Title
The Role of Exploratory Data Analysis and Pre-processing in Omics Studies

Author
Schiffman, Courtney

Publication Date
2019

Peer reviewed|Thesis/dissertation

The Role of Exploratory Data Analysis and Pre-processing in Omics Studies.

by

Courtney Schiffman

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sandrine Dudoit, Co-chair
Professor Haiyan Huang, Co-chair
Emeritus Professor Stephen Rappaport

Spring 2019

The Role of Exploratory Data Analysis and Pre-processing in Omics Studies.

Abstract

The Role of Exploratory Data Analysis and Pre-processing in Omics Studies.

by

Courtney Schiffman

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Sandrine Dudoit, Co-chair

Professor Haiyan Huang, Co-chair

Beginning with microarray data in the 90's, omics technologies have exploded in the last three decades. Proteomic, metabolomic, genomic and epigenomic data are used to understand disease etiology, to detect diseases early on and to identify novel disease therapies. Almost every omics dataset is the result of a complicated experiment and data collection process. The unwanted variation introduced during the experimental process, along with biological complexity and heterogeneity, requires extensive exploratory data analysis and pre-processing to understand the variability within the data.

The goal throughout this dissertation is to demonstrate the need for appropriate exploratory data analysis and pre-processing in various omics data types, and to provide examples of such. Exploratory data analysis refers to extensive visualization and summarization of omics data in order to understand distributional properties of samples and features, to identify unwanted variation, to determine biological patterns, etc. Work done during exploratory data analysis informs subsequent data pre-processing, or a series of steps taken to filter samples and features, to impute missing values, to normalize or transform the data, etc., prior to performing formal statistical analyses.

Here, we first demonstrate exploratory data analysis and pre-processing within the context of single-cell RNA-sequencing data. One goal of single cell RNA-sequencing (scRNA-seq) is to expose possible heterogeneity within cell populations due to meaningful, biological variation. Examining cell-to-cell heterogeneity, and further, identifying subpopulations of cells based on scRNA-seq data has been of common interest in life science research. A key component to successfully identifying cell subpopulations (or clustering cells) is the (dis)similarity measure used to group the cells. We introduce a novel measure, named SIDEseq, to assess cell-to-cell similarity using scRNA-seq data. SIDEseq first identifies a list of putative differentially expressed (DE) genes for each pair of cells. SIDEseq then integrates the information from all the DE gene lists (corresponding to all pairs of cells) to build a similarity measure between two cells. SIDEseq can be implemented in any clustering algorithm that requires a (dis)similarity matrix. This new measure incorporates information from all

cells when evaluating the similarity between any two cells, a characteristic not commonly found in existing (dis)similarity measures. This property is advantageous for two reasons: (a) borrowing information from cells of different subpopulations allows for the investigation of pair-wise cell relationships from a global perspective, and (b) information from other cells of the same subpopulation could help to ensure a robust relationship assessment. We applied SIDEseq to a newly generated human ovarian cancer scRNA-seq dataset, a public human embryo scRNA-seq dataset and several simulated data sets. The clustering results suggest that the SIDEseq measure is capable of uncovering important relationships between cells, and outperforms or at least does as well as several popular (dis)similarity measures when used on these datasets.

We then focus on exploratory data analysis and pre-processing in the context of adductomics data. Metabolism of chemicals from the diet, exposures to xenobiotics, the microbiome, and lifestyle factors (e.g., smoking, alcohol intake) produce reactive electrophiles that react with nucleophilic sites in DNA and proteins. Since many of these reactive intermediates are unknown, we reported an untargeted adductomics method to detect Cys34 modifications of human serum albumin (HSA) in human serum and plasma. Here, we extended that assay to investigate HSA-Cys34 adducts in archived newborn dried blood spots (DBS). As proof-of-principle, we applied the method to 49 archived DBS collected from newborns whose mothers either actively smoked during pregnancy or were nonsmokers. Twenty-six HSA-Cys34 adducts were detected, including Cys34 oxidation products, mixed disulfides with low-molecular-weight thiols (e.g., cysteine, homocysteine, glutathione, cysteinylglycine, etc.), and other modifications. We used careful exploratory data analysis and data preprocessing methods to uncover biological signal in this relatively new omics data type. With an ensemble of statistical approaches, the Cys34 adduct of cyanide was found to consistently discriminate between newborns with smoking versus nonsmoking mothers with a mean fold change (smoking/nonsmoking) of 1.31. Our DBS-based adductomics method is currently being applied to discover in utero exposures to reactive chemicals and metabolites that may influence disease risks later in life.

Finally, we show how exploratory data analysis and pre-processing is essential for the successful analysis of untargeted metabolomics data. Untargeted metabolomics datasets contain large proportions of uninformative features that can impede subsequent statistical analysis such as biomarker discovery and metabolic pathway analysis. Thus, there is a need for versatile and data-adaptive methods for filtering data prior to investigating the underlying biological phenomena. Here, we propose a data-adaptive pipeline for filtering metabolomics data that are generated by liquid chromatography-mass spectrometry (LC-MS) platforms. Our data-adaptive pipeline includes novel methods for filtering features based on blank samples, proportions of missing values, and estimated intra-class correlation coefficients. Using metabolomics datasets that were generated in our laboratory from samples of human blood serum, as well as two public LC-MS datasets, we compared our data-adaptive filtering method with traditional methods that rely on non-method specific thresholds. The data-adaptive approach outperformed traditional approaches in terms of removing noisy features and retaining high quality, biologically informative ones. Our proposed data-adaptive fil-

tering pipeline is intuitive and effectively removes uninformative features from untargeted metabolomics datasets. It is particularly relevant for interrogation of biological phenomena in data derived from complex matrices associated with biospecimens.

To my dad, Dean Schiffman

Who explained two-sample t-tests to me in a restaurant with a napkin and pen, and who taught me how to always make life fun. Wish you were still around to discuss statistics and life with me.

Contents

List of Figures

List of Tables

Acknowledgments

Thanks to all of my collaborators and co-authors for their help with the published manuscripts included in this dissertation. Specifically, thanks to Haiyan Huang, Christina Lin, Lydia Sohn, Funan Shi and Luonan Chen for their help with the manuscript, *SIDEseq: a cell similarity measure defined by shared identified differentially expressed genes for single-cell RNA-sequencing data*, which was published in *Statistics for Biosciences* in 2017. The second chapter of this work is largely composed of this manuscript. Material from the master's thesis work of Courtney Schiffman was preliminary work that laid the foundations for chapter 2.

Thanks to Yukiko Yano, Stephen Rappaport, Sandrine Dudoit, Hasmik Grigoryan, William Edmands, Lauren Petrick, Katie Hall, Todd Whitehead, and Catherine Metayer for their help with the adductomics manuscript, *Untargeted Adductomics of Cys34 Modifications to Human Serum Albumin in Newborn Dried Blood Spots*, published in *ABC* in 2019. This adductomics manuscript constitutes the third chapter of this work.

Thanks to Yukiko Yano, Stephen Rappaport, Sandrine Dudoit, Lauren Petrick, Todd Whitehead, Joise Hayes and Catherine Metayer for their help with the metabolomics manuscript, *Filtering procedures for untargeted LC-MS metabolomics data*, which constitutes the fourth chapter of this work. As of the date of submission of this work, this manuscript was under review in *BMC Bioinformatics*, 2019.

Special thanks to Sandrine Dudoit, Haiyan Huang and Stephen Rappaport for their generous support and mentorship throughout my graduate school career.

Chapter 1

Introduction

1.1 Why is exploratory data analysis and pre-processing necessary for omics data?

Omics data characterize the genetic or molecular profiles of biospecimens in a comprehensive manner, and are often high dimensional [11, 55]. Examples of omics data include genomic, metabolomic, proteomic and lipidomic data. Since the 90's, many high-throughput technologies allowing researchers to study a wide number of omics data types have been developed [11]. Availability of omics data makes it possible to gain a more coherent understanding of biological functions [55]. However, since this explosion of available omics data, numerous researchers have pointed out the inherent noise and heterogeneity present in these large datasets [11, 63, 113, 96, 6]. Therefore, data- and method-dependent exploratory data analysis and pre-processing need to be done in any omics study.

Exploratory data analysis refers to a thorough investigation (often visual) and summary of data that is done before and during data-preprocessing and prior to the primary statistical analysis. This investigation can uncover unwanted and wanted variability, errors in data collection, uninformative features (e.g. genes, metabolites), relationships among covariates/quality control metrics, etc. Exploratory data analysis also demonstrates to the researcher what pre-processing methods are necessary and appropriate for the data set at hand. An example of exploratory data analysis is using box plots of background variability to determine which method of background correction to use for microarray data [104]. Other examples of exploratory data analysis include using principal component analysis (PCA) to determine sources of variation or using box plots of technical replicate samples to visualize batch effects [122].

Data pre-processing is a broad term that can refer to a variety of tasks depending on the omics data type. In general, data pre-processing refers to processing of raw data, quality control analysis, data filtering and data normalization. Data normalization refers to the process of removing unwanted variability and bias in the data to uncover meaningful biological information. Often, researchers refer to data pre-processing as the set of steps done

to transform raw data into a data matrix of features by samples with which to do statistical inference [96, 15, 123, 100]. Examples of pre-processing of omics data include filtering features in RNA-sequencing experiments based on quality scores supplied by software like *FastQC*, normalizing data using methods such as *RUV* or global sample scaling to adjust for unwanted variation, removing poor quality methylation assays based on the abundance of hypermethylated and unmethylated regions, and identifying poor quality samples possibly resulting from sample collection errors [55, 11, 96, 22, 15, 123, 100]. These pre-processing examples demonstrate how exploratory data analysis and data pre-processing go hand-in-hand; it is not enough to simply investigate the data without acting on the information gained, and blind pre-processing without exploratory data analysis will not be data- and method-specific.

It is tempting to put little thought into exploratory data analysis and pre-processing in order to proceed more quickly to the statistical analysis. However, the methods used to pre-process omics data can have a considerable impact on statistical analysis results, as was first pointed out with microarray data [13, 20, 104, 48, 123]. With the advent of microarray data, the topics of exploratory data analysis and data pre-processing began to receive more attention. Numerous researchers demonstrated the effect of the choice of background correction, of normalization procedures and of duplicate spot aggregation on differential expression analysis using microarray data [13, 20, 104]. While true differential expression should be detectable with any appropriate pre-processing method, inadequate or incorrect data pre-processing can result in an abundance of false positives and negatives [104].

Data pre-processing can also have a considerable impact on end results in metabolomics, proteomics and genomics experiments, to name a few [96, 60, 59, 15, 123]. The reason for this is that most omics data suffer from at least one of the following problems: technical variation within and between assays, missing values/data sparsity, abundance of uninformative features, and poor experimental design. An abundance of technical noise can affect measures of association such as p-values and thus mask biological heterogeneity, or make comparisons across assays challenging [104]. Missing values are common in microarray, metabolomics and single-cell RNA-sequencing and epigenomic data [96, 60, 104]. The challenge in pre-processing such data is differentiating between true missing values (i.e. a transcript is not present in a cell or a metabolite abundance is below the limit of detection) and errors in data collection. Incorrectly imputing values during pre-processing of the data can have a considerable effect on subsequent analyses [104, 97].

Most omics data contain vastly more features than samples, and in many cases a substantial portion of the feature are uninformative [48, 122, 60]. While many statistical methods have been developed to analyze high dimensional data, often the success of such methods is limited by the considerable number of uninformative features [55]. Furthermore, due to the complexity of the experiments used to collect omics data, there is ample opportunity for errors in experimental design, such as confounding batch with the biology of interest, failing to randomize samples in an LC-MS run, or arranging samples on a chip by phenotype. The above list of problems encountered in analyzing omics data is incomplete and each

omics dataset is unique in its pre-processing requirements. Thus, even the most experienced researcher must explore and scrutinize their data prior to statistical analysis.

1.2 Managing unwanted variation introduced during data collection

Most omics data suffer from unwanted variation, often introduced during data collection. Examples of unwanted variation include variation due to sample preparation, batch, sample contamination, machine performance, or even within platform biological variation (e.g. GC-content, blood hematocrit levels, etc.) [24, 22, 99]. In order to focus on the biological phenomena of interest, this unwanted variation must be adjusted for. Data normalization, one of the most complex steps within data pre-processing, helps to reduce the unwanted variation present in omics data, and to make samples and features comparable during statistical analysis [24, 22]. Exploratory data analysis prior to data normalization is crucial for identifying the sources of unwanted variation within the data.

For example, PCA, sample box plots and relative log abundance plots are common exploratory data analysis methods used to identify unwanted variation due to batch and machine performance in gene expression and metabolomics studies [24, 22]. Instead of plotting logged abundances of all features for each sample, relative log abundance plots show the logged ratio of each feature to the median abundance of the feature across all samples, for all features. Extensive exploratory data analysis in RNA-sequencing experiments has also shown unwanted variation due to GC-content, transcript size and sequencing depth that is adjusted for with normalization [22]. An increase in exploratory data analysis of untargeted metabolomics data in the recent years has lead to an explosion of normalization techniques developed for this areas of omics research [24, 81, 71]. One of the more complex normalization tasks in untargeted metabolomics is the removal of unwanted variation in untargeted LC-MS metabolomics of neonatal blood spots [89, 88]. Exploratory analysis of such data revealed multiple sources of unwanted variation, including blood volume, LC-MS machine performance, batch effects and age of the neonatal blood spots [89, 88]. Blood hematocrit levels, indicators of blood volume, were verified as a source of unwanted variation by using exploratory analysis. Plotting total feature abundances of samples and estimated factors of unwanted variation against blood hematocrit levels uncovered a striking correlation in several untargeted metabolomics studies [89, 88].

New areas of omics research, such as investigating chromatin interactions, continue to emerge and require their own normalization procedures [16]. One of the most commonly used techniques for investigating chromatin interactions is Hi-C. However, the Hi-C data collection procedure is highly complex and introduces several sources of technical variation, including spurious ligation products, and fragment length and GC-content [16]. Heatmaps are common exploratory data analysis tools used to visualize Hi-C contact matrices, matrices that demonstrate the degree of contact between loci throughout the genome. However, given

the complexity of the technical and biological biases, software has recently been developed to thoroughly visualize and explore Hi-C data with methods that go beyond heatmaps prior to and following normalization for technical artifacts [16]. Open source software such as *GITAR* [16] allows non-computational scientists to visualize their Hi-C data with histograms, heatmaps and bar plots to understand the sources of unwanted variation. Ideally, such software would be available and easily accessible for all kinds of omics data in order for researchers to identify and remove unwanted variation introduced in their experiments.

1.3 Challenges with high dimensional data

Exploratory data analysis and pre-processing is especially important for omics data that contain large proportions of uninformative features, such as untargeted metabolomics, ChIP-Seq, DNA methylation and single-cell epigenomic data [48, 96, 60, 6]. Analysis of these omics data types without proper feature filtering will likely result in an abundance of false positives and negatives in subsequent statistical analysis. For example, due to errors in feature detection, feature matching, retention time alignment and feature integration, untargeted metabolomics data can have large proportions of uninformative features. Some argue that visualizing peak morphology and integration of a few quality control features and then filtering features based on coefficient of variation estimates is sufficient for handling the issue of uninformative features [122, 96]. However, in high throughput settings, these steps often do not sufficiently reduce the number of uninformative features.

Similar to untargeted metabolomics, generating ChIP-seq data involves the challenge of peak calling. A variety of different peak calling methods are available, and each will result in considerably different peak lists [6]. As with metabolomics, visualization of raw data can help with parameter selection within the peak calling algorithms, but is not enough to sufficiently reduce the number of poor quality features in the data [6]. Further exploratory data analysis and pre-processing work is needed for ChIP-seq data in order to better understand background signal and signal profile distributions. One of the highest dimensional omics data types, DNA methylation data, is a perfect example of how large proportions of noisy features can obscure biological variation of interest and also lead to false positives [48, 4]. Researchers have shown that somewhat arbitrary and inflexible feature filtering cutoffs in DNA methylation data can lead to false positives, such as apparent methylation calls in female samples for probes in the Y-chromosome [48]. Non-specific background fluorescence can instead be used to remove uninformative probes from methylation data and uncover strong biological relationships that would otherwise be hidden [48].

Like single-cell RNA-sequencing data, single-cell epigenetic data (e.g. single-cell DNA methylation and single-cell ATAC-seq data) are extremely high dimensional and sparse. For this reason, extensive exploratory data analysis and pre-processing is necessary for this omics data type. For example, similar to untargeted metabolomics, only a small fraction of reads in single-cell ATAC-seq data are used for subsequent analysis [60]. Instead of filtering features to reduce the size of the data, often single-cell epigenetic data are combined across loci

[60, 6]. To decide how to aggregate epigenetic data, researchers use visualization tools like the Integrative Genomics Viewer to view single-cell data [102], and consider the biological motivations behind different aggregation approaches. By doing so, researchers can focus on epigenetic data at the level of promoters/enhancers, repetitive regions, etc.

1.4 Integrating different omics data types

Researchers have recently been looking to the integration of various omics data to gain a comprehensive understanding of disease causes and etiology. Integration of omics data could mean integrating data with shared samples but different feature sets, for example, integrating two datasets with the same n samples, but with one measuring the proteome and the other measuring the transcriptome of the samples [76]. Integration of omics data could also mean combining data with the same feature sets but measured on different samples [76].

Exploratory data analysis and visualization has been used in tools such as *Functional Heatmap* to integrate and identify patterns within omics time-series experiments [130]. *Functional Heatmap* allows researchers to visualize correlations and patterns with heatmaps and line plots (across time) in order to generate hypotheses about relationships between gene expression data and experimental observations. Exploratory data analysis and visualization is especially important when you are trying to understand the relationship between omics data and experimental or clinical data, since the interactions or relationships between exposure and omics data are complex. In fact, many would argue that generation of hypotheses and understanding of variation should begin with the visual identification of patterns and relationships. Although statistical validation is of course needed, if relationships among omics and experimental data cannot be visualized then perhaps progression to statistical validation is futile. Indeed, one sees an emphasis on visualization techniques in the form of dimension reduction methods when integrating omics data [76]. Dimension reduction techniques allow for the simultaneous exploration of multiple omics datasets, and is often the first step in such studies.

Many researchers utilize the abundance of publicly available data when integrating various omics data types. However, with all of the publicly available omics data and interest in data integration, it is becoming more and more important to document how data in public databases are pre-processed. Lakiotaki et al. argue that uniform pre-processing and annotation of omics data in databases is necessary to make omics data comparable [59]. They focus on pre-processing and integration of gene expression and DNA methylation data. They pre-process each of the omics data types in their database in the same way. For example, they use the same normalization method on all microarray datasets to correct for background signal and normalize within array. This uniform pre-processing has several advantages, one being that users of the database do not have the burden of pre-processing each public omics dataset or of documenting their pre-processing methods since they are already clearly documented within the database. Uniform pre-processing of the datasets also makes it easy to perform additional pre-processing and normalization across datasets if desired prior to

integrating the various omics data types.

1.5 The impact of pre-processing on reproducibility and replicability of results.

As previously discussed, exploratory data analysis and pre-processing can have considerable impact on results of omics studies. It is no surprise, then, that researchers have been focusing on how data pre-processing impacts reproducibility and replicability of results in omics research. Analysis results for omics data may fail to reproduce if researchers use different pre-processing methods. Furthermore, results are often not reproducible because studies fail to clearly report the data pre-processing steps used. There is an effort in omics research to more clearly and thoroughly report data pre-processing steps in order to make results reproducible [122, 59, 48]. For example, omics data in public databases are now often uniformly pre-processed and annotated to help with reproducibility and replicability of results [59]. In metabolomics, workflows like *Workflow4Metabolomics* are being created so that researchers can document their own pre-processing workflows for others to use and cite in their own research [36].

Many have argued that uniform data pre-processing will help with replicability in omics studies because uniformly removing spurious values and unwanted variation within studies helps to uncover true, shared biological variability across studies [33, 48]. However, uniform pre-processing will likely only improve replicability if the uniform pre-processing is appropriate across all studies. For example, a filtering cutoff may be appropriate for one metabolomics dataset, but that same cutoff may not sufficiently remove uninformative features in another because of differences in sample preparation, sample size, machine performance, etc. As discussed previously, using the wrong, albeit uniform, pre-processing method can lead to an abundance of uninformative features that mask the biological variability of interest. Therefore, performing uniform pre-processing with a single inflexible filtering threshold could cause the results to fail to replicate across studies. Indeed, widely used inflexible and imprecise pre-processing techniques may be contributing to the challenge of replication in untargeted metabolomics, about which there is growing concern [121, 10].

Exploratory data analysis and data pre-processing can be both data-adaptive and uniform. The same set of exploratory and pre-processing tools can be used across studies, but with data-dependent thresholds and parameters. In this way, uninformative features, unwanted variation, missing values and experimental design challenges within each omics dataset can be appropriately handled, but all under the same framework. It is easier to replicate results if data are correctly explored and pre-processed, which is not always the case with uniform, inflexible methods. Normalization is a perfect example of the replication advantages of using a uniform framework for pre-processing that is also data-dependent. A given normalization framework can be used across multiple datasets, but making adjustments for different factors of unwanted variation, both technical and biological [33, 23, 24]. For

example, using the *functional normalization* framework in DNA methylation pre-processing, which uses control probes to explore and identify sources of unwanted variation within each dataset, has helped with replication in this field [33].

1.6 Benefits and disadvantages of pre-processing platforms.

There has been an effort in recent years to make comprehensive pre-processing of omics data easily accessible, so that scientists do not have to rely on computational biologists, statisticians, or bioinformaticians to properly pre-process data. Web-based pre-processing platforms have been developed for many of the most popular omics data types [1, 36, 21, 75], with the goal of offering a comprehensive set of exploratory data analysis and pre-processing tools to all audiences. For example, the motivation behind developing *eUTOPIA*, an R shiny application for pre-processing and visualizing microarray data, was to allow users with little experience in computer programming to successfully analyze their data. *eUTOPIA* allows users to visualize the effect of the choice of each pre-processing parameter on their microarray data [75].

Platforms that encourage visualization and exploratory data analysis in conjunction with pre-processing allow users to learn about their data and understand the pre-processing steps. Thus, in many cases, comprehensive pre-processing platforms increase reproducibility, replicability and awareness around the importance of pre-processing. However, many pre-processing platforms do not give sufficient guidance on how to appropriately pre-process the data at hand [21, 36]. Users with little or no experience with their omics data may rely too heavily on the platforms without doing enough research into their experiment, and thus may choose methods that are inappropriate for their data. For example, millions of researchers each year utilize the *MetaboAnalyst* server to pre-process their untargeted metabolomics data [21]. However, [21] gives minimal guidance on how to select row- and column-wise normalization, or how to choose the methods or thresholds for feature filtering [21]. Therefore, when using modern, convenient, semi-automated pre-processing platforms, users should be careful to chose those that offer comprehensive exploratory data analysis and data visualization tools.

Chapter 2

Uncovering biological variation with single-cell RNA-sequencing data

2.1 Introduction

RNA-sequencing technologies have allowed researchers to explore the genetic processes underlying many biological phenomena. Typical bulk RNA-sequencing data, by pooling cells together, measures average gene expression. Since many studies require a deeper examination of genetic activities due to the complex and mysterious nature of certain diseases, single-cell experiments quickly became a technological standard in life science research [29, 113]. Focusing in on the single-cell level allows researchers to investigate the meaningful and illuminating heterogeneity among cells of interest and to discover cell-based biologics, e.g., to identify cellular subpopulations and rare cell types, to identify genes differentially expressed within subpopulations, and to examine genetic regulation networks [135].

Numerous clustering methods with varying degrees of complexity have been introduced to study scRNA-seq data [98, 131, 53, 47]. For example, a new algorithm, named GiniClust, was developed to cluster cells using genes with the top normalized Gini indices [53]. SNN-Cliq is a method to identify cell subpopulations by first building a list of the k-nearest-neighbors (k-NN) for each cell using Euclidean distance, and then assessing the similarity between any two cells by examining their k-NN lists [131]. The RaceID (Rare Cell Type Identification) algorithm was designed to differentiate rare, tissue or disease-specific cells among complex populations of cells through two iterations of k-means clustering [47]. The first k-means clustering is done with various specified similarity measures (the default measure is Pearson correlation) in order to identify the measure which results in the most robust clusters. Outlier cells are then identified and clustered separately. In the second k-means clustering, the centers of the original clusters are redefined, and cells are reassigned to the nearest cluster center. The PhenoGraph algorithm focuses on identifying the phenotypes of cells based on signaling proteins, whose expressions are used to construct a k-nearest-neighbor graph (as defined by Euclidean distance) [62]. The Louvain community detection method is then used

to partition the graph in order to find communities of phenotypically similar cells. BackSPIN, a bi-clustering method which seeks to identify subpopulations of cells while simultaneously finding genetic markers of the clusters, has a correlation matrix at the foundation of its complex sorting and splitting algorithm [135]. There are many clustering methods to add to this list, and there are surely more to come. We see that most clustering algorithms rely on some (dis)similarity measure as a basis for clustering regardless of subsequent computational or mathematical complexity. For instance, a key component in the SNN-Cliq or PhenoGraph algorithms is the use of k-NN, derived from Euclidean distances between cells. However, if Euclidean distance was not an appropriate measure to use due to the nature of the data or the study goal, then the k-NN lists as well as the final clustering results would be misleading. Similarly, in other methods, if the employed (dis)similarity measures are not appropriate measures of cell similarity, clustering results from the algorithms may be unreliable. Therefore, the performance and accuracy of many clustering algorithms in the scRNA-seq setting depend on the ability of the used (dis)similarity measures to summarize true, subtle relationships between cells.

In this paper, we focus on introducing a novel measure, named SIDEseq (defined by shared identified differentially expressed genes), to evaluate pair-wise similarities between cells using scRNA-seq data. There are several intriguing and unique ideas behind SIDEseq. Most importantly, the SIDEseq measure incorporates information from all cells in the data set when defining the similarity between just two cells. What kind of information is important to incorporate from all cells when defining cellular relationships? In scRNA-seq data sets, differentially expressed (DE) genes between cells/subpopulations often represent the kinds of relationships and information researchers care about. The SIDEseq measure first identifies the lists of putative DE genes for all pairs of cells and then quantifies the similarity between two cells by examining how much the two cells share in common among their resulting lists of DE genes when they are compared against every other individual cell in the data set. Note that we attempt to evaluate differential expression for a gene based on only two expression values (or between just two cells). This may seem unreasonable at first glance. However, we consider that the DE genes would likely have vague subpopulation-specific information if they were identified across all cells from multiple subpopulations. It is likely that these DE genes would not be as effective at distinguishing between subpopulations as the genes that carry more explicit subpopulation information. SIDEseq attempts to extract and integrate subpopulation-specific information from all cells. Furthermore, since it considers all possible pairwise comparisons of cells, SIDEseq is expected to be robust against noise in any individual list of identified DE genes. The calculation of the SIDEseq measure involves two key quantifications: how to quantify differential expression for a gene between just two cells and how to evaluate consistency among multiple lists of DE genes. To make SIDEseq computationally feasible, we have introduced two simple yet effective statistics to achieve these quantifications.

The development of SIDEseq was motivated in part by our investigation of a scRNA-seq data set consisting of 96 cells from the human epithelial ovarian cancer cell line, CAOV-3. Half of the cells were treated with factors that are hypothesized to be epithelial-to-

mesenchymal (EMT) inducers. There were several motivations behind studying the subpopulations of these cells using their expression profiles. First, such a study could reveal the genetic markers of any subpopulations within the untreated (also referred to as control) or treated cells which could then offer an improved understanding as to why and how they transition to a mesenchymal phenotype. Second, by attempting to cluster cells by treatment status, we could verify whether such treatments could actually induce the cells to transition from epithelial to mesenchymal [42, 136]. Furthermore, the heterogeneous nature of human ovarian cancer cells presented a challenging clustering task which is not only statistically interesting but also biologically interesting in its own right. The cells do not differ by tissue type, cancer type, or other forms of strong biological variation, but they are, by nature, quite heterogeneous. The source of biological variation, which may be the most prominent and noticeable among the cells overall, is their treatment with the two factors. However, when the differences due to treatment are subtle (e.g., when a treatment has only a marginal effect on cells), they could be easily overwhelmed by the cell heterogeneity. This would bring challenges to clustering treated (by different factors) and untreated cells. We explored the human ovarian cancer cell data set using hierarchical clustering paired with Euclidean distance, Pearson and Spearman correlation and the SIDEseq similarity measure, for comparison. The traditional similarity measures were unable to clearly cluster the treated and untreated cells. Hierarchical clustering with the SIDEseq measure was able to cluster the cells by treatment status to a greater extent. Clustering of cells by treatment status was especially challenging for one of the two batches/treatment factors. Therefore, our clustering analysis of the human ovarian cancer cells not only allowed for a useful comparison of measures within a challenging clustering context, but also helped to shed light on the effectiveness (or ineffectiveness) of the two treatment factors in inducing EMT.

For further evaluation of the SIDEseq measure, we studied a public scRNA-seq data set involving human embryo cells [131]. We focused on this public data set because, unlike our human ovarian cancer cell data set, this data set consists of cells from different developmental stages. Therefore, we believed that the cells from this data set could be clustered more successfully and would provide a good comparison of the performance of our proposed measure with current, popular measures. We use both hierarchical clustering and spectral clustering of the data set to compare the similarity measures. The public data set also allows for a comparison of clustering with SIDEseq to a more recent clustering algorithm, called SNN-Cliq, which was originally used to cluster the embryo cells [131].

To further explore the benefits of the SIDEseq similarity measure and its ability to exploit the information found in DE genes to define cell similarity, we also simulated several scRNA-seq data sets. The subpopulations of cells in each data set varied in size, probability of DE genes, mean expression of differentially and non-DE genes, etc. For each simulation, we used the SIDEseq similarity measure with hierarchical clustering to study the measure's ability to react to the various simulation parameters which make clustering of cells into true subpopulations more challenging. We also used the simulation studies to compare the SIDEseq similarity measure with the methods used in the GiniClust algorithm [53]. In the GiniClust algorithm, genes with the top normalized Gini indices are used for clustering [53].

This is similar to how the SIDEseq measure uses largely the top DE genes to define the similarity between two cells, but different in a significant way in that the GiniClust method does not do pairwise comparisons of all cells in the data when defining the similarity between just two cells. To test the importance of this difference, we used the top Gini index genes as identified by the GiniClust algorithm to perform hierarchical clustering with Pearson and Spearman correlation and Euclidean distance and compared the clustering results with those resulting from SIDEseq. In all simulations, SIDEseq outperformed the other measures. For a final comparison, we used the full GiniClust algorithm on the simulated data sets, but found that this algorithm was significantly outperformed by the hierarchical clustering methods described above.

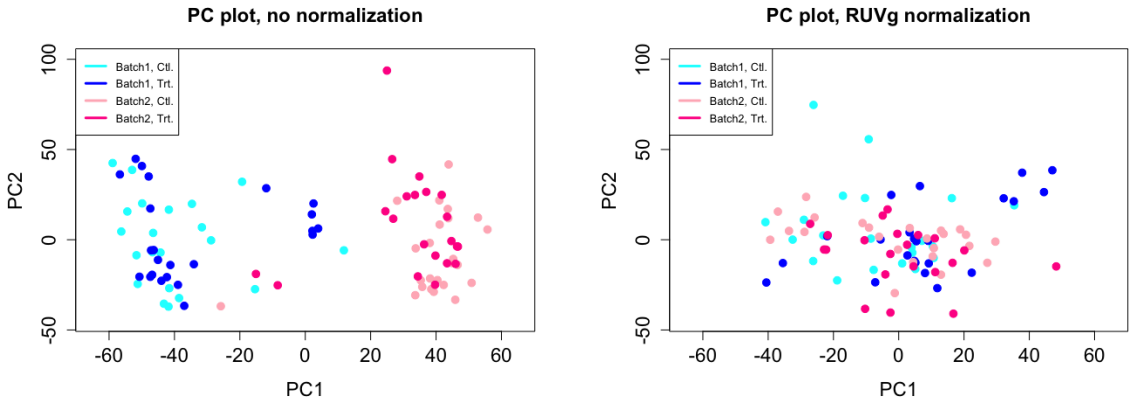
This paper is organized as follows: First, we give a more detailed description of our human ovarian cancer cell data set, and the preprocessing steps we took prior to analysis. We then define our proposed similarity measure, SIDEseq. We use various simulations to compare the methods found in the GiniClust algorithm with the SIDEseq similarity measure and their ability to accurately identify sub-populations. Next, we compare the performance of the SIDEseq measure with common (dis)similarities when used for hierarchical clustering of the human ovarian cancer cell data set. Finally, we compare the performance of the SIDEseq measure with other common measures when used for the clustering of two public, scRNA-seq data sets.

2.2 Methods and materials

The Single-cell RNA-seq data

The novel data set of interest in this study consists of 96 cells from the human epithelial ovarian cancer cells line, CAOV-3 (ATCC, Manassas, VA, USA). CAOV-3 cells were plated on 100 mm tissue culture dishes at a sub-cultivation ratio of 1:5, incubated overnight in supplemented DMEM medium, and then incubated with either thrombin (2.0 U/mL) or TGFβ-1 (5ng/mL) (both from R&D Systems, Minneapolis, MN, USA) for 48 hours. The samples were then prepared per the established protocol for the C1 Single-Cell Auto Prep System (Fluidigm, San Francisco, CA, USA). To prepare the sequencing-ready library for the Bioanalyzer QC and qPCR step, a Nextera XT DNA Sample Preparation Kit was utilized.

The ovarian cancer cells were sequenced in two batches of 48 cells each. Twenty-four of the cells in one batch were treated with TGFβ-1, and 24 of the cells in the second batch were treated with thrombin. The remaining cells in both batches were untreated, control cells. Throughout the paper, the batch containing the 24 cells treated with TGFβ-1 and their corresponding control cells will be referred to as the TGFβ-1 group, and the batch containing the 24 cells treated with thrombin along with their control cells will be referred to as the thrombin group. While TGFβ-1 is a well-established inducer of EMT, there is less evidence to support thrombin's role in EMT [42, 136]. Within the context of cancer, EMT is a process in which cell-cell adhesion and basoapical polarity are lost, EpCAM is down-regulated, and the

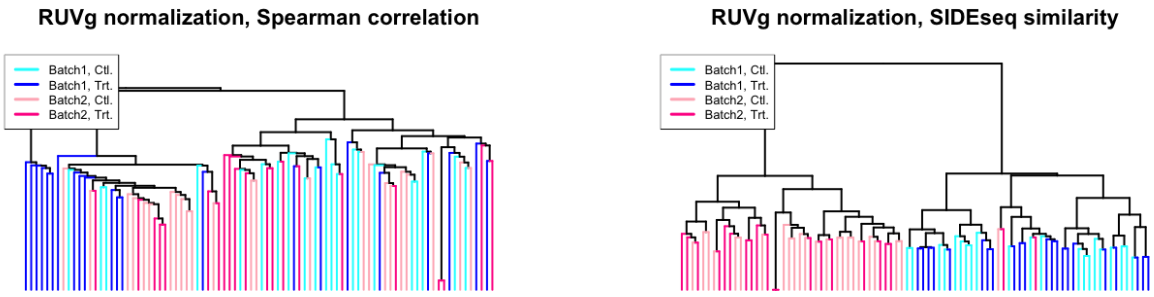


(a) No normalization **(b)** RUVg normalization

Figure 2.1: Principal component plot of ovarian cancer cell dataset.

expression of mesenchymal-associated genes is induced [41, 40]. There is growing evidence that EMT is activated during, and plays a critical role in, cancer invasion and metastasis formation [41, 40, 54, 73, 7, 84]. The heterogeneity of the cellular phenotypes resulting from EMT in ovarian cancer cells is thought to likewise lead to an increased ability to evade early detection [136]. There are several motivations behind studying this data set. Examining the treated ovarian cancer cells and studying whether cells cluster by treatment status can shed light on the effectiveness of the two treatments as EMT inducers, which in turn would lead to a better understanding of the EMT process in ovarian cancer. As previously stated, the TGFβ-1 treatment is a more well-studied inducer of EMT than the thrombin treatment, which requires additional experimental validation [42, 136]. Furthermore, the possibility of a variety of subtle subpopulations within the ovarian cancer cells as a result of EMT brings a statistical challenge of developing sensitive measures for assessing (dis)similarities between cells. If such subpopulations and their associated DE genes could be identified, this would aid in the research of this dangerous, often undetected, gynecologic cancer.

An important source of unwanted biological noise in scRNA-seq experiments, especially pertinent to our human ovarian cancer cell data set, is the variability introduced when cells to be sequenced have different passage numbers [8]. Passage number is defined as the number of times a cell culture was subcultured to maintain continued growth [8]. Passage number has a non-negligible effect on gene expression and regulatory pathways within cell lines [83, 67]. Thus, when cells are sequenced in different batches and the passage numbers are different, cells that were supposed to be biological replicates may become biologically different. This was the case for the human ovarian cancer cell data set, where a difference in passage number (i.e. differed by three) resulted in biologically different cells in the two batches. With a lack of true biological replicates between the two batches, this source of unwanted variation makes normalization across batches very challenging and suggests that within-batch normalization is more appropriate.



(a) Spearman correlation **(b)** SIDEseq similarity

Figure 2.2: Dendrograms of hierarchical clustering of RUVg normalized ovarian cancer cell data with Spearman correlation and SIDEseq similarity.

Indeed, exploratory data analysis reveals that the cells in the two batches vary significantly from each other. Plotting the cells according to their first two principal components through a PCA analysis shows a clear clustering of cells by batch (Fig. 2.1 (a)), especially along the first principal component, which explains roughly ten percent of the variance. Of course, differences between the treated cells in the two batches is expected since the two inducers, TGFβ-1 and thrombin, are different. However, the untreated control cells in both batches should not have significantly different expression profiles. The apparent differences observed between the control cells in the two batches is likely due to technical noise or unwanted biological variability. From this exploratory analysis we see that normalization is needed, though this normalization may ultimately be within-batch.

To demonstrate the challenges behind normalizing across batches, we used a popular normalization method and viewed the resulting data. We used the remove-unwanted-variation technique, *RUVg*, from the *RUVSeq* R package [100]. The *RUVSeq* package provides a few functions to remove unwanted factors of variation from RNA-seq data by using control genes or replicate samples, which are independent of the biological variability of interest, to estimate the factors of unwanted variation using factor analysis. In this case, we try using the *RUVg* method to normalize across batches because we have a set of control genes with which to estimate the hidden factors of unwanted variation in data. In order to do the remove-unwanted-variation normalizations, we used raw read counts, as opposed to TPM or RPKM expression values, following the package instructions.

After *RUVg* normalization of the ovarian cancer cell data across batches, a scatter plot of the first two principal components showed that cells no longer clustered by batch, but they also failed to cluster by treatment status (Fig. 2.1 (b)). It is likely that *RUVg* normalization removed too much variation in this instance, when batches were normalized together and batch and passage number are confounded. As mentioned previously, the different passage

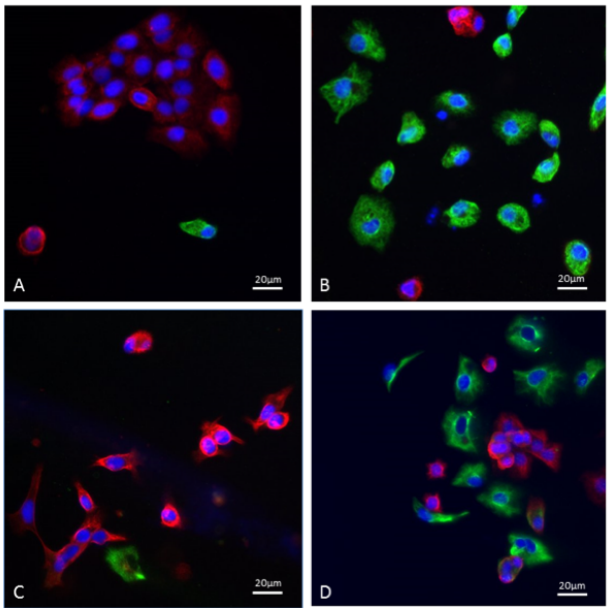


Figure 2.3: Immunostaining image of TGFβ-1 treated (top panels) and thrombin treated (bottom panels) cells.

numbers of the batches likely introduced a large amount of unexpected, biological variation across batches, which the *RUVg* method would determine to be unwanted technical noise [83]. After *RUVg* normalization across batches, we performed hierarchical clustering using Pearson and Spearman correlation, resulting in no apparent clustering of cells by batch or by treatment status (Fig. 2.2 (a)). This agrees with the principal component plot and the concern that *RUVg* normalization might have removed too much variation when used to normalize across batches. However, when we used the SIDEseq measure for clustering, it was interesting to see that cells still clustered almost perfectly by batch, even after *RUVg* normalization across batches (Fig. 2.2 (b)). This suggests that the SIDEseq measure was able to explore the data at a deeper level and brought remaining, subtle differences between batches to the surface.

Since normalizing across batches would always be confounded with the passage number effects, we have focused our analysis within batches. Preliminary clustering of cells within batches 1 and 2 (results not shown) revealed that the thrombin group (batch 2) had much noisier clustering results than what we obtained from the TGFβ-1 group (batch 1). This suggests that the thrombin treatment cells did not differentiate significantly from the untreated cells. Furthermore, this observation is supported by immunostaining images, which show that the thrombin treated cells (top panels of Fig. 2.3, left is untreated, right is treated) have a smaller proportion of cells that have transitioned. This discovery about the thrombin treatment in the data set is an important biological observation and merits further investigation. Due to the apparent ineffectiveness of thrombin as an EMT inducer, we focused on

the TGFβ-1 treated cells in batch 1 for the remainder of the study.

For normalization of the TGFβ-1 treated cells in batch 1, we used the *RUVs* normalization method from the *RUVSeq* package [100], which uses replicate samples that are independent of the biological variability of interest to estimate the factors of unwanted variation using factor analysis. We chose *RUVs* over *RUVg* because we believed a subset of the control cells would act as appropriate replicate samples. Although the remove unwanted variation methods used for normalization were originally developed for bulk RNA-seq data [100], we believe they are still appropriate for removing unwanted factors of variation in scRNA-seq experiments. While it is true that *RUVs* normalization does not take into account the dropout effect in scRNA-seq data, we believe it is still capable of removing unwanted variation from the data set and thus will allow for a meaningful clustering of cells. Furthermore, for the purposes of our clustering task, it may not be necessary to take the dropout effect into account when performing normalization.

We briefly discuss some normalization methods specific to scRNA-seq data. Several scRNA-seq normalization methods require ERCC spike-ins, which are not present in our study. Therefore, we do not apply these normalization techniques [12]. We then explored a normalization method proposed by Lun et al., implemented in the *scran* package in R, who normalize scRNA-seq data by pooling information across cells to create scaling factors that correct for cell-specific biases [70]. We first tried normalizing both batches of cells together using this technique, and then clustered the cells using the normalized data with hierarchical clustering (Spearman correlation). The clustering resulted in almost perfect separation of cells by batch, revealing that the normalization method of Lun et al. did not remove the significant batch effect. When this competing normalization method was used to normalize cells within batch 1 separately, clustering of cells by treatment and control status did not improve over *RUVs* normalization. To quantify this observed difference in normalization methods, we first performed hierarchical clustering with the SIDEseq measure on both the *RUVs* normalized data and on the data normalized by *scran*. We then cut the resulting two dendrograms into four clusters each (four clusters gave the best results) and calculated the Rand Index on all four clusters for each normalization method. We then compared the best two clusters (clusters with the highest Rand Index) from each normalization method, and found that *RUVs* normalization resulted in two clusters each with maximum Rand Index of 1, while the best two clusters from the competing normalization method had Rand Indices of 0.71 and 0.64. Therefore, we concluded that *RUVs* normalization is the preferred normalization method for our human ovarian cancer cell data set.

A new cell similarity measure, SIDEseq

We propose a novel measure, SIDEseq, which is defined by shared identified differentially expressed genes for single-cell RNA-seq data. Introductory work surrounding the SIDEseq measure is explained in "Single Cell RNA-Seq: A Study on Normalization and Sub-Population Identification Techniques" by Courtney Schiffman. The aim of this current work is to fully

explore the statistical properties of the SIDEseq measure and to thoroughly compare it to existing similarity measures.

Method Overview

SIDEseq first chooses DE genes for every cell-pair by only comparing the expression levels of the two cells to produce $N(N-1)/2$ lists of DE genes (one list for one pair of cells; assuming there are N cells in total). Next, SIDEseq assesses the similarity between any two cells by comparing the level of consistency among the relevant lists of DE genes (i.e., to compare cells i and j, SIDEseq evaluates to what level the list of DE genes between cell i and cell t overlaps with the list of DE genes between cell j and cell t, and then integrates such overlapping information across all cells $t \neq i, j$ to define the similarity between cells i and j). The involved integration of multiple DE-gene lists in SIDEseq makes it a quite robust measure against noise in any single list of DE genes.

The ideas behind SIDEseq stem from the belief that various subpopulations likely exist within the data and each has a unique gene activity profile. If two cells come from the same subpopulation, it may be easier to cluster them together by comparing their relationships with other cells in different subpopulations than by comparing their expression profiles in isolation. This might be the case, for example, if the noise in some expression profiles strongly affects the similarity assessment by their expression profiles alone. Another advantage of the SIDEseq measure is that, instead of using all genes, it uses mainly genes evaluated as differentially expressed to build the dissimilarities between cells. This should improve the efficiency of the measure by eliminating noise from uninformative genes.

Method Details

The SIDEseq measure involves two main calculations: the quantification of differential expression for a gene between two cells, and the evaluation of the consistency between multiple lists of DE genes (Fig. 4.1).

The building block for the first calculation in the SIDEseq measure is a simple statistic which is used for a rough evaluation of differential expression between two cells. Suppose one has a matrix of gene expressions of J genes by N cells. We define

$$T_{i,j}^k = \frac{|x_i^k - x_j^k|}{\sqrt{x_i^k + x_j^k}} \tag{2.1}$$

where x_i^k is the expression of gene k in cell i and x_j^k is the expression of gene k in cell j. The result of calculating this statistic over all J genes between cell i and cell j is a vector, $V_{i,j}$, of size J. This vector of statistics is computed for all distinct pairs of the N cells in the data, and roughly indicates the difference in gene expressions between each pair. Each vector is then sorted in decreasing order and truncated to the same length $n \leq J$, so that only the top n genes identified as DE are kept in each vector. As a result, each cell is associated

Flow chart showing the creation of the SIDEseq dissimilarity matrix:

STEP 1

N cells

J genes

$$\begin{bmatrix} x_1^1 & \cdots & x_N^1 \\ \vdots & \ddots & \vdots \\ x_1^J & \cdots & x_N^J \end{bmatrix}_{J\times N} \quad x_j^k = \text{expression of gene } k \text{ in cell } j$$

STEP 2

$$\begin{bmatrix} T_{12}^1 & \cdots & T_{1N}^1 \\ \vdots & \ddots & \vdots \\ T_{12}^J & \cdots & T_{1N}^J \end{bmatrix}_{J\times(N-1)}, \cdots, \begin{bmatrix} T_{21}^1 & \cdots & T_{2N}^1 \\ \vdots & \ddots & \vdots \\ T_{21}^J & \cdots & T_{2N}^J \end{bmatrix}_{J\times(N-1)}, \cdots, \begin{bmatrix} T_{N1}^1 & \cdots & T_{N(N-1)}^1 \\ \vdots & \ddots & \vdots \\ T_{N1}^J & \cdots & T_{N(N-1)}^J \end{bmatrix}_{J\times(N-1)}$$

$T_{i,j}^k$ = differential expression statistic between cell i and cell j for gene k

---

STEP 3

DE Matrix 1 DE Matrix 2 DE Matrix N

$$\begin{bmatrix} G_{11}^{(1)} & \cdots & G_{1N}^{(1)} \\ \vdots & \ddots & \vdots \\ G_{11}^{(n)} & \cdots & G_{1N}^{(n)} \end{bmatrix}_{n\times(N-1)}, \begin{bmatrix} G_{21}^{(1)} & \cdots & G_{2N}^{(1)} \\ \vdots & \ddots & \vdots \\ G_{21}^{(n)} & \cdots & G_{2N}^{(n)} \end{bmatrix}_{n\times(N-1)}, \cdots, \begin{bmatrix} G_{N1}^{(1)} & \cdots & G_{N(N-1)}^{(1)} \\ \vdots & \ddots & \vdots \\ G_{N1}^{(n)} & \cdots & G_{N(N-1)}^{(n)} \end{bmatrix}_{n\times(N-1)}$$

$DE\ Matrix\ i$ = differential expression matrix for cell i.

$G_{i,j}^{(k)}$ = name of gene with the kth largest DE statistic between cell i and cell j ($i.e.\ T_{i,j}^{(k)}$).

STEP 4

$S_{i,j} = \sum_{t=1,\ldots,N,t\neq i,j} \left|\left(G_{it}^{(1)},\ldots,G_{it}^{(n)}\right) \cap \left(G_{jt}^{(1)},\ldots,G_{jt}^{(n)}\right)\right|$ similarity measure between cell i and cell j

---

STEP 5

$$S = \begin{bmatrix} 0 & \cdots & S_{N1} \\ \vdots & \ddots & \vdots \\ S_{1N} & \cdots & 0 \end{bmatrix}_{N\times N}$$ The SIDEseq similarity matrix

---

STEP 6

$$D = \begin{bmatrix} 0 & \cdots & Max - S_{N1} \\ \vdots & \ddots & \vdots \\ Max - S_{1N} & \cdots & 0 \end{bmatrix}_{N\times N}$$ The SIDEseq dissimilarity matrix (Max=maximum in S)

Figure 2.4: Flowchart demonstrating creation of the SIDEseq dissimilarity measure.

with an $n \times (N-1)$ differential expression (DE) matrix, where each column in the matrix is a truncated and sorted vector of statistics comparing the cell's expressions with one of the other $N-1$ cells.

The above procedure to derive lists of DE genes for all pairs of two cells is a key component of SIDEseq, allowing SIDEseq to evaluate the similarity between two cells through examining their relationships with other cells. This is the novel and promising part of the SIDEseq technique which distinguishes it from other methods. We do not consider our statistic T in equation 2.1 to be the best or the only choice to define differential expression using only two expression values. Rather, we consider it a simple yet practical statistic that achieves our analysis goal. Furthermore, it generates satisfactory results. If a better statistic was found, we could replace T by it to further improve the performance of SIDEseq.

The second key calculation in SIDEseq, which produces the final similarity measure, is the evaluation of the consistency among the derived vectors or lists of DE genes that are relevant to every pair of cells (Fig. 4.1). In more detail, for each $t = 1, \ldots, N, t \neq i, j$, the number of genes in the intersection of cell i and cell t's DE gene list and cell j and cell t's DE gene list is found. These numbers are summed across all values of t to get the final SIDEseq measure of similarity between the two cells, which is expected to quantify the level of consistency between the cells' associated differential expression matrices. This measure is the element $S_{i,j}$ (and $S_{j,i}$) of the SIDEseq similarity matrix S. To convert the similarity matrix into a dissimilarity matrix, we take the maximum value in the similarity matrix and subtract every value in the similarity matrix from the maximum value. The diagonal elements of the dissimilarity matrix are set to zero.

Note that an alternative to step four in Fig. 4.1 is to divide the number of genes in the intersection by the number of genes in the union (Eq. 2.2). This alternative similarity measure is related monotonically to the original measure used in SIDEseq, and the two statistics generate almost equivalent results based on what we have observed.

$$S_{i,j} = \sum_{t=1,\ldots,N, t\neq i,j} \frac{|\left(G_{it}^{(1)}, \ldots, G_{it}^{(n)}\right) \cap \left(G_{jt}^{(1)}, \ldots, G_{jt}^{(n)}\right)|}{|\left(G_{it}^{(1)}, \ldots, G_{it}^{(n)}\right) \cup \left(G_{jt}^{(1)}, \ldots, G_{jt}^{(n)}\right)|} \tag{2.2}$$

Selecting n

To determine n for a given data set, one has to determine a number of genes which is large enough to capture the important biological relationships in the data, but small enough so that uninformative, noisy genes are not included. Plotting the values for several vectors of differential statistics is recommended to get an idea of an appropriate range for n in the data set of interest. We found that in all three of the scRNA-seq data sets focused on in this study, there was a range of genes which worked to give optimal clustering results. For the two data sets with RPKM expressions and relatively strong biological signals, anywhere from 150 to 500 genes could be used to get optimal clustering results, corresponding roughly to genes with differential statistics greater than two. More genes may be necessary for data sets with weaker biological variation of interest, such as with the human ovarian cancer cell data set,

where n from 600 to 3000 genes were appropriate. It should be noted that clustering results were stable within a range of genes, providing some flexibility when it comes to selecting this parameter.

It might be suggested that SIDEseq should allow the value of n to change between subpopulations. For example, consider the case where there are three subpopulations of cells: $S1, S2$, and $S3$. Also, suppose there are n_{12} DE genes between $S1$ and $S2$ and n_{23} DE genes between $S2$ and $S3$. Let us consider the case when n_{12} is a lot smaller than n_{23}, and how this would affect the performance of SIDEseq. Suppose that $n_{12} << n << n_{23}$. Since $n_{12} << n$, a list of size n of DE-genes between $S1$ and $S2$ cells would be very noisy (i.e. would contain a lot of non-DE, random genes). However, we wish to point out that no matter if the lists are noisy or not, as long as there is a reasonable mix of noisy and informative DE genes lists (this usually can be achieved with n not too far from the median), all the lists together will provide useful information to help distinguish cells from $S1$, $S2$ and $S3$. To see this point, let us further assume that there are n_{13} true DE genes between $S1$ and $S3$. Without loss of generality, let us assume n_{13} is also very small, that is, both n_{12} and n_{13} are small. Arguments will be similar when n_{13} is large or of reasonable size.

In the above situation, we expect:

- The cells from $S1$ would have the property that all their associated DE gene lists are noisy because n_{12} and n_{13} are both small. Thus, the SIDEseq values between an $S1$ cell and any other cell would be small, because it is unlikely to observe a considerable overlap between DE gene lists if at least one list is noisy.

- The cells from $S2$ would have the property that they have informative DE gene lists against $S3$ cells but noisy DE gene lists against $S1$ cells. Thus, the SIDEseq values between two $S2$ cells would be reasonably large since their associated informative DE gene lists against $S3$ cells would significantly overlap. However, the SIDEseq values between an $S2$ cell and an $S3$ cell would be small since their associated informative DE gene lists are always against different cells (i.e., when compared with $S2$ cells, $S3$ cells will have informative DE gene lists while $S2$ cells will have noisy DE gene lists; When compared with $S3$ cells, $S2$ cells will have informative DE gene lists while $S3$ cells will have noisy DE gene lists).

- The properties of $S3$ cells can be similarly argued as above. In brief, the SIDEseq values between two $S3$ cells would be reasonably large.

In summary, $S2$ and $S3$ subpopulations can be well identified. Cells from $S1$ show different properties from $S2$ and $S3$ cells but it is hard to claim that the $S1$ cells form their own cluster since they have small SIDEseq values among themselves. This however does not seem that unreasonable to us since exceptionally small n_{12} and n_{13} may mean that the subpopulation $S1$ does not possess "convincing characteristics" to form its own subpopulation. Moreover, $S2$ cells and $S3$ cells have stronger subpopulation-specific functional homogeneity compared to the $S1$ cells and thus the smaller SIDEseq measures between $S1$ cells do seem

to make sense to us. Of course this point is debatable. We also note that as n_{12} and n_{13} increase, the SIDEseq values between cells within $S1$ would increase too. Also if there is another subpopulation under consideration, and there is a long list of DE genes between $S1$ and this subpopulation, then the SIDEseq values between cells within $S1$ would become large (that is, SIDEseq would be able to identify $S1$).

Although we have been largely happy with using DE gene lists of the same length, we admit that SIDEseq would be further improved if we could effectively take into account the varied lengths of different DE gene lists. We consider achieving this by adapting the Irreproducible Discovery Rate (IDR) method [63]. The original work on IDR concerns the reproducibility of findings (e.g. identified peaks from ChIP-sep experiments) from replicate experiments. Particularly, the IDR method defines reproducibility as the extent to which the ranks of measures of significance of the findings are consistent across replicates. By jointly modeling the ranks of findings from replicate experiments using a copula mixture model, a score called the IDR (analogous to a false discovery rate) was derived to measure reproducibility. To apply IDR to our analysis, for any two cells, we will first compute our differential statistic, T, for all genes. In this way, we obtain a T-profile for each pair of cells. In the context of our study, we are interested in comparing two T-profiles rather than two lists of findings from replicate experiments. Genes associated with high T-values are likely DE genes, and if there are a lot of common genes with high T-values in both profiles, the two profiles then share a lot of common DE genes. There would also be a stronger dependence among high T-values in the two profiles. Accordingly, we now consider that the bivariate data $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ from two T-profiles consisting of genuine signals (i.e., overlapping DE genes or positively correlated high X_i's and Y_i's) and spurious signals (i.e., non-overlapping DE genes and other random genes or uncorrelated X_i's and Y_i's). Let π_1 and $\pi_0 = 1 - \pi_1$ denote the proportion of overlapping DE genes ($Z_i = 1$) and the rest of the genes ($Z_i = 0$), respectively. We further assume X_i and Y_i are from a continuous bivariate distribution with density h_1 given $Z_i = 1$ (respectively, h_0 given $Z_i = 0$). The mixture copula model can then be expressed as

$$\psi\big(T_1(x), T_2(y), \theta_0, \theta_1\big) = \Big(\pi_0 h_0\big(T_1(x), T_2(y), \theta_0\big) + \pi_1 h_1\big(T_1(x), T_2(y), \theta_1\big)\Big) T_1'(x) T_2'(y) \quad (2.3)$$

with h_1 and h_0 describing different dependence levels between X and Y. $T_1(x)$ and $T_2(y)$ are the unknown scales, which can be estimated empirically. After fitting the copula mixture model, based on the estimates of π_1 and π_0 and the two fitted distributions h_1 and h_0, we can then estimate the chance that a gene is a common DE gene between the two lists (i.e., $P(Z_i = 1 | X_i, Y_i)$) by

$$z(g_x, g_y, i) = \frac{\hat{\pi}_1 \hat{h}_1(\hat{T}_1(x_i), \hat{T}_2(y_i))}{\hat{\pi}_0 \hat{h}_0(\hat{T}_1(x_i), \hat{T}_2(y_i)) + \hat{\pi}_1 \hat{h}_1(\hat{T}_1(x_i), \hat{T}_2(y_i))} \quad (2.4)$$

There are two ways to use the estimated parameters from the above model:

- Since h_1 describes the "dependent" component between X and Y, we can use the estimated dependence parameter associated with h_1 directly to reflect the level of consistency in terms of DE genes between two T-profiles. A SIDEseq measure can then be defined accordingly by replacing the number of intersected DE genes, denoted by $S_{i,k}$ in Fig. 4.1, by the estimated dependence parameter.
- We can classify the genes based on the estimated $z(g_x, g_y, i)$ to obtain a set of common DE genes between two T-profiles. A SIDEseq measure can then be defined accordingly by replacing the number of intersected DE genes, denoted by $S_{i,k}$ in Fig. 4.1, by the size of the inferred set of common DE genes.

Since IDR has an associated R package, a Gaussian copula version of the above method can be easily implemented. However, we note that this approach can be quite time consuming if there are many cells to study, in which case, the simple method based on DE gene lists with the same size would likely be more favorable.

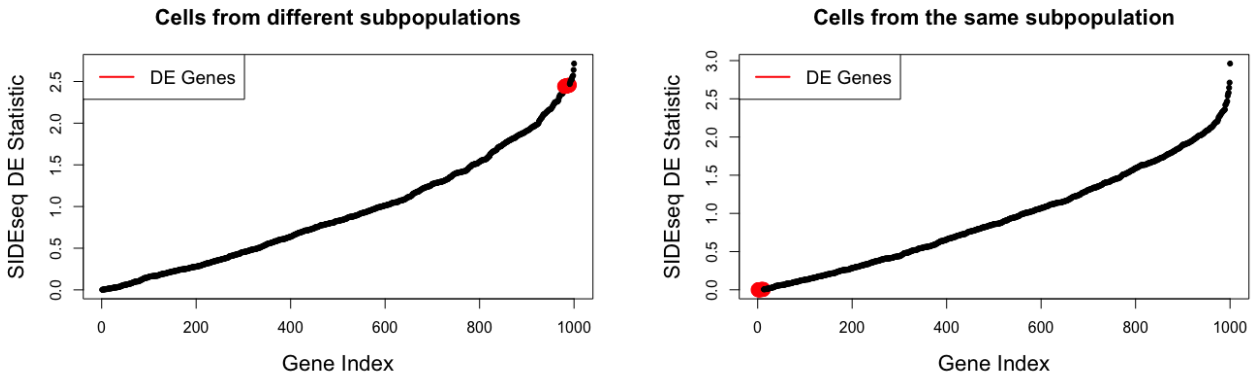
Exploring SIDEseq properties with small simulation studies

In order to better understand the benefits of the SIDEseq measure for sequencing data, we simulated small scRNA-seq data sets and evaluated the ability of the SIDEseq measure, Euclidean distance and Pearson and Spearman correlation to accurately capture the relationships between cells. The simulated data sets consist of 1,000 gene expression measurements for four cells, where two of the cells come from one subpopulation and the other two cells come from another. Each subpopulation is defined by a subset of 10 differentially expressed genes.

$$\begin{aligned}\underline{\text{Subpopulation 1:}}\ \text{Cells a and b, } 10 \text{ genes} &\sim Normal(\mu = 6, \sigma^2 = 0.01^2),\\ 990 \text{ genes} &\sim Normal(\mu = 2, \sigma^2 = 1.7^2)\end{aligned}$$

$$\begin{aligned}\underline{\text{Subpopulation 2:}}\ \text{Cells c and d, } 10 \text{ genes} &\sim Normal(\mu = 0, \sigma^2 = 0.01^2),\\ 990 \text{ genes} &\sim Normal(\mu = 2, \sigma^2 = 1.7^2)\end{aligned}$$

The above data set was generated several times to ensure the robustness of the results. Euclidean distance frequently failed to identify the correct relative similarities between cells and to cluster them correctly by subpopulation. This is because only a small set of genes are differentially expressed between the subpopulations (1%), and the variability in the expressions for the non-differentially expressed genes overwhelmed the difference in expressions for the differentially expressed genes. Pearson correlation (and similarly Spearman correlation) performed worse than Euclidean distance on this data set, for similar reasons. When the



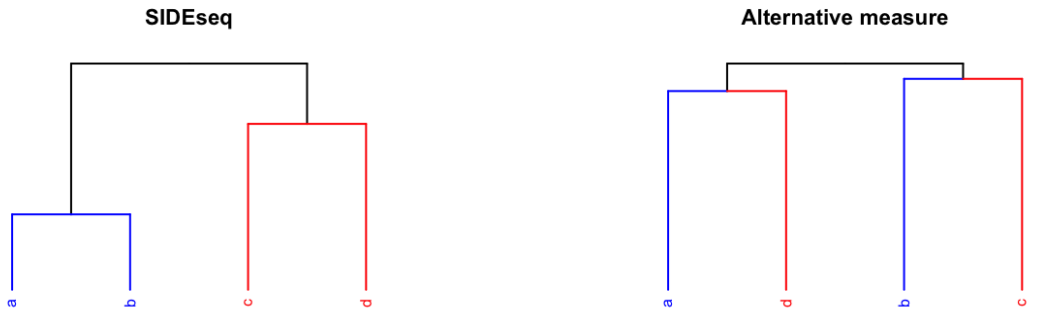
(a) Cells a and c

(b) Cells a and b

Figure 2.5: Sorted SIDEseq DE statistics for cells in different and the same subpopulations.

SIDEseq measure is used, however, it almost always correctly clusters the cells by subpopulation. The SIDEseq measure works because if two cells come from the same subpopulation, they will share many genes in common identified as top differentially expressed genes with the cells in the other subpopulation. A high level of consistency in the top genes identified as differentially expressed with other cells implies a high SIDEseq similarity measure. For example, it is likely that the ten genes which are differentially expressed between cells A and D will be among the top differentially expressed genes as identified by the differential expression statistic in the SIDEseq measure. Many of the genes which are in the top identified differentially expressed genes between cells A and D will also be in the top genes identified as differentially expressed between cells B and D, since cell D is in a different subpopulation. This causes the SIDEseq similarity measure between cells A and B to be high (Fig. 2.5 (b)). Cells from different subpopulations, like cells A and C, will not share a lot of genes which are identified as differentially expressed with the other cells, and will therefore have low SIDEseq similarity measures (Fig. 2.5 (a)).

A suggestion for a similarity measure for scRNA-seq data, which is similar to SIDEseq but may improve upon it, could be the following: to find the similarity between cell i and cell j, separate cell i and j from the population of cells. Call this remaining group of cells which does not include cells i and j the "subpopulation". Find the mean expression level for each gene within the subpopulation. Next, identify the set of genes which are positively and negatively differentially expressed between cell i and the mean expressions of the subpopulation. Do the same for cell j. Then, the similarity between cell i and cell j is the number of genes which they have in common which are positively differentially expressed with the average expression of the subpopulation plus the number of negatively differentially expressed genes they have in common with the mean expressions of the subpopulation. This is similar to the SIDEseq measure with the exception of two big differences: differentially expressed genes are separated by direction of differential expression, and differentially expressed genes are



(a) SIDEseq **(b)** Alternative

Figure 2.6: Dendrograms of hierarchical clustering of small simulation studies with two subpopulations, cells a and b and cells c and d.

found with respect to the average expression levels within the subpopulation.

We used the same simulated data set described previously to compare the performance of SIDEseq with this proposed measure. The data set was generated repeatedly, each time calculating two dissimilarity matrices from each of the two measures and using them to perform hierarchical clustering. The SIDEseq measure outperformed the proposed measure each time, since it was able to group cells A and B together and cells C and D together (Fig. 2.6 (a)). The proposed measure sometimes grouped cells A and B, but failed to group cells C and D or failed altogether to capture the correct relationships (Fig. 2.6 (b)). These results suggest two ideas. One is that splitting differentially expressed genes by sign does not necessarily improve the performance of similarity measures. The other idea for why SIDEseq performs better is because the proposed measure treats the remaining cells as one population and averages their expression levels. In highly heterogeneous data sets with a lot of variability, such as the simulation data set or the human ovarian cancer cell data set, this may cause the similarity measure to perform poorly.

We also provide a small example to help with the intuition behind some of the benefits of the SIDEseq similarity measure over common similarity measures. This small example shows how SIDEseq is able to bypass noise in the expression levels of cells to get at true subpopulations. In the small example, gene 1 and gene 2 are differentially expressed between the two subpopulations of the example, while the other genes are uninformative and simply provide noise (Fig. 2.7). SIDEseq is able to identify the subpopulations, even though there is noise, while the other similarity measures cannot. For example, cells A and C are very similar according to both Pearson and Spearman correlation, even though their expressions for gene 1 and 2 are very different, due to the high proportion of uninformative genes.

(A) Toy Data

		gene 1	gene 2	gene 3	gene 4	gene 5	gene 6	gene 7	gene 8	gene 9	gene 10	gene 11
Subpopulation 1	cell A	5.5	5.5	4.5	4.0	4.5	4.0	4.5	4.0	4.5	4.0	4.5
	cell B	5.5	5.5	4.0	4.5	4.0	4.5	4.0	4.5	4.0	4.5	4.0
Subpopulation 2	cell C	4.5	4.5	4.5	4.0	4.5	4.0	4.5	4.0	4.5	4.0	4.5
	cell D	4.5	4.5	4.0	4.5	4.0	4.5	4.0	4.5	4.0	4.5	4.0

(B) Some commonly used (dis)similarity measures and the SIDEseq measure computed based on the above data

Pearson Correlation

	cell B	cell C	cell D
cell A	0.64	0.72	0
cell B		-0.06	0.77
cell C			-0.69

Spearman Correlation

	cell B	cell C	cell D
cell A	0.05	0.9	-0.31
cell B		-0.39	0.93
cell C			-0.69

Euclidian Distance

	cell B	cell C	cell D
cell A	1.5	1.41	2.06
cell B		2.06	1.41
cell C			1.5

SIDEseq (top 2 DE genes used to define SIDEseq)

	cell B	cell C	cell D
cell A	2	0	0
cell B		0	0
cell C			2

SIDEseq (top 3 DE genes used to define SIDEseq)

	cell B	cell C	cell D
cell A	>=1	<=0.4*	<=0.4
cell B		<=0.4	<=0.4
cell C			>=1

*SIDEseq for cell A and cell C when top 3 DE genes are considered. The computation is based on comparing the DE matrices for cell A and cell C in **(C)**. Looking at the DE gene lists under "A vs. B" and "C vs. B" (cells A and C are separately compared with cell B), the interaction of top 3 in the two lists will have either 0 genes (then 6 genes in the union) or 1 gene (then 5 in the union). There will be similar results when comparing the DE lists under "A vs. D" and " C vs. D". Thus the SIDEseq measure will be at best =1 /5 + 1/5 = 0.4 between cell A and cell C.

(C) Matrices of ordered DE genes between cells

DE matrix for cell A

	A vs. B	A vs. C	A vs. D
Genes ordered by T	genes 3-11 (T=0.16)	genes 1-2 (T=0.32)	genes 1-2 (T=0.32)
	genes 1-2 (T= 0)	genes 3-11 (T=0)	genes 3-11 (T=0.16)

T = |x-y|/sqrt(x+y): differential expression statistic

DE matrix for cell B

	B vs. A	B vs. C	B vs. D
Genes ordered by T	genes 3-11 (T=0.16)	genes 1-2 (T=0.32)	genes 1-2 (T=0.32)
	genes 1-2 (T= 0)	genes 3-11 (T=0.16)	genes 3-11 (T=0)

DE matrix for cell C

	C vs. A	C vs. B	C vs. D
Genes ordered by T	genes 1-2 (T=0.32)	genes 1-2 (T=0.32)	genes 3-11 (T=0.16)
	genes 3-11 (T=0)	genes 3-11 (T=0.16)	genes 1-2 (T= 0)

DE matrix for cell D

	D vs. A	D vs. B	D vs. C
Genes ordered by T	genes 1-2 (T=0.32)	genes 1-2 (T=0.32)	genes 3-11 (T=0.16)
	genes 3-11 (T=0.16)	genes 3-11 (T=0.)	genes 1-2 (T= 0)

Figure 2.7: Properties and advantages of the SIDEseq similarity measure.

2.3 Results and discussion

All of the hierarchical clustering performed in this work uses the *hclust* function in R, specifying the *ward.D2* method [91]. We found that the *ward.D2* method generally resulted in clear clusters for all dissimilarity measures when used on the scRNA-seq datasets, as opposed to the default *complete linkage* method.

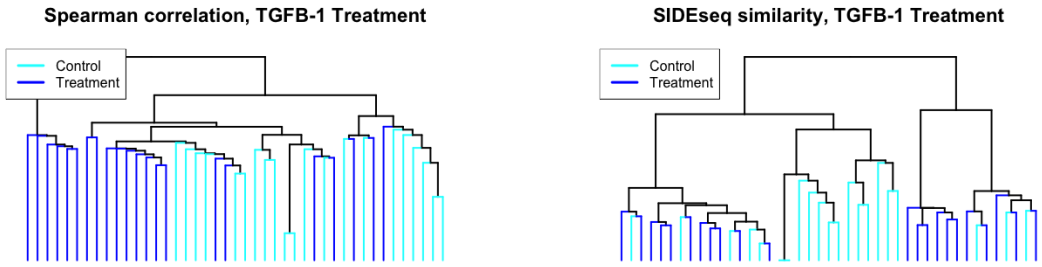
Simulated Data

We used simulation studies to compare the performance of the SIDEseq measure with methods found in the GiniClust algorithm which was designed to detect rare cell types using scRNA-seq data [53]. This is an important comparison because both methods rely on sets of identified DE genes to detect subpopulations or rare cell types, but the ways in which the two methods identify these genes are quite different. The GiniClust algorithm calculates a

normalized Gini index for each gene by looking at the gene's expression across all cells, and then selects the top Gini index genes for clustering and rare cell type identification. The SIDEseq measure, however, identifies DE genes between every cell pair, and then uses the lists of DE genes from all pairwise comparisons to quantify cell similarity. The similarity between two cells is calculated by looking at how consistent the lists of DE genes are that result from their pairwise comparisons with all other cells in the data set. This integration of multiple lists of DE genes into the SIDEseq measure makes this novel similarity measure quite robust to the noise present in any single list of DE genes.

To simulate various single-cell data sets, we used the R package *splatter* [134], with a variety of parameters designed to make the identification of subpopulations more challenging. Each simulated data set consisted of several subpopulations, with different numbers of cells, probabilities of containing DE genes, mean expression of DE genes, etc. We then ran the GiniClust algorithm on each simulated data set, with several variations on the parameters specifying minimum cell number, minimum point number and epsilon. However, regardless of the specified parameters, the GiniClust algorithm only detected "rare cell types," and failed to identify the correct subpopulations of cells, each time clustering all cells not deemed as rare cell types into one large cluster. We believe this is because a relatively small number of genes passed the Gini index cutoff specified in the algorithm, and so there were not enough genes to accurately cluster the cells.

To further compare the GiniClust algorithm with the SIDEseq measure, specifically the way in which they identify and use DE genes, we selected the top Gini index genes (around 150) for each simulation and performed hierarchical clustering of the simulated data with Euclidean distance and Pearson and Spearman correlation. We then used the same number of genes to perform hierarchical clustering with the SIDEseq similarity measure, and compared the clustering results using the Adjusted Rand Index (ARI). Each dendrogram was cut according to the correct number of clusters, and the Adjusted Rand Index was used to compare the resulting clusters with the true subpopulations, with an ARI of one being perfect agreement with the truth and an ARI of zero corresponding to random assignment of cells to clusters. Results are shown in Table 2.1. In simulations 1 through 3, which all contain the same number of cells, genes and subpopulations but vary in the degree and probability of differential expression, the SIDEseq measure outperforms all three common similarity measures. Simulations 4 and 5 increase the number of subpopulations, yet the SIDEseq measure still outperforms all others. In simulation 6, where each sub-population has a different probability for DE genes and is arguably the most realistic model for a scRNA-seq data set, SIDEseq again outperforms all three common similarity measures. The results of the simulation studies suggest several points about the SIDEseq measure: (1.) The method used by the SIDEseq measure of identifying and exploiting DE genes often outperforms methods like those found in GiniClust, where genes are identified as DE based on their expressions over all cells, and (2.) The SIDEseq similarity measure is able to uncover true subpopulations of cells in a variety of scRNA-seq data sets, including those in which subpopulations have different probabilities of their genes being DE or have varying degrees of differential expression.



(a) Spearman Correlation **(b)** SIDEseq similarity

Figure 2.8: Hierarchical clustering of ovarian cancer cell dataset after $RUVg$ normalization, with (a) Spearman correlation and (b) SIDEseq similarity.

Human Ovarian Cancer Cells

The human ovarian cancer cell data set presents more of a challenging clustering task than the simulated data set, due to the uncertain nature of the treatment factors, the passage number effects, the heterogeneous nature of ovarian cancer cells, etc. However, this challenging clustering task is useful for assessing the performance of the SIDEseq similarity measure. We clustered the 'RUVs' normalized counts in the TGFβ-1 group using hierarchical clustering with Spearman correlation and the SIDEseq similarity measure (Fig. 2.8). Since the cells did not cluster well according to treatment status, using the Adjusted Rand Index to compare clustering results from the various similarity measures is not meaningful. Instead, for this data set we rely on visual inspection of the resulting dendrograms. When Spearman correlation is used for hierarchical clustering of the human ovarian cancer cell data set, there are one or two resulting clusters of treatment cells, but cells largely fail to cluster by treatment status (Fig. 2.8 (a)). When the SIDEseq measure is used, three loose clusters of interest can be recognized (Fig. 2.8 (b)). One is a large cluster consisting of only untreated cells. Actually most untreated cells are in this cluster. Another cluster consists of a mix of treated and untreated cells. The third is a large cluster of mostly all treated cells. This cluster is on the outside of the sub-dendrogram formed by the other two clusters. In addition to clearer clusters of cells, the organization of the clusters within the dendrogram is also biologically interesting. The cluster that contains a mix of treated and untreated cells may correspond to a group of cells in the beginning stages of EMT or that have not entirely transitioned to the mesenchymal phenotype. The treated cells within the mixed cluster would then be more biologically similar to the untreated cells.

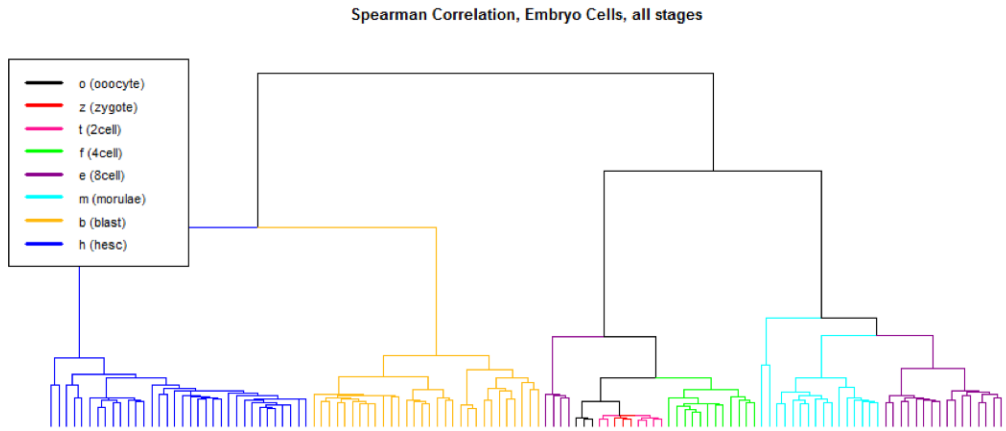


Figure 2.9: Hierarchical clustering of human embryo dataset using Spearman correlation.

Human Embryo Cells

In order to further compare the SIDEseq measure with the common similarity measures, we did hierarchical clustering with different measures on an additional scRNA-seq data set from Yan et al. (2012) [132]. They used a highly sensitive sequencing technique to obtain gene expressions from 124 human embryo cells in various stages of development. The data set covers seven early developmental stages: metaphase II ooocyte (3 cells), zygote (3 cells), 2-cell-stage (6 cells), 4-cell stage (12 cells), 8-cell-stage (20 cells), morula (16 cells) and late blastocyst at hatching stage (30 cells). The data set also includes an eighth stage of development of primary outgrowth during human embryonic stem cell (hESC) derivation (34 cells). Following the filtering method of Xu et al. (2015) for this data set, we used only RefSeq genes with at least one cell with RPKM expression greater than 0.1, resulting in roughly 21 thousand genes [131]. However, while Xu et al. (2015) only used cells from the first seven early developmental stages, we used all 124 cells for the clustering analysis.

Hierarchical clustering using Spearman correlation grouped most cells by developmental stage, with the clusters of cells in the dendrogram following the natural progression of embryonic development (Fig. 2.9). Cells in later developmental stages (late blastocyst and hESC) clustered together on one side of the dendrogram, while cells in the earlier developmental stages (oocyte to morula) clustered on the other. However, Spearman correlation grouped four 8-cell stage cells with the earlier stages. Furthermore, Spearman correlation incorrectly clustered the 2-cell stage cells, separating them into two groups and clustering some of the 2-cell stage cells with the zygote cells. Two morula cells were clustered outside of the 8-cell and morula stage cells. It is interesting to note that simple hierarchical clustering using Spearman and Pearson correlation outperformed or matched the performance of more complex clustering methods for this data set explored by Xu et al. (2015) [131]. They used their proposed clustering algorithm, SNN-Cliq, to cluster 90 cells from this data set (all cells

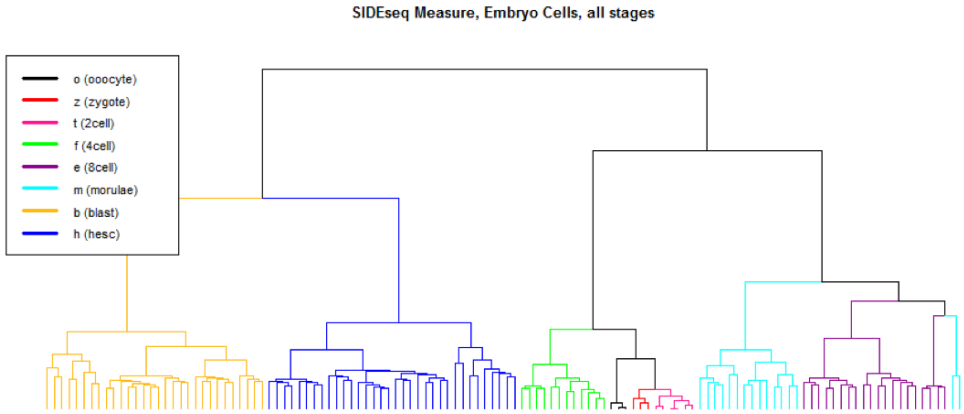


Figure 2.10: Hierarchical clustering of human embryo dataset using SIDEseq similarity.

except the hESC cells) [131]. They also used the k-means and DBSCAN algorithms with Euclidean distance [131]. All methods were matched in performance or outperformed by hierarchical clustering using Spearman correlation.

When we used the SIDEseq measure for hierarchical clustering, it showed a slight improvement over the common similarity measures when clustering the cells in early developmental stages (Fig. 2.10). There was again a split between the early and later developmental stages, but with the SIDEseq measure, all of the 8-cell stage cells were clustered together. Two morula cells broke off from the morula cluster to cluster closer to the 8-cell stage cells, indicating that these may be cells in transition. These are the same two morula cells that were separated by Spearman correlation, but with a different position in the dendrogram. Unlike the traditional similarity measures, the SIDEseq measure successfully clustered all cells in the 8-cell stage together. Furthermore, the SIDEseq measure perfectly clustered the very early stages of oocyte, zygote and 2-cell stage cells. To provide a more quantitative comparison of the similarity measures, we cut each dendrogram into seven, eight (corresponding to the number of cell types) and nine clusters and calculated the ARI for each similarity measure. See Table 2.2 for a full comparison of the Spearman and SIDEseq measures, as well as Pearson correlation and Euclidean distance. SIDEseq outperforms most similarity measures for most cluster number values, except in one case where it is outperformed by Euclidean distance. Here, we note that while the ARI values are informative, they should also be analyzed in the context of the original dendrograms. For example, while Euclidean distance has a higher ARI value than SIDEseq when eight clusters are used, SIDEseq outperforms Euclidean distance in terms of correctly classifying the early developmental stages. These subtle, yet important, clustering details are not taken into account by the ARI when the dendrograms are cut at seven, eight, nine, etc. clusters.

To further explore the subtleties in the clustering of this embryo data set and compare the performance of the different similarity measures, we also used spectral clustering. For each

Dataset index	# sub-populations	# cells, # genes	DE factor	Probability of DE	Pearson w/ Gini genes	Spearman w/ Gini genes	Euclidean w/ Gini genes	SIDEseq
1	5	240,10^4	4	0.1	0.806	0.502	0.148	1
2	5	240,10^4	4	0.05	0.471	0.653	0.404	1
3	5	240,10^4	3	0.1	0.36	0.741	0.198	1
4	7	240,10^4	4	0.05	0.462	0.367	0.222	1
5	7	240,10^4	3	0.05	0.616	0.405	0.064	0.729
6	5	240,10^4	4	(0.15,0.1,0.12,0.05,0.8)	0.446	0.5	0.007	0.552

Table 2.1: ARI values for various similarity measures used to perform hierarchical clustering of simulated datasets.

Public Dataset	# Clusters	Pearson Correlation	Spearman Correlation	Euclidean Distance	SIDEseq Similarity
Embryo Cells, Xu et al.	7	0.740	0.880	0.812	0.880
Embryo Cells, Xu et al.	8	0.659	0.770	0.823	0.770
Embryo Cells, Xu et al.	9	0.740	0.774	0.740	0.828

Table 2.2: ARI values for various similarity measures used to perform hierarchical clustering of the human embryo dataset.

Public Dataset	# Clusters	Pearson Correlation	Spearman Correlation	Euclidean Distance	SIDEseq Similarity
Embryo Cells, Xu et al.	8	0.472 (0.006)	0.718 (0.065)	0.681 (0.069)	0.757 (0.1)
Embryo Cells, Xu et al.	9	0.635 (0.056)	0.745 (0.079)	0.670 (0.073)	0.804 (0.12)
Embryo Cells, Xu et al.	10	0.694 (0.052)	0.747 (0.038)	0.670 (0.035)	0.785 (0.082)

Table 2.3: ARI values for various similarity measures used to perform spectral clustering of the human embryo dataset.

similarity measure, we specified eight, nine and ten clusters, performed spectral clustering 100 times and recorded the average ARI values and their standard deviations (Table 2.3). We chose the number of clusters based on the distributions of the eigenvalues and corresponding eigen-gaps when different values of epsilon where used to build the epsilon graph. The SIDEseq measure outperformed all three traditional similarity measures for all three cluster values, with Spearman correlation being the second best measure. These results suggest that the SIDEseq similarity measure continues to outperform the common similarity measures when used with other clustering algorithms besides hierarchical clustering. Furthermore, when a more principled method is used to choose the number of clusters by using spectral clustering, SIDEseq's performance remains strong, if not improves.

Discussion

Exploratory data analysis is often necessary in scRNA-seq experiments in order to uncover biological heterogeneity. In this work, we demonstrate that the choice of similarity measure used in clustering can have a considerable effect on the success of such exploratory analysis. Exploratory data analysis may also reveal the need for normalization to remove unwanted sources of variation. This was undoubtedly the case for the human ovarian cancer cell data set in this study. There was a clear difference between the cells in the two batches, likely due to both technical variation and induced biological variation as a result of a difference in passage number between batches. We see that in studies where passage number effects are present, the normalization task becomes very challenging. In fact, across-batch normalization may become impossible since this variation can be completely confounded with batch effects. The choice of normalization technique within batch proved to have an effect on the ability of cells to cluster by treatment status. Technical effects such as those observed in this study need to be kept in mind when performing scRNA-seq analysis.

In our study, deriving and integrating lists of DE genes for all pairs of two cells stands as the key component to the proposed similarity measure. This is the novel and promising part of the SIDEseq technique which distinguishes it from other methods. Through studying simulated and real data sets with varying degrees of complexity, we observed the benefits of using the SIDEseq measure. In data sets where there are subtle but important differences between small subpopulations of cells, such as the cells in the early developmental stages of the embryo data set, SIDEseq is able to very accurately identify subpopulations. Furthermore, in data sets where each subpopulation of cells has a different differential expression probability for its genes, SIDEseq seems to outperform traditional similarity measures. Even with data sets where the biological factor of interest is relatively weak, and may be masked by other sources of variability, the SIDEseq measure performs well compared to the commonly used similarity measures. Furthermore, SIDEseq can be utilized in many different clustering methods, like hierarchical clustering and spectral clustering, to accurately identify subpopulations. The success of SIDEseq is due to the novel way in which it uses the consistency among two cells' lists of DE genes (with all other cells) to define their similarity. In this way, SIDEseq is robust to noise in any single list of DE genes, and can investigate the data set at a deeper level than other common similarity measures or clustering algorithms. These novel features of SIDEseq allow it to perform as well as or to outperform more complex clustering methods such as the GiniClust and SNN-Cliq algorithms, even when the measure is paired with a simple method such as hierarchical clustering.

As another interesting observation resulting from the study of the human ovarian cancer cell data, it seems clear that the thrombin treated cells did not differentiate from the untreated cells in their batch as well as the TGFβ-1 treated cells diverged from their respective control cells. The findings from this study support the numerous experimental findings that TGFβ-1 is an inducer of EMT, but they do not provide evidence that thrombin is an EMT inducer. The ability of thrombin to induce EMT merits further investigation.

Chapter 3

Data pre-processing and statistical analysis of untargeted adductomics data

3.1 Introduction

The previous chapter focused on using exploratory data analysis to uncover heterogeneity among cell populations, such as among cancerous and healthy cells. The noise present in the single-cell data had to be understood and managed in order to uncover meaningful biological variability. As discussed previously, many different omics data types exist for exploring biological heterogeneity among human populations. Each data type requires its own set of exploratory data analysis, preprocessing and statistical analysis methods to uncover the biology of interest. In this chapter, another omics data type, adductomics, is discussed. Because this type of data has not been extensively studied by a variety of researchers, meticulous and thorough exploratory data analysis and preprocessing is essential to discovering true differences between phenotypes of interest. Furthermore, the novel application of adductomics to neonatal dried blood spots discussed here provides extra challenges to properly pre-processing and analyzing the data. Like the single-cell ovarian cancer cell dataset, adductomics data suffers from various sources of technical noise. Yet, when proper care is taken to address such noise, the phenotypes of interest can still be studied.

A comparison that is often made using omics data is the difference in exposure to carcinogens between cancer cases and healthy controls [88, 86, 82, 108, 44]. Many carcinogens are reactive electrophiles that are generated through metabolism of chemicals from: the diet and nutrients, exposures to xenobiotics, the microbiome, and lifestyle factors such as smoking and alcohol consumption. Although these reactive intermediates are short-lived *in vivo*, they can be quantified by measuring their reaction products (adducts) with circulating proteins, such as hemoglobin (Hb) and human serum albumin (HSA) [105, 118]. We have focused on HSA adducts bound to the highly nucleophilic sulfhydryl group at Cys34, which is a power-

ful antioxidant and scavenger of reactive electrophiles in the interstitial space [2]. Whereas targeted assays are limited to measurement of particular HSA-Cys34 adducts known a priori, our adductomics methodology motivates discovery and quantitation of unknown HSA modifications of potential health significance [94].

In our Cys34 adductomics pipeline, HSA is isolated from plasma/serum, digested with trypsin, and analyzed via nanoflow liquid chromatography-high resolution mass spectrometry (nLC-HRMS) to pinpoint and quantitate modifications at the third largest tryptic peptide (T3) [45]. In four previous studies, we applied this adductomics pipeline to plasma/serum from healthy smokers and nonsmokers in the U.S. [45], nonsmoking women in China exposed to indoor combustion products and local controls [69], nonsmoking British patients with lung and heart disease and local controls [68], and nonsmoking Chinese workers exposed to benzene and local controls [44]. Here, we modified the adductomics assay to measure Cys34 adducts in newborn dried blood spots (DBS).

Because newborn DBS have been routinely collected at birth to screen for inborn errors of metabolism in the U.S. and worldwide [115], analysis of archived newborn DBS provides an avenue for investigating the etiologies of diseases initiated in utero. Retrospective investigations of chromosomal translocations in DNA from newborn DBS provide direct evidence of the prenatal origin of childhood leukemia, the most common childhood cancer [35, 43, 129]. Chronic diseases in adult life, such as type 2 diabetes mellitus, cardiovascular disease, and the metabolic syndrome, can also have fetal origins [37]. Since HSA has a residence time of 28 days [94], measuring Cys34 adducts in newborn DBS would allow us to investigate exposures to reactive and potentially carcinogenic electrophiles during the last month of gestation.

Here, we describe an untargeted adductomics method to measure HSA-Cys34 adducts in newborn DBS. A major challenge to extending the Cys34 adductomics pipeline to newborn DBS involves the sample matrix, which consists of cellulose and debris from lysed red blood cells and associated proteins, particularly Hb, which are not abundant in serum or plasma [18, 74]. Indeed, Hb is present at a 7-fold higher concentration than HSA in whole blood [18] and interferes with tryptic digestion that releases the T3 peptide and its modifications for analysis [51]. We modified the method to remove Hb and other interfering proteins from DBS extracts prior to digestion. The workflow includes: extracting proteins from DBS, measuring Hb to normalize for blood volume, isolating HSA in solution by precipitating Hb and other proteins, digesting with trypsin, and detecting HSA-Cys34 adducts via nLC-HRMS. As proof-of-principle, we examined HSA-Cys34 adducts in archived DBS collected from 49 newborns with mothers who either actively smoked during pregnancy or did not smoke.

3.2 Methods and materials

Chemicals and Reagents

Acetonitrile (Ultra Chromasolv, LCMS grade), triethylammonium bicarbonate (TEAB) buffer (1 M), ethylenediamine-tetraacetic acid (EDTA, anhydrous), HSA (lyophilized powder, 97-99%), and porcine trypsin were from Sigma-Aldrich (St. Louis, MO). Methanol (Optima, LCMS grade), formic acid (Optima, LCMS grade), and iodoacetamide (IAA) were from Fisher Scientific (Pittsburgh, PA). Purified human Hb was purchased from MP Biomedicals, LLC (Santa Ana, CA). Isotopically labeled T3 (iT3) with sequence $AL-[^{15}N,^{13}C-Val]-LIAFAQYLQQCPFEDH-[^{15}N,^{13}C-Val]-K$ was custom-made ($>95\%$, BioMer Technology, Pleasanton, CA), and the carbamidomethylated iT3 (IAA-iT3)[18] was used as an internal standard to monitor retention time and mass drifts. Water was prepared with a PureLab purification system (18.2 mΩ cm resistivity at $25\,^{\circ}C$; Elga LabWater, Woodridge, IL).

Preparation of capillary DBS for method development

For method development, capillary DBS were collected with informed consent from adult volunteers by finger prick with a sterile safety lancet (Fisher HealthCare, Houston, TX). The first drop of blood was discarded and subsequent drops were collected on Whatman 903 Protein Saver cards (GE Healthcare, Cardiff, UK). Blood spots were air dried for a minimum of 4 days and stored at $-20\,^{\circ}C$ in glassine envelopes (GE Healthcare, Cardiff, UK) prior to use. Punches of 5 and 6-mm diameter were obtained from these DBS with a Biopunch (Ted Pella Inc., Redding, CA).

Archived newborn DBS

Newborn DBS were obtained for 49 healthy control children from the California Childhood Leukemia Study (CCLS) [79]. These newborn DBS had been archived by the California Department of Public Health [26] at $-20\,^{\circ}C$ for 14 to 32 years prior to analysis in the current investigation. Twenty-three of these participants had mothers who actively smoked during pregnancy and the remaining 26 had nonsmoking mothers. Interviews with the biological mother were conducted to collect data on the child's sex, race, and mother's smoking status during pregnancy. A total of 23 smoking/nonsmoking pairs were matched on sex and child's birth year. Smoking/nonsmoking pairs of newborn samples were randomized and then analyzed together to minimize technical variation. Our methodology was developed for 4.7-mm punches from DBS. Because the archived newborn DBS for the present investigation were remnants from previous analyses, they consisted of areas of filter media equivalent to 4.7-mm punches based on size and weight.

Extraction of protein and measurements of Hb and total protein

DBS punches were placed in microcentrifuge tubes and extracted with 55 μL of water at room temperature for 15 min with constant agitation at 1400 rpm (Thermomixer, Eppendorf, Hamburg, Germany). Samples were then centrifuged for 10 s and 5 μL aliquots were diluted with 45 μL of water to measure Hb concentrations (for normalization of blood volumes) with a Cytation 5 microplate spectrophotometer (BioTek Instruments, Winooski, VT) at room temperature. The absorbance of duplicate 2.5 μL sample aliquots was measured at 407 nm, which was the experimentally-determined absorbance maximum corresponding to heme in the ex vivo oxidation state of Hb [127, 77]. Absorbance readings were converted into Hb concentrations with five-point linear calibration curves using Hb standard solutions ranging from 0.5 to 5.0 mg/mL.

Absorbance measurements at 280 nm were used to calculate total protein concentrations in DBS extracts. Total protein was quantified with correction for nucleic acid interferences at 260 nm and background correction at 320 nm.

Sample preparation for adductomics

Various extraction protocols were tested and the method described below was found to be optimal for isolating HSA from DBS while removing Hb and other proteins from the extract (Results and Discussion provides further details). After Hb measurement, 41 μL of methanol was added to the remaining 50 μL of DBS extract (resulting in 45% methanol), vortexed, and mixed at 37 $^{\circ}C$ for 30 min with agitation at 1400 rpm (Thermomixer, Eppendorf, Hamburg, Germany). Samples were then stored at 4 $^{\circ}C$ for 30 min and centrifuged at 14,000×g for 10 min to remove precipitates and cellulose fibers. A 55 μL aliquot of the supernatant was diluted with 95 μL of digestion buffer (50 mM TEAB, 1 mM EDTA, pH 8.0), and the solution was loaded into a Costar Spin-X centrifuge tube filter (0.22 micrometer cellulose acetate, Corning Incorporated, Corning, NY) and centrifuged at 10,000×g for 10 min. A 130 μL aliquot of the filtered solution (containing around 17% methanol to enhance trypsin activity) was transferred into BaroFlex 8-well strips (Pressure Biosciences Inc., South Easton, MA) to which 1 μL of 10 μg/μL trypsin was added (around 1:10 enzyme-to-protein ratio). Pressure-assisted proteolytic digestion was performed with a Barozyme HT48 (Pressure Biosciences Inc., South Easton, MA) instrument, which cycled between ambient pressure (30 s) and 1,380 bar (20 kpsi, 90 s) for 32 min. After digestion, 3 μL of 10% formic acid was added to denature trypsin. Digests were transferred to new tubes and centrifuged for 2 min at 10,000×g. A 100-μL aliquot of the digest was transferred to a 300-μL silanized glass vial (Micosolv Technology Corporation, Leland, NC), and 1 μL of the isotopically labeled internal standard (IAA-iT3, 20 pmol/μL) was added. Samples were stored in liquid nitrogen prior to nLC-HRMS. The 49 newborn DBS were processed daily in four batches of 12 or 13 samples.

nLC-HRMS analysis

Digests were analyzed by an Orbitrap Elite HRMS coupled to a Dionex Ultimate 3000 nanoflow LC system via a Flex Ion nanoelectrospray ionization source (Thermo Fisher Scientific, Waltham, MA), as described previously [45]. Each sample was injected in duplicate, and samples were separated on a Dionex PepSwitft monolithic column (100 μm i.d. $\times$ 25 cm) (Thermo Scientific, Sunnyvale, CA). The mobile phase consisted of 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B). Peptides were separated by gradient elution (2-35 % B, 26 min) at a flow rate of 750 nL/min. Full scan mass spectra were acquired in the positive ion mode with a resolution of 120,000 at m/z 400 in the m/z = 350 – 1500 mass range using the Orbitrap. The MS was operated in data-dependent mode to collect tandem MS (MS2) spectra in the linear ion trap.

Identification, quantitation, and annotation of putative adducts

HSA adducts were identified using the adductomics pipeline with sample preparation modified for DBS as described above [45]. Cys34 adducts are represented by modifications to the triply charged T3 peptide (m/z = 811.7594), and the b^+-series ions prior to b_{14}^+ are shared by all T3 peptides, despite differences in modifications at Cys34. We used in-house R software to screen for this characteristic pattern in MS2 spectra and thereby pinpoint putative T3 adducts when 5 of the 7 most prominent b^+-series fragment ions (i.e., b_3^+ - b_6^+ and b_{11}^+ - b_{13}^+ ions) were detected along with at least 3 of the 5 y_2^+-series ions (y_{142}^+ - y_{182}^+). Adducts were grouped by monoisotopic mass (MIM) within 10 ppm and retention time (RT) within 1.5 min. The means of MIMs and RTs were calculated for each adduct across all samples. Putative adducts were annotated based on the added mass as described previously [45]. Peak picking and integration were performed using the Xcalibur Processing Method (version 3.0, Thermo Fisher Scientific, Waltham, MA) based on the average MIMs (5 ppm mass accuracy) and RTs of the putative adducts. Peaks were integrated with the Genesis algorithm after normalizing the RTs using the internal standard (iT3-IAA) and using a RT window of 60 s.

Exploratory data analysis and pre-processing

Peak areas were log-transformed prior to all exploratory data analysis, pre-processing and statistical analysis. Relative log abundance (RLA) plots [24], which were obtained by standardizing each adduct by the median abundance across samples and logging the resulting ratio, were used to visually inspect the reproducibility of replicate measurements (Fig. 3.1). Two subjects (of nonsmoking mothers) were removed from the analysis and only one measurement was used for one subject due to high variation in adduct abundances based on the RLA plot of duplicate measurements (Fig. 3.1). We also used boxplots of the difference in abundances between injection 1 and 2 for each adduct across all samples to assess the variability of the differences and whether they were centered around zero (Fig. 3.2). With the exception of one adduct (m/z 811.0915), all adducts appear to be relatively reproducible

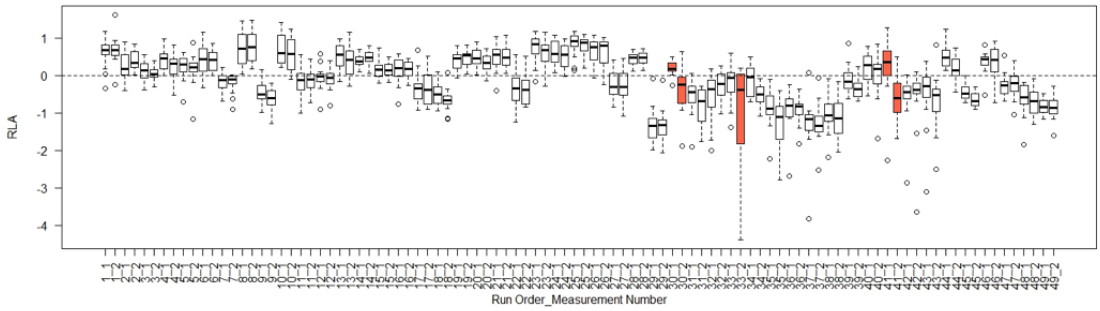


Figure 3.1: Relative log abundance plot of duplicate injections.

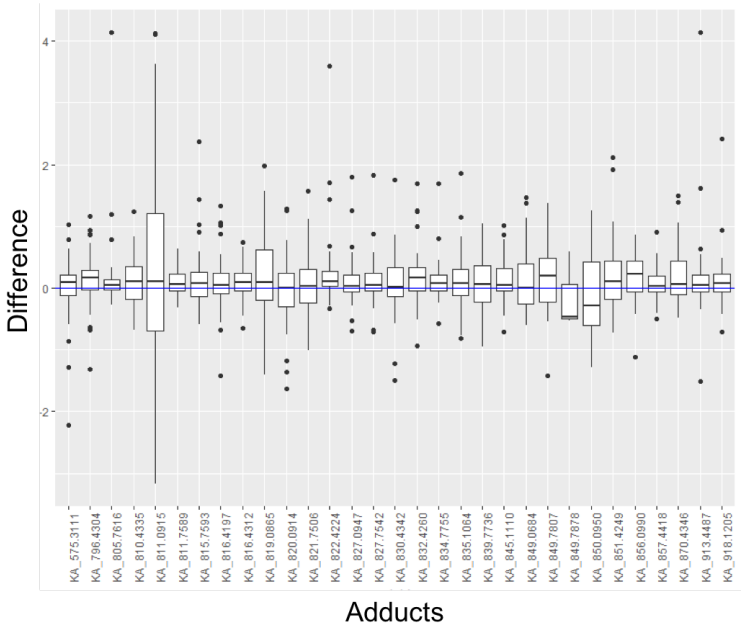


Figure 3.2: Boxplot of difference in duplicate injections for each adduct.

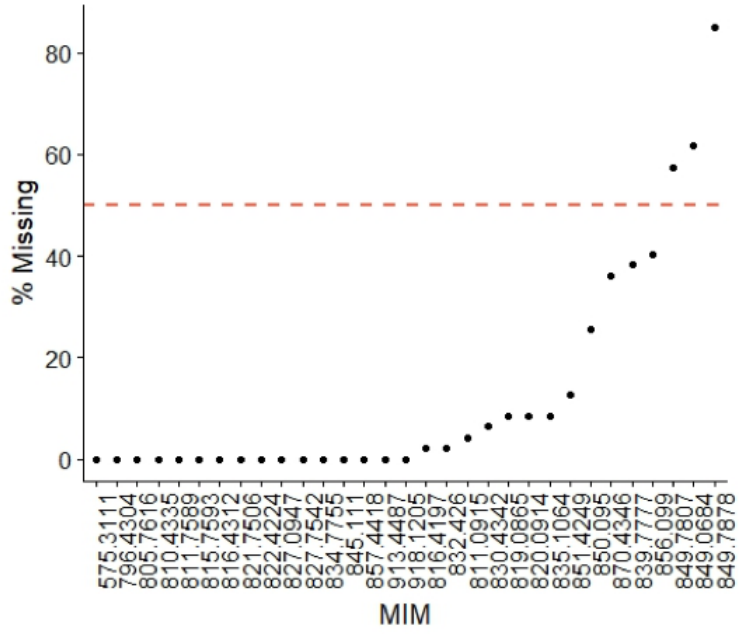


Figure 3.3: Percent missing values across subjects.

across duplicate injections, and no adducts are filtered based on this criterion. However, three adducts were removed from the analysis since they were missing in over half of the samples (including duplicate injections, Fig. 3.3). The means of the duplicate injections were then taken for each subject, ignoring missing values. This resulted in 47 subjects and 26 putative Cys34 adducts for analysis.

Missing values were imputed using a variation of k-nearest-neighbor imputation (k-NN), where the adducts are the neighbors and $k = 4$ [119]. k-NN imputation is a simple and intuitive approach, that was shown to perform well for similar high-dimensional data settings [120]. Pairwise distances between adducts were calculated using the Euclidean distance based on all non-missing values. When an adduct is not detected in a blood spot sample, the abundances of the k nearest adduct neighbors in that sample are averaged to impute the missing value. If, when imputing a missing value for a certain adduct in a certain sample, a nearest adduct neighbor is also missing in the sample, then the next nearest adduct neighbor is found and its value is used, and so on, until all k nearest neighbors have non-missing values in the sample. Adducts with non-detected values, have, on average, lower abundances. Therefore, to choose a suitable value of k for the imputation, low abundance adducts were randomly chosen to be made missing in certain samples, their values were imputed using several values of the k parameter, and the mean squared error was calculated for each k. The k with the smallest mean squared error, $k = 4$, was chosen (Fig. 3.4). Since the lower abundance adducts are the molecules prone to missing values, it is possible that the k-NN imputation method is causing an upward bias in the imputed values. However,

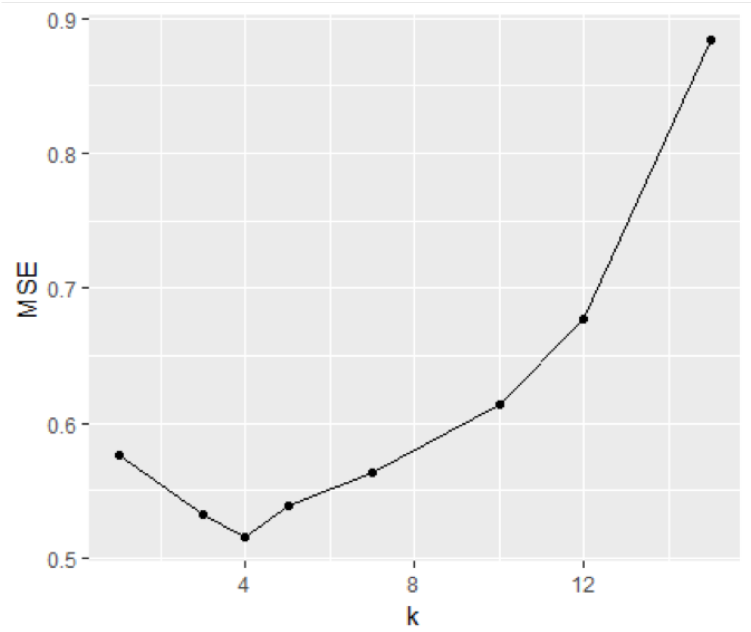


Figure 3.4: Plot of mean squared error values for different k.

there is no known limit of detection to use for imputation. Furthermore, since the nearest neighbors are very likely to have lower abundances, the upward bias should be small.

Adductomics data are complex and often suffer from various sources of unwanted variation. This unwanted variation, known or unknown, can bias subsequent statistical analysis. In the adductomic dataset of interest, there are several known sources of unwanted variation: batch, DBS age, digested HSA, blood volume, instrument performance, etc. (Fig. 3.5). Therefore, normalization is necessary to adjust for the systematic biases in adduct abundances introduced by a variety of sources of unwanted variation. Given the multitude of normalization schemes now available for 'omics' datasets, the task then becomes to assess the impact of each procedure and eventually select an appropriate procedure for the data at hand. To help answer this question, the Bioconductor R package *scone* was used to perform and evaluate a variety of normalization schemes on the dataset of interest [23]. While originally developed for single-cell RNA-Seq, *scone* implements the following normalization procedures that are immediately applicable to adductomic data:

- global-scaling normalization, e.g., upper-quartile, *DESeq* [3], *TMM* [103];
- full-quantile normalization;
- regression of scaled and logged feature abundances on
  - biological covariates of interest (e.g., disease status),
  - known factors of unwanted variation (e.g., batch),

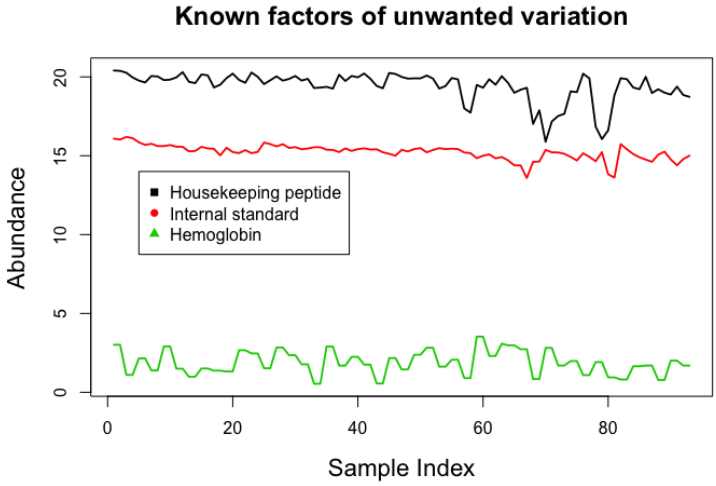


Figure 3.5: Matplot of various quality control metrics.

– estimated unknown factors of unwanted variation, as in *RUV* [101].

The *scone* package then evaluates each candidate scheme with metrics that gauge the removal of unwanted variation and retention of wanted variation. For example, one basis for evaluation in *scone* is the correlation between the first few principle components of the resulting normalized abundances with unwanted factors of variation and the tendency of samples to cluster by batch and the biology of interest after normalization.

For this dataset, the top ranking normalization scheme according to *scone* used *DESeq* scaling and removed unwanted variation due to the following factors: digested HSA, blood volume, DBS age, instrument performance, and batch effects. Here, digested HSA was quantified by the abundance of the tryptic housekeeping peptide adjacent to T3 with sequence 42LVNEVTEFAK51 (doubly charged precursor ion at m/z = 575.3111) [45] (black line in Fig. 3.5). Blood volume was indicated by measurement of Hb in DBS extracts (green line in Fig. 3.5). DBS age (i.e., 2017 – child birth year) was used to adjust for differences in the extraction efficiency due to the age of the DBS [89]. Instrument performance was indicated by the drift in the abundance of the internal standard over time (red line in Fig. 3.5). Batch effect was used to adjust for differences in the four subsets of samples that were prepared on different days (Fig. 3.8 (a)). To help verify the choice of normalization scheme, we examined the relationship between some of the suspected, known sources of unwanted variation and the estimated ones. There was a relatively strong negative correlation (Pearson's r = -0.56, Fig. 3.6) between Hb concentrations and the second estimated factor of unwanted variation using the *RUVg* method [101]. This further suggests that Hb is indeed a factor of unwanted variation in DBS analysis and should be included in the final normalization scheme. Furthermore, the weights of the DBS and Hb concentrations were highly correlated (Pearson's r = 0.93), suggesting that Hb is a good predictor of blood volume in newborn DBS.

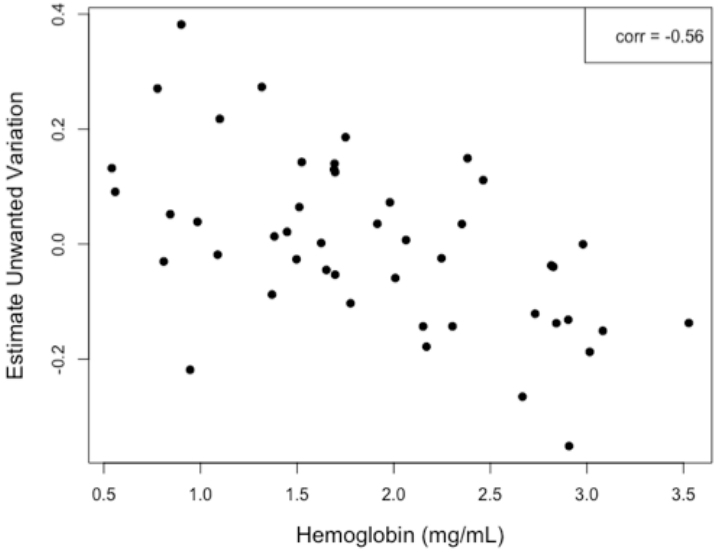


Figure 3.6: Correlation between Hb and second factor of unwanted variation.

Statistical analysis

Variable selection, as referred to here, is the process of identifying features which are associated to an outcome of interest. In this study, we want to identify adducts that are associated with mothers' smoking. A typical approach to variable selection in untargeted omics studies is through multiple hypothesis testing, where the importance of a feature is measured by a (possibly adjusted) p-value for the test of the null hypothesis that the feature is not associated with the outcome and where only the features whose p-values are below some predefined significance threshold are retained [90, 58]. However, commonly-used multiple testing approaches often lead to overly stringent filtering criteria for the features, because features can be highly correlated, etc. Additionally, such testing approaches only consider the marginal association of a feature with the phenotype of interest, rather than the joint effect of sets of features. As an alternative to the traditional multiple hypothesis testing paradigm, we favor variable selection strategies that assess variable importance based on prediction accuracy. We propose branching out from the commonly-used variable selection methods to explore other modern methods, including regularized logistic regression and regression trees and resampling-based aggregates of such methods.

We developed a variable selection procedure which combines three different data-adaptive regression methods, in order to obtain robust variable importance measures that account for both univariate and group-wise associations. Specifically, our variable selection approach relies on three measures of variable importance for the adducts: a p-value for each adduct based on the linear regression of that adduct's abundance measure on mothers' smoking status, a percentage of times each adduct is included in regularized logistic regression of

smoking status on the abundance measures for all adducts for bootstrapped datasets, and a random forest variable importance measure for each adduct when regressing smoking status on the abundance measures for all adducts. Adducts are then ranked by each of the above variable importance measures and the rankings combined to produce a final set of selected adducts. The final step of the variable selection procedure is to estimate the strength of the association of the selected features with the outcome of interest and to quantify the uncertainty of this estimate. A measure of prediction accuracy is found by calculating a cross-validated area under the receiver operating curve (AUC) estimate for smoking status prediction using the set of selected features, as well as a corresponding 95% confidence interval [61]. This cross-validated AUC estimate is likely optimistic, since the data were already used to perform the variable selection. However, given the limited size of the dataset, it was not feasible to split the data into an independent learning set for variable selection and testing set for prediction error estimation, i.e., perform a nested cross-validation.

Variable selection was performed on the logged, normalized peak areas and with the design matrix supplied by *scone*. First, the following multivariate regression model, corresponding to the top ranking normalization in *scone*, was used to find associations between each adduct's abundance and the mothers' smoking status:

$$Y_i = \beta_0 + \beta_1 X_{Smoke} + \beta_2 X_{Sex} + \beta_3 X_{Race} + \beta_4 X_{Batch} + \beta_5 X_{HK} + \beta_6 X_{IS} + \beta_7 X_{Hb} + \beta_8 X_{DBSAge} + \epsilon_i \quad (3.1)$$

where Y_i is a vector of logged, *DESeq* scaled abundances of the i^{th} adduct, X_{Sex} (0 = male, 1 = female) and X_{Race} (0 = other, 1 = white) are binary vectors, X_{Batch} is a four-level categorical variable indicating batch, X_{HK} is the vector of housekeeping peptide abundances, X_{IS} is the vector of internal standard abundances, X_{Hb} is the vector of Hb measurements, and X_{DBSAge} is a vector of DBS sample ages. The coefficients β_1 and estimated p-values were used to rank adducts by their univariate, linear associations with mothers' smoking status. The mean fold change (smoking/nonsmoking) in adduct intensities between newborns of smoking and nonsmoking mothers was calculated as $\exp(\beta_1)$.

Next, a logistic least absolute shrinkage and selection operator (lasso) [117] model was fitted to the logged and normalized adduct abundances, along with the matching variables (i.e., sex, birth year), to select a subset of adducts that best predicted the mothers' smoking status. To increase stability, the logistic lasso regression was performed on 500 bootstrapped data sets [9]. The percentage of times each adduct was selected by the lasso model out of the 500 iterations was used as a measure of variable importance. This process was performed for a range of values of the lasso penalty parameter (lambda range: 0.12-0.20) to ensure that the final variable selections were robust to this choice. Adducts were also ranked in terms of their associations with the mothers' smoking status using random forest variable importance [65]. A random forest with 500 trees was used to predict mothers' smoking status with the normalized, logged adduct abundances and matching variables. Adducts were ranked by the mean decrease in Gini index, which indicates the total decrease in node impurity (as measured by the Gini index when splitting on the adducts within the decision tree averaged over all trees) [17].

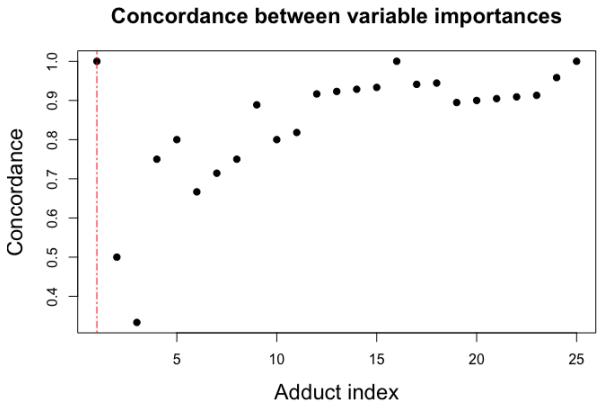


Figure 3.7: Concordance between linear model variable importances.

To select adducts, we first combine the rankings from the two linear models by taking the intersection of the corresponding top J_1 features. To choose the parameter J_1, it is helpful to consider the concordance or agreement between the two rankings. In theory, the two rankings should be in relative agreement for the top ranking features and then eventually begin to disagree more as uninformative features are incorporated into the rankings. To visualize this transition, we suggest using a concordance plot (Figure 3.7). A concordance plot shows, for each $j = 1, \ldots, J$, where J is the total number of features, the size of the intersection of the top j features in each ranking, divided by j, the number of features considered in each list. A cutoff should be selected where the concordance between the two linear methods peaks and/or stabilizes, before decreasing (Figure 3.7). After selecting adducts with a linear association with smoking status, we combine these adducts with those that have strong, non-linear associations as defined by random forest. Adducts that stand out in importance for classifying mothers' smoking status (e.g. have a 25-50% increase in importance) are added to the list of selected variables.

3.3 Results

Measurement of Hb in archived DBS

Our analysis was performed exclusively with newborn DBS that had been maintained in freezers at $-20\,^{\circ}C$ for 14 to 32 y prior to analysis. Quantitative analysis of Hb in DBS stored at room temperature can be problematic because of oxidation of Hb [14]. Indeed, when DBS were stored at room temperature for several months compared to storage at $-20\,^{\circ}C$, we observed that the color changed from deep red to dark brown indicating oxidation of Hb [14], and decreased water solubility of Hb (data not shown). The absorption spectra in the 250 – 750 nm range for the Hb calibration curve and Hb measured from extracts of

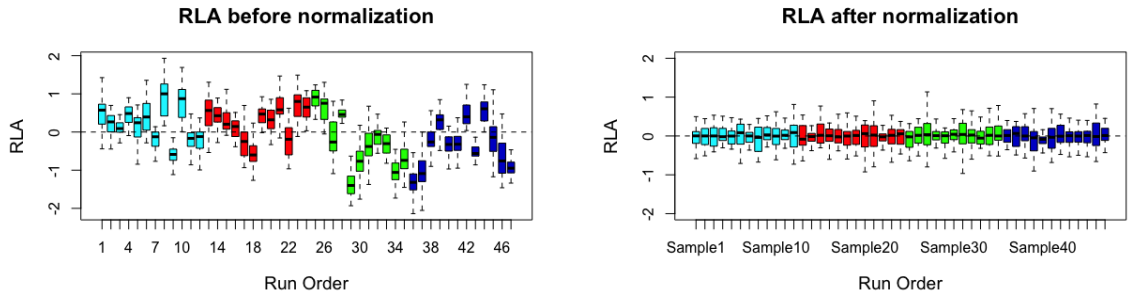
10 randomly selected 4.7-mm punches from archived newborn DBS both showed maximum absorbance at 407 nm. The Hb calibration curves measured for each of the four batches of newborn DBS showed a strong linear relationship between Hb concentrations and absorbance measurements at 407 nm ($r^2 > 0.99$).

Adductomics analysis of DBS

In preparing DBS for adductomics, HSA must first be extracted from the filter paper and isolated from whole blood. Previous analyses of proteins extracted from DBS have used mixed aqueous-organic solutions to selectively precipitate proteins in solution [34, 27]. Since Hb is one of the most prominent interfering proteins in whole blood, we tested various mixtures of organic solvents (ethanol, methanol, acetonitrile, and 1-propanol) to precipitate Hb while retaining HSA in solution (data not shown), and found ethanol and methanol to be most effective. When the concentration of ethanol and methanol were gradually increased from 30 to 60% (v/v), HSA remained in solution at concentrations less than 40% ethanol or 45% methanol and increasingly precipitated at concentrations up to 60% for both ethanol and methanol. Methanol was more effective at precipitating Hb with a 95% decrease in concentration at 45% methanol when compared to DBS extracted with water.

The recovery of HSA was also influenced by the total protein concentration of the DBS extract, where higher total protein concentrations led to a lower recovery of HSA after precipitation. Based on preliminary analysis of ten 4.7-mm punches from archived newborn DBS, we found that the average total protein concentration for newborns (4.98 mg/mL) was approximately equivalent to a 6-mm punch from an adult DBS. The observed higher total protein concentration of newborn blood reflects the higher Hb concentrations in newborns compared to adults [5]. Therefore, we used 6-mm punches from adult DBS to find the optimal concentration of methanol in the extraction mixture to isolate HSA. In addition, the isoelectric points (pI) of fetal and adult Hb differ (fetal Hb: pI 6.98, adult Hb: pI 6.87),41 and fetal Hb precipitates more readily at neutral pH. When comparing 40, 45, 48, and 50% methanol, we observed that Hb did not precipitate with 40% methanol and that there was a 40% loss of HSA when the methanol concentration was increased to 50%. Based on this result, we chose 45% methanol to isolate HSA in the DBS extract. We also found that incubating the samples at $37\,^{\circ}C$ (as opposed to room temperature) after addition of methanol to the aqueous DBS extract was essential for denaturing and precipitating Hb. When we tested extraction with 45% methanol on four 4.7-mm punches from archived newborn DBS, there was no loss of HSA and the residual Hb concentration was 0.02 mg/mL (1.2% of the initial concentration).

Digestion of HSA was optimized by testing various digestion programs using the pressure-cycling technology and by adjusting the proteolytic enzyme-to-protein ratio (E:P). While conventional proteomics approaches perform reduction and alkylation of proteins prior to digestion[52], we did not apply these techniques in order to preserve Cys34 disulfides and to prevent the formation of artifacts. The digestion time was tested at 16, 32 and 64 min to determine the optimal time needed for a high yield of digestion. Both the total ion



(a) Before normalization **(b)** After normalization

Figure 3.8: Relative log abundance plot of samples before and after normalization, colored by batch.

chromatogram and base peak chromatogram were examined for the presence of undigested proteins and yield of peptides [51]. While chromatograms from 32 and 64 min digests were comparable, there were fewer peptides and a more prominent peak for undigested proteins with the 16 min digestion, suggesting that 32 min was sufficient. Undigested protein was observed despite longer digestions, and probably reflects the lack of denaturation and reduction of disulfide bonds. It may also be due to the presence of residual non-HSA proteins, including Hb, which increase competition for trypsin cleavage sites and thereby interfere with the digestion of HSA [52]. We tested various 30 min digestion programs consisting of shorter and longer cycles at high pressure, but there was little difference in the resulting chromatograms (data not shown). The E:P was optimized to ensure an amount of trypsin that was sufficient for digestion while preventing autolysis [56]. When the E:P was increased from 1:18 to 1:3, trypsin activity showed a plateau at about 1:10, after which a further increase in trypsin did not improve the digestion. Increasing trypsin to a ratio of 1:5 resulted in incomplete digestion and more trypsin autolysis products, which could lead to ion suppression during MS detection.

Analysis of archived newborn DBS

The distribution of adduct peak areas in each sample before and after normalization for unwanted factors of variation (i.e., Hb concentration, DBS age, housekeeping peptide, internal standard, and batch effects) is shown in Figure 3.8. By comparing the RLA plots from before (Fig. 3.8 (a)) and after (Fig. 3.8 (b)) normalization, it can be seen that this scheme effectively removed unwanted variation.

Nineteen of the 26 adducts have been previously reported, including truncations, unmodified T3, methylated T3, Cys34 sulfoxidation products (e.g., sulfinic and sulfonic acids), a cyanide adduct, and Cys34 disulfides of low-molecular-weight thiols [45, 69, 44]. Only three of the remaining 7 adducts had putative annotations, i.e., the Cys34 sulfenamide (811.09), a CH2 crosslink (815.76), and the sodium adduct (819.09). Aside from the unmodified T3

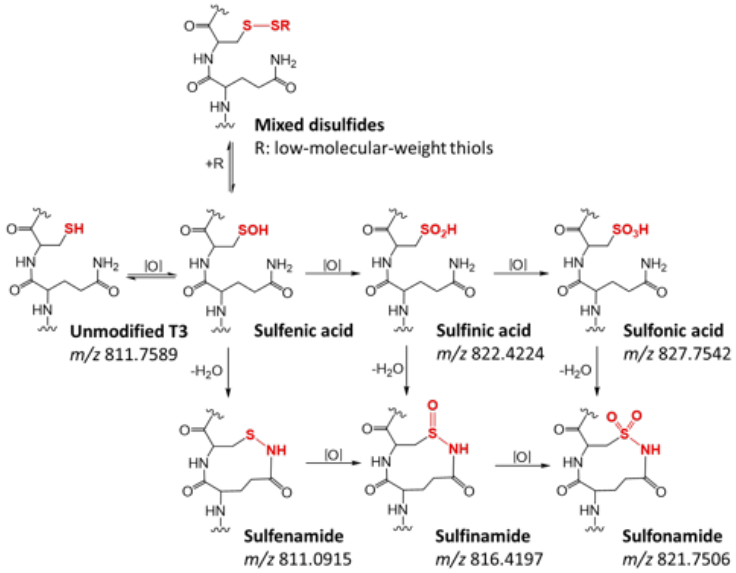
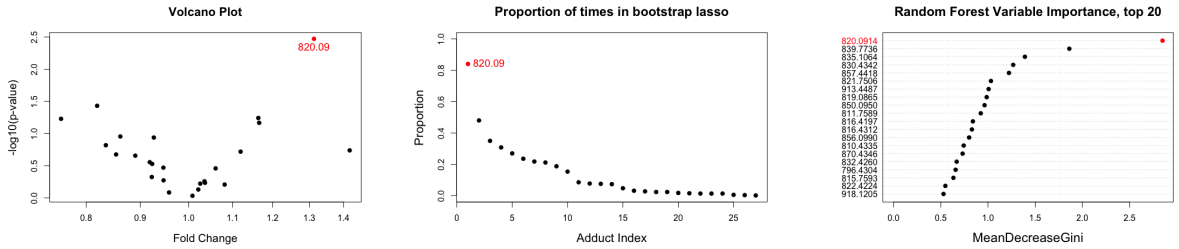


Figure 3.9: Reaction pathways proposed for the formation of Cys34 oxidation products.

peptide (811.76), the Cys34 sulfinic acid (822.42) and the S-glutathione (GSH) disulfide (913.45) were the most abundant adducts across all samples.

In studying reaction pathways leading to Cys34 sulfoxidation products, Grigoryan et al. reported an intramolecular cyclic sulfinamide adduct (816.42) with the added mass $(+O, -H_2)$, which results from the formation of a cross-link between Cys34 and the amide group of the adjacent Gln33 [46]. Two different pathways were proposed for the formation of the sulfinamide adduct: (1) from dehydration of Cys34 sulfenic acid (Cys34-SOH) resulting in the sul-fenamide adduct (mass difference $[-H_2]$) with the Cys34-Gln33 cross-link, which is then oxidized to the sulfinamide adduct; (2) from oxidation of the sulfenic acid to the sulfin-ic acid (822.42, Cys34-SO2H), from which loss of water results in the sulfinamide adduct (Fig. 3.9). The second reaction pathway appeared to be more likely because the intermediate sulfenamide adduct had not been detected in plasma/serum samples [45, 69, 46]. However, in newborn DBS we detected both the sulfenamide (811.09) and sulfinamide (816.42) adducts, suggesting that formation of the sulfina-mide adduct is possible via both pathways. In addition, we detected the sulfonamide adduct (821.75, added mass $[+O_2, -H_2]$), which results from oxidation of the sulfinamide adduct (Fig. 3.9) [38]. The formation of these intramolecular cyclic adducts may have been promoted by the drying of DBS which could have led to the dehydration of sulfenic, sulfinic, and sulfonic acids to produce the sulfenamide, sulfinamide, and sulfonamide adducts, respectively (Fig. 3.9). It is also possible that these intramolecular cyclic adducts (particularly sulfenamide) were detected in the present analysis due to an increased stability of these analytes in DBS compared to plasma and serum. Analytes in DBS are typically less reactive than in liquid blood because they are stabilized through adsorption onto a solid cellulose matrix (i.e., the filter paper)



(a) Volcano plot **(b)** Lasso **(c)** Random Forest

Figure 3.10: Variable selection results for mothers' smoking status.

[125]. Proteins commonly degrade in aqueous solutions due to aggregation, oxidation, and hydrolysis48 that appear to be minimized during long-term storage of DBS in a freezer. In fact, we did not observe the T3 dimer in our analysis of newborn DBS although this dimer is routinely detected in plasma/serum samples [45, 69, 68, 44]. Furthermore, adducts in DBS may be less susceptible to formation of artifacts because HSA is immobilized by the filter paper and less likely to interact with other molecules during storage.

Adducts that discriminated newborns of smoking and nonsmoking mothers

We had anticipated that Cys34 adducts of two toxic contaminants of cigarette smoke, i.e., ethylene oxide and acrylonitrile, might be more abundant in newborns of smoking mothers given our previous detection of these adducts in plasma from adult smokers and their absence in plasma from nonsmokers [45]. However, these adducts were not detected in the newborn DBS, possibly due to low concentrations of the precursor molecules in newborn blood. Mothers may have stopped smoking during the third trimester or before the last month of pregnancy, and this may explain why we did not see all of the expected adducts in the present analysis.

One adduct, the Cys34 adduct of cyanide (820.09), was selected by the ensemble method as predictive of mothers' smoking status (Fig. 3.7 and Fig. 3.10). Of the 26 adducts that were tested, the Cys34 adduct of cyanide (820.09) was ranked the highest by all three statistical methods. Although a feature does not have to be top-ranking for all *three* methods in order to be selected, the relationship between cyanide and mothers' smoking status is strong enough that this adduct is top ranking for all three methods. As seen from the volcano plot, which shows the relationship between the smoker/nonsmoker fold change of a given adduct and the statistical significance of the difference in adduct abundance obtained from Equation 3.1, the cyanide adduct showed a marked difference between newborns of smoking and nonsmoking mothers, with a smoker/nonsmoker mean fold change of 1.31 (nominal p-value = 0.0017, Fig. 3.10 (a)). The cyanide adduct was also top-ranked by both the lasso model (Fig. 3.10 (b)) and random forest (Fig. 3.10 (c)).

To test whether the cyanide adduct could be used to distinguish between newborns based on maternal smoking status, we performed a ROC analysis using logistic regression with the cyanide adduct as the predictor. The cyanide adduct alone provided a cross-validated AUC of 0.79 (95% CI: 0.65, 0.93). Although the AUC estimate is likely to be optimistic since we did not have an independent test set for the ROC analysis, this indicates good discrimination between newborns of smoking and nonsmoking mothers. The elevated levels of the cyanide adduct among newborns of smoking mothers are consistent with inhalation of hydrogen cyanide from tobacco smoke [114]. The half-life of cyanide in blood is less than one hour, which makes it difficult to obtain accurate measurements of cyanide exposure from the direct analysis of blood from smokers [124]. While more stable metabolites of cyanide, such as thiocyanate, are often used as surrogate measures of cyanide exposure, pairwise correlations are small between such metabolites and cyanide exposures [124, 31]. Since the residence time of HSA is about 1 month [94], the cyanide adduct of Cys34 arguably represents a more accurate measure of chronic low-level exposure to cyanide.

Using adductomics to discover biomarkers of *in utero* exposures

Adducts of HSA represent biomarkers of in utero exposures during the last month of gestation. A good example of such exposures is maternal smoking during pregnancy, which has been consistently associated with increased risks of adverse birth outcomes (e.g., low birth weight, preterm birth) [28] and has also been suggested to increase the risk of diseases later in life, including various types of cancer [57, 108, 78, 39]. However, the long term effects of in utero tobacco-smoke exposures on the risk of childhood cancer have been inconsistent, with many studies reporting null associations [82]. One limitation of these epidemiological investigations has been reliance on maternal self-reports to retrospectively characterize fetal exposures to tobacco smoke [82, 128]. Exposure misclassification due to recall and reporting bias is a particular concern among pregnant women, who may feel uncomfortable discussing their smoking histories during pregnancy, and can result in underestimation of fetal health effects from smoking mothers [32]. Biomarkers complement interview-based exposure assessment by providing objective measures of exposure that are not susceptible to recall bias. Nicotine and its metabolite cotinine are commonly measured in biological fluids (e.g., urine, blood, saliva) to assess tobacco smoke exposures [32]. For retrospective analyses of fetal exposures, archived newborn DBS are particularly attractive because they are readily available in California's repository that is maintained at $-20\,^{\circ}C$. In addition, newborn DBS enable direct measurement of fetal exposures that can accumulate in the placenta and exceed those of the mother [50]. Metabolites of nicotine, mainly cotinine, have been measured in newborn DBS to improve smoking surveillance among pregnant women [111, 133]. However, the half-life of cotinine is only about 28 h in infants, and cotinine may only be detected in newborns of heavy smokers who smoke throughout pregnancy [110]. Since the residence time of HSA is 28 days [94], Cys34 adducts detected in newborn DBS represent exposures received during the last month of gestation and are only marginally affected by the day-to-day variability in exposure [93]. In the present study, the Cys34 cyanide adduct discriminated between

mothers who reported smoking during pregnancy vs. those who did not, suggesting that maternal self-reported smoking was reliable in the 47 subjects tested.

3.4 Discussion

With careful data exploration and visualization, various sources of technical noise were identified and managed in this study. Extensive plotting of sample and adduct characteristics, as well as quality control metrics, uncovered then challenges (and their respective solutions) behind analyzing DBS adductomics data. This study helped to lay the groundwork for future analysis of DBS adductomics data. For example, it proved the utility of the *scone* framework for normalizing adductomic data, and discovered sources of unwanted variation to be aware of in future work. Indeed, the *scone* framework has also been successfully applied to untargeted metabolomics data for normalization of considerably higher dimensional data [88, 86]. Proper data pre-processing of the adductomics dataset allowed for discovery of a biomarker of mothers' smoking status that has considerable predictive ability for a single molecule. The technical noise present in the data was reduced sufficiently to uncover this relationship in all three measures of variable importance.

Chapter 4

Data-adaptive Filtering in Untargeted Metabolomics

Many of the exploratory and pre-processing data analysis methods for adductomics are also applicable to untargeted metabolomics studies. For example, the ensemble variable selection method discussed in the previous section has also been used to identify metabolites that discriminate between incident childhood Leukemia cases and controls [88]. The *scone* normalization framework has also been used to remove unwanted variation due to sample contaminants, batch effects, machine performance, blood volume, etc. in untargeted metabolomics [86, 88, 89]. Here, we focus on a pre-processing issue in untargeted metabolomics that is not present in adductomics, feature filtering.

4.1 Background

Metabolomics represents the small-molecule phenotype that can be objectively and quantitatively measured in biofluids such as blood serum/plasma, urine, saliva, or tissue/cellular extracts [95, 126, 19, 30]. Untargeted metabolomics studies allow researchers to characterize the totality of small molecules in a set of biospecimens and thereby discover metabolites that discriminate across phenotypes [19, 95, 106]. Among the techniques employed for untargeted metabolomics, liquid chromatography-high-resolution mass spectrometry (LC-HRMS) has become the analytical tool of choice due to its high sensitivity, simple sample preparation, and broad coverage of small molecules [126, 112]. However, many of the thousands of features detected by untargeted metabolomics are not biologically interesting because they represent background signals from sample processing or multiple signals arising from the same analyte (adducts, isotopes, in-source fragmentation) [72]. Furthermore, feature detection and integration with software such as *XCMS* [109] is imperfect, in that noise can erroneously be identified as a peak group, the domain of integration can be incorrect, etc. Thus, large metabolomics datasets can contain thousands of falsely identified features or features with imperfect integration (e.g., incorrect integration regions and missing values).

Inadequate feature filtering can affect subsequent statistical analysis. For example, if high quality features are erroneously filtered, they will not be considered as candidate biomarkers in univariate tests of significance for association with biological factors of interest or in metabolic pathway analysis. Furthermore, if one performs univariate tests of significance and ranks features based on p-values, biologically meaningful features could be lost in an abundance of noise without adequate feature filtering. Failure to filter noise could also result in false positives when assessing the significance of metabolic pathways with software such as *Mummichog*, which relies on sampling features from the entire dataset to create null distributions of pathway statistics [64].

Therefore, untargeted metabolomic data require a set of filtering methods to remove noise prior to investigating the biological phenomena of interest. Data normalization has received a lot of recent attention in untargeted metabolomics [36, 25, 21, 81, 71]. Feature filtering, however, remains a fairly automated, indelicate, and brief step in the preprocessing of untargeted metabolomic data. Many studies rely on valuable preprocessing pipelines offered from programs like *Metaboanalyst* and *Workflow4Metabolomics* to process their raw data. Such programs have greatly advanced the field of untargeted metabolomics and have improved data pre-processing and analysis and replication of results. However, many users of these programs rely heavily on the provided, default cutoffs for feature filtering that are largely independent of their data, and do not attempt to identify appropriate, data-specific filtering cutoffs. Thus, improper feature filtering in untargeted metabolomics is in part due to user error in pre-processing pipelines.

For example, *MetaboAnalyst* allows users to filter features based on mean/median value across samples, as wells as variability across biological samples and quality control (QC) samples. While these are indeed useful filtering metrics, most users do not determine the filtering thresholds appropriate for their specific data. *Metaboanalyst* suggests removing the lowest k percent of features based on the size of the dataset (e.g., lowest 40% of features for a dataset with more than one thousand features based on mean/median abundance across samples), and a relative standard deviation (RSD, the same as a coefficient of variation or CV) cutoff of 25% for LC-MS data [21]. While these are helpful guidelines for selecting cutoffs, users often fail to investigate if they are appropriate for their data. Similarly, *Workflow4Metabolomics*, for good reasons, allows users to filter features based on variability across replicates and sample mean vs. blank mean ratios, but many users select default or commonly used cutoffs. We recognize that it is tempting for users to rely on default filtering cutoffs without consulting their data, and we aim to assist researchers in selecting more appropriate cutoffs.

We argue that filtering methods should be data-adaptive. A data-adaptive pipeline is one which tailors filtering to the specific characteristics of a given dataset, rather than using predefined methods. In what follows, we present a series of steps (Fig. 4.1) representing a data-adaptive pipeline for filtering untargeted metabolomics data prior to discovering metabolites and metabolic pathways of interest. Our data-adaptive filtering approach contains novel methods for removing features based on blank sample abundances, proportions of missing values, and estimated intra-class correlation coefficients (ICC). To create data-

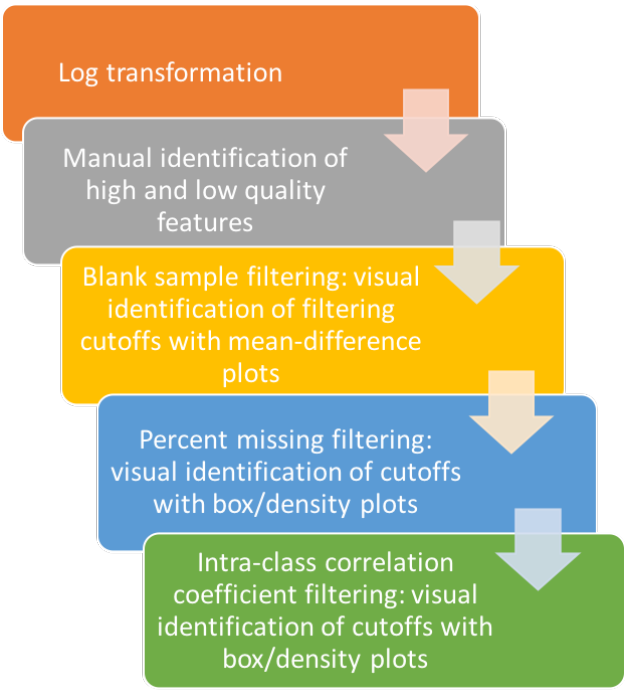


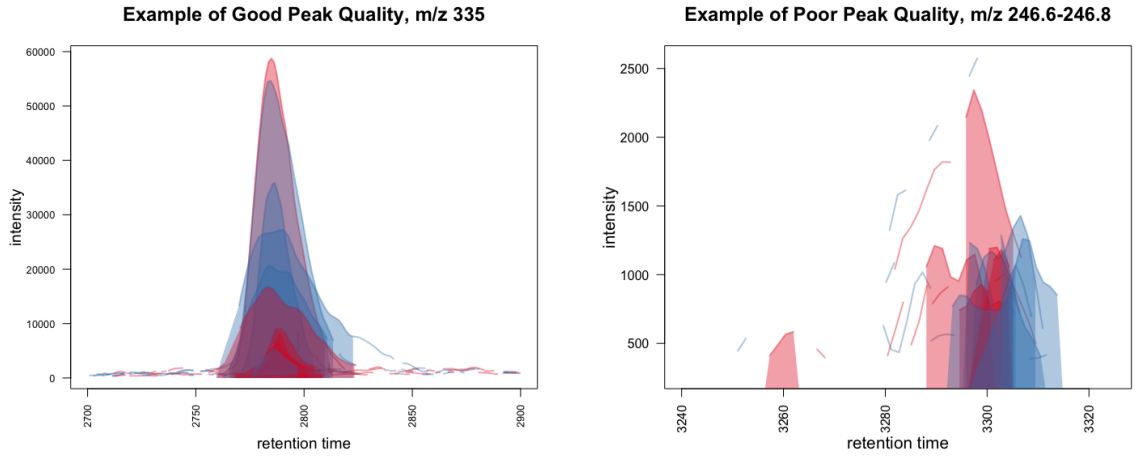
Figure 4.1: Flowchart of a data-adaptive filtering pipeline for untargeted metabolomics data.

dependent thresholds for the above three feature characteristics, we propose visualizing the differences in the characteristics between known high and low quality features. By examining such differences for each dataset, one can minimize noise without compromising biological signal. Once this is done for several datasets generated from a given laboratory, the determined filtering cutoffs can likely be applied to all such similar datasets. Properly filtered untargeted metabolomic data can then be used as input into valuable processing pipelines such as *MetaboAnalyst* and *Workflow4Metabolomics* for further preprocessing such as data normalization. We compare our data-adaptive filtering method to common filtering methods using an untargeted LC-HRMS dataset that was generated in our laboratory and two public LC-MS datasets. To compare the methods, we identified hundreds of high and low quality peaks in each dataset. We then showed how our data-adaptive pipeline surpasses workflows that use default cutoffs at removing the low quality features and retaining high quality features.

4.2 Methods

Visualizing high and low quality features

When working with untargeted LC-MS data, visualization of extracted ion chromatograms (EIC) of features can be used to optimize peak detection, peak quantification, and biomarker



(a) High Quality Feature **(b)** Low Quality Feature

Figure 4.2: Example of a high and low quality peak group.

discovery [88, 86, 109]. We propose randomly sampling several hundred EICs after peak detection and quantification to visualize peak morphology and integration. The EICs can then be classified by the user as "high" or "low" quality (see Fig. 4.2). A high quality peak has good morphology (e.g. bell-shaped), the correct region of integration across all samples, and proper retention time alignment. Such visualization is made easy with plotting functions from peak detection software such as the 'highlightChromPeaks' function within *XCMS* [109]. In almost all cases, we find the distinction between high and low quality peaks to be clear, and classify any ambiguous peaks as low quality to be conservative. Once features are classified as high or low quality across samples, their characteristics such as average blank and biological sample abundance, percent missing, and ICC can be compared and used to perform feature filtering. We recognize that the identification of high and low quality peaks is the most time intensive step of the proposed filtering pipeline. However, with parallelization of the plotting task, we have found that visualization and quality inspection of hundreds of features takes between 1-2 hours. Moreover, after feature visualization, executing the remaining steps of the filtering pipeline should take no more than 1 hour. Compared to the time spent struggling to uncover biological signal with improperly filtered data, we find this step well worth the added work.

Data-adaptive feature filtering

Example datasets

To help present and visualize our data-adaptive feature filtering methods, we introduce an untargeted LC-HRMS dataset generated in our laboratory on a platform consisting of an Agilent 1100 series LC coupled to an Agilent 6550 QToF mass spectrometer. The dataset

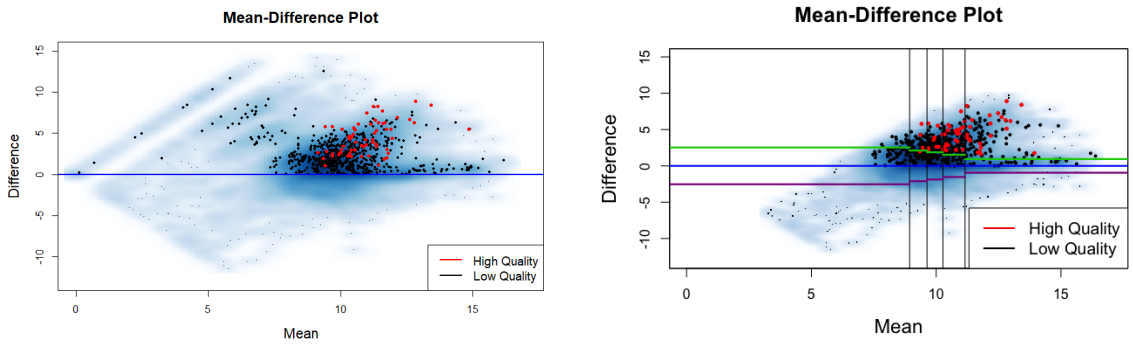
contains the metabolomes of 36 serum samples from incident colorectal cancer (CRC) case-control pairs as described in [86, 87]. Over 21,000 features were detected in the 36 serum samples that were analyzed in only one batch [86, 87]. We randomly sampled over 900 features from the dataset and classified these as "high" or "low" quality according to their peak morphology and integration quality. To demonstrate the performance of our data-adaptive pipeline, we split the known high and low quality features into a training set (60%) and a test set (40%). Features in the training set are used to visualize appropriate, data-dependent cutoffs, whereas features in the test set will be used to evaluate the effectiveness of the selected cutoffs. At each stage of the data-adaptive filtering, we compared our method to more traditional filtering methods by examining what proportion of high and low quality features in the test set were removed.

We also visualized and classified over 200 features in each of two public LC-MS datasets. One of the public datasets was generated on a platform consisting of an Accela liquid chromatographic system (Thermo Fisher Scientific, Villebon-sur-Yvette, France) coupled to an LTQ-Orbitrap Discovery (Thermo Fisher Scientific, Villebon-sur-Yvette, France). This dataset contains the metabolomes of 189 human urine samples. We took a subset of 45 of the urine samples in the first batch, along with 14 pooled QC samples and 5 blank samples. We processed this dataset using the original *xcms* functions and parameters used by the authors (W4M00002_Sacurine-comprehensive) [36, 116]. The second public dataset was generated on a platform consisting of an Accela II HPLC system (Thermo Fisher Scientific, Bremen, Germany) coupled to an Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific) [92]. This dataset contains the metabolomes of epithelial cell lines treated with low and high concentrations of chloroacetaldehyde. We used all 27 cell line samples in negative mode treated with low concentrations, as well as 6 pooled QC and 11 blank samples. The original work did not use *xcms* to process the raw data, so we used the R package *IPO* to determine the *xcms* parameters [66].

Filtering features based on blank samples

Blank control samples, which are obtained from the solvents and media used to prepare biological samples, can help to pinpoint background features that contribute to technical variation [85, 19, 126, 49, 36]. A common filtering method is to use a fold-change (biological signal/blank signal) cutoff to remove features that are not sufficiently abundant in biological samples [19, 36, 21]. Rarely does the user examine the data to determine a suitable cutoff. We employ a data-adaptive procedure that takes into account the mean abundance of features in blank and biological samples, the difference between mean abundances in blank and biological samples, and the number of blank samples in which each feature is detected. Our method then assigns cutoffs according to the background noise and average level of abundance. If the dataset contains several batches, filtering is performed batch-wise.

We use a mean-difference plot (MD-plot) to visualize the relationship between feature abundances in the blank and biological samples and assess background noise (Fig. 4.3). First, abundances are log transformed prior to all data pre-processing and visualization. The mean

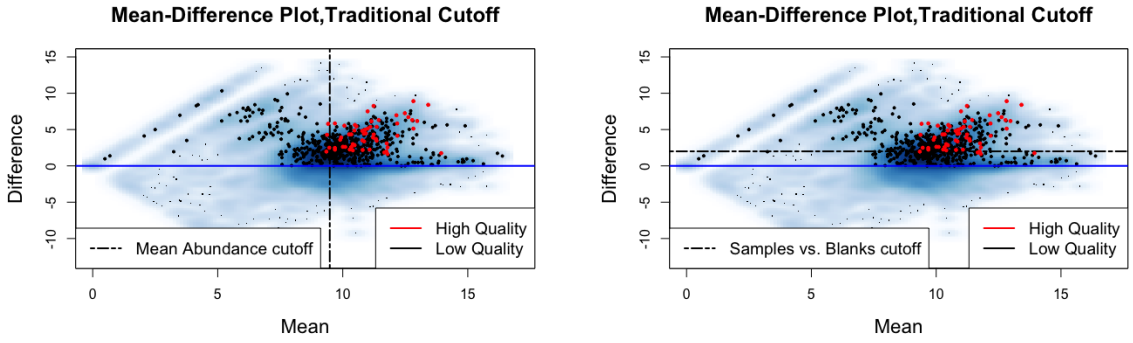


(a) All Features **(b)** Features Detected in 3 Blanks

Figure 4.3: MD-plot for the CRC dataset.

log abundances of each feature in the biological and blank samples are then calculated and the average of and difference between these two means are then plotted on the x- and y-axes, respectively. The horizontal zero-difference line (blue lines in Fig. 4.3) represents the cutoff between features having higher mean abundances in the blank samples and those having higher mean abundances in the biological samples. If there are n blank samples in a batch, then $n + 1$ clusters of features will typically be visually identifiable in the MD-plot, where cluster $i = 0, \ldots, n$ is composed of features that are detected in i blank samples. For example, because three blank samples per batch were used in the example dataset, four clusters are identifiable in Fig. 4.3 (a). Similar clusters can be identified in all datasets generated from our laboratory and in the public datasets. Filtering is then performed separately for each cluster. If a cluster contains no high quality features, as is often the case with clusters that contain lower abundance features, that cluster can be removed entirely.

The cluster corresponding to features detected in all n blank samples tends to have the highest number of features (around 95% of the total number of features), features with higher average abundances, and the highest number of high quality features. Therefore, careful, data-dependent filtering of this cluster is crucial for the success of subsequent analyses. This cluster also has a non-uniform distribution of mean feature abundances (Fig. 4.3 (b)). This cluster is thus partitioned based on quantiles (20th, 40th, 60th, and 80th percentiles) of the empirical distribution of mean abundances (x-axis). This ensures that each partition has the same number of features and that the features are uniformly distributed throughout the dynamic range. Within each partition, the empirical distribution of abundances below the zero-difference line is used to estimate the technical variation above that line. The absolute value (green lines in Fig. 4.3 (b)) of an appropriately identified percentile of the negative mean differences (purple lines in Fig. 4.3 (b)) is used as a cutoff to remove uninformative features. Users may identify appropriate percentiles of the negative mean differences (purple lines) based on how many high quality features would be removed if the absolute values of those percentiles (green lines) were used as cutoffs. We find percentiles between the lower



(a) Average Value Cutoff **(b)** Blank Sample Cutoff

Figure 4.4: Two traditional filtering cutoffs.

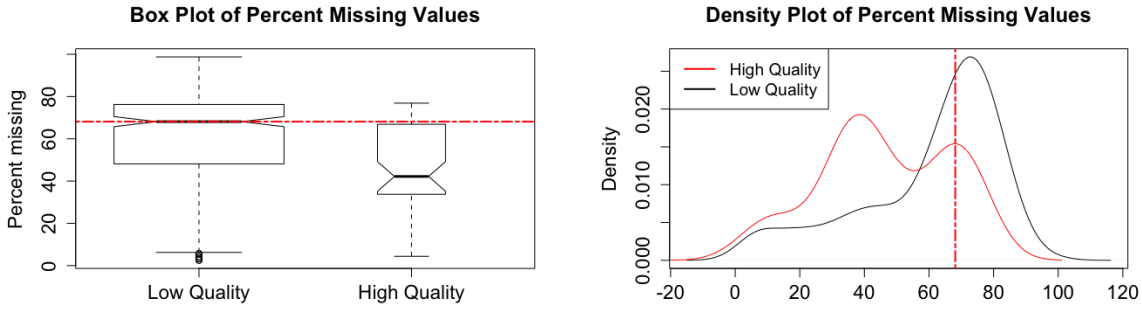
quartile and median to be appropriate for this cluster of features, because they remove as many low quality features as possible without removing high quality ones. Feature filtering in the remaining clusters can be performed in a similar manner, but without the need to partition features based on average abundance.

Using MD-plots to filter features allows for the simultaneous filtering of features by both the difference in abundance in blank and biological samples (y-axis) and average abundance (x-axis). Average abundance of features across biological samples is a commonly used filtering characteristic, but the filtering is often done using pre-specified cutoffs (e.g., lowest forty percent for datasets with more than one thousand features) (Fig. 4.4 (a)) [21, 36]. Although we advocate for the filtering approach described previously, if users prefer to filter by just average abundance, the MD-plot allows for easy visualization of a data-dependent cutoff that removes as many low quality features as possible without removing high quality ones. The same can be said for identifying a data-adaptive fold-change (biological signal/blank signal) cutoff, rather than using default cutoffs provided in preprocessing workflows (Fig. 4.4 (b)) [36]. We note that, although it is possible for background signal to modify biological signal (e.g., via ion suppression), we do not consider this source of variability.

Filtering features by percent missing

As mentioned above, low-abundance metabolomic features tend to have a high proportion of undetected values across samples. In addition, when using software such as *XCMS* for peak detection and quantification, oftentimes peaks can be missed by the first round of peak detection and integration. Functions such as 'fillChromPeaks' in *XCMS* are often used to integrate signals for samples for which no chromatographic peak was initially detected [109, 21]. Low quality peaks tend to have higher proportions of missing values on average after initial peak identification and integration (Fig 4.5).

To determine the appropriate filtering cutoff for percent missing, we create side-by-side



(a) **(b)**

Figure 4.5: Distributions of percent missing for high and how quality peaks in the training set

boxplots of percent missing values for the high and low quality features classified as such by EIC (Fig. 4.5 (a)). The boxplots help to compare the percentiles of the distributions of percent missing values for the high and low quality features, and to select an appropriate cutoff based on these percentiles. Density plots of percent missing values can also be used to visualize the modes and percentiles of the distributions for high and low quality features (Fig. 4.5 (b)), and cutoffs can be determined based on these distributional properties. For example, appropriate cutoffs would be those that discriminate between the modes of the two distributions, that remove long tails of distributions of low quality features, that correspond to extreme percentiles of one distribution but intermediate percentiles of another, etc. To ensure that we do not remove features that are differentially missing between biological groups of interest (e.g., mostly missing in cases but not controls), we perform a Fisher exact test for each feature, comparing the number of missing and non-missing values against the biological groups of interest. A small p-value for a given feature would indicate that there is a significant dependence between the phenotype of interest and missing values. Features with a percent missing below the identified threshold or with a Fisher exact p-value less than some threshold (we recommend a small value such as the one hundredth percentile of the p-value distribution) are retained. This test of association between the phenotype of interest and missing values can easily be extended to studies where the biological factor of interest is a multilevel categorical variable or a continuous variable by using, for example, a Chi-Square test or a Wilcoxon rank-sum test, respectively.

Filtering features by ICC

High quality and informative features have relatively high variability across subjects (biological samples) and low variability across replicate samples [36, 21] (Fig. 4.7). Typically, the coefficient of variation (CV) is calculated across pooled QC samples for each feature and those with a CV above a predetermined cutoff (e.g., 20–30%) are removed [95, 126, 85, 21,

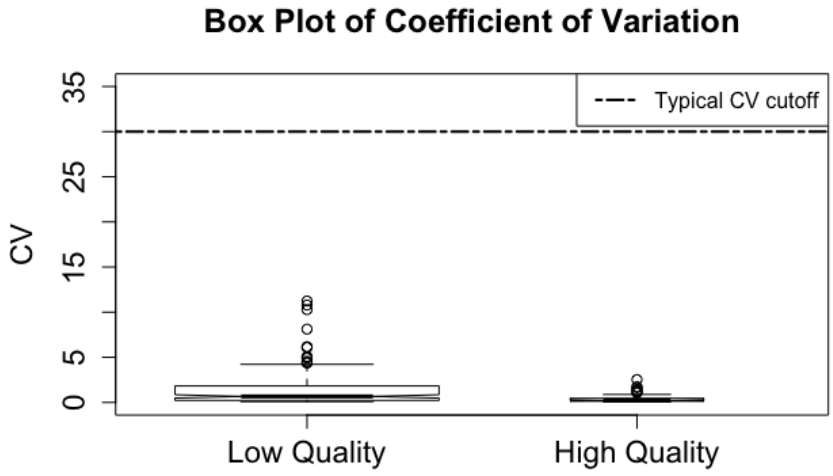


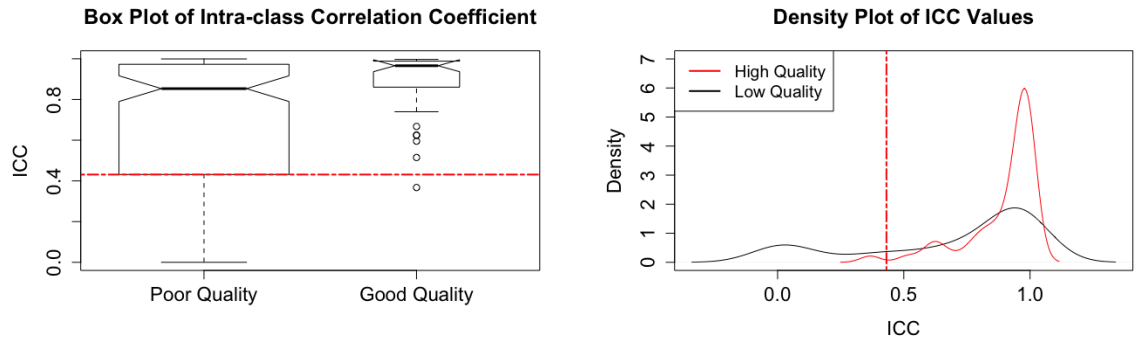
Figure 4.6: Box plot of CV values in the CRC dataset.

36]. However, we find that the CV is often a poor predictor of feature quality (Fig. 4.6) because it only assesses variability across technical replicates, without considering biologically meaningful variability across subjects. Instead, we propose examining the proportion of between-subject variation to total variation, otherwise known as the intra-class correlation coefficient (ICC) [107], as a characteristic for filtering. Since the ICC simultaneously considers both technical and biological variability, a large ICC for a given feature indicates that much of the total variation is due to biological variability regardless of the magnitude of the CV.

Our method for estimation of the ICC employs the following random effects model:

$$Y_{i,j} = \mu_j + b_{i,j} + \epsilon_{i,j,k}, \tag{4.1}$$

where $Y_{i,j}$ is the abundance of feature j in subject i, μ_j is the overall mean abundance of feature j, $b_{i,j}$ is a random effect for feature j in subject i, and $\epsilon_{i,j,k}$ is a random error for replicate measurement k for feature j in subject i. The ICC is estimated by taking the ratio of the estimated variance of $b_{i,j}$ (between-subject variance) to the estimated variance of $b_{i,j} + \epsilon_{i,j,k}$ (total variance). If replicate specimens or LC-MS injections are analyzed for each subject, then application of Equation (4.1) is straightforward. However, since metabolomics data are often collected with single measurements of each biospecimen and employ repeated measurements of pooled QC samples to estimate precision, then Equation (4.1) can be fit by treating the pooled QC samples as repeated measures from a 'pseudo-subject'. As with percent missing, density plots and boxplots of the estimated ICC values for high and low quality features can be compared to determine a data-specific filtering cutoff (Fig. 4.7).



(a) **(b)**

Figure 4.7: Distributions of estimated ICC values for high and low quality peaks in the training set

Again, we look to the modes and percentiles of the distributions of the high and low quality features to select an appropriate cutoff that strikes a balance between removing low quality peaks and retaining high quality ones. If multiple batches are involved, the final feature list represents the intersection of features from all batches.

4.3 Results and discussion

The MD-plot shows that all high quality features in the training set are in the same cluster corresponding to features detected in all three blank samples (Fig. 4.3). Because features in this cluster have higher average abundances and lower percent missing than those in the other three clusters, it is not surprising that this cluster is comprised of many high quality peaks. We therefore remove features in the other three clusters for this dataset, and focus on the data-adaptive filtering of the cluster containing the high quality features (Fig. 4.3 (b)). We use the lower-quartile of noisy features below the zero difference line to estimate the noise above the zero difference line because this cutoff removes a considerable number of low quality features without removing many of the high quality features (Fig. 4.3 (b)). In fact, this filtering step removes 68% of the 21,000 features, and 41% of the identified low quality features in the test set (Fig. 4.8). Almost all (95%) of the high quality features in the test set are retained (Fig. 4.8). A common approach to filtering would be to remove features based on their mean abundance, such as removing the lowest 40% [21]. If this threshold was used to filter the CRC dataset, only 31% of the identified low quality features in the test set would be removed, and many remaining features would have higher average abundance in the blank samples (Fig 4.4). Another traditional approach is to arbitrarily select a cutoff (2–5) for the ratio between average biological and blank sample abundances. A similar cutoff applied to the CRC dataset (a cutoff of two for the difference between average log abundances in biological and blank samples) would remove only 36% of the low

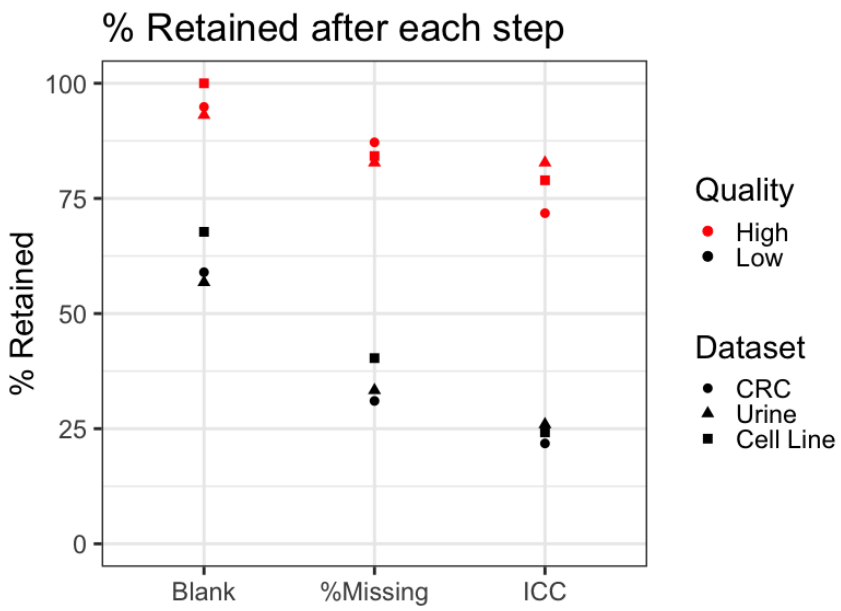


Figure 4.8: Percent of high and low quality features in the test set remaining after each filtering step.

quality features in the test set and 10% of the high quality features in the test set, and fails to remove many of the low quality features in the clusters removed by our data-adaptive filtering. Utilizing blank samples in filtering certainly helps to reduce the number of low quality features. Furthermore, utilizing data visualization helps to ensure that filtering is done appropriately, i.e. that an appropriate balance is struck between removing low quality features and retaining high quality ones.

The next step in the data-adaptive filtering is to visualize differences in percent missing among the remaining high and low quality features (Fig. 4.5). Using the information on distribution modes and percentiles provided by boxplots and density plots of the data in the training set, we chose to remove features with more than 68% missing values (median of percent missing for low quality features). When a Fisher exact test was used for each feature to detect significant associations between missing values and the biological factor of interest (CRC), 68 features had p-values less than 0.027 (the one hundredth percentile of the p-values) and were retained regardless of their percent missing values. Combining these two filtering criteria removed 47% of the remaining low quality features and only 11% of the remaining high quality features in the test set (Fig. 4.8).

We used the 12 QC samples from the CRC dataset to calculate ICC values for each of the remaining features. Using the information provided by the density and boxplots, we chose to remove features with ICC values less than 0.43 (the lower hinge of the box plot for low quality features in the training set) (Fig. 4.7). This removed 23% of the remaining low

quality features in the test set and only 15% of the remaining high quality ones (Fig. 4.8). Compare this to using CV values to perform filtering, where a typical CV cutoff of 30% or even 20% (Fig. 4.6) [36] results in no further filtering of the remaining low quality features in the test set. With all steps of the data-adaptive pipeline, the CRC dataset was reduced to just $3,009$ features. The data-adaptive filtering removed 76% of features identified as low quality and retained 72% of those identified as high quality in the test set (Fig. 4.8).

When the data-adaptive pipeline was applied to the public urine dataset [116], 83% of the high quality features in the test set were retained and 74% of the low quality features in the test set were removed. We used a percent missing cutoff of 69% (median of percent missing in the low quality feature training set) and an ICC cutoff of 0.35 (lower whisker of the box plot of ICC values for low quality features in the training set). When the data-adaptive pipeline was applied to the public cell line dataset [92], 79% of the high quality features in the test set were retained and 76% of the low quality features in the test set were removed. We used a percent missing cutoff of 27% (median of percent missing values in the low quality feature training set) and an ICC cutoff of 3.8×10^{-9} (median of ICC values for low quality features in the training set).

We recognize that our data-adaptive pipeline involves several steps of manual work, such as the visual identification of high and low quality features and the selection of filtering cutoffs. Such methods do present the opportunity for user error, but we argue that such error will not effect the end results of a study. To our knowledge, *xcms* does not provide peak quality scores for an automated identification of high and low quality peaks. Furthermore, as stated previously, in the vast majority of cases the contrast between images of high and low quality features is striking. Occasional miss-classification of features as high or low quality will not considerably affect the distributions of the feature characteristics used to select the cutoffs, and therefore will not have a large impact on final filtering results. We see the manual selection of filtering cutoffs based on thorough data visualization as an advantage of our proposed pipeline. Researchers may likely have specific requirements for the balance between removing low quality and retaining high quality features depending on their scientific question of interest, their analysis plan or the size of their data. Manual selection of filtering cutoffs, as opposed to using pre-determined cutoffs, allows researchers to adjust the stringency of their feature filtering to fit the needs of their study.

4.4 Conclusions

Pipelines such as *Workflow4Metabolomics* and *MetaboAnalyst* have been crucial for advancing LC-MS based untargeted metabolomics. However, we find that users of these pipelines often rely heavily on default filtering parameters that are less than optimal for all analytical platforms and methods. The aim of our work was to assist users in understanding appropriate filtering methods for their specific data. Given the inherent heterogeneity of metabolomic studies, we argue that feature filtering of such data should be data-adaptive. Here, we provide filtering criteria for each step in a metabolomic pipeline and discuss how

to choose cutoffs based on data visualization and distributional properties of high and low quality features. Because of the random noise present in untargeted LC-MS data, we also encourage investigators to visually inspect features of interest for peak morphology and integration prior to inclusion in reports of biomarker discovery and pathway analysis results. We appreciate that our data-adaptive filtering method requires more effort than selecting default or common cutoffs, but argue that the improved data quality will greatly improve statistical analyses performed in applications involving biomarker discovery and pathway characterization.

Chapter 5

Conclusion and future directions

As omics technologies continue to grow and develop, it is becoming increasingly important to not lose sight of the importance of appropriate exploratory data analysis and data-preprocessing. While it is true that omics technologies provide a more in-depth investigation of health and disease, in many cases this is only possible once exploratory data analysis and pre-processing have sufficiently reduced the variability and bias within the resulting data. Because of the considerable amount of bias, noise and variability sometimes present in omics data, exploratory analysis and pre-processing can often be time intensive. However, the effort spent in these initial stages of analyses will increase the success and reliability of the ultimate findings.

The general goal of the work presented here was to demonstrate the importance of exploratory data analysis and data preprocessing on different kinds of omics data. We began with a popular area of omics research, single-cell RNA-sequencing, and demonstrated how early exploratory analysis and pre-processing allowed us to better understand the variability in the data. Careful investigation of the experimental design, data collection and sources of unwanted variability lead us to understand how the two batches of cells should ultimately be analyzed. Exploratory analysis of the single-cell RNA-sequencing dataset revealed to us that we needed to focus on developing a single-cell similarity measure that could effectively uncover cellular relationships amid considerable noise.

Adductomics is a relatively new area of research and thus requires thorough exploratory data analysis in order to understand the behavior of the data. Due to the lower dimensionality of adductomics data, there is less of a need to utilize more complex feature filtering methods, especially because the integration of each adduct can usually be manually verified. However, the complexity of the experimental protocol in adductomics makes data normalization all the more necessary. In the adductomics dataset discussed here, and in all adductomics datasets generated in our laboratory, we have found that including an internal standard and running duplicate injections for each subject are useful for reducing variability in adductomics. However, including duplicate injections may cause more harm than good when hundreds of subjects are included in a study, since doubling the amount of samples considerably increases the run time of the data collection and thus affects the machine per-

formance. It is an interesting experimental design question to consider the variability trade off between running duplicate injections and minimizing machine performance issues. However, running duplicate injections was not an issue with the adductomics dataset discussed here, which contained only a small number of samples.

The final aspect of this work was to consider the challenging task of filtering features from untargeted LC-MS data. We discussed visualization and filtering techniques that considerably reduced the number of low quality features in metabolomics data, regardless of the sample collection/extraction techniques, chromatography, mass spectrometry, etc. In our own lab, we have found that using the proposed data-adaptive filtering pipeline, has considerably increased the efficiency and accuracy of our subsequent statistical analysis. After data-adaptive feature filtering, we observe fewer false positives and false negatives in biomarker discovery and pathway analysis. In the beginning of this work, we posed that data-adaptive pre-processing may help with result replication. For example, it could be that data-adaptive feature filtering in metabolomics would help biomarker and pathway analysis results to replicate across studies. This idea could be studied in future work with several independent datasets. Both data adaptive and traditional filtering methods that rely on pre-determined/common filtering thresholds could be used on the datasets. Then, one could determine whether results are replicated with only one or both of the filtering techniques.

A very challenging direction of future work that was discussed briefly in the introduction is the integration of various omics data types. An interesting example of data integration is combining metabolomic, adductomic, genetic and epigenetic data from the archived neonatal blood spots (NBS) that are part of the California Childhood Leukemia Study [80]. Multiple punches have been taken from the Guthrie cards on which the NBS are stored and used for these different omics studies. Integrating this data is an example of combining measurements of a variety of omics features (e.g. genes, metabolites, etc.) that are all taken on the same set of subjects. Extensive exploratory data analysis and data pre-processing would most certainly be required before attempting to integrate such data. However, due to the complexity and heterogeneity of childhood leukemia, and the relative mystery behind its causes, many would argue that an integrative omics approach is necessary for studying this disease.

Bibliography

[1] Enis Afgan et al. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update". In: *Nucleic Acids Res.* 44.W1 (2016), W3–W10.

[2] G. Aldini et al. "Albumin Is the Main Nucleophilic Target of Human Plasma: A Protective Role Against Pro-Atherogenic Electrophilic Reactive Carbonyl Species?" In: *Chem Res. Toxicol.* 21.4 (2008), pp. 824–835.

[3] S. Anders and W. Huber. "Differential expression analysis for sequence count data." In: *Genome Biology* 11.106 (2010).

[4] Simon Anders and Wolfgang Huber. "Differential expression analysis for sequence count data". In: *Genome Biology* 11.10 (2010).

[5] D. B. Andropoulos. "Pediatric Normal Laboratory Values". In: *Gregory's Pediatric Anesthesia; Wiley-Blackwell: Oxford, UK* (2011), pp. 1300–1314.

[6] V.E. Angarica and A. Del Sol. "Bioinformatics Tools for Genome-Wide Epigenetic Research." In: *Adv Exp Med Biol.* (2017).

[7] Simon Asiedu et al. "AXL induces epithelial-to-mesenchymal transition and regulates the function of breast cancer stem cells". In: *Oncogene* 33.10 (2014).

[8] ATCC. "Passage number effects in cell lines". In: (2010). URL: `https://www.atcc.org/~/media/PDFs/Technical`.

[9] F. R. Bolasso: Bach. "Model Consistent Lasso Estimation through the Bootstrap". In: *In Proceedings of the International Conference on Machine Learning; ACM Press* (2008), pp. 33–40.

[10] F. Baig et al. "Caveats of Untargeted Medtabolomics for Biomarker Discovery". In: *JACC* 68 (2016).

[11] B. Berger, J. Peng, and M. Singh. "Computational solutions for omics data." In: *Nature Review* 14 (2013).

[12] D. Bing et al. "Normalization and noise reduction for single cell RNA-seq experiments." In: *Bioinformatics* 31 (2015).

[13] B.M. Bolstad et al. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". In: *Bioinformatics* 19 (2003).

[14] R. H. Bremmer et al. "Age Estimation of Blood Stains by Hemoglobin Derivative Determination Using Reflectance Spectroscopy". In: *Forensic Sci* 206 (2011), p. 1.

[15] James Bullard et al. "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments". In: *BMC Bioinformatics* 11.94 (2010).

[16] R. Calandrelli et al. "GITAR: An Open Source Tool for Analysis and Visualization of Hi-C Data." In: *Genomics Proteomics Bioinformatics* 16 (2018).

[17] M. L. Calle and V. Urrea. "Letter to the Editor: Stability of Random Forest Importance Measures". In: *Brief. Bioinform* 12 (2011), pp. 86–89.

[18] A. G. Chambers et al. "Comparison of Proteins in Whole Blood and Dried Blood Spot Samples by LC/MS/MS." In: *J. Am. Soc Mass Spectrom* 24.9 (2013), pp. 1338–1345.

[19] L. Chen et al. "Characterization of The Human Tear Metabolome by LC-MS/MS." In: *Journal of proteome research* 10 (2011), pp. 4876–4882.

[20] S.E. Choe et al. "Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset". In: *Genome Biol* 6 (2005).

[21] J. Chong et al. "MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis." In: *Nucleic Acids Research* 46 (2018).

[22] H.A. Chowdhury, D.K. Bhattacharyya, and J.K. Kalita. "(Differential) Co-Expression Analysis of Gene Expression: A Survey of Best Practices." In: *IEEE/ACM Trans Comput Biol Bioinform.* (2019).

[23] Michael Cole and Davide Risso. *scone: Single Cell Overview of Normalized Expression data*. R package version 1.0.0. 2017.

[24] A. M. De Livera et al. "Normalizing and Integrating Metabolomics Data". In: *Anal Chem* 84.24 (2012), pp. 10768–10776.

[25] A.M. De Livera et al. "Statistical methods for handling unwanted variation in metabolomics data." In: *Anal. Chem* 87.7 (2015), pp. 3606–3615.

[26] California Department. "of Public Health (CDPH)". In: *Newborn Screening Specimens Use and Storage* (). URL: `https://www.cdph.ca.gov/Programs/CFH/DGDS/Pages/nbs/NBSDBS-Storage.aspx%20(accessed%20May%2011,%202018)`.

[27] A. DeWilde et al. "Tryptic Peptide Analysis of Ceruloplasmin in Dried Blood Spots Using Liquid Chromatography-Tandem Mass Spectrometry: Application to Newborn Screening". In: *Clin. Chem* 54.12 (2008), pp. 1961–1968.

[28] J. R. DiFranza, C. A. Aligne, and M. Weitzman. "Prenatal and Postnatal Environmental Tobacco Smoke Exposure and Children's Health". In: *Pediatrics* 113.4 (2004), pp. 1007–1015.

[29] J. Eberwine et al. "The promise of single-cell sequencing". In: *Nature Methods* 11 (2014).

[30] L. Escriva et al. “Mycotoxin Analysis of Human Urine by LC-MS/MS: A Comparative Extraction Study.” In: *Toxins* 9.10 (2017), pp. 1–15.

[31] M. J. Fasco et al. “Unique Cyanide Adduct in Human Serum Albumin: Potential as a Surrogate Exposure Marker”. In: *Chem Res. Toxicol* 24 (2011), pp. 505–514.

[32] A. Florescu et al. “Methods for Quantification of Exposure to Cigarette Smoking and Environmental Tobacco Smoke: Focus on Developmental Toxicology”. In: *Ther Drug Monit* 31 (2009), pp. 14–30.

[33] J.P. Fortin et al. “Functional normalization of 450k methylation array data improves replication in large cancer studies.” In: *Genome Biol.* 15.12 (2014).

[34] W. E. Funk et al. “Hemoglobin Adducts of Benzene Oxide in Neonatal and Adult Dried Blood Spots.” In: *Cancer Epidemiol Biomarkers Prev* 17 (2008), pp. 1896–1901.

[35] K. B. Gale et al. “Backtracking Leukemia to Birth: Identification of Clonotypic Gene Fusion Sequences in Neonatal Blood Spots.” In: *Proc. Natl Acad Sci.* 94.25 (1997), pp. 13950–13954.

[36] F. Giacomoni et al. “Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics.” In: *Bioinformatics* (2014).

[37] P. D. Gluckman et al. “Effect of In Utero and Early-Life Conditions on Adult Health and Disease”. In: *N. Engl. J. Med* 359.1 (2008), pp. 61–73.

[38] Y.-m. Go, J. D. Chandler, and D. P. Jones. “The Cysteine Proteome”. In: *Free Radic Biol. Med* 84 (2015), pp. 227–245.

[39] S. Gonseth et al. “Genetic Contribution to Variation in DNA Methylation at Maternal Smoking-Sensitive Loci in Exposed Neonates”. In: *Epigenetics* 11.9 (2016), pp. 664–673.

[40] T.M. Gorges, K. Pantel, et al. “Circulating tumor cells as therapy-related biomarkers in cancer patients”. In: *Cancer Immunol. Immunother.* (2013).

[41] T.M. Gorges et al. “Circulating tumour cells escape from EpCAM-based detection due to epithelial-to-mesenchymal transition”. In: *BMC Cancer* 12 (2012).

[42] Wen Feng Gou et al. “The role of RhoC in epithelial-to-mesenchymal transition of ovarian carcinoma cells”. In: *BMC Cancer* 14.477 (2014).

[43] M. Greaves. “In Utero Origins of Childhood Leukaemia”. In: *Early Hum Dev* 81.1 (2005), pp. 123–129.

[44] H. Grigoryan et al. “Adductomic Signatures of Benzene Exposure Provide Insights into Cancer Induction”. In: *Carcinogenesis* 39.5 (2018), pp. 661–668.

[45] H. Grigoryan et al. “Adductomics Pipeline for Untargeted Analysis of Modifications to Cys34 of Human Serum Albumin”. In: *Anal. Chem* 88.21 (2016), pp. 10504–10512.

[46] H. Grigoryan et al. “Cys34 Adducts of Reactive Oxygen Species in Human Serum Albumin.” In: *Chem Res Toxicol* 25.8 (2012), pp. 1633–1642.

[47] Dominic Grun et al. "Digital synthesis of plucked-string and drum timbres". In: *Nature* 525 (2015).

[48] J.A. Heiss and A.C. Just. "Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses". In: *Clinical Epigenetics* 11.15 (2019).

[49] S. Herman et al. "Mass spectrometry based metabolomics for in vitro systems pharmacology: pitfalls, challenges, and computational solutions." In: *Metabolomics* 13.79 (2017).

[50] B. D. Holbrook. "The Effects of Nicotine on Human Fetal Development". In: *Birth Defects Res Part C* 108.2 (2016), pp. 181–192.

[51] H. K. Hustoft et al. "Critical Assessment of Accelerating Trypsination Methods." In: *J. Pharm Biomed. Anal* 56.5 (2011), pp. 1069–1078.

[52] H. K. Hustoft et al. "Critical Review of Trypsin Digestion for LC-MS Based Proteomics". In: *Leung, H.-C., Ed.; InTech* 2012 ().

[53] Lan Jiang et al. "GiniClust: detecting rare cell types from single-cell gene expression data with Gini Index". In: *Genome Biology* 17.144 (2016).

[54] S. Kasimir-Bauer et al. "Expression of stem cell and epithelial-mesenchymal transition markers in primary breast cancer patients with circulating tumor cells". In: *Breast Cancer Res.* 14.1 (2012).

[55] M. Kebschull et al. "Differential Expression and Functional Analysis of HighThroughput -Omics Data Using Open Source Tools." In: *Methods Mol Biology* 1537 (2017).

[56] A. A. Klammer and M. J. MacCoss. "Effects of Modified Digestion Schemes on the Identification of Proteins from Complex Mixtures". In: *J. Proteome Res* 5.3 (2006), pp. 695–700.

[57] V. S. Knopik et al. "The Epigenetics of Maternal Cigarette Smoking during Pregnancy and Effects on Child Development." In: *Dev Psychopathol* 24.4 (2012), pp. 1377–1390.

[58] A. Koulman et al. "The development and validation of a fast and robust dried blood spot based lipid profiling method to study infant metabolism". In: *Metabolomics* 1.8 (2014).

[59] K. Lakiotaki et al. "BioDataome: a collection of uniformly preprocessed and automatically annotated datasets for data-driven biology." In: *Database* (2018).

[60] C. Lareau et al. "Preprocessing and Computational Analysis of Single-Cell Epigenomic Datasets." In: *Methods Mol Biol* 1935 (2019).

[61] Erin LeDell, Maya Petersen, and Mark van der Laan. *cvAUC: Cross-Validated Area Under the ROC Curve Confidence Intervals*. R package version 1.1.0. 2014. URL: `https://CRAN.R-project.org/package=cvAUC`.

[62] Jacob Levine et al. “Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis”. In: *Cell* 162 (2015).

[63] Qunhua Li et al. “Measuring reproducibility of high-throughput experiments”. In: *The Annals of Applied Statistics* 5.3 (2011).

[64] S. Li et al. “Predicting network activity from high throughput metabolomics.” In: *PLoS computational biology* (2013).

[65] A. Liaw and M. Wiener. “Classification and Regression by RandomForest”. In: *R News* 2 (2002), p. 3.

[66] G. Libiseller et al. “IPO: a tool for automated optimization of XCMS parameters”. In: *BMC Bioinformatics* 16 (2015), p. 118.

[67] Hui-Kuan Lin et al. “Suppression Versus Induction of Androgen Receptor Functions by the Phosphatidylinositol 3-Kinase/Akt Pathway in Prostate Cancer LNCaP Cells with Different Passage Numbers”. In: *Journal of Biological Chemistry* 51 (2003).

[68] S. Liu et al. “Cys34 Adductomes Differ between Patients with Chronic Lung or Heart Disease and Healthy Controls in Central London. Environ”. In: *Sci. Technol* 52.4 (2018), pp. 2307–2313.

[69] S. S. Lu et al. “Profiling the Serum Albumin Cys34 Adductome of Solid Fuel Users in Xuanwei and Fuyuan, China”. In: *Environ. Sci* 51.1 (2017), pp. 46–57.

[70] A.T.L. Lun et al. “Pooling across cells to normalize single-cell RNA sequencing data with many zero counts.” In: *Genome Biology* 17 (2016).

[71] Sysi-Aho M et al. “Normalization method for metabolomics data using optimal selection of multiple internal standards”. In: *BMC Bioinformatics* 8.93 (2007).

[72] N.G. Mahieu and G.J. Patti. “Systems-Level Annotation of a Metabolomics Data Set Reduces 25000 Features to Fewer than 1000 Unique Metabolites.” In: *Analytical Chemistry* 89.19 (2017), pp. 10397–10406.

[73] S.A. Mani et al. “The epithelial-mesenchymal transition generates cells with properties of stem cells”. In: *Journal of Biological Chemistry* 133.4 (2008).

[74] N. J. Martin and H. J. Cooper. “Challenges and Opportunities in Mass Spectrometric Analysis of Proteins from Dried Blood Spots”. In: *Expert Rev Proteomics* 11.6 (2014), pp. 685–695.

[75] V.S. Marwah et al. “eUTOPIA: solUTion for Omics data PreprocessIng and Analysis”. In: *Source Code for Biology and Medicine* 14.1 (2019).

[76] C. Meng et al. “Dimension reduction techniques for the integrative analysis of multi-omics data”. In: *Brief Bioninform* 17 (2016).

[77] F. Meng and A. I. Alayash. “Determination of Extinction Coefficients of Human Hemoglobin in Various Redox States”. In: *Anal. Biochem* 521 (2017), pp. 11–19.

[78] C. Metayer et al. “Parental Tobacco Smoking and Acute Myeloid Leukemia.” In: *Am J. Epidemiol* 184.4 (2016), pp. 261–273.

[79] C. Metayer et al. “Tobacco Smoke Exposure and the Risk of Childhood Acute Lymphoblastic and Myeloid Leukemias by Cytogenetic Subtype.” In: *Cancer Epidemiol Biomarkers Prev* 22.9 (2013), pp. 1600–1611.

[80] C. Metayer et al. “Tobacco smoke exposure and the risk of childhood acute lymphoblastic and myeloid leukemias by cytogenetic subtype.” In: *Cancer Epidemiol Biomarkers Prev* 22.9 (2013).

[81] Hajime Mizuno et al. “The great importance of normalization of LC–MS data for highly-accurate non-targeted metabolomics”. In: *Biomedical Chromatography* 31.1 (2017). e3864 BMC-16-0509.R1, e3864–n/a. ISSN: 1099-0801. DOI: 10.1002/bmc.3864. URL: http://dx.doi.org/10.1002/bmc.3864.

[82] N. C. Momen et al. “Exposure to Maternal Smoking during Pregnancy and Risk of Childhood Cancer: A Study Using the Danish National Registers”. In: *Cancer Causes Control* 27.3 (2016), pp. 341–349.

[83] Lorraine O’Driscoll et al. “Phenotypic and global gene expression profile changes between low passage and high passage MIN-6 cells”. In: *Journal of Biological Chemistry* 191 (2006).

[84] E. Ozkumur et al. “Inertial focusing for tumor antigen-dependent and -independent sorting of rare circulating tumor cells”. In: *Sci Transl Med* 5.179 (2013).

[85] R.E. Patterson et al. “Improved Experimental data processing for UHPLC-HRMS/MS lipidomics applied to nonalcoholic fatty liver disease.” In: *Metabolomics* 12.89 (2016).

[86] K. Perttula et al. “Evaluating Ultra-long-Chain Fatty Acids as Biomarkers of Colorectal Cancer Risk”. In: *Cancer Epidemiology, Biomarkers and Prevention.* 25.8 (2016).

[87] K. Perttula et al. “Untargeted lipidomic features associated with colorectal cancer in a prospective cohort”. In: *BMC Cancer* (2018).

[88] L. Petrick et al. “Metabolomics of Neonatal Blood Spots Reveal Distinct Phenotypes of Pediatric Acute Lymphoblastic Leukemia and Potential Effects of Early-life Nutrition.” In: *Cancer Letters* (2019).

[89] L. Petrick et al. “An Untargeted Metabolomics Method for Archived Newborn Dried Blood Spots in Epidemiologic Studies”. In: *Metabolomics* 13.3 (2017), p. 27.

[90] P. Prentice. “Lipidomic analyses, breast and formula feeding, and growth in infants.” In: *Journal of Pediatrics* 166.2 (2015), pp. 276–281.

[91] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.R-project.org/.

[92] C. Ranninger et al. “Nephron Toxicity Profiling via Untargeted Metabolome Analysis Employing a High Performance Liquid Chromatography-Mass Spectrometry-based Experimental and Computational Pipeline.” In: *The Journal of Biological Chemistry* 290 (2015).

[93] S. M. Rappaport et al. “Albumin Adducts of Benzene Oxide and 1,4-Benzoquinone as Measures of Human Benzene Metabolism”. In: *Cancer Res* 62.5 (2002), pp. 1330–1337.

[94] S. M. Rappaport et al. “Characterizing Exposures to Reactive Electrophiles”. In: *Toxicol. Lett* 213.1 (2012), pp. 83–90.

[95] S. Reinke et al. “Metabolomics analysis identifies different metabotypes of asthma severity.” In: *Asthma* 49 (2017).

[96] S. Riccadona and P. Franceschi. “Data Treatment for LC-MS Untargeted Analysis.” In: *Methods Mol Biol* 1738 (2018).

[97] D. Risso et al. “A general and flexible method for signal extraction from single-cell RNA-seq data.” In: *Nature Communications* 9 (2018).

[98] D. Risso et al. “clusterExperiment and RSEC: A Bioconductor package and framework for clustering of single-cell and other large gene expression datasets.” In: *PLOS Computational Biology* (2018).

[99] D. Risso et al. “GC-Content Normalization for RNA-Seq Data.” In: *BMC Bioinformatics* 12 (2011).

[100] Davide Risso et al. “Normalization of RNA-seq data using factor analysis of control genes or samples”. In: *Nature biotechnology* 32.9 (2014).

[101] D. Risso et al. “Normalization of RNA-Seq Data Using Factor Analysis of Control Genes or Samples”. In: *Nat Biotechnol* 32.9 (2014), pp. 896–902.

[102] J.T. Robinson et al. “Integrative Genomics Viewer”. In: *Nature Biotechnology* 29 (2011).

[103] M. Robinson and Alicia Oshlack. “A scaling normalization method for differential expression analysis of RNA-seq data.” In: *Genome Biology* 11.25 (2010).

[104] A. Rotter et al. “Finding Differentially Expressed Genes in Two-Channel DNA Microarray Datasets: How to Increase Reliability of Data Preprocessing.” In: *Omics* 12 (2008).

[105] F. M. Rubino et al. “An ”Omic” Physiopathology of Reactive Chemicals: Thirty Years of Mass Spectrometric Study of the Protein Adducts with Endogenous and Xenobiotic Compounds”. In: *Mass Spectrom. Rev* 28.5 (2009), pp. 725–784.

[106] E. Scoville et al. “Alterations in lipid, amino acid, and energy metabolism distinguish Crohn’s disease from ulcerative colities and control subjects by serum metabolomic profiling.” In: *Metabolomics* 14.17 (2018).

[107] S.R. Searle, G. Casella, and C.E. McCulloch. "Introduction". In: *Variance Components*. New Jersey: John Wiley and Sons, 2006.

[108] A. J. de Smith et al. "Correlates of Prenatal and Early-Life Tobacco Smoke Exposure and Frequency of Common Gene Deletions in Childhood Acute Lymphoblastic Leukemia". In: *Cancer Res* 77.7 (2017), pp. 1674–1683.

[109] C.A. Smith et al. "XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification." In: *Analytical Chemistry* 78 (2006).

[110] L. G. Spector et al. "Detection of Cotinine in Newborn Dried Blood Spots." In: *Cancer Epidemiol Biomarkers Prev* 16.9 (2007), pp. 1902–1905.

[111] L. G. Spector et al. "Prenatal Tobacco Exposure and Cotinine in Newborn Dried Blood Spots". In: *Pediatrics* 133 (2014), p. 6.

[112] R. Spicer et al. "Navigating freely-available software tools for metabolomics analysis." In: *Metabolomics* 13.106 (2017).

[113] Oliver Stegle, Sarah Teichmann, John C Marioni, et al. "Computational and analytical challenges in single-cell transcriptomics". In: *Genetics* 16 (2015).

[114] R. Talhout et al. "Hazardous Compounds in Tobacco Smoke. Int. J. Environ". In: *Res Public Health* 8.2 (2011), pp. 613–628.

[115] B. L. Therrell et al. "Current Status of Newborn Screening Worldwide: 2015". In: *Semin. Perinatol* 39.3 (2015), pp. 171–187.

[116] E.A. Thevenot et al. "Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses." In: *Journal of Proteome Research* 14 (2015).

[117] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso." In: *J R Stat Soc. Ser.* 58 (1996), pp. 267–288.

[118] M. Tornqvist et al. "Protein Adducts: Quantitative and Qualitative Aspects of Their Formation". In: *Analysis and Applications. J Chromatogr. B* 778.1 (2002), pp. 279–308.

[119] O. Troyanskaya et al. "Missing Value Estimation Methods for DNA Microarrays". In: *Bioinformatics* 17.6 (2001), pp. 520–525.

[120] O. Troyanskaya et al. "Missing value estimation methods for DNA microarrays". In: *Bioinformatics* 17.6 (2001), pp. 520–5.

[121] I. Tzoulaki et al. "Design and Analysis of Metabolomics Studies in Epidemiologic Research: A Primer on -Omic Technologies". In: *American Journal of Epidemiology* 180 (2014).

[122] I. Tzoulaki et al. "Design and Analysis of Metabolomics Studies in Epidemiologic Research: A Primer on -Omic Technologies." In: *American Journal of Epidemiology* 180 (2014).

[123] C.A. Vallejos et al. "Normalizing single-cell RNA sequencing data: challenges and opportunities." In: *Nature Methods* 14 (2017).

[124] C. V. Vinnakota et al. "Comparison of Cyanide Exposure Markers in the Biofluids of Smokers and Non-Smokers". In: *Biomarkers* 17.7 (2012), pp. 625–633.

[125] M. Wagner et al. "The Use of Mass Spectrometry to Analyze Dried Blood Spots". In: *Mass Spectrom. Rev* 35 (2016), p. 3.

[126] E.J. Want et al. "Global metabolic profiling of animal and human tissues via UPLC-MS." In: *Nature Protocols* 18.1 (2013).

[127] M. R. Waterman. "Spectral Characterization of Human Hemoglobin and Its Derivatives". In: *Methods in Enzymology;* 703 (1978), pp. 456–463.

[128] T. P. Whitehead et al. "Childhood Leukemia and Primary Prevention". In: *Curr. Probl. Adolesc. Health Care* 46.10 (2016), pp. 317–352.

[129] J. L. Wiemels et al. "GWAS in Childhood Acute Lymphoblastic Leukemia Reveals Novel Genetic Associations at Chromosomes 17q12 and 8q24.21". In: *Nat Commun* 9.1 (2018).

[130] J.R. Williams et al. "Functional Heatmap: an automated and interactive pattern recognition tool to integrate time with multi-omics assays". In: *BMC Bioinformatics* 20 (2019).

[131] Chen Xu and Zhengchang Sui. "Identification of cell types from single-cell transcriptomes using a novel clustering method". In: *Bioinformatics Advance Access* 31.12 (2015).

[132] L. Yan et al. "Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells". In: *Nature Structural and Molecular Biology* 20 (2013).

[133] J. Yang et al. "Levels of Cotinine in Dried Blood Specimens from Newborns as a Biomarker of Maternal Smoking Close to the Time of Delivery." In: *Am. J Epidemiol* 178 (2013), p. 11.

[134] L. Zappia, B. Phipson, and A. Oshlack. "splatter: Simple Simulation of Single-cell RNA Sequencing Data". In: (2017). URL: `https://github.com/Oshlack/splatter`.

[135] Amit Zeisel et al. "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq". In: *Science* 347.6226 (2015).

[136] Yi-Cun Zhong et al. "Thrombin promotes epithelial ovarian cancer cell invasion by inducing epithelial-mesenchymal transition". In: *Journal of Gynecologic Oncology* 24.3 (2013).