Atlas Research N: **Comprehensive molecular characterization of urothelial bladder carcinoma**. *Nature* 2014, **507**(7492):315-322.
36. Knowles MA, Hurst CD: **Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity**. *Nat Rev Cancer* 2015, **15**(1):25-41.
37. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G: **GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers**. *Genome Biol* 2011, **12**(4):R41.
38. Lindgren D, Frigyesi A, Gudjonsson S, Sjodahl G, Hallden C, Chebil G, Veerla S, Ryden T, Mansson W, Liedberg F *et al*: **Combined gene expression and genomic profiling define two intrinsic molecular subtypes of urothelial carcinoma and gene signatures for molecular grading and outcome**. *Cancer Res* 2010, **70**(9):3463-3472.
39. Biton A, Bernard-Pierrot I, Lou Y, Krucker C, Chapeaublanc E, Rubio-Perez C, Lopez-Bigas N, Kamoun A, Neuzillet Y, Gestraud P *et al*: **Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes**. *Cell Rep* 2014, **9**(4):1235-1245.

40. W-C C: **On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions**. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 1983, **32**(3):9.
41. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling**. *Proc Natl Acad Sci U S A* 2000, **97**(18):10101-10106.
42. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P: **'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns**. *Genome Biol* 2000, **1**(2):RESEARCH0003.
43. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV: **Fundamental patterns underlying gene expression profiles: simplicity from complexity**. *Proc Natl Acad Sci U S A* 2000, **97**(15):8409-8414.
44. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA: **Tackling the widespread and critical impact of batch effects in high-throughput data**. *Nat Rev Genet* 2010, **11**(10):733-739.

CHAPTER V
General Discussion

With the development of omics technologies, biological systems are being investigated at an unprecedented heterogeneous fashion. The methods that enable the integrative analysis of multiple omics data nowadays attract more attentions of bioinformaticians or even the entire life science community. The European Commission funded two FP7 projects, MIMOmics (http://www.mimomics.eu) and STATegra (http://www.stategra.eu), which aimed to develop statistical methods and software for the integration of omics data. This thesis describes some applications of multivariate methods for the integrative omics data analysis – MCIA, an attractive method for integrative exploratory and visualization of multi-omics data which contain the same samples; the moCluster solves integrative clustering problem and moGSA provides a powerful yet simple tool to perform integrative gene-set analysis. All these approaches make use of different multivariate methods such as multiple co-inertia analysis (MCIA), consensus principal analysis (CPCA) and multiple factorial analysis (MFA). One of the advantages of these methods is that they do not require pre-filtering of features or mapping features from different datasets to a set of common features, therefore, these methods can extract important features even when they are not presented across all datasets. During the period of this study, Tenenhaus et al. proposed regularized generalized canonical correlation analysis (RGCCA), which provides a unified framework of multiple-table multivariate methods [1] and includes many such methods as its special cases, including CPCA and MCIA. The RGCCA introduces two extra parameters – a shrinkage and a linkage parameter. The linkage parameter is defined so that the connection between matrices could be customized. As a special case, MCIA and CPCA (used in Chapters II and III) utilize the full linkage between matrices, that is, the covariance between all pairs of matrices is considered and optimized. On the other hand, the shrinkage parameter ranges from 0 to 1. Setting this parameter to 0 will force the correlation criterion whereas a shrinkage parameter of 1 will apply the covariance criterion (used by MCIA and CPCA). A value in between leads to a compromise between the two options. In practice, the correlation criterion is better in explaining the correlated structure across datasets, thus discarding the variance within each individual dataset. The introduction of the above two parameters makes RGCCA highly versatile and has great potential to be modified and applied to different study purposes and data types (further discussed later). Despite the great potential, the application of this method is just rising. In one of its applications, RGCCA has been combined with a feature selection procedure and is applied to integrate

and select markers from gene expression, comparative genomic hybridization and qualitative phenotype data measured on 53 children with glioma [2].

The challenges of data integration analysis are from both conceptual and practical aspects. The most fundamental conceptual or statistical challenge is inherited from the high dimension of omics data – a single omics dataset often consists of thousands of variables (dimensions), and the combination of them exacerbates the problem further. Although the research studies hope to derive a more accurate and comprehensive picture of the system under study by adding more layers of data, in some particular cases, the inflated dimension and noise make the overall prediction weaker [3]. Conventional univariate analysis methods, such as t-test or analysis of variance (ANOVA), are well accepted and widely used in omics data analysis. However, increasing the number of variables may also reduce the power of univariate analysis [4-6]. In contrast, the multivariate analysis methods used in this thesis circumvent this problem by dimension reduction. These methods identify the coherent or complementary pattern across datasets and represent main variance within a lower dimensional space, whereas the variance potentially reflecting the artifact factors or noise could be removed. In the moGSA method described in Chapter IV, this is particularly important in increasing the sensitivity of integrative gene-set analysis.

In another family of un-supervised data analysis methods, cluster analysis has been widely applied to omics data analysis [7]. Cluster analysis generally investigates pairwise distances or similarities between objects. However, cluster methods have limitations when applied to data without discrete clusters, including unimodal data, data with complex phenotypes, gradients, or overlapping clusters. For example, Şenbabaoğlu et al. (2014) applied consensus clustering to the randomly generated unimodal data (i.e. no cluster structure) and found that consensus clustering divides the data into apparently stable sub-groups. However, principal component analysis (PCA) was unable to identify the subtypes found by the consensus clustering [8], which suggests that this clustering method is prone to discover clusters even if these are not well separated. Moreover, overlapping clusters have been identified in many tumor types analyzed by the TCGA including glioblastoma and serous ovarian cancer [8, 9]. However, most widely used clustering methods, including k-means and hierarchical cluster analysis (HCA), fail to uncover complex data structures such as sub-clusters or overlapping clusters, which are intrinsic to multiple omics datasets. Gusenleitner and colleagues applied k-means and HCA to simulated data with seven overlapping clusters of different sizes. Neither of the approaches could identify the correct cluster structure. Because HCA cannot assign an observation to more than one clusters, it identified either two large clusters or five smaller clusters depending on how the dendrogram was cut. These limitations were observed recently when the cluster-of-cluster assignment algorithm was applied to TCGA Pan-cancer multiple omics data of 3,527 specimens from 12 types of cancer sources [10]. The resulted cluster model was driven largely by anatomical origin and failed to identify any cluster associated with one of the many known cancer pathways. In contrast, dimension reduction or latent variable methods consider the global variance of datasets and will thus highlight general gradients or patterns in the data [11]. The resultant latent variables capture different variation sources of the data. The moCluster approach (Chapter III) combines the clustering analysis

and multivariate dimension reduction analysis. On the one hand, the clustering analysis considers the joint pattern of omics data and each patient has a unique cluster assignment, which facilitates the stratification of cancer patients; on the other hand, the latent variables discover multiple sources of variation, which is often associated with direct biological mechanisms, between different clusters.

Despite these advantages, multivariate method is less applied in the omics data analysis, in part because most of them are descriptive approaches, that is, they do not test any hypothesis and do not return a significance level (*p*-value), which is familiar to scientists and has an universal interpretation. One of the solutions of this problem is to include parse or penalty factors in the multivariate methods. Therefore, the dimension reduction is integrated with a feature selection function, that is, only a small subset of features are associated with each latent variables. In this thesis, moCluster includes a sparse factor in the CPCA algorithm. The selected features can be passed to gene-set analysis to facilitate the interpretation of latent variables. Other studies also introduce methods combining multivariate analysis and feature selection procedures, such as penalized CCA [12], sparse CCA [13], CCA-l1, CCA elastic net (CCA-EN) [14] and CCA-group sparse [15]. Witten et al. [16] provided an elegant comparison of various canonical correlation analysis (CCA) extensions accompanied by a unified approach to compute both penalized CCA and sparse PCA. They used a fast and efficient implementation of the regularized SVD [16]. In addition, Witten and Tibshirani [17] extended the sparse CCA into a supervised framework. The supervised CCA selects variables from two datasets that are not only highly correlated but also associated with a dependent variable. This method is useful to integrate two datasets with a quantitative phenotype, for example, selecting variables from both genomics and transcriptomics data and linking them to drug sensitivity data.

Furthermore, the problem of dimensionality while integrating multiple omics data could also be relieved by using prior biological knowledge [18]. For example, moGSA combines multiple factorial analysis (MFA) with gene-set information so that the high dimensional feature space is reduced to the lower dimensional gene-set space. The uniformed framework RGCCA has a great potential to include extra biological knowledge. For instance, while integrating genomics, transcriptomics and proteomics data, a linkage parameter could be specified to describe the known relationships between different layers of data (i.e., central dogma). Therefore, one can specifically model the correlation of "genome to transcriptome" and "transcriptomic to proteomic" whereas the correlation of "genome to proteins" is excluded. Consequently, the complexity of the model is lower than modeling all pairwise relationships and the result is more intuitively understandable.

One of the challenges of omics data integration from a practical perspective is the missing value problem – more than thousands of features are potentially targeted in one measurement of an omics experiment, but the actually detected features are always not perfectly matched across measurements because of either technological or biological reasons. Therefore, the combination of multiple measurements results in missing values in the matrix. Simply removing missing values ignores their biological significance since missing values may indicate a low expression, particularly lower than the detection threshold, of such features. Therefore, features with some missing values are often the ones of much interest, that is, up- or down-regulated features across experimental conditions.

Unfortunately, statistician community largely overlooked this problem; and most multivariate methods cannot handle missing values directly. In practice, the missing values can be replaced with a low constant (the method used in this thesis). Imputation methods, underlying either on the global or local structure, have been developed to solve this problem. A recent review comparing different imputation methods has suggested that local structure-based approaches generally yield better results than others do. However, no single method has an overwhelming superiority than others. Because of the complex underlying mechanism of missing values, the selection of the imputation method should be dataset-dependent [19].

In addition, visualization is another challenge that needs to be solved for the interpretation of results of multivariate analysis. Dimension reduction methods have the advantage of data visualization in nature. Nevertheless, most traditional visualization approaches are suitable for datasets with fewer variables. The visualization of thousands of variables and observations simultaneously is complex. For example, the interpretation of MCIA plot (Chapter II) would be significantly faster and easier if an observation and its associated variables can be interactively linked, highlighted and selected through visualization. The interactive plot provides a promising solution to this problem. There are several projects, such as D3.js (http://d3js.org), which aim to provide an easy framework to create interactive plots. In addition, some new R package “rCharts” and “ggvis” provide tools to create and share beautiful, interactive plots online. It can be foreseen that the user-friendly interactive figures will greatly facilitate the application of multivariate analysis and the interpretation of result in the near future.

Last, but not least, the development of new methods is an active area in omics data analysis, yet there is a lack of golden standard data that can be used to evaluate newly created methods. In this thesis, I used simulated data to benchmark the methods, but the simulated data cannot represent the complexity of real biological data. In an attempt to solve this problem, the STATegra project generated multiple omics data, including RNA-seq, proteomics, metabolomics and other sequencing data (http://www.stategra.eu/the-b3-system), under well-controlled conditions. The generation of such dataset would accelerate computational method development and comparison. However, the systematical evaluation of methods designed for different experimental types and study purposes requires more benchmarking data, which would need the long-term efforts in the future.

ABBREVIATIONS

CCA	Canonical Correlation Analysis
CPCA	Consensus Principal Component Analysis
HCA	Hierarchical Clustering Analysis
MCIA	Multiple Co-Inertia Analysis
PCA	Principal Component Analysis
RGCCA	Regularized Generalized Canonical Correlation Analysis
TCGA	The Cancer Genome Atlas

REFERENCES

1. Tenenhaus A, Tenenhaus M: **Regularized generalized canonical correlation analysis**. *Psychometrika* 2011.
2. Tenenhaus A, Philippe C, Guillemot V, Le Cao KA, Grill J, Frouin V: **Variable selection for generalized canonical correlation analysis**. *Biostatistics* 2014, **15**(3):569-583.
3. Kristensen VN, Lingjaerde OC, Russnes HG, Vollan HK, Frigessi A, Borresen-Dale AL: **Principles and methods of integrative genomic analyses in cancer**. *Nat Rev Cancer* 2014, **14**(5):299-313.
4. Hackstadt AJ, Hess AM: **Filtering for increased power for microarray data analysis**. *BMC Bioinformatics* 2009, **10**:11.
5. Bourgon R, Gentleman R, Huber W: **Independent filtering increases detection power for high-throughput experiments**. *Proc Natl Acad Sci U S A* 2010, **107**(21):9546-9551.
6. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc Natl Acad Sci U S A* 2003, **100**(16):9440-9445.
7. Brazma A, Culhane AC: **Algorithms for gene expression analysis**. *Encyclopedia of Genetics* 2005.
8. Senbabaoglu Y, Michailidis G, Li JZ: **Critical limitations of consensus clustering in class discovery**. *Sci Rep* 2014, **4**:6207.
9. Verhaak RG, Tamayo P, Yang JY, Hubbard D, Zhang H, Creighton CJ, Fereday S, Lawrence M, Carter SL, Mermel CH *et al*: **Prognostically relevant gene signatures of high-grade serous ovarian carcinoma**. *J Clin Invest* 2013, **123**(1):517-525.
10. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V *et al*: **Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin**. *Cell* 2014, **158**(4):929-944.
11. Legendre P, Legendre LFJ: **Numerical ecology**. *Numerical ecology* 2012.
12. Waaijenborg S, Zwinderman AH: **Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks**. *BMC Bioinformatics* 2009, **10**:315.
13. Parkhomenko E, Tritchler D, Beyene J: **Sparse canonical correlation analysis with application to genomic data integration**. *Stat Appl Genet Mol Biol* 2009, **8**:Article 1.
14. Le Cao KA, Martin PG, Robert-Granie C, Besse P: **Sparse canonical methods for biological data integration: application to a cross-platform study**. *BMC Bioinformatics* 2009, **10**:34.
15. Lin D, Zhang J, Li J, Calhoun VD, Deng HW, Wang YP: **Group sparse canonical correlation analysis for genomic data integration**. *BMC Bioinformatics* 2013, **14**:245.
16. Witten DM, Tibshirani R, Hastie T: **A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis**. *Biostatistics* 2009, **10**(3):515-534.
17. Witten DM, Tibshirani RJ: **Extensions of sparse canonical correlation analysis with applications to genomic data**. *Stat Appl Genet Mol Biol* 2009, **8**:Article28.
18. Reshetova P, Smilde AK, van Kampen AH, Westerhuis JA: **Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data**. *BMC Syst Biol* 2014, **8 Suppl 2**:S2.
19. Webb-Robertson BJ, Wiberg HK, Matzke MM, Brown JN, Wang J, McDermott JE, Smith RD, Rodland KD, Metz TO, Pounds JG *et al*: **Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics**. *J Proteome Res* 2015, **14**(5):1993-2001.

Acknowledgements

During the past four years of my PhD studies, I received a lot of help and support from many people. Therefore, at the end of the thesis, I would like to express my sincere gratitude to all of them.

First, I would like to express my sincere thanks to my supervisor Prof. Bernhard Kuster for providing me the opportunity to do my research in his laboratory. Thanks for his continuous support of my PhD studies and related research. During the most difficult times in my studies, he gave me immense support and freedom that I needed to move on. Meanwhile I would also like to thank Prof. Dr. Hans-Werner Mewes and Prof. Dr. Dmitrij Frishman for being part of my thesis committee.

My sincere thanks also goes to Dr. Amin Moghaddas Gholami and Dr. Aedin Culhane. Dr. Moghaddas Gholami helped me come up with the thesis topic and guided me over the first two and a half years of my PhD studies. He also helped while starting the work in the laboratory. Although she is in the United States, Dr. Aedin Culhane provided me valuable suggestions and comments that widen my knowledge in this field.

I also want to thank our secretaries Gabriele Kröppelt and Silvia Rötzer. As a foreign student, I really appreciate their support in administrative issues and many others. I wish to express my sincere thanks to all the laboratory members. They all created a friendly research environment. I can always have an insightful and stimulation discussion there. I enjoy working together with all of them. In particular, I am grateful to Dr. Dominic Helm and Dr. Hannes Hanne who helped me to read the thesis.

Last but not least, I would like to thank my family: my wife and my parents for their unquestioning and unconditional love and support throughout this study and my whole life. To my wife, 军功章有我的一半，也有你的一半. When I am completing my thesis, my son Renke Meng comes to this world, who brings me the happiness that I never had before, thanks.

List of Publications

Meng C, Zeleznik O, Thallinger G, Kuster B, Moghaddas Gholam A, Culhane A: Dimension reduction techniques for the integrative analysis of multi-omics data. Accepted. Briefing in Bioinformatics 2015.

Meng C, Kuster B, Peters B, Culhane AC, Moghaddas Gholami A: moGSA: integrative single sample gene-set analysis of multiple omics data. Under reviewer

Meng C, Helm D, Frejno M, Kuster B: moCluster: Identifying joint patterns across multiple omics datasets. Accepted. Journal of Proteome Research.

Meng C, Kuster B, Culhane AC, Moghaddas Gholam A: A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics 2014, 15:162.

Moghaddas Gholami A, Hahne H, Wu Z, Auer FJ, **Meng C**, Wilhelm M, Kuster B: Global proteome analysis of the NCI-60 cell line panel. Cell Rep 2013, 4(3):609-620.

Zheng L, Meng Y, Ma J, Zhao X, Cheng T, Ji J, Chang E, **Meng C**, Deng N, Chen L et al: Transcriptomic analysis reveals importance of ROS and phytohormones in response to short-term salinity stress in Populus tomentosa. Front Plant Sci 2015, 6:678.

Adhikari P, Upadhyaya B, **Meng C**, Hollmén J: Gene Selection in Time-Series Gene Expression Data. PRIB'11 Proceedings of the 6th IAPR international conference on Pattern recognition in bioinformatics 2011:12.