Understanding Tumor Composition and Evolution Through Geometric Models

Theodore Roman

CMU-CB-17-100

May 4, 2017

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Russell Schwartz, Chair
Robin E. C. Lee
Adrian Lee
Jian Ma
Jessica Zhang

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

This research was supported in part by the U.S. National Institutes of Health via awards 1R01CA140214 and T32EB009403. Additionally, this research was supported in part by a Carnegie Mellon University GuSH grant. Further, this research was supported in part by usage of the Data Exacell, which is supported by National Science Foundation award number ACI-1261721 at the Pittsburgh Supercomputing Center (PSC). The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, donors, or the U.S. Government.

The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations, or conclusions.

Keywords: Tumor Progression, Tumor Evolution, Tumor Phylogenetics, Copy Number Variation, Simplicial Complex, Maximum Parsimony, Tumor Genomics, Cancer Biology

To my wife, Danielle, my parents, grandparents, and many other family members, whose support and love have enabled me and stretched across the miles and years.

Abstract

As part of the rapid progress in the understanding and treatment of disease over the course of the past 100 years, diagnosis and treatment of cancer has become a focal point for basic science research. As a result, advances have been made in quantifying the myriad changes in tumor genomes, transcriptomes, epigenomes, and metagenomes relative to healthy tissue. Specific to the work of this thesis, technical advances have led to more robust quantification of RNA expression states via RNA-seq, and DNA copy number quantification via DNA-seq. These approaches allow for the measurement of the state of tens of thousands of genes in a sample. Moreover, the enhanced quantification has led to better understanding of molecular heterogeneity among and within tumors. Previous efforts used methods from computational geometry in order to explain tumors in terms of component subpopulations and the mutations associated with those subpopulations. Prior to the work contained in this thesis, however, there remained a knowledge gap in applying suitably identified methods to data in such a way as to identify branching points in the evolution and progression of tumors.

The two paths of focus for this thesis are advancement in methodology and advances in biology. In this thesis, we develop models appropriate for identifying tumor subpopulations across a large number of tumor samples and a large number of sites within a sample. We then apply the methods to large-scale cancer experiments based on RNA-seq data. This model introduces the geometric concept of simplicial complexes for deconvolution of tumor data. The data considered are a selection of breast tumor samples, for which we demonstrate the correlation of metrics from the method to existing clinical and basic science subtypes. We next enhance clustering approaches used to decrease the amount of expert knowledge required to operate and interpret the estimated deconvolution. We apply the enhanced clustering approach to DNA-seq tumor data from two tumor types — ovarian and lung squamous small cell cancers — and demonstrate the separability of prognoses based on the resultant partitions. Lastly, we integrate the enhanced clustering approach into the initial framework of tumor unmixing, and further add a layer of weighting. The layer of weighting allows for a softer partitioning of data into branches of the tumor phylogeny. Such an approach provides improvement over the initial approach in terms of better capturing uncertainty in the biological data in a quantifiable and interpretable way in terms of the probability model underlying the approach. We then apply the approach to three cancer data sets — one of DNA-seq breast tumor data, one of RNA-seq breast tumor data, and one consisting of a heterogeneous combination of both the DNA-seq and RNA-seq data. We are able to correlate favorably with state-of-the-art approaches based on other data types and statistical models, while providing additional flexibility with our framework with respect to data type and quality.

Acknowledgments

Professional support has comes in many forms for me. My advisor, Dr. Russell Schwartz, provided an environment where each lab member continually improves their research, and becomes aware of the good done by science. He has allowed me to leverage my strengths, and to improve my areas of weakness to develop into a better scientist.

My committee has provided helpful feedback and fresh eyes upon my thesis work. Thank you to Drs. Adrian Lee, Robin E.C. Lee, Jessica Zhang, and Jian Ma.

A special thanks to my undergraduate advisor, Dr. Mehmet Koyuturk, for fostering my interest in graduate studies, and encouraging me to aim high in my research.

Thank you also to former labmates who provided feedback and suggestions on my research — Drs. Lu Xie, John Kang, and Greg Smith, as well as current labmates Marcus Thomas and Alan Shteyman. The collaborators on the thesis work — Drs. Gary Miller, Brittany Terese Fasy, Amir Nayyeri, and Lu Xie — have contributed their perspective, and have bettered the work contained herein.

Certainly not least, I acknowledge the tremendous support, love, willpower, and positive environment provided by my wife, Danielle Kory Roman. Your flexibility and kindness provided the conditions to be the best scientist I could. It is difficult to imagine working with vigor and zest without your sacrifices. Similarly, I struggle to envision the fortunate opportunity to write a thesis without the love and encouragement of my mother Dr. Karen Ferrick-Roman and my late father R. Daniel Roman, who instilled the value of perseverance and fostered the self-confidence to pursue lofty goals.

Contents

List of Figures

List of Tables

Acronyms

aCGH Array Comparative Genome Hybridization. 7, 11, 47, 60
ARI Adjusted Rand Index. 55

BIC Bayesian Information Criterion. 64, 84
BRCA Breast Cancer. 11, 15–17, 22, 31, 35, 66, 67

CNV Copy Number Variation. 2, 16, 43, 47, 59, 60, 66, 78–81

ER Estrogen Receptor. 24, 35, 67, 68, 70, 79, 80

FISH Fluorescence in Situ Hybridization. 7, 9, 20

GBM Glioblastoma Multiforme. 11, 15–18, 47, 48, 52, 53
GMM Gaussian Mixture Model. 30–34, 38

HER2 Human Epidermal Growth Factor 2. 24, 35, 67, 68, 70, 79, 80

ICA Independent Components Analysis. 26

KNN K-Nearest Neighbors. 59

LUSC Lung Squamous Small Cell Lung Cancer. 11, 15, 16, 47, 48, 50–53, 56

MST Minimum Spanning Tree. 9, 20, 38
MVES Minimum Volume Enclosing Simplex. 28

OV Ovarian Cancer. 11, 15, 16, 47, 48, 50–53, 56

PAM50 Prediction Analysis for Microarray 50 Gene Panel. 35
PCA Principal Components Analysis. 6, 26, 27, 29, 30, 47, 48, 58, 59, 61, 63, 84, 85
PCR Polymerase Chain Reaction. 2, 3
PR Progesterone Receptor. 24, 35, 67, 68, 70, 79, 80

RAM Random Access Memory. 44, 78

RMSD Root Mean Squared Deviation. 31

SC-seq Single Cell Sequencing. 2, 3, 17

SNP Single Nucleotide Polymorphism. 85, 86

SNR Signal-to-Noise Ratio. 64

SV Structural Variant. 3, 24

SVD Singular Value Decomposition. 30

TCGA The Cancer Genome Atlas. 3, 11, 16, 22, 30, 31, 34, 35, 47, 48, 50, 56, 59, 66, 67, 69, 70, 78–81, 83

TNBC Triple Negative Breast Cancer. 67, 68, 70, 79, 80

Chapter 1

Background and Introduction

Cancer lies at the intersection of seemingly paradoxical conditions. For instance, cellular growth is unencumbered, yet ultimately the individual is harmed. The cells of the tumor find ways to avoid the mechanisms typical of cell death, but at the expense of the individual experiencing the disease [39, 40]. Despite these paradoxical conditions, substantive progress has been made in understanding the mechanisms and basic biology of cancer [39, 40, 96]. Cancer can be characterized by the sets of acquired mutations to normally healthy cells; but translating this molecular characterization into actionable treatment remains a stumbling block for many cancer types [96]. As a result, better detection, treatment, and outcomes for later-stage cancer has become a priority.

Cancer has been characterized by a number of hallmarks [39, 40]. Among these hallmarks are avoidance of cell death, uncontrolled growth, and recruitment of vascular tissue. Chronologically, cancer can be modeled as cells acquiring key mutations that allow the cells to escape the normally-present set of checkpoints that control the cell cycle, proliferation, cell death, and the other notable hallmarks of cancer. Further, after the cells exhibit the characteristics of cancer, further mutations can more easily pass to daughter cells, allowing for additional diversity and heterogenity of cell genomes in a tumor [26, 67, 96]. Over time, these changes may lead to tumors metastasizing, using the recruited vascular tissue.

1.1 Biological and Technological Limitations

In this section, we provide an overview of the steps of quantifying tumor biology data. Broadly speaking, the quantitative tumor data considered in this thesis may come from either DNA, RNA, or combinations thereof. In the DNA case, we consider Copy Number Variation (CNV), where the number of copies of a particular gene is altered. In the case of RNA data, we consider expression of genes that are dysregulated compared to some expected amount of expression. Although terminology associated with genomics is used, the same basic steps are followed for quantification of RNA as well. With the exclustion of the still-developing technology of Single Cell Sequencing (SC-seq), the sequencing of tumor genomes has traditionally been obtained through the use of bulk sequencing technologies [8, 46, 50, 53, 64, 65, 104, 105]. In bulk sequencing tumors, a tumor is excised from a patient. Following excision, the tumor is put through a Polymerase Chain Reaction (PCR) set of cycles, in which the DNA present in the sample is ligated and amplified in a controlled and cyclical manner. As a consequence of the PCR cycles, many copies of the tumor genomes of individual cells are freely mixed in the tube containing the sample. These strands are then run through a sequencing platform, the result of which is called raw reads. The raw reads are mapped using software tools to either a reference genome or assembled *de novo* [4, 18, 78, 110]. After assembly, a piece of software known as a caller provides information as to which areas of the mapped reads exhibit unexpected amplification or deletion [22, 52, 60], also known as a call. These calls are then arranged by the caller into a list. It is this list of calls that functions as the input to the models in this thesis.

The bulk sequencing pipeline requires models capable of performing inference by the nature of its setup. Earlier, we discussed that tumors are evolving at the level of single cells — that is, each cell may acquire mutations that are passed on to daughter cells, which then acquire additional mutations, and pass them to the grand-daughter cells of the original cell, and so on and so forth. However, bulk sequencing measures many cells at the same time, and genomes — and the alterations manifest at a single-cell level — are mixed freely. As a result, the reads that

are output from a sequencer to the aligner then to the caller is a type of average of all mutations acquired by cells in the sample. Cruicially, previous work has demonstrated that this type of average is not representative of individual cells in the tumor, in terms of prediction [7, 29, 30, 62].

Figure 1.1 shows heterogeneity in bulk sampling may yield non-representative resulting genomes.

A proposed workaround to this limitation is the application of SC-seq technologies. However, SC-seq technologies remain out of reach in several key metrics. Although SC-seq technologies provide insights as to the genomes of individual cells, the loss of some genetic information during PCR cycling [64] creates more substantive gaps in aligning genomes obtained through SC-seq, which in turn may reduce the efficacy of calling tools [46, 50]. Further, SC-seq technologies at present are not cost competitive on a per-base-pair basis with bulk sequencing technologies [50].

1.2 Prior Work

1.2.1 Methods in Cancer Genomics Modeling

As a result of the limitations of SC-seq, bulk sequencing remains the underlying technology for the majority of available tumor datasets. Due to the limitations of bulk sequencing with respect to resolving related subpopulations within an individual bulk sample, numerous computational techniques have been developed. These computational techniques can be specific to a type of data, such as Structural Variant (SV) [2]. Others use knowledge based in clustering variant alleles [49, 55, 83]. However, estimation of some requisite parameters of these models may be difficult [24, 69]. Other requirements of the models may require substantial read depth [49, 83] that is unlikely to be achieved on large-sample efforts such as The Cancer Genome Atlas (TCGA) [45]. Other methods that leverage DNA copy number data have modeled deconvolution, but do not consider how the phylogenetic aspects of tumor biology may play a role in better understanding the tumor composition [11, 36]. Others permit only one data type per run [68, 69,

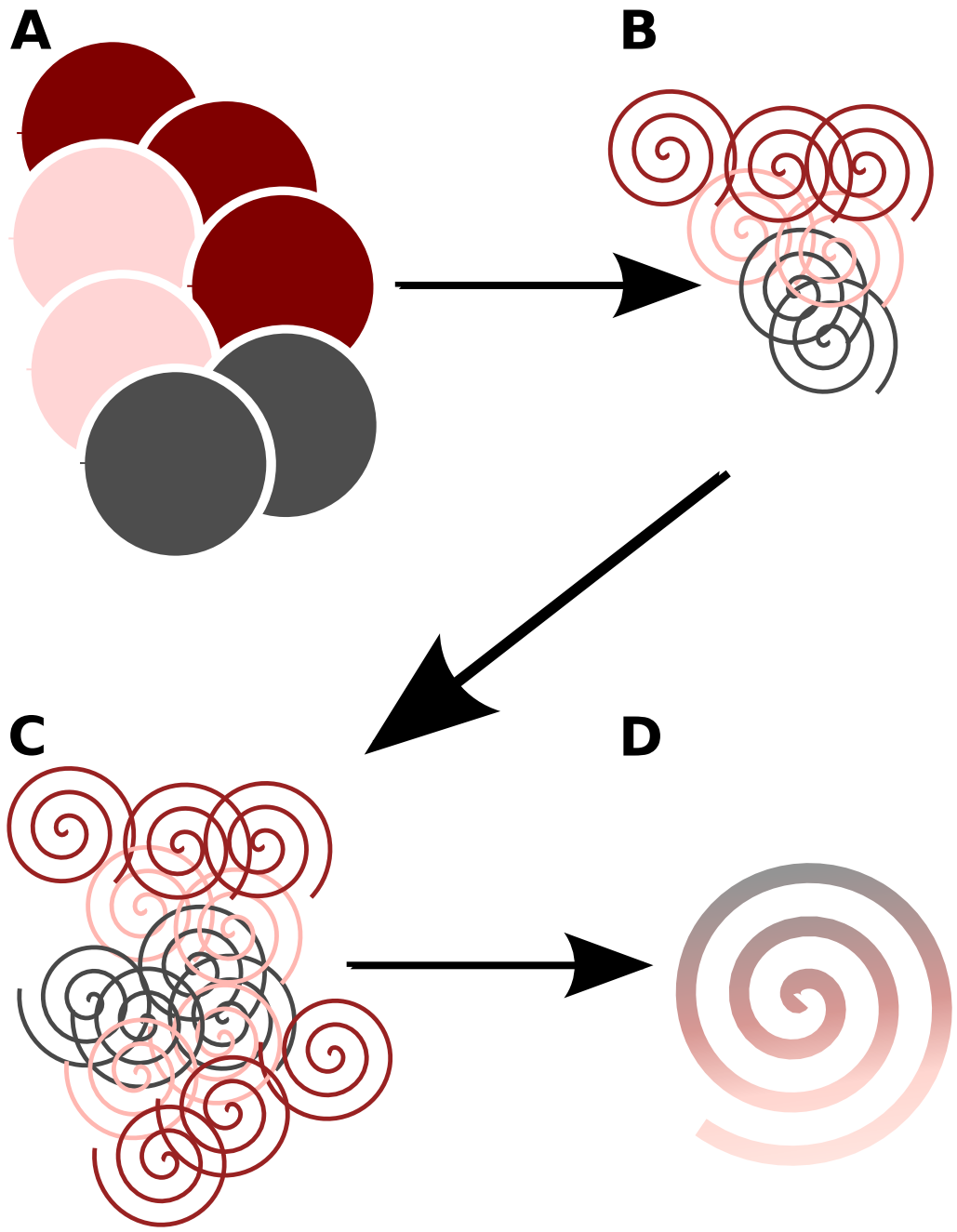

Figure 1.1: Heterogeneity in tumor samples creates challenges in interpreting sequences of samples. Multiple subpopuations (panel A) lead to multiple genomes in a bulk sample (panel B). When these genomes are amplified (panel C), and assembled, the resulting genome may not be representative of any one of the individual subgroups (panel D).

88]. However, each of these prior works has contributed to the understanding of tumor biology, as well as tumor deconvolution modeling. We can further partition the existing approaches prior to the thesis work by the computational methods employed. Some approaches use probabilistic graphical models as the core of the approach [83]. Others instead use hidden Markov models [37]. Foundational works in the field used approaches such as conjunctive Bayesian networks [31]. Yet others use hierarchical clustering as the approach to building a tumor phylogeny [21]. Other efforts have focused on constrained linear programming as the means to obtain a tree [74, 75]. Others still have preferred to use Markov Chain Monte Carlo sampling [49]. Our work has built upon the geometric approaches pioneered by Schwartz and Shackey [86]. Concurrent work to this thesis has focused on techniques such as non-negative matrix factorization [70], graphical models [83], and maximimum likelihood models [68, 69]. Below we provide a brief timeline of the development of prior methods on which our work is constructed, as well as concurrent approaches developed for related inference tasks.

Ehrlich and Full first applied principles of geometry to deconvolution [23]. In their work, as reviewed by Size [90], the authors identify the relationship between samples containing a mixture of numerous populations, and a geometric body called a simplex. More specifically, a simplex can be formulated as a complete graph embedded in a higher dimensional space — for instance, triangles, tetrahedra, and their higher-dimensional equivalents. Points within the volume enclosed by the vertices of the simplex can then be formulated as a convex combination of the vertices. Because the points can be formulated as a convex combination of the vertices, each point in the volume can be assigned specific mixture components (mixture fractions) from each of the vertices. Although Ehrlich and Full [23] applied the approach to deconvolve soil samples from a discrete number of geological subpopulations, Schwartz and Shackney [86] applied the principles to tumor samples. The application methods used by [86] stemmed from methods by Chan et al. [13], who were able to efficiently code the simplex enclosing algorithm.

In the foundational work of [86], the authors note that a small number of related groups of

cells, termed subtypes, can well characterize tumors. Additionally, the same constraint of convex combinations apply as in [23], given that the subtypes are sufficient to explain the composition of the tumor. Furthermore, the authors assert that based on how data points are distributed throughout the surface of the simplex, inferences can be made as to which subtypes are ancestral to other subtypes. In particular, the paper brought forth the idea that because tumor progression can be modeled as an evolutionary process, related subtypes ought to be observed as mixed in the samples, while subtypes on different lineages of a tumor phylogeny would not be observed to intermix. The data also needed to exhibit geometric patterns with respect to the distribution of points in order for any analysis based on geometry to have application. It is key to note that [86] demonstrated that bulk tumor genomic data filtered with Principal Components Analysis (PCA) [73] exhibited this property.

Building on the work of [86], Tolliver et al. [98] provided an improvement to better handle outliers and noise. Whereas the initial work of [86] held a strict requirement that all data points be contained in the simplex, [98] relaxed this constraint via modifications to the objective function. Interpreted biologically, a strict requirement would imply that each mixed sample could be explained as some mixture of subpopulations, whereas a soft constraint would imply that each mixed sample could be explained as some mixture of subpopulations, plus some allowance to explain technical and biological noise. [98] accomplished this expansion through a crucial change in description of the problem of geometric unmixing. Rather than the previous efforts, which focused on the application of minimum-volume enclosing simplices to point clouds, as outlined in [13], [98] proposed a probability framework that in the limit of no noise would reduce to a minimum volume simplex. Using a novel formulation of the probability of the data fitting a proposed solution, coupled with Bayes' Rule [6], [98] decomposed the probability of a simplex best fitting some data into terms proportional to the fit of data to that simplex, and the volume of the simplex. The key conceptual contribution to this decomposition was that as distance to the surface of the body increased linearly, the goodness of fit decreased exponentially. Similarly,

in the prior portion of their decomposition, as the volume increased linearly, the goodness of fit decreased exponentially.

1.2.2 Tumor Phylogenetics

Early and influential work in understanding phylogenetic models of tumor progression can be attributed to Desper et al. [21]. Desper et al. used the technology of Array Comparative Genome Hybridization (aCGH) as the basis for the models they constructed, which they termed "oncogenetic trees".

The aCGH platform works by attaching fluorescent dye to reference and tumor DNA — green fluorescent dye is attached to hybridized tumor DNA, while red fluorscent dye is tagged to reference DNA. Images of the cells are then taken, and a ratio of the quantified fluorscence describes the ratio of copy numbers of the tumor cells to the normal cells. For full details on the procedure, see Kallioniemi et al. [51].

Once the quantified aCGH data were acquired, [21] proposed a tree-oriented model of tumor progression. The models proposed center around heuristic solutions to construct maximally likely trees based on the evolutionary distance between a healthy cell and tumor cell, with the distance weighted per site. One of the key assumptions of the approach was that the average measurements from a bulk tumor would be representative of the tumor as a whole. Nevertheless, the work pioneered the idea that tumor progression shared common properties with species evolution.

Podlaha et al. [77] later advanced the understanding of tumor phylogenetics by bringing to light the interplay of intratumor heterogeneity in the context of tumor evolution with selective pressure and potentially changing rates of mutation as a function of time. Greaves and Maley [35] reviewed the state of understanding how tumor clones evolve in that same year, focusing their review on the similarities of Darwinian evolution and tumor development, with the role of the tumor microenvironment providing selective pressure for some subpopulations. Importantly, the

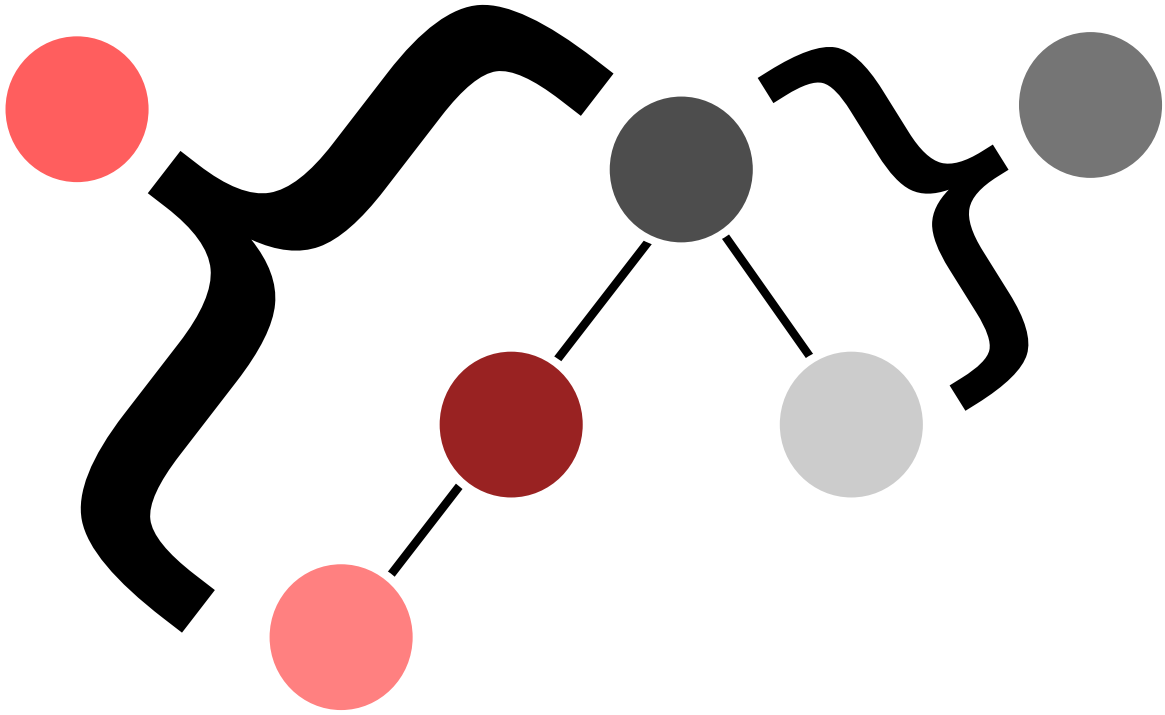

Figure 1.2: Coupled with heterogeneity, the evolutionary processes can lead to mixed tumor samples based on the lineage of the particular sample. Mixed tumor samples can differ substantially in genomic profile. Despite coming from the same phylogeny, mixed samples (far left and far right, above brackets) can substantially differ due to phylogenetic branch and tumor heterogeneity, creating a need for computational approaches.

review pointed out that some mutations are viewed as driving the development of tumors, while others simply serve as passengers. Distinguishing which mutations are more vital to the progression of disease (driver mutations) versus those that are observed but less relevant to progression (passenger mutations) remains a question in building models. Brocks et al. [9] demonstrated the correlation among DNA methlyation patterns and DNA CNV patterns with respect to estimated evolutionary time, highlighting the need for methods capable of multiple data types.

Figure 1.2 illustrates that due to the nature of heterogeneous tumor samples and tumor evo-

lution, samples from the same tumor phylogeny may exhibit seemingly different properties. As a result, computational approaches to understand and unify the samples are necessary.

To combat the heterogeneity within a tumor, Pennington et al. [74, 75] contributed models built upon cell-by-cell analysis of tumors. In the approach, tumor progression is modeled based on Fluorescence in Situ Hybridization (FISH) data. Compared to bulk genomics approaches, FISH holds the key advantage that samples can be resolved to a single-cell level. However, the technology imposes limits on the number of markers that can be tested per cell to a handful of pre-selected probes rather than the entire genome [100]. Leveraging the single-cell resolution of FISH data, [74, 75] created a model in which individual events of copy number alteration both within a patient, and across patients are considered. In this way, [74, 75] represents an key step forward in tackling both intratumor heterogeneity and intertumor heterogeneity. Furthermore, the phylogenies produced from [74, 75] provided ordered copy number alteration events and the relative frequencies in the cell populations studied. In terms of translating the model to a clinical application, by ordering the copy number alteration events, the model provided a guideline for therapies that may be applicable earlier in the life of the tumor.

This work in the application of tree-based models to understanding tumor progression was built upon by [86] and [98]. New models of optimal trees were proposed. The models were trees relating to the most likely lineages of tumor progression, but with different sets of assumptions.

In [86] the authors aimed to combine the resolution of FISH data with the broad feature capture of bulk sequencing to produce tumor phylogenies with novel putative prognostic and diagnostic markers. Based on the formulation of a tree as an optimization of a constrained system, in a fashion related to [21], [86] added a geometric mixture element in the set of constraints to incorporate a measurement of tumor heterogeneity in the bulk samples. The phylogeny inference portion of the work was framed as a maximum likelihood model, where each subpopulation's frequency depends on its ancestral population, and is independent of other populations. In practice, to recover the tree from the geometric body, a MST was used. The phylogenetic results from

[86] provided evidence demonstrating putative evolution of clinical subtypes of lung cancer.

1.2.3 Tumor Heterogeneity

Although tumor heterogeneity has been recognized as a key characterization of cancer for over a decade [26], we recognize specifically the pioneering studies of tumor heterogeneity and decomposition posited by Etzioni et al. [25]. In the work, the authors demonstrated that differences in the composition of tumors created a statistically significant difference between patients with and without recurrence. The data set used were prostate tumor samples, stained with immunohistochemical antibody stain. Conceptually speaking, the staining process works by applying an antibody (e.g., syndecan-1) to a tissue specimen. Alterations of the gene assoicated with the antibody (SDC1) have been demonstrated as dysregulated in prostate cancer cells [92]. Binding to the antibody is detected through the use of a reactive tag that triggers color change in the stain where localization occurs.

[25] observed within-specimen heterogeneity in prostate tumor samples. To quantify the observed heterogeneity in the samples, [25] used a method termed compositional data analysis. Within their framework, samples were categorized into fractional components based on pure subpopulations corresponding to discrete staining levels. In the study, the authors categorized stain intensity into mild, moderate, and intense, and a corresponding mixture component from each profile for each heterogeneous tumor sample. The authors then correlated the estimated composition to recurrent or non-recurrent status. As it relates to the work of this thesis, the authors introduced the idea of quantifying tumors with respect to composition of subgroups, and further, correlating those subgroups to clinical markers.

Additionally, although not explicitly stated, [25] introduced a geometric element to the mixture analysis. That is, the authors state that a 3-state system such as theirs — having mild, moderate, and intense profiles — corresponds to a ternary diagram (i.e., a system with three subpopulations corresponds to composition of mixtures of a triangle).

[86] made explicit both the intratumor and intertumor heterogeneity considered by their model, as well as the ties to convex combination and geometry. As a result, an explicit optimization problem was stated. Later work emphasizing the biological interpretability of the optimization was substantiated. [98] enhanced the model proposed by [86] by making explicit allowances for noise in the model. As a result, fits to synthetic datasets demonstrated a higher level of robustness in the models. It is reasonable to conclude that the models of real aCGH data are also more robust than the fits proposed by [86]. As a result of the enhanced robustness, we presume the inferences as to the biology of the subpopulations of tumors made by [98] represent a closer approximation to the biological truth.

1.3 The Cancer Genome Atlas

TCGA [45] provides a substantial amount of the real data used in our experiments. This section provides a brief description of TCGA in terms of data made available, and data used. TCGA has served to provide a centralized repository of tumor data with support from both the National Institutes of Health and the National Cancer Institute. The data are highly multimodal, consisting of DNA data including copy number variation and methlyation data; and RNA data including RNA-Seq quantitative gene expression data; clinical data including demographic, drug, and longevity data. The data come from multiple organ sites (26 primary organs). Although those data that contain germline mutations have protected access, substantial portions of the data are freely accessible [45]. The large number of samples in selected organ sites — for instance, over 1000 samples in the breast tumor dataset — make the repository ideal for our work. The data used as part of this thesis include DNA copy number data, RNA-Seq quantitative gene expression data, and clinical outcomes data for BRCA, OV, LUSC, and Glioblastoma Multiforme (GBM).

1.4 Contributions of the Thesis in Brief

We advance the prior work through further development of geometric models of tumor evolution and heterogeneity. The work contained in this thesis validates the methods through a combination of synthetic measurement and correlation with results from other methods. Although prior work has made significant contributions toward a geometric understanding of tumor evolution, progression and heterogeneity, key assumptions remain that we seek to better address. The use of simplices in modeling tumor evolution encodes the assumption that for all tumor samples it is reasonable to expect a mixture of all subpopulations. However, given that cancer is an evolutionary process, we expect that structured relationships among subsets of all possible subpopulations give rise to the observed heterogeneous samples, rather than mixtures from the set of all possible subpopulations [80]. Prior to the work of this thesis, there is no method known to us imposing constraints of subsets of the mixed tumor sample subpopulations.

Additionally, previous geometric models parameterize the number of subpopuations in a tumor panel. This poses a particularly cumbersome challenge to investigators. Although cell-by-cell studies demonstrated the uniqueness of tumor cells, genetically similar collections of cells have been grouped together for the purpose of better understanding the composition of tumors [80, 86, 98]. As the diversity of the subpopulations increases, the number of groups present will decrease, but choosing a particular number without prior expert knowledge or an assumed model of probability remains challenging [80, 81]. From a technical perspective, the computational resources required to run geometric models has remained a consistent challenge. Many geometric algorithms suffer from high runtime as the number of dimensions increase. In our application, the number of dimensions are propotional to the number of subpopulations. Practically, runtime is an important consideration for deconvolution algorithms, features may number in the 10,000s, and the number of data points may number in the 1,000s. The number of samples would be expected to grow with public efforts to collect tumor samples. Therefore, methods to reduce the runtime of deconvolution algorithms represent a key step in applying more sophisticated unmix-

ing models.

Further challenges remained in relating the geometric approach pioneered in [86] and [98]. Although the models have produced putative tumor subtypes and tumor phylogenies, there have not been published correlations of those results to the deconvolution of tumors produced with other methods. Agreement between some other method [3, 37, 68, 69, 83] and geometric approaches would lend further validity to the geometric approach as an appropriate framework.

Many of the statistical models proposed prior to the work of the thesis, and concurrent with the work of the thesis — both geometric and using other frameworks [3, 37, 68, 69, 83] — are limited to a single type of data. In other cases, specific information about the platform is not required, but using both DNA and RNA data simultaneously is not demonstrated [3, 37, 83]. As efforts emerge to sequence select tumors at a single-cell level, a platform capable of leveraging the data present in bulk sequencing, as well as the flexibility to capture additional information from orthogonal data types becomes imperative.

In order to address the concerns outlined above, we contribute several key studies. To allow for more explicit determination of tumor phylogenies from a geometric perspective, we develop and implement a model of tumor deconvolution based on simplicial complexes. Simplicial complexes fit more closely the assumptions associated with deconvolution of an evolutionary process. Geometrically, a simplicial complex is a union of simplices such that the intersection of any of the components is itself a simplex. In practical terms, the simplicial complex framework developed for this thesis constrains solutions to be one connected component, and there is at least a shared point among all sub-simplices. Phylogenetically, the shared componentry — whether a point, line, triangle, tetrahedron, or hyper-tetrahedron — corresponds to a shared lineage before branching events in the tumor evolution, while vertices unique to the sub-simplices correspond to vertices unique to branches of the inferred phylogeny. The development of such models introduce new parameters to the model-building. The thesis addresses the inference of these parameters, and demonstrates the usefulness of the newly-developed model under the simplicial

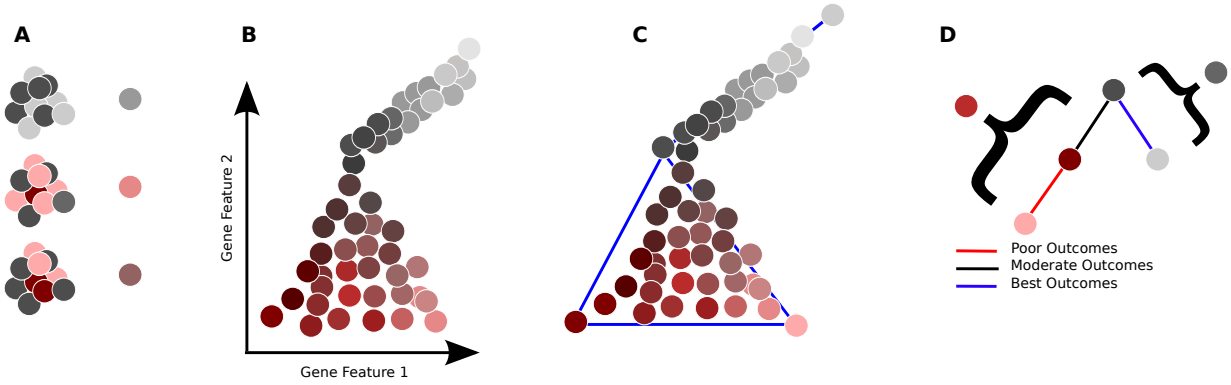

Figure 1.3: Tumor samples contain numerous subpopulations, which are mixed and have mixed representation when sequenced (panel A). Quantifying the similarities and differences in similarly evolved samples over a patient population yields geometric structure (panel B). Our contribution is to leverage the geometric properties to fit models (panel C), which can be used to recapitulate the tumor phylogeny, and to make predictions as to outcomes based on the composition of samples (panel D).

complex assumptions, as compared to the assumptions made in the simplex and robust simplex cases [86, 98].

Figure 1.3 demonstrates our overall workflow, where we model biological truths by leveraging the present geometric structure implied by the assumed evolutionary process and heterogeneity of samples.

A second key contribution of the thesis improves the clustering approach used to link the previous simplex-based approach to the simplicial complex framework of the thesis. We demonstrate improvement in both runtime and accuracy of clustering across realistic synthetic scenarios through the application of medoidshift-based clustering to the data. This non-parametric clustering approach has asymptotic runtime benefits for clustering. In addition to the novel application of the improved clustering in the simplicial complex framework, we demonstrated the ability of the improved clustering approach to meaningfully partition real-world cancer datasets into groups of differenentiated survival.

Another substantial contribution of this thesis is improvement of the simplicial complex framework to reduce parameters, allow for heterogeneous data types — such as DNA and RNA at once — and automated detection of partial membership of tumor samples in multiple phylogenetic trajectories. We reduce the number of parameters required for the unmixing framework by introducing the improvements in clustering outlined in our second key contribution to the general framework outlined in simplicial complex unmixing. The mixed membership of tumor samples across phylogenetic trajectories allows for a better capture of uncertainty in data very near to shared inferred subpopulations.

1.5 Datasets Used in This Thesis

In this thesis, we applied our methods to DNA and RNA datasets for several types of cancer: BRCA, OV, GBM, and LUSC. A practical challenge remains in the application of the approach for DNA copy number data. Input data may be formatted as a spreadsheet of calls containing the chromosome, start position, end position, and value of the aberration. However, our method requires a matrix as input. In order to derive the matrix, we implemented a lossless compression procedure we called blocking. For a static data set, we considered chromosome location pairs for which a change was observed in at least one data point in the data set at that site. We then store the state of the data set (the copy number values for each data point)prior to that change site, and store the chromosome and position pair. The chromosome and position pair then describes a genomic feature (column) in the matrix, and the values for each data point define the rows in that column. The procedure of finding positions of change is then repeated across the genome, with increasing numbers of columns as the dataset becomes more complex in some areas, and fewer features for when there are fewer changes observed in copy number. We then use the vector of stored chromosome and start positions to map back from gene features used in our method to genome positions, and ultimately to gene names or symbols. For each of the tumor subtypes, clinical data was also obtained. More specifically, used the following datasets:

- A BRCA RNA-Seq data set [45] containing 1,100 tumors at 20,531 gene sites. Each of the samples contained a bulk quantified transcriptome for a tumor sample. The dataset was accessed via `cancergenome.nih.gov`. We note this access portal has been modified since the time the data set was accessed.
- a BRCA clinical data set [45] for the 1,100 tumors above was also collected. This data included survival information, as well as clinical subtype information.
- An OV DNA CNV data set [45] containing 472 samples was accessed on 15 July, 2015. This data contained gene copy number alteration data. Each of the samples were blocked, and the top 10 principal components were used in our analysis.
- An OV clinical data set [45] consisting of 472 samples, including survival information.
- An LUSC DNA CNV data set [45] containing 408 samples. This data contained gene copy number alteration data. Each of the samples were blocked and compressed in such as way that the top 10 principal components were analyzed.
- A LUSC clinical data set [45] consisting of 408 samples. This data contained survival information.
- A BRCA DNA CNV data set [45] containing 1,079 tumor samples, at 20,531 genes.[1]
- a BRCA RNA-Seq data set [45] containing genomic tumor data for 1,041 samples at 20,531 genes. [1]
- A BRCA clincial data set [45] containing clinical samples, tumor subtype, and survival information. [1]
- A GBM genome data set accessed from TCGA [45] containing DNA CNV data.
- A GBM clinical data set accessed from TCGA [45] containing survival and demographic information corresponding to the genome data from the GBM CNV data set.

Generally speaking, prognosis for breast tumors under early detection is good, but falls based

[1]The dataset was accessed as part of the Exacel project, funded via NSF.

on clinical subtype and clinical stage [1, 71]. The BRCA datasets are variations of the same TCGA dataset. The TCGA data was not mature at the time of first access, whereas it had been curated further later through the thesis process. The BRCA genomic and transcriptomic data is used extensively due to the large number of samples relative to other organs, as well as having well-established clinical subtypes. The genomic and transcriptomic data is used in both the inferences of subtypes across the BRCA population, as well as the prediction as to the percentage-wise makeup of samples contained in the dataset. The problem of heterogeneity prediction presents an elusive and cruicial problem to better understanding cancer biology. Increased heterogeneity has been shown to correlate with poorer outcomes [1, 71]. Better estimates of heterogeneity and tumor progression may provide impactful information toward improved prognosis and diagnosis of breast tumors. By using the DNA and RNA datasets together, we demonstrated the applicability of our framework to heterogeneous sets with respect to data type, which will become increasingly requisite as small numbers of SC-seq samples are added to large respositories of bulk sequencing samples.

Ovarian data used in this study came from both genomic and clinical sources [45]. Previous work has suggested there may be a basis for molecular segregation of ovarian tumor patients correspondent to outcomes [99]. Although some subtyping in ovarian cancer has been presented [66], wide adoption of clinically substantiated subtyping of ovarian cancer remained out of reach for the timeframe of this thesis. The usage of our techniques results in clusters of samples separated by expectations of survival, which may provide a basis for a future subtyping.

Lung cancer data used in the study included both genomic and clinical sources. Some molecular markers of outcomes for lung cancer have been proposed [43], suggesting that grouping by outcome from our clustering approach could be possible. We found outcomes that had weakly significant differences; however, we concluded that these differences may be due in large part to the imbalance in inferred survival class sizes.

The GBM dataset included both clinical and genomic data [45]. Outcomes in many scenarios

for those with GBM is universally poor. Indeed, in demonstrating a limit of the applicability of our method, we were unable to segregate groups based on survival. Our clustering approach found one group, which dovetails with the clinical data suggesting all patients died shortly after biopsy [45]. The outcomes data suggests agreement with the literature in terms of difficulty finding long-term survivors of GBM [12, 95].

1.6 Organization of This Thesis

Chapter 2 is based on text originally published as [80], and provides a detailed description of the inference of tumor subpopulations and tumor phylogenies using the lens of simplicial complexes. The model presented in Chapter 2 demonstrates a highly-parameterized version of the model bridging the gap between simplex-based unmixing and simplicial complex-based unmixing by first clustering data into putative simplicies, then applying a modified form of the simplex-based unmixing, after which a post-processing stage of reconciliation into a unified simplicial complex is applied. Machine learning algorithms are then applied to correlate the results to subtypes and outcomes.

Chapter 3 is based on text originally published as [81], and improves the methods used in the clustering stage of the simplicial complex-based unmixing approach by eliminating parameters, and providing a novel kernel function well-suited to the particular geometric scenarios modeled. Simulations are performed to validate the choice of kernel function, and to demonstrate the suitability of the method for assumed scenarios.

Chapter 4 is based on text currently submitted, and describes methods used for dimension estimation in tumor deconvolution, as well as relative areas of strength and drawback. Additionally, it includes a discussion of the approach integrated into our framework.

Chapter 4 also integrates the approach outlined in Chapter 3 into the simplicial complex-based framework.

Chapter 5 summarizes the key findings and contributions, and outlines additional steps that

may arise moving beyond the scope of this thesis in terms of topics in tumor deconvolution, phylogeny estimation, and progression.

Chapter 2

A Simplicial Complex-Based Approach to Unmixing Tumor Progression Data[1]

Studies of the genetic diversity of tumors demonstrate that bulk tumor samples exhibit both intratumor heterogeneity and intertumor heterogeneity [7, 26]. As discussed in Chapter 1, this means that individual cells within a tumor sample exhibit key genetic differences from one another (intratumor heterogeneity), and also, that on the aggregate, there are genetic differences among different tumor samples taken from a patient panel (intertumor heterogeneity). Based on the microenvironments that give rise to tumors, selective pressure has been identified to act on tumors in a similar fashion to speciation [7, 26]. In particular, checkpoint pathways regulating cell development, tumor growth, and invasion have been common areas of mutatation [39, 40]. The combination of ideas — that despite individual developmental differences traceable to the microenvironments of each tumor, common pathways were observed to be perturbed — led to the development of modeling tumors as phylogenies [21].

The procedure was also dubbed oncogenetic tree formation [21]. The Desper et al. [21] approach of oncogenetic trees relied on the intertumor heterogeneity of samples, coupled with

[1]Work in this chapter is based on material originally published as Roman et al. (2015). A Simplicial Complex-Based Approach to Unmixing Tumor Progression Data *BMC Bioinformatics*, *16*(1), 254. [80]

similarities to use methods analogous to inference of phylogenetic trees for species. Following [21], [74, 75] proposed the use of intertumor heterogeneity and single-cell studies in order to increase the resolution of tumor phylogenetics. Although both approaches advanced the field of tumor progression study, both the approaches of [21] and [74, 75] carried disadvantages. The bulk tumor approach outlined in Desper et al. [21] did not account for intratumor heterogeneity, and as a result could not model the development of tumors within a single patient. This stemmed from the fact that one sample was limited to one output, which caused an output representative of an average genome, rather than any particular genome in the sample [67]. Like other previous studies [21] the work of Pennington et al. [74, 75] relied on FISH technology. The FISH technology employed by these previous works is most notably limited in the number of markers that can be measured per cell. In principle, single-cell sequencing technologies can combine the single-cell resolution of FISH with the breadth of whole-genome sequencing, but technological and cost challenges remain [63]. In an effort to mitigate the disadvantages of previous approaches, [86] integrated the ideas of geometric unmixing introduced in [23] into tumor phylogenetics. This approach has also been called "mixture modeling", although it is more appropriate to describe the sort of model as a "mixed membership model" [80]. In a technical sense mixture models refer to models in which data points come from exactly one class, and in the case of geometric unmixing for tumor data, the samples contain multiple points, therefore a mixture of different classes of subpopulations. In previous work, similar mixed membership models have been applied to determine the level of contamination from stromal cells or normal cells [11, 94]. In addition to the novel application setting, [86] provided a phylogenetic interpretation of the geometric model based on a MST. That is, the minimum span in genome space is interpreted as the minimum evolution tree. [98] extended the idea of geometric unmixing to explicitly model for noise in the data. In both [86] and [98], the data were whole-genome data. In the interim between [98] and [80] other advances in deconvolving biological data were also made. As covered in detail in Chapter 1, these advances included [37, 68, 69, 83], among others.

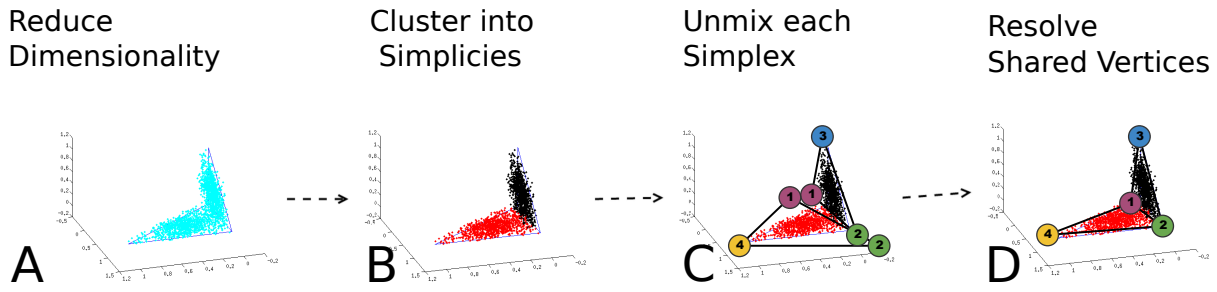

Figure 2.1: Overview of the simplicial complex unmixing pipeline. Data are reduced in dimensionality (panel A), clustered (panel B), unmixed (panel C), and resolved to a unified simplicial complex (panel D). This figure was originally published as Figure 2 of [80].

Whole-genome data most often comes in the form of bulk sampling, which contains multitudes of cells. As a result, the heterogeneous sampling contributes to the output of the sample being representative of a mixture of the subpopulations present, rather than of an indivdual subpopulation or set of subpopulations. We advance the geometric approach to leverage the phylogenetic nature of tumors as well as intratumor heterogeneity. More specifically, we model the branched nature of tumor development and progression, as well as noise in the data. The use of parametric modeling used are extended in future contributions to provide better interpretability and usability. Stated explicitly, our contributions in this chapter are as follows:

1. Extend models of geometric unmixing from simplicies to simplicial complexes and provide biological interpretation for the model extension.
2. Design and implement a per-simplex model of unmixing consistent with minimum evolution phylogenetic trees for bulk tumor samples.
3. Test and evaluate the performance of the simplicial complex model on tumor data, and interpret the results.

2.1 Methods

In this section, we describe the methodology behind extending simplex models of tumor unmixing to simplicial complex models of tumor unmixing. Further, we provide a description of the models via code, and runtime analysis of the algorithm. We then propose conditions under which the algorithm achieves optimized performance. The code for the approach is available at

`https://static-content.springer.com/esm/art%3A10.1186%2Fs12859-015-0694-x/MediaObjects/12859_2015_694_MOESM3_ESM.zip` .

Figure 2.1 shows visually the basic steps of the method.

2.1.1 Data Sets

Here we provide a brief description of the data sets used in testing and evaluating the simplicial complex method presented in this chapter. The data sets used consist of both synthetic and real data sets. The synthetic data sets were data sampled uniformly from the surface and interior of simplices. These sampled data were then concatenated to provide samples from various simplicial complexes. We joined the data to form the following simplicial complexes:

1. Two lines sharing a point
2. Two tetrahedra sharing a point
3. Two triangles sharing a point
4. Two triangles sharing an edge

A visualization of the synthetic data is provided in Figure 2.2. After the data were generated and conjoined, we added 0-mean Gaussian noise to reflect the technological limits and biological variance of the data.

We also used real tumor data in the method covered in this chapter. The real tumor data in this chapter is BRCA RNA-Seq data from TCGA [45]. The data consists of 1,100 tumors at 20,531 gene locations. Although we controlled the range of values for synthetic data, we provided pre-

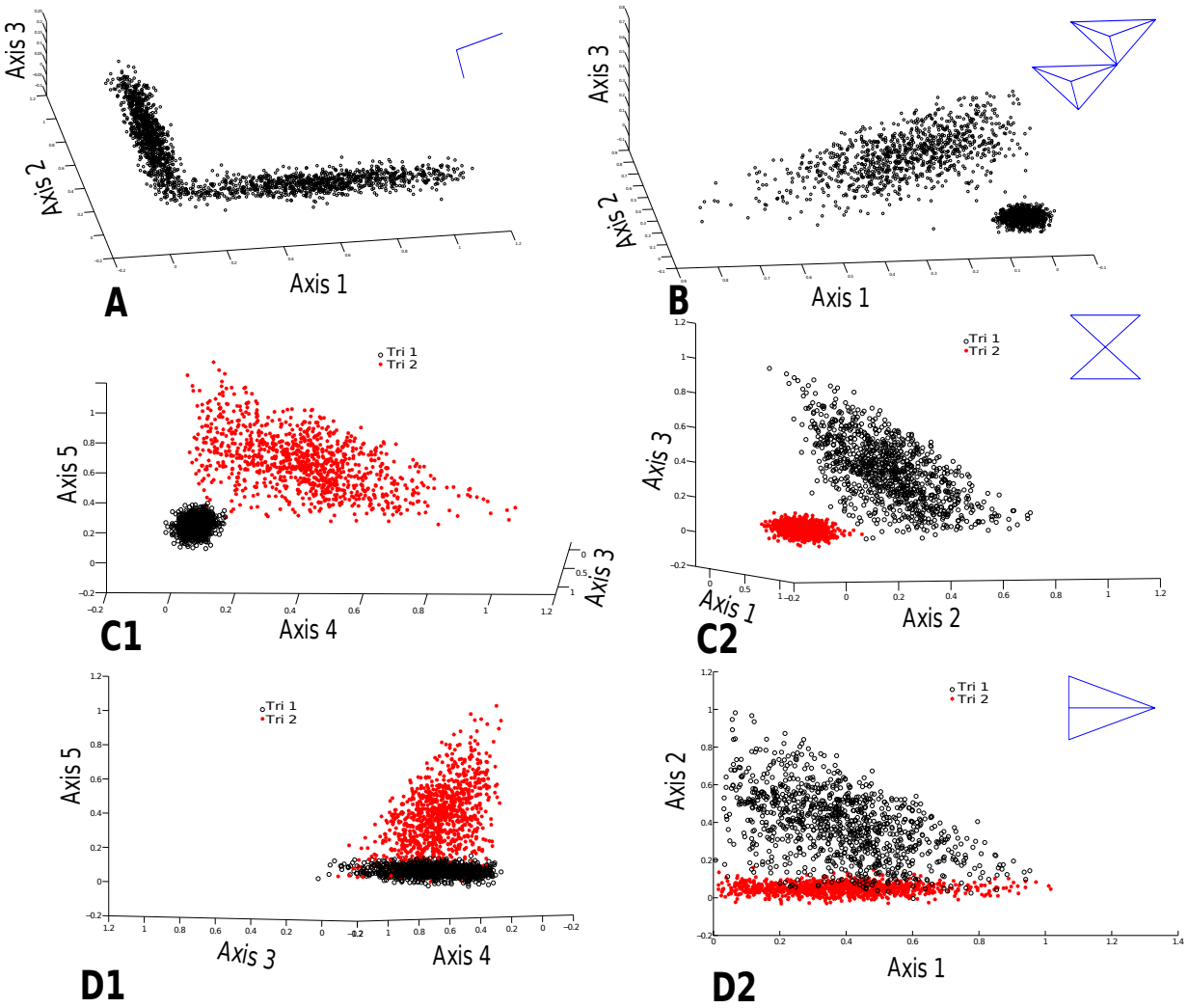

Figure 2.2: Visualization of synthetic simpicial complex data used in [80]. Scenarios considered include lines conjoined at a point (panel A) , tetrahedra conjoined at a point (panel B), and triangles conjoined at both a point and an edge (panels C and D). These correspond to phylogenies of one branching event that may have more than one common ancestor. This figure was originally published as Figure 3 of [80].

processing for the real tumor data in order to achieve the same end. The normalization procedure is described in detail in further sections. Notably, the real tumor data used in this chapter contains expression data for clinically-relevant subtyping genes: Human Epidermal Growth Factor 2 (HER2), Estrogen Receptor (ER), and, Progesterone Receptor (PR).

2.1.2 Algorithm for Simplicial Complex Approach

In this section, we describe a pipeline used for deconvoling data using the simplicial complex framework outlined in [80]. We break down the procedure into several key steps:

- Dimensionality Reduction
- Cluster Points into Simplices
- Unmix Each Simplex
- Merge Vertices to Form Simplicial Complex
- Map Components to Full Dimensional Space

In each of the following subsections, we provide more complete details as to the nature of each step.

Mathematical Model

In this section we describe the mathematical theory behind this model. We presume that a dataset we would like to analyze can be represented as a matrix. Suppose we call the matrix $M \in \mathbb{R}^{g \times s}$, where there are g gene measurements and s samples. Although we suppose the measurements are associated with genes, in principle the scope of measurement is not relevant to the mathematical model. That is, the measurements could be chromosome arms or SVs. Further, we call the samples tumor samples, although other variants of interest may be substituted. We assume that the data has been normalized by taking the Z-score of raw measurements, so that M has mean 0 and variance 1 across each gene.

The goal of the method is to infer mixture components $C = \overrightarrow{c_1}, \ldots, \overrightarrow{c_k}$. For each i, $\overrightarrow{c_i} = c_{i,1}, \ldots, c_{i,g}$ represents the profile of a subpopulation in the biology. That is, from a mathematical perspective, each $\overrightarrow{c_i}$ is a vertex of the simplicial complex. This can also be conceptualized as the basis for a convex combination to explain the data points. That is, for each j, if a sample is denoted by $\overrightarrow{s_j}$, then $\overrightarrow{s_j} = f_{j,1}\overrightarrow{c_1} + f_{j,2}\overrightarrow{c_2}, \ldots, +f_{j,k}\overrightarrow{c_k} + \epsilon_j$, where ϵ_j represents a small noise factor, and $f_{j,i}$ is a mixture fraction. Because $f_{j,i}$ is a mixture fraction, it can be constrained to be part of a convex combination. That is, we can constrain the system to those f for which $\sum_{i=1}^{k} f_{j,i} = 1$ and $f_{j,i} \geq 0 : \forall j, i$. We can then collect all of the mixture fractions in a matrix, which can be denoted F. So then, the goal becomes in the inference of C and F such that $M = CF + \epsilon$, where ϵ is an aggregated noise term. By phrasing the system as such, we mirror the goals of other unmixing or decomposition approaches.

Probability Model

The objective function we chose was driven by a probability model that enveloped two key factors: an assessment of the quality of fit of the model to the data, and the biological plausibility of the model. If we call our model Θ, and data X, we can maximize a Bayesian likelihood function:

$$P(\Theta|X) \propto P(X|\Theta)P(\Theta) \tag{2.1}$$

The functional form has a basis in the prior work in [98]. For the prior function, we chose an exponential function form for both the conditional and prior distributions to reflect strong penalties for a model failing to explain the observed data, and for increasingly large structures, correspondant to increasingly large biological evolutionary time. Plugging our chosen functional forms into Equation 2.1 produces

$$P(X|\Theta)P(\Theta) = exp(-\sum |x_i - KF_i|_p)mst(K)^{-\gamma} \tag{2.2}$$

In Equation 2.2, x_i is a data point, KF_i is the estimate of the data point based on the vertices K and mixture fractions F, p is the value of the chosen p-norm, $mst(\ldots)$ is the minimum spanning

tree function, and γ controls the relative weight given to evolutionary time as compared to the ability of the model to explain data. In practice, it is more convenient to minimize the negative log likelihood:

$$\sum(|x_i - KF_i|_p) + \gamma \log(mst(K)) \tag{2.3}$$

The choices of the $p = 1$ and $\gamma = 10$ hyperparameters are based on the in the model assumptions that the influence of one gene set to another gene set is unknown ($p = 1$), and that the noise level is about 5-40 percent of the signal level ($\gamma = 10$). The noise level assumptions are made from [19]. Significant sensitivity to the γ parameter was not observed in our simulation study.

2.1.3 Experimental Pipeline

Dimensionality Reduction

In order to pre-process real tumor data used in this chapter, we reduced the dimensionality of the original dataset to a more computationally-feasible size. In terms of the previously-specified mathematical model, we reduce the dimensionality of the input matrix M from g to g', where $g' << g$. To do so, the user sets a parameter k as the maximum number of subpopulations of any branch, usually chosen based on computational constraints, and we then choose $g' = k - 1$. By specifying this relationship between g' and k, we are using the relationship between the number of vertices of a simplex and the dimension of the simplex — that is, the number of vertices of a simplex is one greater than the dimensionality of that simplex.

In order to realize our goal of dimensionality reduction, we used PCA [73]. It is understood that there are numerous mechanisms to achieve this goal of dimensionality reduction — among them, Independent Components Analysis (ICA) [47], kernel PCA [61], or other dimensionality reduction techniques such as local PCA [107] — PCA has a straightforward implementation, and it is unclear how differences in dimesionality reduction techniques might lead to advantages in terms of implementation or interpretation.

The PCA decomposes the original matrix M into approximately $M'V + A$, where M' is the

closest representation of M in the lower-dimensional space, V is a matrix consisting of $k-1$ orthogonal basis vectors, and A is a matrix of offsets. A can also be conceptualized as the set of translations that shift $M'V$ back to M. Granted, PCA is a lossy compression technique when used as outlined above, so there is a noise factor to have true equality: $M = M'V + A + \epsilon$ — again, in a bit of an abuse of notation, ϵ represents a noise factor, which may differ from the noise factors outlined above.

Cluster Points into Simplices

After the data were compressed, we needed to have determined subgroups of the data set (that is, groups of tumor samples) that collectively lie within the span of unique low-dimensional simplices. In a certain sense, this is the task of a clustering problem; however, the nature of the clustering problem is rather unusual. Specifically, standard clustering approaches look for points where the spatial distance within the cluster is small, and between two clusters is large. Rather than that spatial marking, our clusters must have small distance from a simplex that we later specify, and distinct dimensionality from other simplices later specified. In other words, the set of dimensions sufficient to explain one cluster should be small and distinct from the set of dimensions necessary to explain other clusters. In order to meet the needs of the situation, we use a distance measure with a basis in manifold learning [84].

Algorithmically, we first form an $s \times s$ matrix D where the entries of the matrix are the square of the Euclidean distance between samples. That is $D_{i,j}$ is the square of the Euclidean distance between the i^{th} and j^{th} sample. We then create an updated distance matrix D', where $D'_{i,j}$ is the shortest path distance from the i^{th} to the j^{th} sample using the edge weights found in D. This procedure approximates the geodesic distance that fits the assumptions described earlier in this section [84, 97]. The geodesic distance is the distance within the manifolds necessary to go from one point to another, rather than the Euclidean distance that can cut through different manifolds.

After this distance metric was established, we applied k-medoids clustering [41]. Our goal

using this algorithm was to create $k \in \mathbb{N}$ clusters where the reprsentatives of the clusters were the medoids of the clusters — that is, the data points nearest to the mean in Euclidean distance. The small difference with the original k-means technique outlined in [41] prevents the inference of cluster representatives outside of the span of the clusters, which leads to no natural biological interpretation.

Unmix Each Simplex

After the data were partitioned into putative clusters, we fit a simplex to each of the clusters that we later merge into a simplicial complex. In order to fit the simplicies to each of the clusters, we optimize the negative log likelihood described in the probability model subsection, subject to the constraints of convex combination on a per simplex basis. That is, we seek

$$\begin{aligned} \underset{K,F}{\operatorname{argmin}} \sum |x_i - KF_i|_p + \gamma \log(mst(K)) \\ \texttt{s.t.} \quad \forall F_i : \sum F_i^t = 1, F_i \geq 0 \end{aligned} \tag{2.4}$$

where $\sum F_i^t$ is the sum over all inferred components (components of K) is 1, and each of the mixture fraction entries is non-negative. x_i, K, F, p, and $mst(\ldots)$ are as in the probability model section. By using a minimum spanning tree function rather that the volume prior in the previous works, we are able to sum priors over each of the clusters / subsimplices. As a result, there is a unified measure of evolution across the entire simplicial complex, even in the situation where different subsimplices have different dimensionalities.

In order to implement the code, we use an initial guess based on the Minimum Volume Enclosing Simplex (MVES) code outlined in Chan et al. [13]. Then, we alternate optimization over the F and K hyperparameters by holding the other constant. When the K are fixed, the F are computed by using the `lsqlin` function built into Matlab, and while F are fixed, the K are computed with `fmincon` using the Matlab platform. The optimization terminates when a fixed number of iterations are exceeded, or the objective function does not improve within a tolerance parameter. In later steps, we use a measurement of simplex uncertainty to yield fits to

the unified simplicial complex. To simulate this uncertainty, we use bootstrapped replicates with 10 replicates. For an example, see [106].

Merge Vertices to Form Simplicial Complex

In order to form a unified simplicial complex, we use the data from previous steps that infer mean positions and variance of the positions of vertices for subsimplices. In order to identify pairs of vertices that may be representative of the same ground truth subpopulation, we look for overlap in the surfaces from which we assume the trials are sampled. That is, if the Euclidean distance between two inferred vertices is less than the sum of the individual variances, we merge the vertices. Formally, we merge a vertex v_1 with a vertex v_2 when

$$\begin{aligned} &\exists v_1, v_2 \in V : v_1 \neq v_2 \cap \exists i \in \mathbb{N} \cap [1, |v_1|] \\ &\texttt{s.t.} \sum \frac{||v_1 - v_2||}{\max_i(\sigma(v_{1,i})\sigma(v_{2,i}))|v_1|} \leq 1 \end{aligned} \tag{2.5}$$

where

- V is the set of vertices
- $||\ldots||$ is Euclidean distance
- $\sigma(\ldots)$ is the standard deviation
- $v_{j,i}$ is the j^{th} subsimplex in the i^{th} index of vertex for that subsimplex.

Map Components to Full Dimensional Space

In our pipeline, the solved vertices, correspondent adjacency matrix, and mixture fractions determined in the merging step of the procedure are solved for in PCA space. In order to determine the genes of interest related to the numerical inference, we transformed the vertices back into gene space. This is accomplished because the coefficient matrix (loading matrix), first determined in the pre-processing step, can be used to map the vertices back to gene space. The coefficient matrix describes for each principal component what the linear combination of features is that

composes a principal component. So, we can use algebra to determine the inverse operations, and perform those to the principal components to yield values in the gene space.

2.1.4 Complexity Analysis

There are a substantial number of parameters influencing the performance of the code. Additionally, some aspects of the code — for example, the k-medoids clustering portion — are iterative, so the overall complexity expression can be complicated to express. However, a piecewise breakdown on an upper bound of the runtime is reasonable to express. The PCA to pre-process the code uses Singular Value Decomposition (SVD) on the backend in Matlab [33]. The runtime for SVD is $O(mn^2)$ for m data points and n dimensions, which in our case would be the number of genes. The clustering phase of our protocol uses k-medoids [41]. The complexity of runtime for this phase is $O(km)$. The simplex fitting for each cluster applies a similar procedure to the previous work outlined in [98], with a modified objective function for each cluster. So, for T rounds of iternation, k clusters, d_p simplex vertices for the p^{th} simplex, and P PCs, the runtime is $O(Tkn \max_p(2^{d_p+1}, nP^2))$ for each of the b bootstrapped replicates. The resolution phase is computed by learning a Gaussian Mixture Model (GMM) for each of the $\sum_p(d_p) + k$ vertices [101], which would contribute a factor of $O(b(\sum_p(d_p)+k))$. So, the runtime of the entire pipeline can be upper-bounded by $O(\max_{m,n,k,m,b,T,k,d_p,P}(mn^2, km, bTkn, \max_p(2^{d_p+1}, nP^2), b(\sum_p(d_p)+k))$. In practical terms, the bottleneck occurs in the optimization / simplex inference stage, given that there is an exponentially increasing cost for increasing the dimensionality of the simplices. However, this can be an improvement due to a dependence on the maximum on a sub-simplex, rather than the sum total of vertices in the prior work.

2.1.5 Experimental Application

In order to test the usage of the pipeline, we tested the procedure on both synthetically generated data, as well as real tumor data acquired from TCGA. Because it is difficult to envisage an

accurate ground truth model for how the heterogeneous combinations of cells and their progression interact, we base our synthetic model on sampling uniform mixtures of cells in space. The scenarios tested (see Figure 2.2) correspond to multiple realistic phylogenetic scenarios for our model. In addition to testing against the previous robust simplex approach in the synthetic cases, we also tested against a GMM, part of the architecture used by the PyClone [83] approach.

Real tumor data were taken from TCGA. BRCA RNASeq data were downloaded from the TCGA data portal, and preprocessed to obtain Z-scores. The BRCA RNASeq panel was chosen due to the large number of samples available publically compared to other patient panels, as well as the clinical subtypes. As part of the validation of the approach, we correlated the mixture fraction inferences made by the simplicial complex approach outlined in this chapter to the clinical labels supplied by TCGA.

2.2 Results

2.2.1 Synthetic Data

We tested the ability of the method on key tasks: inference of mixture fractions, and inference of vertex positions. The mixture fraction inference task represents the inference of the composition of a tumor, and the vertex inference task represents the inference of the genomic profile of a subpopulation. In order to test the method, we used Root Mean Squared Deviation (RMSD) per point per dimension for mixture fraction inference, and RMSD per model unit per dimension for the task of vertex inference. Our method compares favorably to the simplex-based approach outlined in [98] in both tasks, and depending on the noise level, compares favorably to the GMM. The results are displayed in the figure below. Increasing values along the x-axis in both figures correspond to increasing standard deviations of the noise added, while increasing values on the y-axis indicate increasing RMSD.

An examination of the results and methods led us to expect the simplicial complex approach

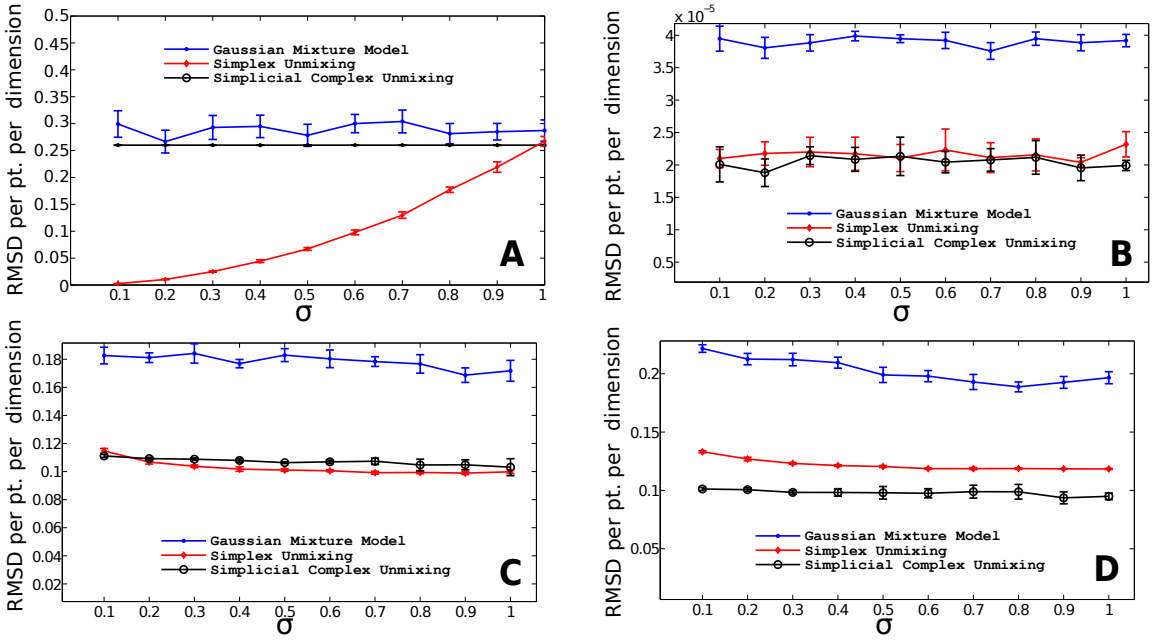

Figure 2.3: Comparison of the ability of methods to infer mixture fractions in multiple synthetic scenarios. Included are two lines at a point (panel A), two tetrahedra at a point (panel B), two triangles at a point (panel C) and two triangles at an edge (panel D). This figure was originally published as Figure 4 in [80]. In the figure, σ represents the standard deviation of noise added to the synthetic data.

ought to have performed better than the other approaches, particularly at low noise levels. This increased performance was observed; however, it was unexpected to observe that the GMM approach performed better in the vertex resolution task in some scenarios at higher noise levels. Although at the noise level $\sigma = 1$, corresponding to a signal-to-noise-ratio of 1, we would expect the data would be inferred better by a GMM, we did not expect the result to hold at lower values of σ. In general, though, for the mixture fraction inference task, across all scenarios, the geometric methods performed better than the GMM.

We came to a few points of understanding. The explicit nature of the geometric models may provide an advantage in the point cloud structure in these cases. The dimensional aspects of the simplicial complex model in particular would drive a greater advantage in cases where the

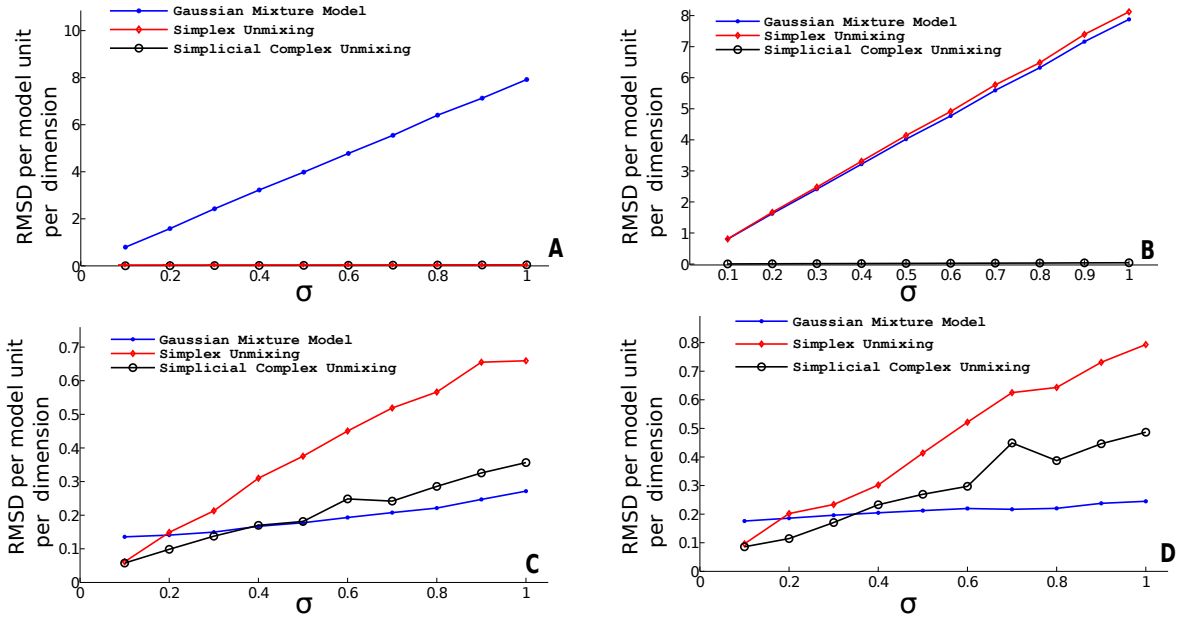

Figure 2.4: Comparison of the ability of methods to infer vertices in multiple synthetic scenarios. Included are two lines at a point (panel A), two tetrahedra at a point (panel B), two triangles at a point (panel C), and two triangles at an edge (panel D). This figure was originally published as Figure 5 in [80]. In the figure, σ represents the standard deviation of the noise addded to the synthetic data.

low-dimensional clouds are distinct, and embedded in a relatively high dimensional space — as this would cause GMMs to make inferences based on the entire ambient dimension only, and the simplex method would perform inferences based on the sum of the low-dimensional spaces, rather than on each low-dimensional space in turn. That geometry corrsponds to situations where tumors have branching events early in the progression of the tumors, so that the spaces spanned by the subtypes are generally distinct. This understanding is supported by the higher relative performance of the simplex approach in scenario 1, and the simpicial complex approach in scenario 4. In terms of vertex inference, a more fuzzy picture emerged. Although the GMM performed best in high-noise scenarios, prior work on RNASeq [19] showed that a realistic estimate for the noise is to the lower-middle end of the tested noise levels.

Further, although real-time runtime was not the primary focal point of the method, we presented real-time runtime data below:

Scenario	Simplicial Complex	Simplex	GMM
Two Lines	173s	0.794s	0.0955s
Two Tetrahedra	297s	4.62s	0.112s
Two triangles at a point	829s	1.22s	0.133s
Two triangles at a line	554s	5.94s	0.147s

Table 2.1: Runtime in seconds based on a run of 400 data points in each synthetic scenario

The results of Table 2.1 demonstate the additional sophistication of the simplicial complex approach carries a true cost in terms of real-time running time. However, the algorithm needs to be run only once per data set, mitigating the increased cost of runtime.

2.2.2 Validation on Real Tumor Data

We ran the described algorithm on 1,100 tumor samples from TCGA [45]. The clustering step found two clusters in the data. The unmixing phase identified seven total subpopulations, two of which were determined to be overlapping based on the merging criterion. As a result, six subpopulations in total were extracted from the data. The six subpopulations formed a simplicial complex with a triangle as one sub-simplex, and a tetrahedron as the other sub-simplex, joined together by a point. A visual representation of these findings is presented in Figure 2.5:

In order to ensure the method had extracted clinically-relevant findings, we sought correlation between the results of our approach, and known breast tumor subtypes. First, we performed correlation between the vertex with the highest mixture fraction — that is, which vertex number was presumed to contribute most to a sample — and clinical label. Based on a chi-squared test of independence for the contingency table of subtype distributions over vertex labels, we recorded a chi-squared score of 542.5002 ($p < 0.001$). Next, we sought a correlation between the distribu-

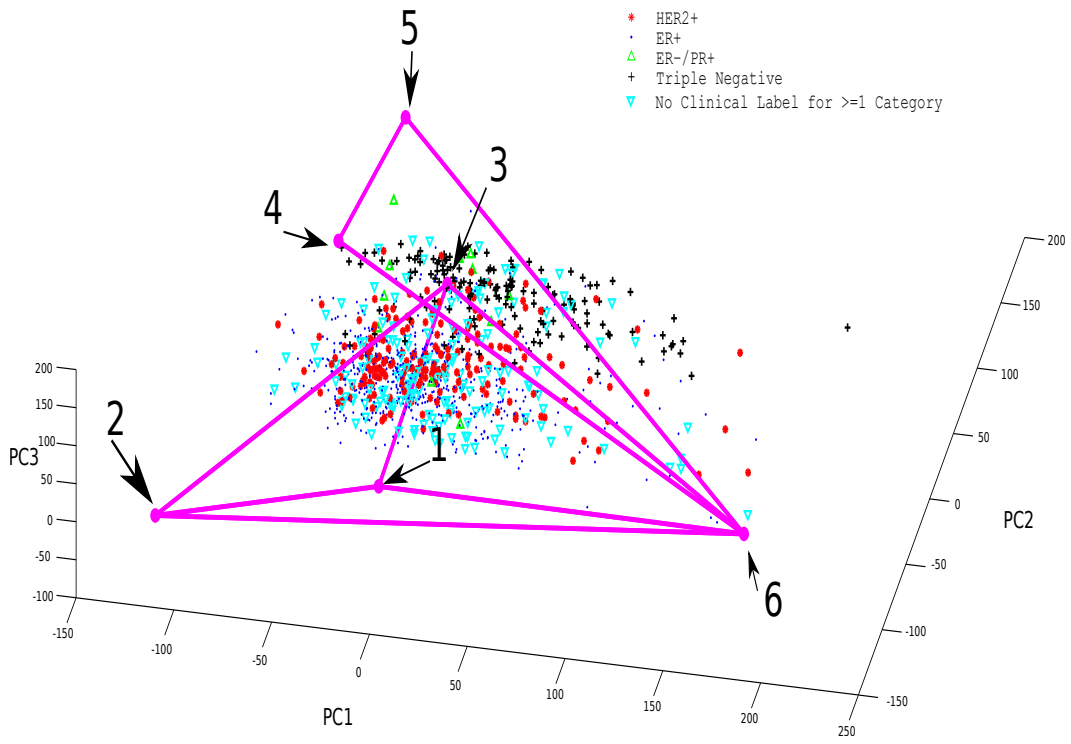

Figure 2.5: Simplicial Complex Fitting to TCGA RNASeq BRCA data [45]. Data are plotted as a projection in PC space. Subtype labels are inferences of clinical labels (HER2+, HER2-;ER/PR+; Triple Negative, or missing label). This figure was originally published as Figure 6 of [80].

tion of subtyping labels from a standardized 50 gene breast cancer panel based on the Prediction Analysis for Microarray 50 Gene Panel (PAM50) [72], and our simplex assignments (that is, sub-simplex 1 or sub-simplex 2). Using a chi-squared test for independence for a contingency table of subtype distribution over sub-simplex labels, we recorded a test statistic of 775.2952 ($p < 0.001$). In the labeling, if the assignment of genes names were ambiguous from PAM50, we used the first suggested gene name from the BioDB gene ID conversion tool [48]. We further

sought a less granular comparison of the PAM50 [72] subtyping to our subtyping. To do so, we peformed a chi-squared test of independence between the predominant mixture fraction outlined in the clinical label case, and the PAM50 labeling. We found a strongly significant chi-squared test statistic of 1128.6668 ($p < 0.001$). The result suggests that goemetric structure in the PC space can be related to clinical measures.

In order to better understand the annotations of pathways, we leveraged the knowledge base of the DAVID ontology analysis tool [20]. For each of the inferred subpopulations of our model, we took those genes with Z-score ≥ 3 or ≤ -3. We used these lists of genes as input to the DAVID platform, and recorded the top ontological terms associated with each. The tool was able to identify terms and pathways of interest, which we present in Tables 2.2 and 2.3

	Z-score ≥ 3
Vertex 1	Defense Response, Inflammatory Response, Response to Wounding
Vertex 2	Myofibril, Contractile Fiber Part, Sarcomere, I band, Actin Cytoskeleton, Z Disk
Vertex 3	$\emptyset$
Vertex 4	RNA recognition motif (and assoc. terms), RNKP-1, alpha-beta plait
Vertex 5	Cell cycle (and assoc. terms), M phase, organielle fission, nuclear division
Vertex 6	Ribosome (and assoc. terms), translational elongation, cytosolic part

Table 2.2: Increased Z-score DAVID [20] ontological results.

Tables 2.2 and 2.3 were originally presented as Tables 1 and 2 in [80]. We did not find breast cancer specific roles; however, the terms could be related to pathways connected in many cases to the hallmarks of cancer [39].

Additionally, we constructed a minimum spanning tree of the simplicial complex in order to make inferences as to how the subtypes associated with each of the vertices may have evolved. The phylogeny is presented in Figure 2.6:

We attribute the fulcrum of our simplicial complex to be the most recent common ancestor of the data points, and given that tumor cells progress from normal tissue, we assign this label.

	Z-score ≤ -3
Vertex 1	$\emptyset$
Vertex 2	Chemokine (and assoc. terms), Cytokine (and assoc. terms), interleukin-8-like
Vertex 3	Alternative splicing, splice variant, phosphoprotein, polymorphism, sequence variant
Vertex 4	Signal, signal peptide, glycoprotein, glycosylation site: N-linked, disulfide bond
Vertex 5	icosanoid, unsaturated fatty acid, alkene, leukotriene, transmembrane protein, lipid
Vertex 6	Zinc (and assoc. terms), c2h2-type

Table 2.3: Decreased Z-score DAVID [20] ontological results.

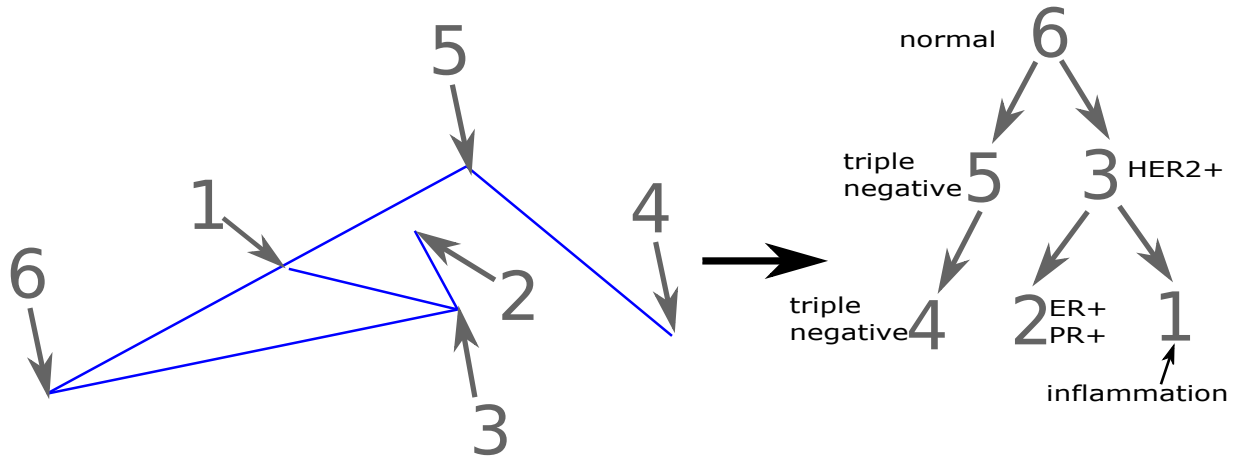

Figure 2.6: The MST (left side) relates to the phylogeny presented (right side). Clinical labels associated with each vertex are used to label the phylogeny inferred from the data. This figure was originally published as Figure 7 in [80].

We assign the inflammation label based on the DAVID terms associated with vetex 1.

The most direct possible comparison with existing tools was with the purity estimation tool called ESTIMATE [108]. Although our approach is designed for a more general problem, we sought to compare the estimates derived from our approach with the method. By considering all non-vertex 6 (normal) tissue as tumor tissue, we were able to derive purity fractional estimates on our samples. While the method assigned only a fraction of samples to contain a portion of vertex 1, those that were assigned showed correlation to the ESTIMATE [108] samples at 0.264

Spearman correlation ($p < 0.00001$). Other correlations did not prove as statistically significant. The lack of coherence may reflect a misidentification of component 6 or incompatibility of the scoring schemes.

2.3 Conclusions

The key contribution outlined in this chapter represents an improvement in the state of the art of geometric tumor unmixing. By allowing for a more realistic model of tumor subpopulation evolution, the model is better able to fit mixture fractions and subpopulation genomes simultaneously when compared to classical models (GMM), and to the prior state of the art (simplex unmixing). Additionally, clinical insights derived from the model provide a validation of the proof-of-concept. However, gaps still remained at the conclusion of this work. Specifically, the decreased performance in some tasks of the simplicial complex model to the GMM suggested that improved performance might be obtained by integrating aspects of the GMM into the geometric framework. Additionally, less noisy data types — such as DNAseq — would in general provide a decreased noise profile when compared to RNASeq. Ideally, a complete MST optimization would be computed over the entire simplicial complex, rather that piecewise; however, this may provide substantial algorithmic challenges. Nevertheless, significant improvement could be made to model the optimization as over the entire simplicial complex.

Chapter 3

Medoidshift Clustering Applied to Genomic Bulk Tumor Data[1]

3.1 Introduction

As outlined in Chapter 2, within the simplicial complex unmixing framework, pre-clustering of data points into putative sub-structures remained an important facet of the research after publishing the framework [80]. This chapter focuses on improvements made to the clustering approach that are provided in [81]. The k-medoids based approach in [80] is a parametric clustering approach. That is, the user to the simplicial complex platform must pre-specify the number of clusters. The biological interpretation of this parameter is the number of putative paths within a tumor phylogeny. As a result, except for users expert in both biology and computation, it remains a challenge to use the platform.

In order to make the platform more user-friendly, we created a novel clustering approach more specific to the application setting. The approach was designed to eliminate parameters (i.e., be non-parametric), to be scalable to larger data sets in terms of runtime, and to leverage

[1]This chapter corresponds to work originally published as Roman et al., (2016). Medoidshift clustering applied to genomic bulk tumor data. *BMC Genomics*, *17*(1), 6.

the geometry specified in the problem of tumor deconvolution inference.

A novel variant of medoidshift clustering provides the requisite qualities outlined above [89]. Medoidshift clustering is a non-parametric clustering approach with the option to apply kernel functions to the distance metric used. Additionally, it does not suffer from the so-called ”curse of dimensionality” that affects other clustering approaches [41, 101]. Additionally, Medoidshift clustering allows for a flexible kernel function [89], including the use of kernel functions such as ISOMAP [97] that approximate the geodesic distance well-suited to the phylogenetic scenarios we encounter.

Chapters 1 and 2 contain additional details on how the dimensional and mathematical properties relate well to the biological properties of the system of study. The constraints of evolution in tumors coupled with the high dimensionality of high-throughput biological experiments, and tumor heterogeneity creates data sets where each sample has data a large number of probes, but each of the paths in the tumor phylogeny can be defined by a relatively small number of features. These paths corresponds to clusters of the data in a genomic sense. Further, the intrinsic dimensions used to characterize each of the paths may be partially overlapping.

As a consequence of the constraints of the system, medoidshift clustering [89] fits our system quite well. In this chapter, we discuss a focal point of our contribution by demonstrating the applicability and limits of medoidshift clustering to genomics data. We present an application to our method on synthetic data consistent with the constraints of tumor phylogeneitcs, and assess our method in comparison to other kernel approaches in medoidshift [89].

3.2 Methods

Medoidshift falls within a class of clustering approaches known as mode-seeking clustering [89]. A closely related mode-seeking clustering approach on which Sheikh et al. [89] based their approach was meanshift clustering [17]. The key modification of medoidshift clustering when comparing to meanshift clustering is that in medoidshift clustering, cluster representatives are

constrained to be themeselves data points, whereas the cluster representatives in meanshift clustering need not be contained in the data set. Additionally, because of the constraints of medoidshift clustering, recomputation with the addition of more data to a pre-computed data set is less computationally costly than in meanshift clustering [89].

We now provide a brief description of the workflow model of medoidshift clustering. Data are input as a matrix $X \in \mathbb{R}^{m \times n}$, where there are m data points, each with n ambient dimensions. With a shadow kernel function $\Phi(\ldots)$, and a neighborhood hyperparameter $h \in \mathbb{R}$, with a scaling constant $c_0 \in \mathbb{R}$, then, we can define a weight that describes a contribution of a given point in the data set to a centroid for a given cluster as follows:

$$f(x) = c_o \sum_{i=1}^{m} \left(||\frac{x - x_i}{h}||^2 \right) \tag{3.1}$$

as described in [89]. In Equation 3.1, c_o is the scaling constant outlined above, $\Phi(\ldots)$ is a shadow kernel function, and $||\ldots||$ is a distance function.

The algorithm works by assigning the medoid — that is the data point nearest to the center-of-mass of a cluster — as the cluster representative of a neighborhood of points. We can then describe how a data point and its representative relate to one another in the following way, using an extension of the notation above.

$$y_{k+1} = \underset{y \in N(x)}{\arg\min} \sum_{i=1}^{m} ||x_i - y|| \phi \left(||\frac{x_i - y_i}{h}||^2 \right) \tag{3.2}$$

where y is the representative, subscripts designate index of the step of the iterative procedure, and $N(x)$ is the set of points within the neighborhood of x — that is, the neighborhood h designates a radius, and the points within that radius are the composition of $N(x)$. To further clarify, $\phi = -\Phi'$;. In theory, y_0 can be any point within $N(x)$ to form a base case, although in practice we chose the point nearest to x in kernel space.

Shiekh et al. [89] offers a proof of convergence for the approach, and details in the theory of medoidshift. In the remainder of the chapter, we reference "no kernel" or "kernel-free" approaches where Euclidean distance is used; however, we clarify that strictly speaking the Eu-

clidean distance function itself is a kernel function. The usage of the terms are a matter of parlance and simplicity when referencing the outcomes of each of the various methods.

Based on the way we normalized data in our experiments, a neighborhood hyperparameter value of 1 was chosen ($h = 1$). A scaling factor of 1 ($c_o = 1$) was also used. This stemmed from the fact that data can be normalized be be within the $[0, 1] \in \mathbb{R}$ range in each dimension. The choice of hyperparameter values encoded the belief that each cluster is approximately unit length and influential based on the distance in the kernel space. Additional pre-processing was performed on the real tumor data used in this study, and is described in the real tumor data subsection of the data section of this chapter.

3.2.1 Two-stage medoidshift clustering

Several of the key features of our data set demand modifications to the basic medoidshift procedure outlined above. In particular, our data contains biological and technical variation, which produces noise in the data. Additionally, a denoised version of the data set lies in distinct but overlapping low-dimesional subspaces relative to the ambient dimensionality of the data. Details as to how the technical conditions of the data set relate to the biological systems of study can be found in Chapters 1 and 2. Supposing the data are denoised, the geodesic distance remains a widespread and applicable measure of distance between points, and the ISOMAP kernel [97] encodes the geodesic distance into an algorithm. To do so, data points are conceptualized as a complete graph, where the edge weights are the distances between any two points. The ISOMAP distance is then the shortest path length between pairs of points in the graph. In general, the tumor data used in our studies would be expected to align with the situations for which ISOMAP is most appropriate [84, 97]. Nevertheless, the noise produced by real-world biological data sets does not conform well to the situations for which ISOMAP was developed; and it has been noted that in the presence of factors of noise, the performance of ISOMAP in identifying distinct low-dimensional sub-space drops off considerably [5]. It is exactly the case that we seek to resolve

the low-dimensional sub-spaces in the prescence of noise [69, 76].

As a result, we have sought to balance the ISOMAP [97] algorithm's ability to identify sub-spaces in noiseless cases with an earlier stage to suppress noise. In the first iteration of medoidshift in our approach, we used a kernel-free Euclidean squared distance to determine a compressed and lower-noise representation of a data set. We use the square of Euclidean distance to save computational resources compared to Euclidean distance. Because medoidshift uses mode-seeking, the representation of the data set after the first round will be a compression of the original data, including a compression of the noise present [89]. By using this distance measure for the first round, we built in the assumption that the signal-to-noise ratio is relatively high, and the noise does not exhibit meaningful bias. Based on prior work, these are reasonable assumptions for DNA CNV data [19].

For all rounds following, we used the ISOMAP [97] distance, with a kernel function K:

$$K = 1 - e^{D/h} \tag{3.3}$$

where D is the matrix of ISOMAP-like [97] distances between points, and $h = 1$ is the neighborhood hyperparameter. Distances in D are computed as shortest paths, where edge weights are squared Euclidean distances. The points used in computing D, however, are those representatives from the first round, rather that a computation on the raw points of the data set. We continued to use the kernel function K for all rounds following the first until convergence. As a result of our choice of a negative exponential kernel function, representatives were chosen as extrema in each of the clusters — that is, data points that were most outside the span of the sub-spaces that were descriptive of other clusters. This choice was meaningful for our data, given that tumor data consists of distinct evolving subpopulations [26]. As a result, extrema ought to be more characteristic of potential evolutionary trajectories than data points that are reasonably represented by other putative evolutionary trajectories. Here, the ability to represent a data point by using another sub-space corresponds biologically to the ability to represent a data point by using a different combination of subpopuations. By combining these two stages by altering the

kernel function based on the stage of the clustering, we created a method that clustered data from the low-dimensional sub-spaces representative of our biological setup that was also able to accomodate the noise inherent with biological experimentation.

The method was run on a desktop workstation equipped with an Intel Core i7-4770K processor at 3.5 GHz per core with 32GB of Random Access Memory (RAM), and Matlab running in 64-bit mode.

3.2.2 Validation on Synthetic Data

We considered application to seven unique synthetic scenarios, to imitate seven realistic phylogenetic scenarios. For a given phylogenetic scenario, each path from root to leaf in the phylogeny corresponded with one sub-space or sub-simplex in a simplicial complex. The number of nodes in the path correspond to the number of nodes in the sub-simplex. The number of shared nodes with alternate paths correspond to the number of shared vertices with other sub-simplices. For full details on the relationship between tumor heterogeneity and simplicial complexes, see Chapters 1 and 2. In Figure 3.1 is a visual representation of each of the scenarios, from both a geometric and phylogenetic perspective.

Each synthetic scenario began with a description of a tumor phylogeny — that is, an evolutionary tree describing how subpopulations present in a tumor may have evolved. Given the inter- and intratumor heterogeneity present in tumor samples [21, 26, 64, 67, 85], mixtures of subpopulations would be expected to be observed in samples. As a result of the convex combination constraint, for a given phylogenetic path, a simplex can be constructed [98]. By combining multiple paths, the simplicies can be stitched together to form simplicial complexes [80]. For instance, Figure 3.1 panel A presumes there is an early state that progresses to a progeneitor state that splits into two differentiated tumor subtypes. The result of these two paths, each with three subpopulations, and with two ancestors in common, would be two triangles conjoined by an edge. Similar interpretations can be had throughout the rest of Figure 3.1. The seven scenarios

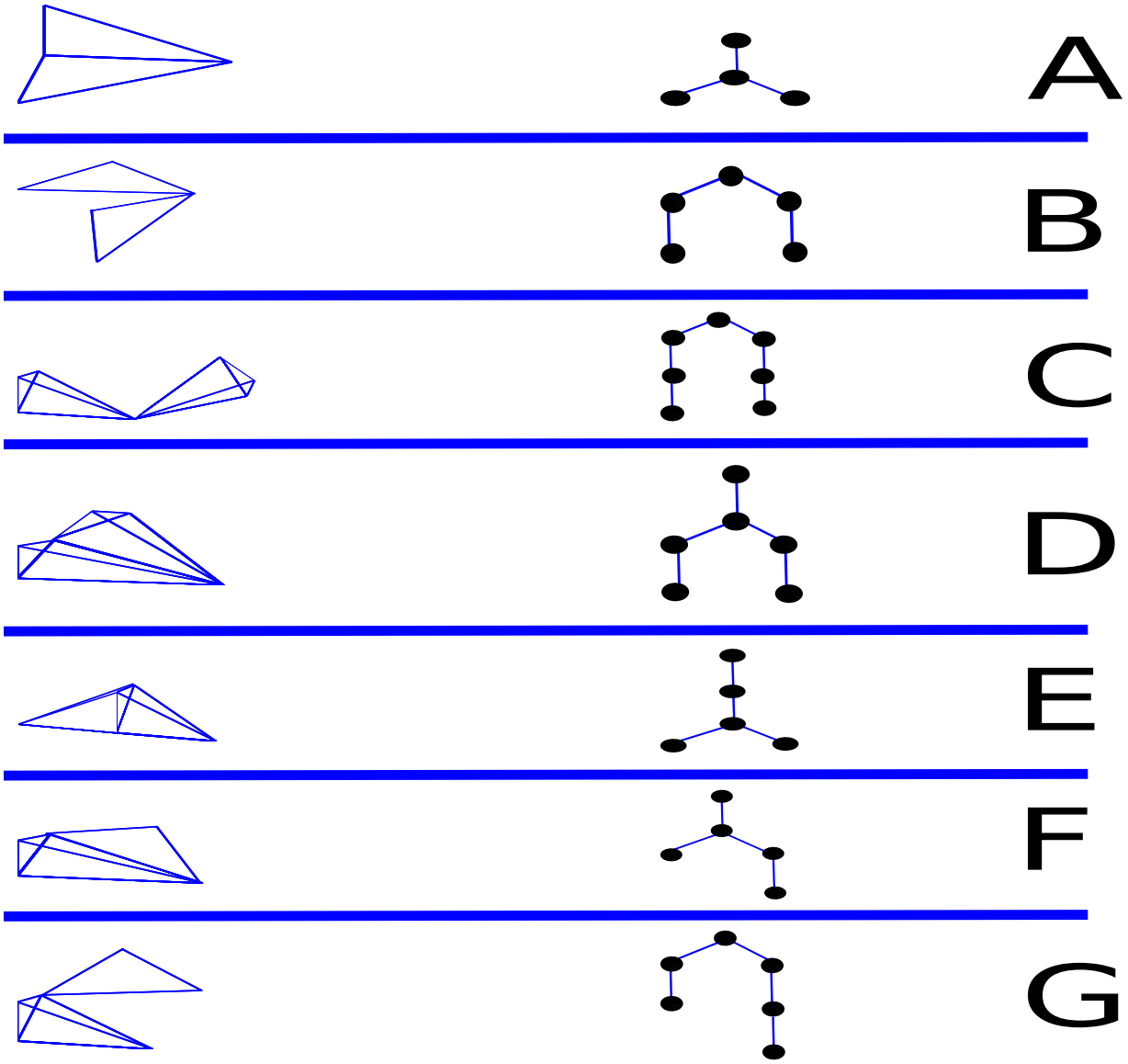

Figure 3.1: Visual representation from geometric (left column) and phylogenetic perspective (middle column) of each of the seven synthetic scenarios considered in the content of this chapter. Labels (right column) designate each row / scenario. This figure was originally published as Figure 1 of [81].

of study can be geometrically described as follows:

1. Two triangles joined at an edge
2. Two triangles joined at a point

3. Two tetrahedra joined at a point
4. Two tetrahedra joined at an edge
5. Two tetrahedra joined at a face
6. A triangle and a tetrahedron joined at a point
7. A triangle and a tetrahedron joined at an edge

In each of the cases, we generated 100 replicates. In each of the replicates, 500 data points were embedded in a 10 dimensional space. The 10 dimensions approximated the PCs of the data used in our previous contribution. Unbiased Gaussian noise in 10 dimensions was then added, with standard deviation 0 (no noise) up to standard deviation of 0.2 in increments of 0.05. Alterations in the standard deviation of the noise allowed us to probe how the performance of alternative approaches would vary with respect to changes in noise. We considered three approaches to the application of medoidshift clustering: using a kernel-free Euclidean-based approach for all stages, using the negative exponential kernel function from Equation 3.3 in all stages, and using the 2-stage approach where the first stage is Euclidean-based distance, and the remaining stages employ the kernel function from Equation 3.3.

We simulated points in a p-dimensional PC space by first defining a set of k basis vectors. We called this set $B \subset \mathbb{R}^p$. Then, a given sub-simplex can be conceptualized as $b = (b_1, \ldots, b_{k'}) \subset B$, representing $k' \leq k$ of the possible basis vectors. To generate a mixed sample, we then chose raw propsensities from a uniform distribution between 0 and 1 inclusive, and normalized by the sum of these propensities to derive mixture fractions. A mixed sample then consisted of an element-wise sum of the mixture fractions multiplied by the vectors in b. To add noise, we sampled from a multivariate Gaussian with 0 bias, and added it to the mixed sample. The standard deviation varied in the manner described in the earlier section to assess sensitivity to noise.

In order to assess the goodness of fit from the clusters produced by the medoidshift clustering approach to the ground truth clusters, we used the adjusted Rand index [44, 79, 103]. The

adjusted Rand index can be computed in the following way. The index assumes we have two partitions of a data set. Say partition $X = X_1, X_2, \ldots, X_r$, where each X_i is a set of points comprising a cluster in X for each $i \in [1, r] \cap \mathbb{N}$. Further, suppose another clustering is $Y = Y_1, Y_2, \ldots Y_s$, where each Y_j is a set of points comprising a cluster in Y, for each $j \in [1, s] \cap \mathbb{N}$. Further, we designate a matrix N to describe the amount of intersectionality in the clusters of X and of Y. Then, we let $n_{i,j} \in N = |X_i \cap Y_j|$ where $n_{i,j}$ is the i^{th} row, j^{th} column of N. Although such a matrix would measure the amount of overlap in clusters, it does not adjust for the size of clusters. To do so, we denote vectors describing these sizes. Let $a_i = \sum_{j=1}^{r}(n_{i,j}); \forall i \in [1, r] \cap \mathbb{N}$. Similarly, let $b_j = \sum_{i=1}^{s}(n_{i,j}); \forall j \in [1, s] \cap \mathbb{N}$. Then the adjusted Rand index for X and Y is as follows, from [44, 79, 103]:

$$ARI(X, Y) = \frac{\sum_{i,j}\binom{n_{i,j}}{2} - (\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2})/\binom{n}{2}}{1/2(\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}) - (\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2})/\binom{n}{2}} \tag{3.4}$$

As a result of the mechanics of measuring adjusted Rand index, the highest value of the $ARI(\ldots)$ function is 1, which would mean that the partitions are identical. The lower the value for the adjusted Rand index, the more dissimilar the paritions are. To assess performance on the synthetic data, we computed the adjusted Rand index comparing the clustering produced by a given experimental method to the known ground truth for the synthetic data.

3.2.3 Application to Real Tumor Data

Real tumor data for these experiments were collected from TCGA. Several varieties of DNA CNV data were collected: ovarian tumors (OV), lung squamous small cell cancers (LUSC), and GBM. In addition to the genomic data, clinical data were also extracted for each of the tumor types. The genomic data consisted of aCGH, which provides data in the form of $log2$ ratios of the copy numbers of tumor samples compared to normal. We chose aCGH data in this case, as it was generally perceived as lower noise and less susceptible to contamination from cell-to-cell communication (including immune response) than RNA data [98]. Data were presented in a set of records of CN alterations. So, to transform the data into a matrix compatible with the

medoidshift approach, we pre-processed the data in a method we termed blocking. That is, for a fixed data set, whenever a change in CN for one of the samples occurs, we track the location, and begin a new feature. The result is a fixed state of the data set within a feature, and changing values for at least one sample in the data set for adjacent features. The procedure allowed a compression of the data, given that we tracked the starting and ending positions of each feature. We then termed the blocked matrix $M \in \mathbb{R}^{m \times n}$ with m samples and n blocks. We then used PCA to compress the data into the top 10 PCs of the blocks. PCA also had the effect of increasing the density of the point cloud, requisite for mode-seeking approaches such as medoidshift [89]. As a matter of procedure, blocks of adjacent chromosomes were concatenated as additional features. Sex chromosomes were not considered in these experiments.

Earlier in our description of the methods, we said data were normalized to be in the $[0, 1]$ range in each PC. To perform that normalization, we performed the following computation:

$$\hat{M_{i,j}} = \frac{M_{i,j} - \min\limits_{s} M_{s,j}}{\max\limits_{q} M_{q,i} - \min\limits_{r} M_{r,j}} \tag{3.5}$$

where $\hat{M_{i,j}}$ is the 0-1 normalized version of an entry in M, and $M_{i,j}$ is the i^{th} row, j^{th} column entry of M.

The OV data set consisted of 472 samples, while the LUSC data consisted of 408 samples. The GBM data was not tracked in the resulting publication from this work [81] due to negative results demonstrating the limits of applicability of the method; however, the GBM dataset currently contains 406 samples. After blocking and PCA, each data set contained 10 PC features. In order to assess the usefulness of the medoidshift clustering approach, we conducted survival analysis. We constructed Kaplan-Meier curves and used log-rank as a statistic to test the significance of the correlation between cluster assignment and censored survival for each of the real data sets [59]. The OV data set had 4 subtypes reported to TCGA based on RNA data: differentiated, imunoreactive, mesenchymal, and proliferative [45, 66, 99]. For that data set, we further looked at a chi-squared test of independence for the distribution of survival based on RNA subtype and cluster membership. Additionally, we conducted functional analysis using the DAVID

bioinformatics platform [20]. In order to generate the gene names that DAVID uses as input, we looked for blocks demonstrating 4-fold amplification. We considered base pairs within 10kb of the feature, and used Biomart [91] to produce gene names consistent with the range of base pair values. For a given cluster, all gene names were pooled and entered into DAVID [20], using the default parameters. The annotations belonging to the top functional cluster were recorded.

3.3 Results

3.3.1 Synthetic Data

After applying the kernel-free, negative exponential kernel, and 2-stage approaches outlined in the methods sections to the synthetic data generated as part of the experiment, the adjusted Rand index was computed in comparison to a known ground truth partition. The results of the computation are displayed in Figure 3.2:

Based on the results in Figure 3.2, we found that in many cases, the performance of the kernel-free approach was infereior in the average and worst case than the kernel-based approaches. Additionally, although the ISOMAP kernel function performed better than the 2-stage approach in instances where the noise level was low or the data were noiseless, in cases where realistic noise levels [19] were added, the average-case performance of the 2-stage approach had outshined the ISOMAP-like kernel only.

We also tracked the real-time running time of the approaches. The mean runtimes were 1.2854s, 1.2914s, and 0.0121s for the kernel-free, ISOMAP kernel, and 2-stage approaches respectively. The reduced runtime we attributed to the need to compute fewer paths, as the ISOMAP-like kernel in the 2-stage approach only was used on a compressed representation of the data, rather than the full data set.

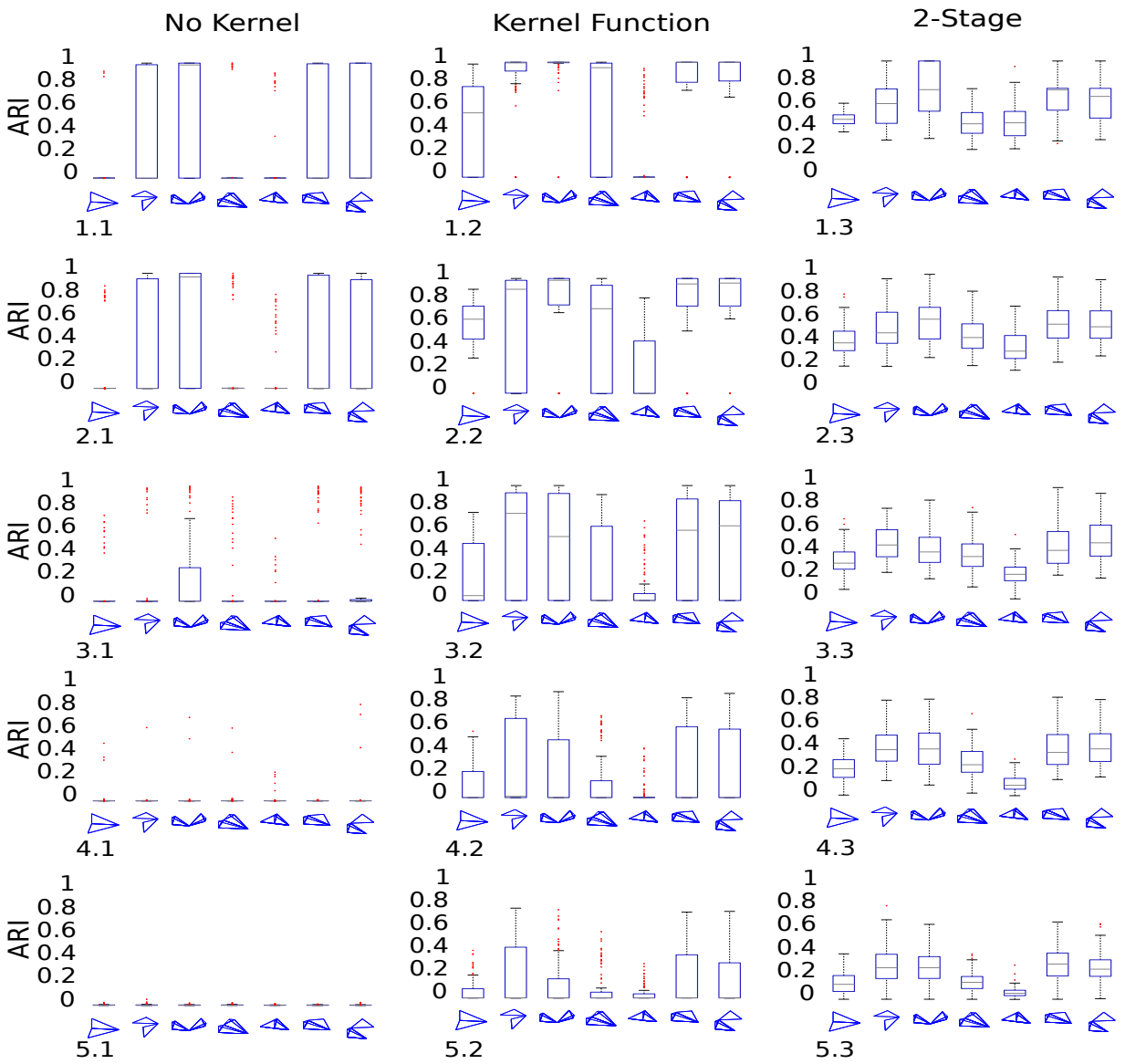

Figure 3.2: Distributions of ARI, which functions as a goodness of fit, for synthetic data experiments. For each subfigure, row designates standard deviation of noise, and column designates the approach employed. For example, 4.2 is the fourth noise level (0.15 standard deviation), second approach (ISOMAP-like kernel). Within a subfigure, each column represents the geometry labeled at the bottom of the subfigure. This figure was originally published as Figure 2 in [81].

3.3.2 Real Tumor Data

For each of the TCGA data sets used in the real tumor experimentation [45], we performed the 2-stage medoidshift clustering. The OV data set used provided an expected simplicial complex structure, with three lines radiating outward from a central point. Similarly, the LUSC data also demonstrated a simplicial complex structure, with what appeared to be two planar objects conjoined at an edge. The OV data were clustered into three clusters, roughly corresponding to the three arms radiating from the central point. The LUSC data clustered into two clusters; however, these clusters visually did not tightly correspond to the perceived planar structures. A visualization of the data in PC space, as well as the clusters found by running the medoidshift clustering approach is provided in Figure 3.3.

In order to quantitate the results observed in Figure 3.3, we performed Kaplan-Meier survival analysis [59] to test for differences in the distribution of censored survival time obtained from the clinical data sets associated with the genomic data sets analyzed. For the OV data, a comparison of cluster 1 to cluster 2 log rank test statistic had a chi-squared value of 12.557. Cluster 2 to cluster 3 yielded a test statistic of 50.9565, and cluster 1 to cluster 3 yielded a chi-squared test statistic of 180.9646. All these values were significant at the $p < 0.01$ level. As a result, the clustering in the OV case demonstrates statistically significant separability with respect to survival times.

The Kaplan-Meier analysis for the LUSC data set showed weakly significant differences in the survival times ($p = 0.0.0339$) based on a test statistic value of 3.6413. However, the differences may be related to the relatively small size of the "fringe" cluster, or the grouping of those data points as very far from dense regions of the data cloud.

Also in the OV case, we sought a comparison to existing proposed subtypes based on RNA data. The subtypes by the RNA data were published by Verhaak et al. [102]. The four subtypes — differentiated, immunoreactive, mesenchymal, and proliferative — identified by [102] were computed based on a non-negative matrix factorization of RNA expression data, which assigns

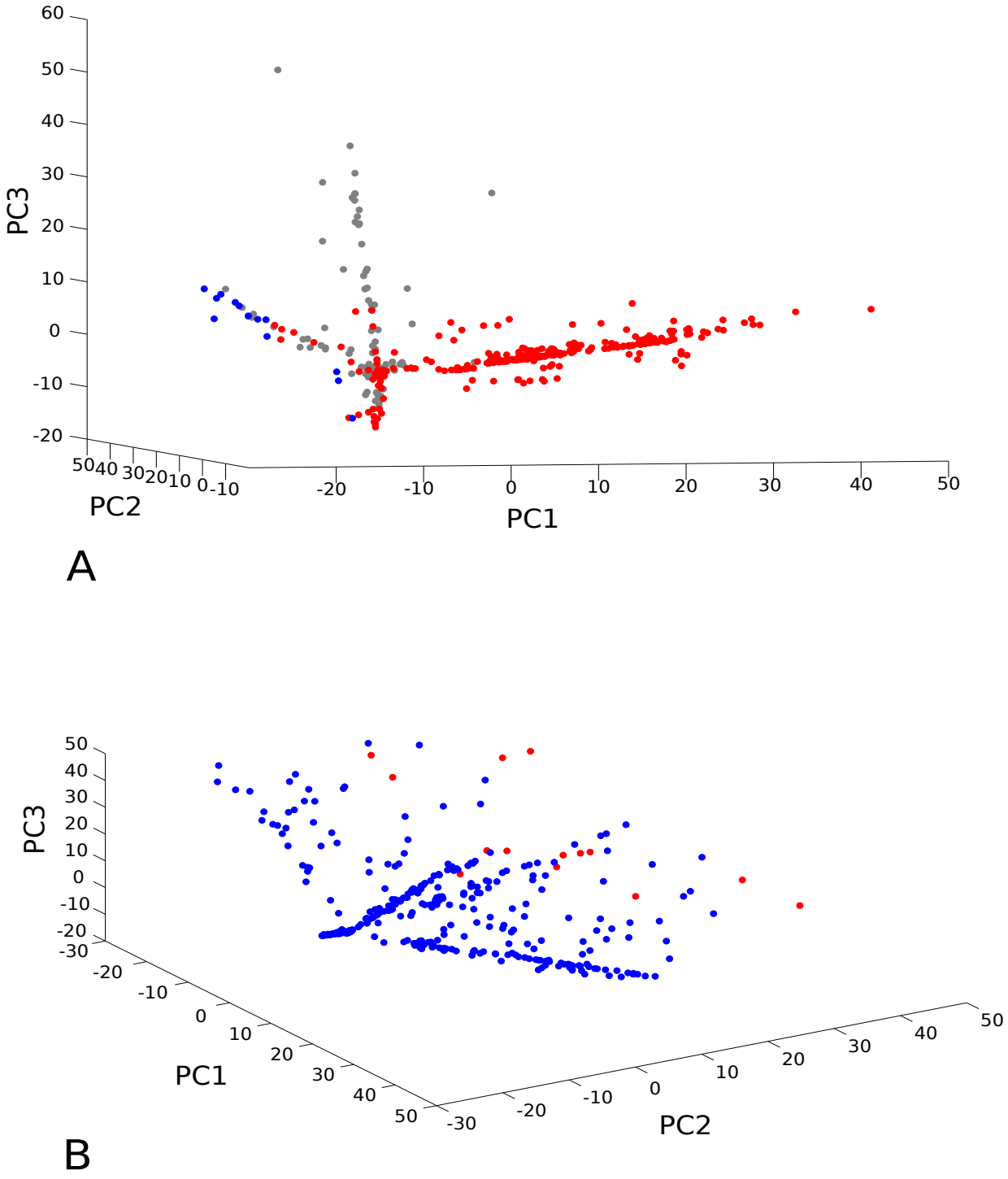

Figure 3.3: Visualization of clustering of TCGA CN data by color in PC space for OV data (panel A), and LUSC data (panel B). This figure was originally published as Figure 3 in [81].

non-negative weights to profiles in a latent space, in a similar fashion to the contribution outlined in Chapter 2 [80]. In order to assess how the OV results compared with the results of [102], we set up a contingency table to run a chi-squared test of independence. The result was a chi-squared value of 15.48091566, which suggests weak significance ($p = 0.0168$). We found one of our clusters to associate with mesenchymal, one to associate with proliferative, and the third cluster to have no distinguishable association.

We also performed ontological analysis via the DAVID tool [20]. Table 3.1 shows top functional enrichment terms for both the OV and LUSC clusters.

Cancer type	Cluster number	Terms
OV	1	Keratinization, small proline-rich, epidermal cell differentiation, epithelial cell differentiation
OV	2	Antigen processing, MHC class II, asthma allograft rejection, type I diabetes mellitus, cell adhesion
OV	3	Keratin, coil 1a/b/2/12 intermediate filament, cytoskeleton, non-membrane-bound organelle
LUSC	1	Zinc finger, KRAB, C2H2 transcriptional regulation, DNA-binding, metal binding
LUSC	2	Keratin, peripherin, intermediate filament family orphan 1

Table 3.1: Summary of top DAVID terms for both OV and LUSC data sets used in medoidshift experiments. **NB:** highly similar terms have been merged. This table was originally published as Table 1 in [81].

Based on the ontological results, we sought links in the literature as to the identified terms and pathways and known mechanisms of disease. In the OV set, cluster 1 shows terms associated with differentiation of epithelial tissue, a known pathway pertubed in tumors [32, 39]. Other clusters in OV may have varying degrees of ties to the literature, but it is difficult to quantify how meaningful those ties may be. In the LUSC case, we found terms relating to the zinc and keratin

families, which may have a role in tumor development [10, 28]. However, it remained difficult to quantify these results with respect to how significant the findings may be.

We achieved a negative result on the GBM data set. The clustering approach found only one cluster, so differentiable survival analysis was not possible. We looked for reasons as to why this observation may have occurred, and noticed that there were no long-term (multi-year) survivors for the GBM data set used, indicating that perhaps only one cluster — a poorly surviving cluster — could be supported for that data set. We also observed poor differentiability with respect to geometry. That is, many of the data points appeared to have been sampled from some high-dimensional multivariate Gaussian, indicating further that we would not expect to see multiple clusters given our approach.

3.4 Conclusions

The contribution outlined here demonstrates improvment in the state of clustering approaches applied to bulk tumor genomic data. The 2-stage approach may hold broader implications for mitigating the effects of noise in other domains where mode-seeking clustering is applicable, in addition to the contributions outlined here.

However, there is room to improve further given the suboptimal results in the LUSC cases, and especially in the GBM cases. As a result of these stumbling blocks, additional normalization or pre-processing may assist. Gross inspection of the OV data set illustrates a more dense point cloud locally than the LUSC case, which indicates we may have been hitting the limit of applicability of such a clustering approach. Similarly, in the GBM case, the improvement demonstrated here is limited when data do not cluster well into differentiable groups with respect to clinical variables. As single-cell data continues to flourish and become more cost feasible, we might expect further modifications to the approach to capture the unmixed nature of single-cell data that could lead to additional clarity in the clustering procedure. Further, a combination of RNA and DNA data coupled with integrating this clustering approach into the unmixing framework out-

lined in Chapter 2 may yield significant new insights into tumor deconvolution, and specifically into the geometric approach.

Chapter 4

Toward Automated Deconvolution of Bulk Tumors with Simplicial Complexes[1]

4.1 Introduction

Although work prior to this thesis and contributions from previous chapters have advanced the understanding the interplay of intratumor heterogeneity, intertumor heterogeneity, tumor evolution, and challenges stemming from biological and technical limits, opportunities for improvement remain. In short summary form, tumors exhibit properties similar to species evolution [26, 67]. In addition to evolutionary properties, the tumors exhibit both intra- and intertumor heterogeneity that we model here, and has been modeled in numerous methods previously [3, 15, 16, 21, 27, 29, 34, 42, 49, 54, 55, 57, 58, 62, 68, 69, 74, 75, 80, 81, 82, 83, 86, 98, 109]. In the work outlined in Chapter 2, several parameters to the model were challenging for users to estimate. The model required input of both the number of clusters in the data, as well as the dimensionality of each of the clusters. The work outlined in Chapter 3 provides evidence

[1]Work from this chapter corresponds to work originally presented in Roman et al. (2017) Automated Deconvolution of Structured Mixtures from Heterogeneous Tumor Genomic Data, *Submitted*. Access to a pre-print of a previous version of the work is available at `https://arxiv.org/pdf/1604.02487.pdf`

that non-parametric clustering could be applied to the pipeline of simplicial complex unmixing [81]. Integrating the approach into the simplicial complex framework would eliminate the need for the detection of the number of clusters. In addition to integrating the medoidshift clustering into the simplicial complex unmixing framework, other key opportunities for improvement to the simplicial complex framework remained.

In [80] it was observed that allowing for mixed membership in putative evolutionary trajectories may better allow for uncertainty in the data. Furthermore, another parameter controlled the dimensionality of each of the clusters in [80]. In this chapter, we outline an approach to estimate dimensionality in an automated fashion from the data, and integrate that module into the simplicial complex framework. Lastly, we address the opportunity for heterogeneous data types in [81] by considering data sets of both DNA and RNA simulataneously. Although the DNA data has relatively less noise than RNA data [19], combinations may prove fruitful in instances where additional replication is gained, or where only RNA data are available for some samples.

In brief, the steps to the simplicial complex unmixing framework were as follows:

1. Pre-processing / filtering
2. Dimensionality Reduction via PCA
3. Pre-clustering (partitioning) via k-medoids clustering into submixtures
4. Unmixing substructures
5. Reconciling substructures into a simplicial complex

In this chapter, we outline several key improvement to the simplicial complex framework — we integrate the medoidshift clustering from Chapter 2 into the simplicial complex framework we develop automated dimensionality estimation modules, and integrate those modules into the framework; and, we address the challenge of simplicial complex deconvolution on data sets of combinations of DNA and RNA data.

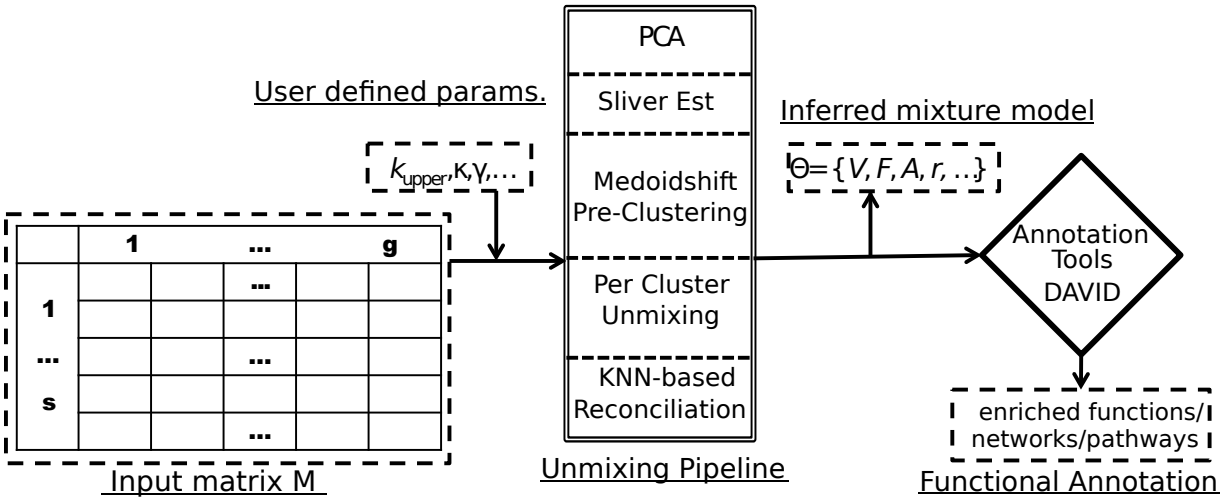

Figure 4.1: Overview of our analysis pipeline. Input are represented as call records. These CNV records are converted into a matrix. Following conversion, data are passed to the inference code. The inference code includes compression via PCA, dimension estimation, with pre-clustering based on a KNN computation. The subsimplices are merged using a maximum likelihood model. The inferences are then analyzed using downstream analysis tools to identify pathways and term association.

4.2 Methods and Data

In the following section, we outline the improvements made to the simplicial complex unmixing pipeline, as well as providing application to tumor data from TCGA, and comparison with state-of-the-art methods. The figure below provides an overview of how we segment the inference problem with respect to the bulk tumor data in a DNA case.

We modeled the data as a matrix $M \in \mathbb{R}^{s \times g}$, where there are s samples and g genes in the data set. Although strictly speaking the features may be smaller than genes — such as single nucleotide variants or larger such as chromosome arms — we use the term genes for the sake of interpretability of the method. The samples could be multiple samples from a single patient, or samples spread across a patient population. We preferred the matrix format, as data from multiple sources, including both RNA and DNA, could be transformed into this generic format.

Pointedly, the data could stem from various DNA platforms as well, including both aCGH or DNASeq read depth. Although the approach could be applied to a variety of genomic data, CNV data would be most closely expected to exhibit the geometric constraints of our model [80, 98]. The method we employ to satisfy the constraints of the model while capturing notions of tumor phylogenetics and heterogeneity can be generally modeled as

$$M = FV + \epsilon \tag{4.1}$$

where M is as before, $F \in \mathbb{R}^{s \times k}$ represents fractional composition of data points, $V \in \mathbb{R}^{k \times g}$ are the unmixed, or "pure" subtypes of the tumor, and $\epsilon \in \mathbb{R}^{s \times g}$ is an error matrix. We chose V as the designation for the unmixed subtypes, given that from the geometric perspective, the unmixed subtypes are vertices of a simpicial complex. Because in our case the F are mixture components, we restricted them to have the properties of a convex combination:

$$\begin{gathered} \sum_i F_{i,j} = 1 \; \forall i, \text{for fixed j} \\ \forall i, j, 0 \leq F_{i,j} \leq 1. \end{gathered} \tag{4.2}$$

Another way to state our goal is to compute F and V given M, with a needed intermediate step of determining k.

4.2.1 Pre-processing

We first compute the Z-scores of our matrix:

$$M_z = \frac{M - \mu_M}{\sigma_M} \tag{4.3}$$

where M_z is the matrix of Z-scores, μ is the vector of means for each feature and σ is the vector of standard deviations for each feature. **NB**: In the case of using both DNA and RNA data together, because we expect different distributions of copy number verus read count for DNA and RNA, the features are normalized independenly, then pooled together. That is, we compute

the Z-scores for the DNA and Z-scores for the RNA separately, where all features are pooled into one distribution, then concatenate the results together into a combined Z-score matrix.

Next, we compress the data using PCA [73], as in prior chapters [80, 81], and work prior to this thesis [86, 98]. While different approaches may be used for dimensionality reduction, PCA presents a straightforward and fast-running method to compress the data. We chose a maximum of 12 components, but to generalize the approach, we deemed this parameter $k_{upper} = 12$. We implemented the PCA using Matlab's `pca` function, with economy mode on. For further sections, we call the representation of the data in PC space $S_M \in \mathbb{R}^{s \times k_{upper}}$.

In order to further whittle the dimensionality of the data in a more automated fashion, we applied the sliver-based dimensionality estimation of [14] to the data. The core of the model suggested by Cheng and Chiu [14] was that dimension of a point cloud should be increased, until the surrounding body forms a sliver. A sliver was defined in mathematical terms as follows:

$$\begin{aligned} \nu &< \delta^j r \\ r &= \frac{L^j}{j!} \end{aligned} \tag{4.4}$$

In the sliver equations, ν is the volume of an enclosing body to the point cloud, j is the current estimate of the instrinsic dimensionality, increasing by 1 each time until the top line is false (up to a limit of k_{upper}), δ is a noise parameter between 0 and 1; and L is the length of the longest edge. The guesses for enclosing structures were made using the approach outlined in [86] for speed purposes.

After the dimensionality estimation is completed, and we take the top $\min(j, k_{upper})$ PCs, we compute a [0,1] normalization of the data to conform with the parameter assumptions made in [81]. We compute that normalization as

$$S_{[0,1]} = \frac{S_M - \min S_M}{\max S_M - \min S_M} \tag{4.5}$$

Where the minima and maxima are computed over all samples on a per-PC basis.

4.2.2 Pre-Clustering

We next applied the pre-clustering approach outlined in [81] and discussed in detail in Chapter 3. The goal of the method was to identify the low-dimensional substructures that compose the point cloud, with the caveat that these substructures may be partially overlapping in dimensionality. As a result of having distinct dimensionality from one another, they were expected to be lower dimensionality than j determined in the prior step. For example, two triangles overlapping at a point but having different subspaces might appear 3-dimensional on the aggregate, but each substructure is 2-dimensional. Because of the likely distinct subspaces the clusters may have occupied, we used an ISOMAP-like distance [97] on the later stages of the clustering, with Euclidean-based distance for noise suppression in the first round [81]. Further, we use a negative exponential kernel on the ISOMAP-like distance in order that very distinct data points are chosen as representatives, rather than points more similar to points from other clusters, as the data are distributed differently based on planes and hyper-planes, rather than a Gaussian distribution. Chapter 3 contains full details on the deployment of the pre-clustering approach. The result of the pre-clustering was a set of cluster representatives and clusters:

$$M_{2-stage} = \cup_i M_{N(x_i)} \tag{4.6}$$

where $N(x_i)$ is the set of points within the neighborhood of a data point $x_i \in S_{[0,1]}$, and $M_{N(x_i)}$ represents the points in the cluster corresponding to that neighborhood. Over the set of all of these neighborhoods, we define a partition labeled $C = C_1, \ldots, C_r$, and $S_{[0,1]} = \cup_{C_i \in C} C_i$.

Based on the partition C, we can then define an uncertainty in the clustering. We used a weight function of a multivariate Gaussian with 0 mean vector and orthogonal covariance where the entries were the distance from each cluster center to the mean of all cluster centers, using the fact that cluster centers are extrema. We called this the raw weight, and sought to normalize it on a 0-1 scale. For a raw weight for a point R_i, the normalized weight was defined

$$W_i = \frac{R_i - \min\limits_{C_j \in C} R_i}{\max\limits_{C_j \in C} R_i - \min\limits_{C_j \in C} R_i} \tag{4.7}$$

where W_i is the new normalized weight term. Another interpretation of the equation is that we normalize weight to be a percentage of the total density granted to a point from all clusters for a particular cluster.

4.2.3 Dimensionality Estimation

Although a pre-processing step of sliver-based dimensionality estimation was applied [14], each cluster may have a unique intrinsic dimension. However, given that the number of data points in a cluster may be substantially less than the number of data points total, the approach offerred in [14] was not appropriate. As a result, we performed a PCA-based instrinsic dimensionality detection. The approach works as follows: first, we built a noise model by sampling the percentage of variance explained by principal components of Gaussian noise with identity covariance and 0 mean. From those experiments, we derived a sample mean $\mu_{G(i)}$ and standard deviation $\sigma_{G(i)}$ for each increasing dimension $i \in \mathbb{N} \cap [1, \ldots, \min(j, k_{upper})]$. Next, we take the PCs of the data, and find the smallest dimension i such that the variance explained by the i^{th} PC of the data set is less than $\mu_{G(i)} + \kappa\sigma_{G(i)}$, where $\kappa \in \mathbb{R}$ scales the significance threshold. In our experiments, we chose $\kappa = 3$ to correspond with $p < 0.01$ as the uncertainty in rejecting the hypothesis that the variation were due to noise.

As a result of this procedure, we were able to consstruct a vector of inferred dimensions for each cluster $D \in (1, \ldots, \max(j, k_{upper})^r$, where there are r clusters total as above. The test is designed to be conservative in its approach, but improves as the quality of data improves.

4.2.4 Per-Cluster Unmixing

In order to provide an estimate of the subpopulations within each cluster, we pose the problem of unmixing each cluster as performing optimization of a function describing the probability of observing data we see given the model choices we make. Our key choices are that the prior distribution of the enclosing structure should be exponential on the size of the minimum spanning

tree, which corresponds to the amount of evolutionary times that passes, and we encorporate a Bayesian Information Criterion (BIC) penalty to reduce the number of subpopulations. The conditional model we choose is exponential in terms of genomic distance that is not explained by the model.

In other words, we suppose, as in [80, 98] that we pose the statement of probability as follows:

$$\begin{aligned} P(\Theta|X) &\propto P(X|\Theta)P(\Theta) \\ &\propto \prod_{i=1}^{r} \left(exp(-\sum_{j=1}^{s}(|x_i - F_j^i V_j^i|W_j^i) \right) mst(V_j, A_j)^{-\gamma}\beta) \end{aligned} \tag{4.8}$$

where

- γ is a regularization parameter based on the Signal-to-Noise Ratio (SNR) [80]
- V are inferred vertices
- A is the adjacency matrix of the simplicial complex
- $mst(\ldots)$ is the minimum spanning tree function
- W is the relative weight computed earlier
- F are the inferred mixture components
- x_i is the i^{th} data point
- β is a BIC-like penalty [87]
- $|\ldots|$ is L_1 distance

We implemented the optimization code using minimization of the negative log likelihood in order to both prevent underflow and for ease of implementation. We can re-pose the optimization problem in negative log likelihood as follows:

$$\begin{aligned} -log(P(\Theta|X)) &\propto -log(P(X|\Theta) - log(P(\Theta)) \\ &\propto \sum_{j=1}^{s}(|x_i - F_j^i V_j^i|W_j^i) + \gamma \log(mst(V_j, A_j) + \beta) \end{aligned} \tag{4.9}$$

We implemented a solution using the `fmincon` function in Matlab, with initial guesses for the enclosing strucutres as in [86], and a generalized expectation maximization-like procedure,

where V and F were alternately held constant. To compute F, the `lsqlin` function in Matlab was used.

4.2.5 Reconciliation into a Simplicial Complex

After various different sets of vertices are found for each substructure, we needed a way to reconcile the bodies into a simplicial complex. We followed a two-stage approach. First, we considered a hypergeometric distribution on the nearest neighbors of each of the vertices. In practice we found little sensitivity to the number of nearest neighbors chosen, and chose 15 nearest neighbors for our experiment. Based on the background distribution, we would expect $\frac{|N_1||N_2|}{N}$ nearest neighbors in common where there are N data points total, and N_1, N_2 represent the sets of nearest neighbor points. For our procedure, if the number of observed nearest neighbors in common were greater than the amount expected at random, the vertices were joined.

In those instances where the nearest neighbor approach did not yield one connected component — which is cruicial for the biological interpretability of the model — we looked at each candidate pair of vertices, then merged vertices such that the objective function outlined in Equation 4.9 was kept as low as possible — that is, that the likelihood was maximized — until one connected component was observed. In the case that vertices were merged, whether from the nearest neighbor or maximum likelihood phases, the new vertex was set as the mean of the two old vertices. In Figure 4.2, we outline the algorithm in pseudocode.

4.3 Results

We applied the improvements outlined above to several data sets from TCGA [45]. These data sets included BRCA DNA CNV data, BRCA RNASeq data, and the clinical data corresponding to the genomic and transcriptomic samples. The copy number data included is level 4 data, which includes mapping copy numbers to genes. The RNASeq data used was level 3 data, which also

1: $initialization$

2: **while** $notSingleConnectedComponent(A_{cand})$ **do**

3: $merge_{candIdx} \leftarrow \binom{size(V_{cand})}{2}$ ▷ Generate a list of all possible pairs.

4: **for** $i = 1 : length(merge_{candIdx})$ **do**

5: $A_{candList}(i) \leftarrow merge(A_{merge_{candIdx}(i)}, A)$ ▷ Merge the candidates.

6: $V_{candList}(i) \leftarrow merge(V_{merge_{candIdx}(i)}, V)$

7: $likelihood(i) \leftarrow computeLikelihood(A_{candList}(i), V_{candList}(i))$

8: **end for**

9: $singleConnCompIdx \leftarrow find(numConnComp(A_{candList}) == 1)$

10: **if** !isEmpty(singleConnCompIdx) **then**

11: $A_{candList} \leftarrow A_{candList}(singleConnCompIdx)$

12: $likelihood(i) \leftarrow likelihood(singleConnCompIdx)$

13: $V_{candList} \leftarrow V_{candList}(singleConnCompIdx)$

14: **end if**

15: $A_{cand} \leftarrow A_{candList}(\arg\max likelihood)$

16: $V_{cand} \leftarrow V_{candList}(\arg\max likelihood)$ ▷ Choose the most likely

17: ▷ Repeat if not 1 connected component

18: **end while**

19: $return : V_{cand}, A_{cand}$

Figure 4.2: Pseudocode for merging protocol to select most likely from a set of candidate models provided none are simplicial complexes.

included gene lists for the normalized read counts provided. As a result of the gene lists being provided, we did not need to include a blocking procedure like the one that was used in [81].

We applied the pipeline outlined in the methods section to the TCGA data using the following parameters: $k_{upper} = 12$; 1000 replicates for the simulated Gaussians, neighborhood size 1; 15 nearest neighbors for vertex merger; $\kappa = 3$; the maximum number of iterations for the `fmincon`

optimization: 1000. The choices reflect realistic computational limits coupled with desirable values for uncertainty.

We applied the method in three separate contexts: DNA data only, RNA data only, and DNA and RNA data coupled together. The BRCA data set has the desirable property for this application in that clinically relevant subtypes are defined: HER2+, HER2-;ER/PR+, Triple Negative Breast Cancer (TNBC). This subtyping is useful in the validation phase of the results.

4.3.1 RNA Data

Although numerous simplicial and simplicial complex models of unmixing the tumor data were considered by our algorithm, ultimately, the tetrahedron visualized in Figure 4.3 was judged the most likely based on our model.

4.3.2 DNA Data

The DNA data from TCGA holds the benefit of a reduced anticipated noise profile relative to RNA, potntially making it less suspect in terms of contamination from infiltrating stromal and immune cells [19, 45]. The decreased noise resulted in a sharply defined 3-armed simplicial complex that was recapitulated visually based on our approach, visualized in Figure 4.4.

4.3.3 RNA and DNA Combined

We also constructed a representation of both RNA and DNA data together. We normalized RNA and DNA data separately, as outlined in the methods section. Figure 4.5 is a plot of the results of the experiment, using the same color coding system as earlier plots.

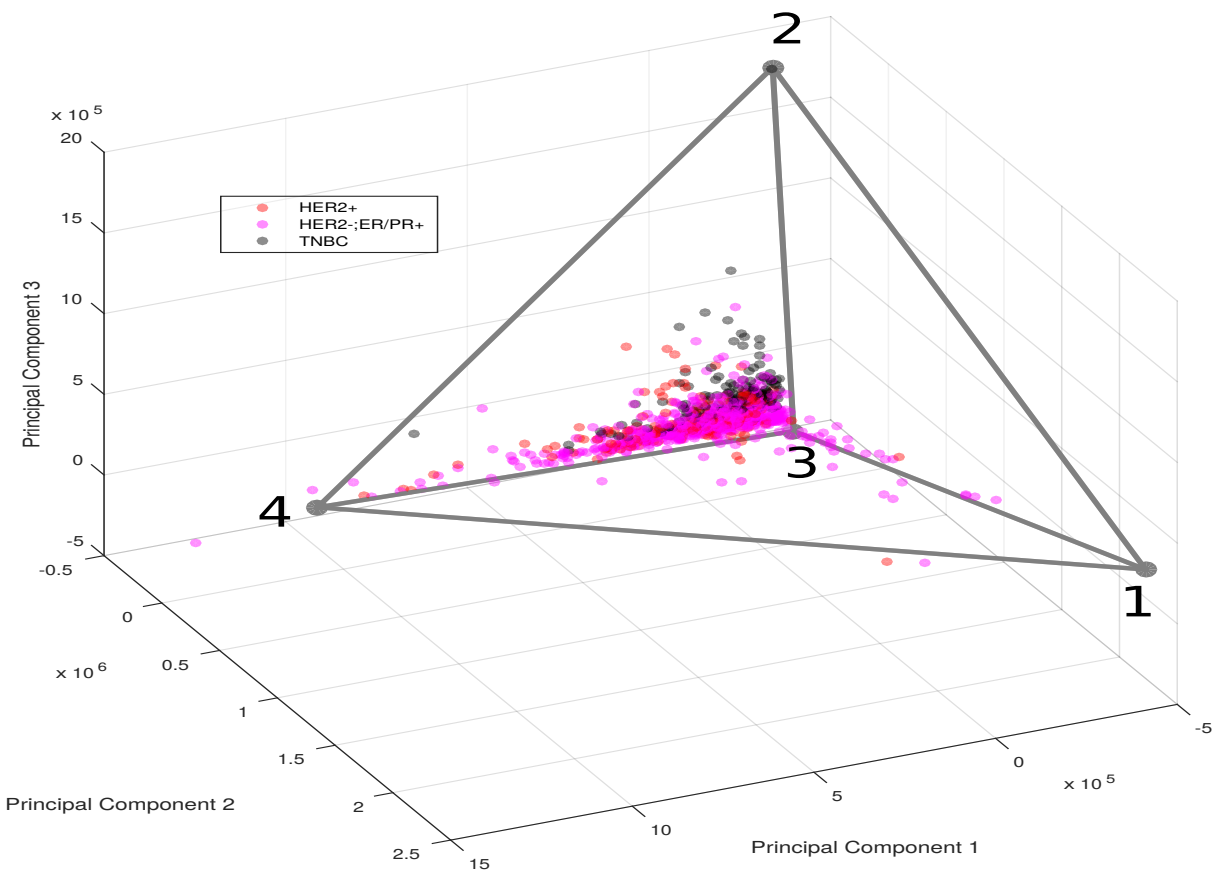

Figure 4.3: Visualization of the unmixing of the RNA data. Data are colored based on clinical subtype in PC space: HER2+ is red, purple for ER/PR+, and black for TNBC.

4.3.4 Ontological Analysis

To validate the functional and biological significance of the findings, we also projected the inferred vertices from PC space back into the 'omics space, and looked for statistically significantly perturbed genes. We identified those genes that had Bonferonni-corrected Z-scores at significant levels. For RNA and the combined RNA and DNA levels, we used $p = 0.01$ as the significance

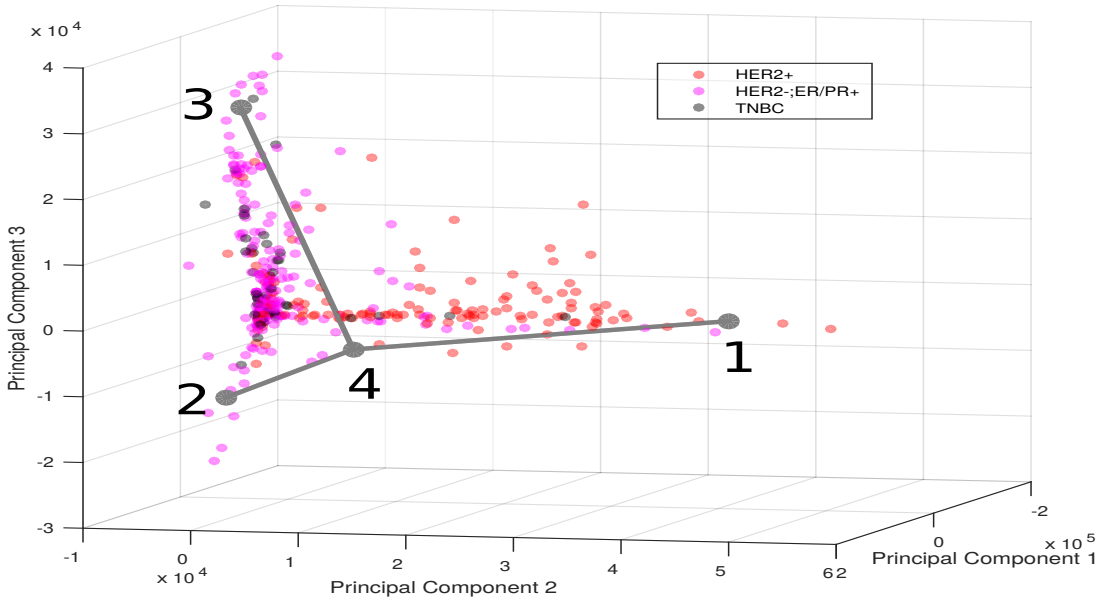

Figure 4.4: Visualization of unmixing of DNA TCGA data. Data are visualized based on clinical subtype: red for HER2+, purple for ER/PR+, and black for TNBC.

threshold; however, in the DNA case, so many genes were marked as significant at that level that the ontology analysis tool would not run, so we used a more stringent level of $p = 2.1905e - 11$. At levels smaller than that p-value, the inverse density function used to produce the resulting Z-score in Matlab underflowed. The significant gene lists were then run through DAVID [20], a meta-ontology tool. We evaluated the upregulated genes in each of the cases. As we expected,

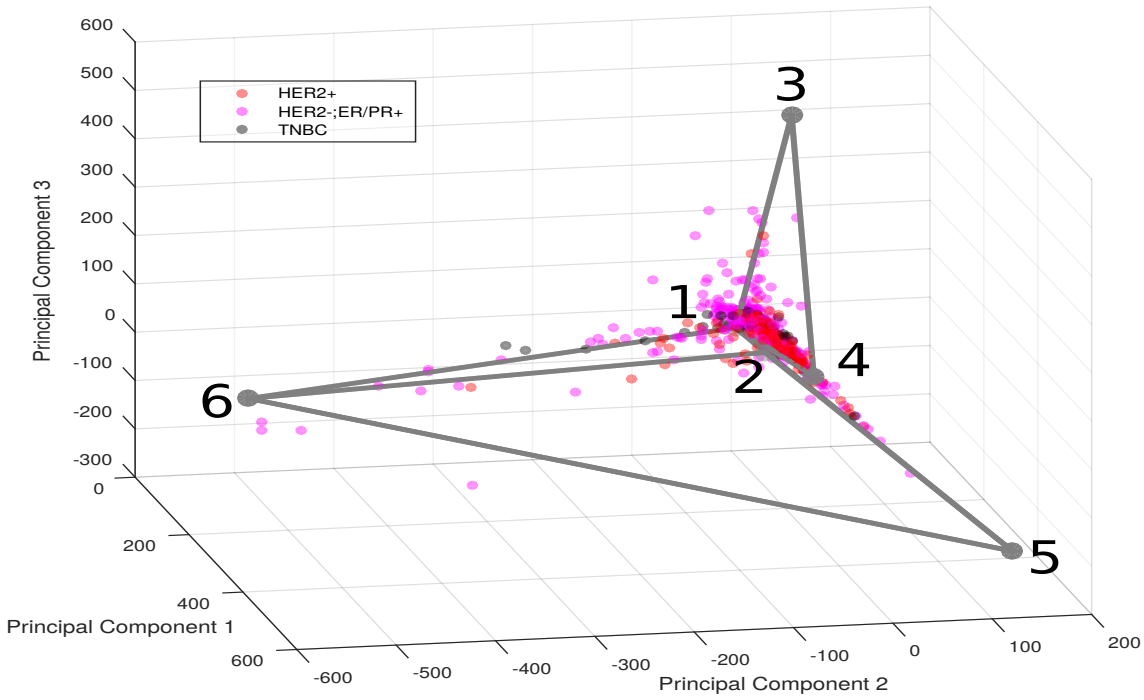

Figure 4.5: Visualized unmixing of DNA and RNA data combined, using the colors based on clinical labels as in Figures 4.4 and 4.3

when we investigated the tissues and pathways that had significant enrichment at a Benjamini corrected weak significance or lower ($p \leq 0.05$), we encountered breast-specific terms, as well as cancer-specific terms. The full list of results are presented in Tables 4.1, 4.2, and 4.3

Source	Term	Benjamini Score	Vertex
GAD DISEASE CLASS	Immune	3E-010	3
GAD DISEASE CLASS	Infection	0.00052	3
BIOCARTA	Antigen Processing and Presentation	0.019	3
BIOCARTA	CTL mediated immune response against target cells	0.0057	3
KEGG PATHWAY	Allograft Rejection	4.7E-009	3
KEGG PATHWAY	Cell adhesion molecules	2.4E-009	3
KEGG PATHWAY	Graft-versus-host disease	3.8E-009	3
KEGG PATHWAY	Antigen Processing and Presentation	3.3E-009	3
KEGG PATHWAY	Type I diabetes mellitus	5.1E-009	3
KEGG PATHWAY	Viral myocarditis	4.7E-009	3
KEGG PATHWAY	Autoimmune thyroid disease	2.8E-008	3
KEGG PATHWAY	Intestinal immune netowkr for IgA production	2.9E-007	3
KEGG PATHWAY	Asthma	2.7E-005	3
KEGG PATHWAY	Systemic lupus erythematosus	0.00011	3
KEGG PATHWAY	cytokine-cytokine reception interaction	0.00084	3
KEGG PATHWAY	natural killer cell mediated cytotoxicity	0.00092	3
KEGG PATHWAY	hematopoietic cell lineage	0.0093	3
KEGG PATHWAY	chemokine signaling pathway	0.0094	3
Panther Pathway	T cell activation	0.00093	3
Reactome pathway	Signaling in Immune System	2.8E-014	3
UP TISSUE	Spleen	1.9E-017	3
UP TISSUE	Blood	1.4E-012	3
UP TISSUE	B-Cell	1.1E-011	3

UP TISSUE	Lymph	3.1E-011	3
UP TISSUE	Lymph Node	1.1E-005	3
UP TISSUE	Peripheral Blood	0.0001	3
UP TISSUE	T-Cell	0.00025	3
UP TISSUE	Leukocyte	0.00047	3

Table 4.1: RNA Data

Source	**Term**	**Benjamini Score**	**Vertex**
CGAP SAGE	stomach Adenocarcinoma 3rd	1.6E-3	1
CGAP SAGE	cartilage Dedifferentiated chondrosarcoma lung metastatis	5E-3	1
KEGG PATHWAY	Systemic lupus erythematosus	9.1E-012	1
KEGG PATHWAY	Alcoholism	1.4E-010	1
UP TISSUE	Blood	1.7E-005	1
UP TISSUE	Spinal cord	1.4E-005	1
UP TISSUE	Pancreas	0.0003	1
UP TISSUE	Ovary	0.0047	1
UP TISSUE	Mammary gland	0.004	1
UNIGENE	Breast (mammary gland) cancer disease 3rd	0.00057	1
UNIGENE	Blood normal 3rd	0.00034	1
UNIGENE	Ear normal 3rd	0.00043	1
UNIGENE	mammary gland normal 3rd	0.00034	1

UNIGENE	prostate normal 3rd	0.00043	1
UNIGENE	placenta normal 3rd	0.0013	1
UNIGENE	oral tumor disease	0.002	1
UNIGENE	spleen normal 3rd	0.0027	1
UNIGENE	respiratory tract tumor disease	0.0028	1
UNIGENE	bone normal 3rd	0.0032	1
UNIGENE	thymus normal 3rd	0.003	1
UNIGENE	uterus normal 3rd	0.0031	1
UNIGENE	colon normal	0.0043	1
UNIGENE	chondrosarcoma disease	0.0049	1
UNIGENE	adrenal gland normal	0.0056	1
UNIGENE	muscle tissue tumor	0.0073	1
UNIGENE	uterine tumor disese	0.005	1
UNIGENE	ovarian tumor 3rd	0.00046	2
UNIGENE	ovary normal	0.00046	2
CGAP Sage	mammary gland breast carcinoma	0.0048	2
CGAP SAGE	colon adenocarcinoma	0.0058	2
CGAP SAGE	mammary gland carcinoma metastasis to lung	0.0072	2
CGAP SAGE	liver poorly differentiated adenocarcinoma	0.0076	2
UNIGENE	ear normal 3rd	7.8E-005	3
UNIGENE	ovarian tumor 3rd	0.0012	3
UNIGENE	bone marrow normal	0.001	3
UNIGENE	tongue normal 3rd	0.0038	3
CGAP SAGE	mammary gland breast carcinoma	4.4E-016	3
CGAP SAGE	liver poorly differentiated adenocarcinoma	8.8E-006	3

CGAP SAGE	mammary gland carcinoma	0.0028	3
CGAP SAGE	prostate carcinoma	0.00037	3
CGAP SAGE	stomach, poorly differentiated carcinoma by surgery	0.0022	3
CGAP SAGE	thyroid follicular adenoma	0.0027	3
CGAP SAGE	liver	0.0032	3

Table 4.2: DNA Data Table

Source	**Term**	**Benjamini Score**	**Vertex**
UNIGENE	Breast (mammary gland) cancer_disease_3rd	0.0083	2
GAD DISEASE CLASS	Cancer	0.0022	4
GAD DISEASE	breast cancer	5.1E-006	4
GAD DISEASE	Breast Cancer	1.1E-005	4
GAD DISEASE	Asthma — Autoimmune disease	0.0013	4
KEGG PATHWAY	Melanoma	0.036	4
UNIGENE EST	Breast (mammary gland) cancer_disease_3rd	0.02	4

Table 4.3: Table of combined DNA and RNA terminology

Although we expected the DNA and RNA data combined to yield the most clear simplicial complex and term structure, the specificity of terms in that data set relative to the others coupled with a few number of vertices demonstrating significant terms may reflect a tradeoff in sensitivity and specificity based on the data types used for our model.

4.3.5 Comparison to Existing Methods

Although direct comparison with competitor methods was challenging due to competing requirements on data type, platform, data format, and requirements of paired tumor-normal samples, we were able to allow our data set to be run on PyClone [83], a widely-cited approach, by making some assumptions. We also were able to compare our results with those of [70]. The methods from [70] rely on methylation data from TCGA, yet another type of data, while PyClone ran on the DNA portion of the data used in our experiment. In order to fulfill the data format constraints of PyClone [83], we assumed read length of 300 and baseline copy number of two. We note also that PyClone is designed for deeply sequenced (1000X+) targeted data points, while the data in TCGA does not have that level of coverage. Further, although we attempted to run all genes through the PyClone pipeline, it did not complete in approximately one week of runtime on our workstation with an Intel i7-4770K processor, 32GB of RAM, and operating at 3.5 GHz per core. We omit allele frequency data, as it is not publicly available for our data set [45]. We found by reading the documentation that PyClone's designation as running on targeted data implied an upper limit on the order of 30 genes per sample, so we trimmed our data list by using a published list of driver genes only [93].

We present the comparison using Spearman correlation of our mixture estimates with those derived from running PyClone [83] on the reduced feature data set. We found significant ($p < 0.01$) positive correlation between three of our four vertices and three of the four clusters inferred by PyClone [83], demonstrating general agreement between the methods. Table 4.4 shows the results from the comparison.

We also compared the PyClone [83] results and our results in a correlative fashion to the clinical labels supplied by TCGA [45]. The Spearman correlations for those experiments are in Table 4.5. We found at least weak significance $p \leq 0.05$ with our approach to a number of the clinical labels, at a similar rate to the PyClone results.

We also validated our approach using a different approach outlined in [70] based on non-

	Py1	Py2	Py3	Py4
V1	-0.1269 (0.0012)	0.1305 (0.0009)	0.1137 (0.0037)	0.0312 (0.4283)
V2	-0.0914 (0.199)	-0.0994 (0.0113)	-0.0869 (0.0269)	0.209 (<0.00001)
V3	-0.0265 (0.5002)	-0.026 (0.5083)	-0.0237 (0.547)	0.039 (0.3212)
V4	0.204 (<0.0001)	-0.0743 (0.0588)	-0.0383 (0.3298)	-0.1501 (<0.00001)

Table 4.4: Spearman Correlation values (ρ values) among inferred vertices from simplicial complex unmixing by our method and subpopulation clusters derived from PyClone applied to TCGA breast cancer CNV data. The Py prefix is used for PyClone clusters. For our estimates, we use the V prefix. P-values for the comparisons appear in parentheses.

	HER2+	HER2-;ER/PR+	TNBC
Py1	-0.0756 (0.0543)	0.1244 (0.0015)	-0.0207 (0.5986)
Py2	-0.1702 (<0.0001)	0.0349 (0.3754)	0.0566 (0.1498)
Py3	-0.0383 (0.3307)	0.0709 (0.0711)	0.0343 (0.3835)
Py4	0.2451 (<0.0001)	-0.2551 (<0.0001)	0.0032 (0.9345)

Table 4.5: Spearman correlation from PyClone clusters derived from TCGA breast cancer CNV data to clinical labels of samples. P-values for the correlations appear in parentheses

	HER2+	HER2-;ER/PR+	TNBC
V1	0.1864 (<0.0001)	-0.1830 (<0.0001)	-0.0533 (0.1752)
V2	0.0117 (0.7670)	0.0252 (0.5222)	-0.0026 (0.9467)
V3	-0.0782 (0.0467)	0.0215 (0.5853)	0.003 (0.9395)
V4	-0.2027 (<0.0001)	0.1566 (0.0001)	0.0711 (0.0706)

Table 4.6: Spearman correlation of simplicial complex mixture fractions derived from TCGA breast CNV cancer data to clinical labels of samples.

negative matrix factorization of DNA methylation data. The results of Onuchic et al. [70] categorized five cancer subgroups present in samples, as well as stromal, immune, and normal

subgroups. We found significant $p \leq 0.01$ positive correlations between our estimated fulcrum, which corresponded to the most recent ancestor in our data, to the stromal, immune, and normal composition of Onuchic et al. [70]. The full results are displayed in Table 4.7.

	Cancer_1	Cancer_2	Cancer_3	Cancer_4
V1	0.1026 (0.0009)	-0.0151 (0.6261)	-0.0137 (0.6586)	0.0411 (0.1855)
V2	-0.0196 (0.5298)	0.03 (0.3341)	-0.0076 (0.8074)	0.0007 (0.982)
V3	0.047 (0.1306)	0.0133 (0.6697)	-0.0003 (0.9932)	0.0171 (0.5822)
V4	-0.1562 (<0.0001)	0.0018 (0.9545)	0.0118 (0.7041)	-0.0768 (0.0134)
	Cancer_5	Stromal	Immune	Normal
V1	0.091 (0.0034)	-0.0666 (0.0321)	-0.0467 (0.1332)	-0.0399 (0.1989)
V2	0.0207 (0.5057)	-0.0489 (0.1156)	-0.0326 (0.2946)	-0.0199 (0.5224)
V3	0.0171 (0.5822)	-0.0726 (0.0194)	-0.0607 (0.0506)	-0.0433 (0.164)
V4	-0.0855 (0.0059)	0.1197 (0.0001)	0.103 (0.0009)	0.1008 (0.0012)

Table 4.7: Spearman correlation values for simplicial complex unmixing fractional estimates from TCGA breast cancer CNV data to fractional estimates from Onuchic et al [70]. Parentheses enclose p-values.

We also performed correlation analysis on the RNAseq data from our experiments to the clinical labels. PyClone does not perform analysis on RNAseq data, so we cannot provide those results. The data are presented in Table 4.8. We note several statistically significant ($p \leq 0.01$) correlations from our data to the clinical labels in the RNASeq data.

While our method does not provide a perfect head-to-head comparison with PyClone or the Onuchic et al. [70] approach, we nevertheless demonstrate significant correlation among our results, the results of those approaches, and the clinical labels supplied by TCGA [45]. We further note several key advantages our method provides over existing methods:

1. The option to run on expression data, CNV data, protein expression data, or heterogeneous combinations thereof

	HER2+	HER2-;ER/PR+	TNBC
V1	-0.1304 (<0.0001)	0.2235 (<0.0001)	-0.2618 (<0.0001)
V2	0.004 (0.8969)	-0.2397 (<0.0001)	0.3131 (<0.0001)
V3	-0.1029 (0.0009)	0.0033 (0.915)	0.0267 (0.3888)
V4	0.1636 (<0.0001)	0.0979 (0.0016)	-0.1451 (<0.0001)

Table 4.8: Spearman correlation values for simplicial complex unmixing fractional estimates from TCGA breast cancer RNAseq data with TCGA-provided clinical subtypes. Parentheses enclose p-values.

2. The option to run on matched tumor-normal data or unmatched data
3. Amenability of our method for whole-genome or whole-exome sequencing data (i.e., data with many features)

In the case where the constraints of an approach such as PyClone are well-satisfied, that approach may provide advantages of being specfically designed for the data based on the underlying statistical model. However, in the case where those constraints are not met, the contribution presented here may provide a reasonable alternative. As an added note, we attempted to run our method on the trimmed data set used by PyClone in this experiment; however, the fractured nature of the pathways did not allow for the presumed continuous geometry, and therefore conflicted with the geometric assumptions underlying our model.

4.4 Conclusions

The contribution presented here allows for improvement in the space of geometric models for tumor deconvolution. The assumptions of the model mirror those in Chapter 2: a reliance on intra- and intertumor heteroegenity coupled with tumor phylogenetics in order to better make inferences over a large number of samples, given that these traits encode nicely as geometric properties. We advanced the models presented in Chapter 2 by eliminating nusiance parameters,

which allowed for greater usability. We found our tool to correlate well with a state-of-the-art set of approaches using both CNV and methylation data, demonstrating the clinical and molecular efficacy of the tool.

Nonetheless, there remain areas for future contributions, to be detailed in the following chapter. These areas of contribution include the use of integrated data about data type in the case of heterogeneous data — for instance, an expression of higher confidence in DNA data versus RNA data. Further, models that can integrate single-cell data as it becomes widely available will be of great use in prognostic and diagnostic tools.

Chapter 5

Conclusion and Future Directions

5.1 Conclusions

Better methods to model the interaction among tumor evolution, intratumor heterogeneity, and intertumor heterogeneity have remained challenging problems in the field of tumor biology. Although prior work has created geometric models for tumor heterogeneity, to our knowledge, this thesis represents the first efforts toward applications of simplicial complex models of tumor progression and heterogeneity. The extension from a simplex to a simplicial complex represents several key advantages in the study of tumor biology. Through the use of a simplicial complex, each branch of a tumor phylogeny can be modeled separately. The interactive nature of tumor heterogeneity and evolution implies that modeling each branch of the phylogeny specifically yields greater understanding of the heterogeneity of the branches, which in turn may permit better models of survival.

In Chapter 2, we presented extensions of the simplicial approach to tumor deconvolution to a mixed membership model for tumors based on simplicial complexes. In addition, we presented an objective function consistent with a minimum evolution interpretation of tumor phylogenies. We applied the framework to a collection of synthetic data sets for validation versus previous approaches, and real tumor data from TCGA in order to yield additional insights, ontology-

related validation, and survival-related validation.

In Chapter 3, we improved the clustering used as a portion of the framework presented in Chapter 2. Rather than the parametric clustering approach used in Chapter 2's framework where cluster representatives were in the middle of the subsimplices, the nonparameteric approach outlined in Chapter 3 coupled with the novel kernel function and distance metric permitted cluster representatives to be extrema of the clusters. As a result, the cluster representatives had correspondence to the most extreme examples of subpopulations in a tumor, rather than medoid points within a branch. By changing from a parametric to non-parametric approach, we eliminated a difficult to estimate parameter of the previous model. We validated the approach on several representative synthetic data sets, and applied the method to real tumor copy number data from ovarian, lung, and brain tumors. We demonstrated highly positive results on the ovarian tumors, as well as the limits of the approach for functional application on the brain tumor data set. We then were able to interpret the results on the ovarian and lung data to derive significant survival differences and putative subtyping of the tumors.

In Chapter 4, we applied the improvements made in Chapter 3 to the framework presented in Chapter 2. We further improved the framework by adding feature selection via BIC into the objective function. Additionally, we developed and implemented dimensionality detection both via slivers [14] and by a PCA [73] oriented approach, depending on the problem context. We applied the improved pipeline to several real tumor data sets from breast tumors. The data included both RNASeq and DNA copy number data, as well as heterogeneous combinations of DNA and RNA data. We validated the approach through statistically significant correlation to two state-of-the-art methods [70, 83]. One of those approaches used DNA variant data, while the other used DNA methylation data. We further validated the approach through the use of a meta-ontological tool [20], demonstrating enrichment for breast tumor terms. The framework further demonstrated statistically significant correlation to clinically-validated markers.

5.2 Future Directions

This thesis provides important and significant strides toward using geometric features of genomic point clouds to better understand tumor biology. The developments of theory, algorithms, and validation demonstrate improvement in relating more subtle aspects of the tumor genome to clinically-relevant markers. Nonetheless, further contributions may increase the ability of the algorithm to make inferences, and to translate those inferences into actionable clinical statements.

One such important future direction remains the relationship of the first round of dimensionality reduction via PCA [73]. Initial work has demonstrated the efficacy of more advanced approaches based on spectral clustering to compression of genomic data with the end goal of deconvolution [38, 56]. Importantly, however, the initial clustering step cannot perturb the geometry uncovered through our current approaches. An approach that might implement a more exotic reduced representation of the data might employ some approach, such as those outlined in [38, 56], then use PCA [73] on the reduced representation. The learned relationships among features from the first stage would be coupled with the orthogonality of the reduced representation in PCA [73]. Otherwise, if the geometry of the point cloud is not preserved, the assumptions of point cloud geometry leveraged in this thesis will not hold and the approach will fall short.

Further, we consider heterogeneous data types in terms of RNA and DNA quantified expression and copy number data combined into a single data set in Chapter 4, only DNA data in Chapter 3, and only RNA data in Chapter 2. A true integration of specific noise models for multiple platforms and data types emboided in the objective function may yield a more realistic model of the data. For example, future work may seek to expand the platform established here to include data from both copy number, structural variant, and single-nucleotide polymorphism sources. One potential implementation for this sort of technology might be to incorporate Single Nucleotide Polymorphism (SNP) mutation rates — that is, the rate of change from a particular base pair to another base pair along the genome — in inferring subpopulations. Additionally, distinct noise models associated with complimentary single-cell sequencing data may further

increase the specificity of the model, and move toward unifying models of individual tumor evolution and the models of bulk tumor evolution. If we assume data were present for samples that identified new genomic features, such as SNP mutation rates and structural variant rates, these could be incorporated into the existing data matrix outlined in earlier chapters of the thesis. One contemporary work that embodies incorporating structural variant data into a deconvolution model is THeTA [68, 69], which may be useful as a foundational work to future directions in this space.

Additionally, if we suppose single-cell whole-genome data were available, these ought to be placed in a different data matrix, due to the fact the unmixing would handle single-cell data differently. For instance, the conditional distribution part of the probability statement places a penalty on bulk data points from the surface of a sub-simplex, which corresponds to a penalty for not being well-represented by some mixture of the proposed subtypes. On the other hand, a distance-based penalty for single-cell data to the nearest inferred population might make more sense, because the single-cell data ought not to be represented by a mixture of inferred subpopulations, but rather by a specific subpopulation. Further, by using different matrices, different noise parameters would be able to be used for both the single-cell and bulk data.

A specific step that can be taken to implement such an extension is to expand the parameter set for weight or noise to be platform-specific as the data permits, and otherwise to learn a noise model from the data. Currently, we penalize data points in probability as a function of distance from the surface of a subsimplex. If we suppose we knew the data point were a single-cell sequence data point, we would presume that the data ought to be near a vertex representing a subpopulation of the data. Rather than distance to the surface of the subsimplex, which corrsponds to the area well-represented by bulk samples, we could penalize in probability for distance from an inferred vertex. To address the different noise models for DNA and RNA data, we could explicitly model the parameter that weighs the conditional part of our probability model relative to the prior (called γ in earlier chapters) as a vector, with each element corresponding to noise

for a data type, rather than a scalar. In regards to further automation of these parameters, as computational resources increase, the number of experiments that could be done ought also to increase. As a result, the neighborhood parameter used by the medoidshift clustering aspect [81] might be learned by trying multiple values, and choosing the value that leads to the maximum likelihood model. In a similar fashion, the number of iterations parameter could be transformed into a maximum runtime for the unmixing parameter, analogous to a walltime for other compute jobs. As a result, the model could be reduced to a choice of walltime and p-value, with the other parameters learned from the data or chosen in a maximum likelihood fashion.

The many challenges associated with modeling tumor progression and making clinically-relevant predictions based on the models of tumor progression continue to provide a barrier to progress in tumor biology. Nevertheless, scientific inquiry has pressed against these hurdles in order to improve understanding. This thesis represents a step toward a more realistic understanding of tumor biology based on improvements in understanding how geometric structure in tumor point clouds may influence outcomes, and may be influenced by tumor progression and heterogeneity. We encapsulate these phenomena through the use of simplicial complex modeling, which provides greater specificity that previous work. In short, the work presented in this thesis represents an important step in a quantitative understanding in how tumors evolve, and how the evolution of tumors relate to likely outcomes for patients. It serves as a crucial step toward an improved deployment of quantitation and automation in personalized medicine.

Bibliography

[1] D Craig Allred, Yun Wu, Sufeng Mao, Iris D Nagtegaal, Sangjun Lee, Charles M Perou, Syed K Mohsin, Peter O'Connell, Anna Tsimelzon, and Dan Medina. Ductal carcinoma in situ and the emergence of diversity during breast cancer evolution. *Clinical Cancer Research*, 14(2):370–378, 2008. 1.5

[2] Noemi Andor, Julie V Harness, Sabine Mueller, Hans W Mewes, and Claudia Petritsch. Expands: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics*, 30(1):50–60, 2014. 1.2.1

[3] Dvir Aran, Marina Sirota, and Atul J Butte. Systematic pan-cancer analysis of tumour purity. *Nature communications*, 6, 2015. 1.4, 4.1

[4] Monya Baker. Functional genomics: the changes that count. *Nature*, 482(7384):257–262, 2012. 1.1

[5] Mukund Balasubramanian and Eric L Schwartz. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002. 3.2.1

[6] Thomas Bayes, Richard Price, and John Canton. *An essay towards solving a problem in the doctrine of chances*. C. Davis, Printer to the Royal Society of London, 1763. 1.2.1

[7] Philippe L Bedard, Aaron R Hansen, Mark J Ratain, and Lillian L Siu. Tumour heterogeneity in the clinic. *Nature*, 501(7467):355–364, 2013. 1.1, 2

[8] Martin Bengtsson, Martin Hemberg, Patrik Rorsman, and Anders Ståhlberg. Quantifi-

cation of mrna in single cells and modelling of rt-qpcr induced noise. *BMC molecular biology*, 9(1):63, 2008. 1.1

[9] David Brocks, Yassen Assenov, Sarah Minner, Olga Bogatyrova, Ronald Simon, Christina Koop, Christopher Oakes, Manuela Zucknick, Daniel Bernhard Lipka, Joachim Weischenfeldt, et al. Intratumor dna methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell reports*, 8(3):798–806, 2014. 1.2.2

[10] Ricardo Camilo, Vera Luíza Capelozzi, Sheila Aparecida Coelho Siqueira, and Fabíola Del Carlo Bernardi. Expression of p63, keratin 5/6, keratin 7, and surfactant-a in non–small cell lung carcinomas. *Human pathology*, 37(5):542–546, 2006. 3.3.2

[11] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, et al. Absolute quantification of somatic dna alterations in human cancer. *Nature biotechnology*, 30(5): 413–421, 2012. 1.2.1, 2

[12] Kaisorn L Chaichana, Khan K Chaichana, Alessandro Olivi, Jon D Weingart, Richard Bennett, Henry Brem, and Alfredo Quiñones-Hinojosa. Surgical outcomes for older patients with glioblastoma multiforme: preoperative factors associated with decreased survival: clinical article. *Journal of neurosurgery*, 114(3):587–594, 2011. 1.5

[13] Tsung-Han Chan, Chong-Yung Chi, Yu-Min Huang, and Wing-Kin Ma. A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing. *IEEE Transactions on Signal Processing*, 57(11):4418–4432, 2009. 1.2.1, 2.1.3

[14] Siu-Wing Cheng and Man-Kwun Chiu. Dimension detection via slivers. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1001–1010. Society for Industrial and Applied Mathematics, 2009. 4.2.1, 4.2.3, 5.1

[15] Salim Akhter Chowdhury, Stanley E Shackney, Kerstin Heselmeyer-Haddad, Thomas Ried, Alejandro A Schäffer, and Russell Schwartz. Phylogenetic analysis of multiprobe

fluorescence in situ hybridization data from tumor cell populations. *Bioinformatics*, 29 (13):i189–i198, 2013. 4.1

[16] Salim Akhter Chowdhury, Stanley E Shackney, Kerstin Heselmeyer-Haddad, Thomas Ried, Alejandro A Schäffer, and Russell Schwartz. Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics. *PLoS Comput Biol*, 10(7):e1003740, 2014. 4.1

[17] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002. 3.2

[18] Phillip EC Compeau, Pavel A Pevzner, and Glenn Tesler. How to apply de bruijn graphs to genome assembly. *Nature biotechnology*, 29(11):987–991, 2011. 1.1

[19] Seqc/Maqc-Iii Consortium et al. A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature biotechnology*, 32(9):903–914, 2014. 2.1.2, 2.2.1, 3.2.1, 3.3.1, 4.1, 4.3.2

[20] Glynn Dennis, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. David: database for annotation, visualization, and integrated discovery. *Genome biology*, 4(9):R60, 2003. (document), 2.2.2, 2.2, 2.3, 3.2.3, 3.3.2, 4.3.4, 5.1

[21] Richard Desper, Feng Jiang, Olli-P Kallioniemi, Holger Moch, Christos H Papadimitriou, and Alejandro A Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of computational biology*, 6(1):37–51, 1999. 1.2.1, 1.2.2, 1.2.2, 2, 3.2.2, 4.1

[22] Li Ding, Michael C Wendl, Daniel C Koboldt, and Elaine R Mardis. Analysis of next generation genomic data in cancer: accomplishments and challenges. *Human molecular genetics*, page ddq391, 2010. 1.1

[23] Robert Ehrlich and William E Full. Sorting out geologyunmixing mixtures. *Use and Abuse of Statistical Methods in the Earth Sciences*, pages 33–46, 1987. 1.2.1, 2

[24] Mohammed El-Kebir, Gryte Satas, Layla Oesper, and Benjamin J Raphael. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Systems*, 3(1):43–53, 2016. 1.2.1

[25] Ruth Etzioni, Sarah Hawley, Dean Billheimer, Lawrence D True, and Beatrice Knudsen. Analyzing patterns of staining in immunohistochemical studies: application to a study of prostate cancer recurrence. *Cancer Epidemiology and Prevention Biomarkers*, 14(5): 1040–1046, 2005. 1.2.3

[26] Isaiah J Fidler. Tumor heterogeneity and the biology of cancer invasion and metastasis. *Cancer research*, 38(9):2651–2660, 1978. 1, 1.2.3, 2, 3.2.1, 3.2.2, 4.1

[27] Andrej Fischer, Ignacio Vázquez-García, Christopher JR Illingworth, and Ville Mustonen. High-definition reconstruction of clonal composition in cancer. *Cell reports*, 7(5):1740–1752, 2014. 4.1

[28] Judah Folkman. Antiangiogenesis in cancer therapyendostatin and its mechanisms of action. *Experimental cell research*, 312(5):594–607, 2006. 3.3.2

[29] Marco Gerlinger, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl j Med*, 2012(366):883–892, 2012. 1.1, 4.1

[30] Marco Gerlinger, Stuart Horswell, James Larkin, Andrew J Rowan, Max P Salm, Ignacio Varela, Rosalie Fisher, Nicholas McGranahan, Nicholas Matthews, Claudio R Santos, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature genetics*, 46(3):225–233, 2014. 1.1

[31] Moritz Gerstung and Niko Beerenwinkel. Waiting time models of cancer progression.

Mathematical Population Studies, 17(3):115–135, 2010. 1.2.1

[32] Andrew K Godwin, Joseph R Testa, and Thomas C Hamilton. The biology of ovarian cancer development. *Cancer*, 71(S2):530–536, 1993. 3.3.2

[33] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. *Numerische mathematik*, 14(5):403–420, 1970. 2.1.4

[34] Rodrigo Goya, Mark GF Sun, Ryan D Morin, Gillian Leung, Gavin Ha, Kimberley C Wiegand, Janine Senz, Anamaria Crisan, Marco A Marra, Martin Hirst, et al. Snvmix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 26(6):730–736, 2010. 4.1

[35] Mel Greaves and Carlo C Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012. 1.2.2

[36] Arief Gusnanto, Henry M Wood, Yudi Pawitan, Pamela Rabbitts, and Stefano Berri. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics*, 28(1):40–47, 2012. 1.2.1

[37] Gavin Ha, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M Prentice, Nataliya Melnyk, Andrew McPherson, Ali Bashashati, Emma Laks, et al. Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome research*, 24(11):1881–1893, 2014. 1.2.1, 1.4, 2

[38] Eunjung Han, Peter Carbonetto, Ross E Curtis, Yong Wang, Julie M Granka, Jake Byrnes, Keith Noto, Amir R Kermany, Natalie M Myres, Mathew J Barber, et al. Clustering of 770,000 genomes reveals post-colonial population structure of north america. *Nature Communications*, 8:14238, 2017. 5.2

[39] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000. 1, 2, 2.2.2, 3.3.2

[40] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011. 1, 2

[41] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1): 100–108, 1979. 2.1.3, 2.1.4, 3.1

[42] Kerstin M Heselmeyer-Haddad, Lissa Y Berroa Garcia, Amanda Bradley, Leanora Hernandez, Yue Hu, Jens K Habermann, Christoph Dumke, Christoph Thorns, Sven Perner, Ekaterina Pestova, et al. Single-cell genetic analysis reveals insights into clonal development of prostate cancers and indicates loss of pten as a marker of poor prognosis. *The American journal of pathology*, 184(10):2671–2686, 2014. 4.1

[43] Yoshitsugu Horio, Takashi Takahashi, Tetsuo Kuroishi, Kenji Hibi, Motokazu Suyama, Takao Niimi, Kaoru Shimokata, Kazuhiro Yamakawa, Yusuke Nakamura, Ryuzo Ueda, et al. Prognostic significance of p53 mutations and 3p deletions in primary resected non-small cell lung cancer. *Cancer research*, 53(1):1–4, 1993. 1.5

[44] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2 (1):193–218, 1985. 3.2.2

[45] Thomas J Hudson, Warwick Anderson, Axel Aretz, Anna D Barker, Cindy Bell, Rosa R Bernabé, MK Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, et al. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010. 1.2.1, 1.3, 1.5, 2.1.1, 2.2.2, 2.5, 3.2.3, 3.3.2, 4.3, 4.3.2, 4.3.5, 4.3.5, 4.3.5

[46] Clyde A Hutchison III and J Craig Venter. Single-cell genomics. *Nature biotechnology*, 24(6):657–659, 2006. 1.1, 1.1

[47] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004. 2.1.3

[48] Tadashi Imanishi and Hajime Nakaoka. Hyperlink management system and id converter

system: enabling maintenance-free hyperlinks among major biological databases. *Nucleic acids research*, 37(suppl 2):W17–W22, 2009. 2.2.2

[49] Wei Jiao, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics*, 15(1):35, 2014. 1.2.1, 4.1

[50] Tomer Kalisky and Stephen R Quake. Single-cell genomics. *Nature methods*, 8(4):311–314, 2011. 1.1, 1.1

[51] Anne Kallioniemi, Olli-P Kallioniemi, Damir Sudar, Denis Rutovitz, Joe W Gray, Fred Waldman, and Dan Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5083):818–822, 1992. 1.2.2

[52] Razi Khaja, Junjun Zhang, Jeffrey R MacDonald, Yongshu He, Ann M Joseph-George, John Wei, Muhammad A Rafiq, Cheng Qian, Mary Shago, Lorena Pantano, et al. Genome assembly comparison identifies structural variants in the human genome. *Nature genetics*, 38(12):1413–1418, 2006. 1.1

[53] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015. 1.1

[54] Philippe Lamy, Claus L Andersen, Lars Dyrskjot, Niels Torring, and Carsten Wiuf. A hidden markov model to estimate population mixture and allelic copy-numbers in cancers using affymetrix snp arrays. *BMC bioinformatics*, 8(1):434, 2007. 4.1

[55] Nicholas B Larson and Brooke L Fridley. Purbayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics*, 29(15):1888–1889, 2013. 1.2.1, 4.1

[56] Ann B Lee, Diana Luca, Lambertus Klei, Bernie Devlin, and Kathryn Roeder. Discovering genetic ancestry using spectral graph theory. *Genetic epidemiology*, 34(1):51–59, 2010.

5.2

[57] Ao Li, Zongzhi Liu, Kimberly Lezon-Geyda, Sudipa Sarkar, Donald Lannin, Vincent Schulz, Ian Krop, Eric Winer, Lyndsay Harris, and David Tuck. Gphmm: an integrated hidden markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome snp arrays. *Nucleic acids research*, page gkr014, 2011. 4.1

[58] Yi Li and Xiaohui Xie. Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity. *Bioinformatics*, page btu174, 2014. 4.1

[59] Herwig-Ulf Meier-Kriesche, Jesse D Schold, and Bruce Kaplan. Long-term renal allograft survival: Have we made significant progress or is it time to rethink our analytic and therapeutic strategies? *American Journal of Transplantation*, 4(8):1289–1295, 2004. 3.2.3, 3.3.2

[60] Matthew Meyerson, Stacey Gabriel, and Gad Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, 11(10):685–696, 2010. 1.1

[61] Sebastian Mika, Bernhard Schölkopf, Alexander J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel pca and de-noising in feature spaces. In *NIPS*, volume 11, pages 536–542, 1998. 2.1.3

[62] Nicholas Navin, Alexander Krasnitz, Linda Rodgers, Kerry Cook, Jennifer Meth, Jude Kendall, Michael Riggs, Yvonne Eberling, Jennifer Troge, Vladimir Grubor, et al. Inferring tumor progression from genomic heterogeneity. *Genome research*, 20(1):68–80, 2010. 1.1, 4.1

[63] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 2011. 2

[64] Nicholas E Navin. Cancer genomics: one cell at a time. *Genome biology*, 15(8):452, 2014. 1.1, 1.1, 3.2.2

[65] Nicholas E Navin. Delineating cancer evolution with single-cell sequencing. *Science translational medicine*, 7(296):296fs29–296fs29, 2015. 1.1

[66] Cancer Genome Atlas Research Network et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 2011. 1.5, 3.2.3

[67] Serena Nik-Zainal, Peter Van Loo, David C Wedge, Ludmil B Alexandrov, Christopher D Greenman, King Wai Lau, Keiran Raine, David Jones, John Marshall, Manasa Ramakrishna, et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012. 1, 2, 3.2.2, 4.1

[68] Layla Oesper, Ahmad Mahmoody, and Benjamin J Raphael. Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome biology*, 14(7):R80, 2013. 1.2.1, 1.4, 2, 4.1, 5.2

[69] Layla Oesper, Gryte Satas, and Benjamin J Raphael. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, 30(24):3532–3540, 2014. 1.2.1, 1.4, 2, 3.2.1, 4.1, 5.2

[70] Vitor Onuchic, Ryan J Hartmaier, David N Boone, Michael L Samuels, Ronak Y Patel, Wendy M White, Vesna D Garovic, Steffi Oesterreich, Matt E Roth, Adrian V Lee, et al. Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types. *Cell reports*, 17(8):2075–2086, 2016. (document), 1.2.1, 4.3.5, 4.3.5, 4.7, 4.3.5, 5.1

[71] So Yeon Park, Mithat Gönen, Hee Jung Kim, Franziska Michor, and Kornelia Polyak. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *The Journal of clinical investigation*, 120(2):636–644, 2010. 1.5

[72] Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi

Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8): 1160–1167, 2009. 2.2.2

[73] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901. 1.2.1, 2.1.3, 4.2.1, 5.1, 5.2

[74] Gregory Pennington, Stanley Shackney, and Russell Schwartz. Cancer phylogenetics from single-cell assays. *Dept. Comput. Sci., Pittsburgh, PA, USA, Carnegie Mellon Univ., Tech. Rep. CMU-CS-06-103*, 2006. 1.2.1, 1.2.2, 2, 4.1

[75] Gregory Pennington, Charles A Smith, Stanley Shackney, and Russell Schwartz. Reconstructing tumor phylogenies from heterogeneous single-cell data. *Journal of bioinformatics and computational biology*, 5(02a):407–427, 2007. 1.2.1, 1.2.2, 2, 4.1

[76] Dalila Pinto, Katayoon Darvishi, Xinghua Shi, Diana Rajan, Diane Rigler, Tom Fitzgerald, Anath C Lionel, Bhooma Thiruvahindrapuram, Jeffrey R MacDonald, Ryan Mills, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature biotechnology*, 29(6):512–520, 2011. 3.2.1

[77] Ondrej Podlaha, Markus Riester, Subhajyoti De, and Franziska Michor. Evolution of the cancer genome. *Trends in Genetics*, 28(4):155–163, 2012. 1.2.2

[78] Mihai Pop, Adam Phillippy, Arthur L Delcher, and Steven L Salzberg. Comparative genome assembly. *Briefings in bioinformatics*, 5(3):237–248, 2004. 1.1

[79] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971. 3.2.2

[80] Theodore Roman, Amir Nayyeri, Brittany Terese Fasy, and Russell Schwartz. A simplicial complex-based approach to unmixing tumor progression data. *BMC bioinformatics*, 16(1): 254, 2015. (document), 1.4, 1.6, 1, 2, 2.1, 2.2, 2.1.2, 2.3, 2.4, 2.5, 2.2.2, 2.6, 3.1, 3.2.2,

3.3.2, 4.1, 4.2, 4.2.1, 4.2.4, 4.2.4

[81] Theodore Roman, Lu Xie, and Russell Schwartz. Medoidshift clustering applied to genomic bulk tumor data. *BMC genomics*, 17(1):6, 2016. 1.4, 1.6, 3.1, 3.1, 3.2.3, 3.2, 3.3, 3.1, 4.1, 4.2.1, 4.2.1, 4.2.2, 4.3, 5.2

[82] Andrew Roth, Jiarui Ding, Ryan Morin, Anamaria Crisan, Gavin Ha, Ryan Giuliany, Ali Bashashati, Martin Hirst, Gulisa Turashvili, Arusha Oloumi, et al. Jointsnvmix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, 28(7):907–913, 2012. 4.1

[83] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. Pyclone: statistical inference of clonal population structure in cancer. *Nature methods*, 11(4):396–398, 2014. 1.2.1, 1.4, 2, 2.1.5, 4.1, 4.3.5, 4.3.5, 5.1

[84] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000. 2.1.3, 3.2.1

[85] Hege G Russnes, Nicholas Navin, James Hicks, and Anne-Lise Borresen-Dale. Insight into the heterogeneity of breast cancer through next-generation sequencing. *The Journal of clinical investigation*, 121(10):3810–3818, 2011. 3.2.2

[86] Russell Schwartz and Stanley E Shackney. Applying unmixing to gene expression data for tumor phylogeny inference. *BMC bioinformatics*, 11(1):42, 2010. 1.2.1, 1.2.2, 1.2.3, 1.4, 2, 4.1, 4.2.1, 4.2.1, 4.2.4

[87] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6 (2):461–464, 1978. 4.2.4

[88] Subhajit Sengupta, Jing Wang, Juhee Lee, Peter Müller, Kamalakar Gulukota, Arunava Banerjee, and Yuan Ji. Bayclone: Bayesian nonparametric inference of tumor subclones using ngs data. In *Pacific Symposium on Biocomputing*, volume 20, page 467, 2015. 1.2.1

[89] Yaser Ajmal Sheikh, Erum Arif Khan, and Takeo Kanade. Mode-seeking by medoidshifts. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. 3.1, 3.2, 3.2, 3.2, 3.2.1, 3.2.3

[90] William B Size. Use and abuse of statistical methods in the earth sciences. 1987. 1.2.1

[91] Damian Smedley, Syed Haider, Steffen Durinck, Luca Pandini, Paolo Provero, James Allen, Olivier Arnaiz, Mohammad Hamza Awedh, Richard Baldock, Giulia Barbiera, et al. The biomart community portal: an innovative alternative to large, centralized data repositories. *Nucleic acids research*, page gkv350, 2015. 3.2.3

[92] Michael Stanbrough, Glenn J Bubley, Kenneth Ross, Todd R Golub, Mark A Rubin, Trevor M Penning, Phillip G Febbo, and Steven P Balk. Increased expression of genes converting adrenal androgens to testosterone in androgen-independent prostate cancer. *Cancer research*, 66(5):2815–2825, 2006. 1.2.3

[93] Philip J Stephens, Patrick S Tarpey, Helen Davies, Peter Van Loo, Chris Greenman, David C Wedge, Serena Nik-Zainal, Sancha Martin, Ignacio Varela, Graham R Bignell, et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403):400–404, 2012. 4.3.5

[94] Xiaoping Su, Li Zhang, Jianping Zhang, Funda Meric-Bernstam, and John N Weinstein. Purityest: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics*, 28(17):2265–2266, 2012. 2

[95] Michael E Sughrue, Tyson Sheean, Phillip A Bonney, Adrian J Maurer, and Charles Teo. Aggressive repeat surgery for focally recurrent primary glioblastoma: outcomes and theoretical framework. *Neurosurgical focus*, 38(3):E11, 2015. 1.5

[96] James E Talmadge and Isaiah J Fidler. Aacr centennial series: the biology of cancer metastasis: historical perspective. *Cancer research*, 70(14):5649–5669, 2010. 1

[97] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework

for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000. 2.1.3, 3.1, 3.2.1, 3.2.1, 4.2.2

[98] David Tolliver, Charalampos Tsourakakis, Ayshwarya Subramanian, Stanley Shackney, and Russell Schwartz. Robust unmixing of tumor states in array comparative genomic hybridization data. *Bioinformatics*, 26(12):i106–i114, 2010. 1.2.1, 1.2.2, 1.2.3, 1.4, 2, 2.1.2, 2.1.4, 2.2.1, 3.2.2, 3.2.3, 4.1, 4.2, 4.2.1, 4.2.4

[99] Richard W Tothill, Anna V Tinker, Joshy George, Robert Brown, Stephen B Fox, Stephen Lade, Daryl S Johnson, Melanie K Trivett, Dariush Etemadmoghadam, Bianca Locandro, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical cancer research*, 14(16):5198–5208, 2008. 1.5, 3.2.3

[100] Barbara J Trask. Fluorescence in situ hybridization: applications in cytogenetics and gene mapping. *Trends in Genetics*, 7(5):149–154, 1991. 1.2.2

[101] Jakob J Verbeek, Nikos Vlassis, and Ben Kröse. Efficient greedy learning of gaussian mixture models. *Neural computation*, 15(2):469–485, 2003. 2.1.4, 3.1

[102] Roel GW Verhaak, Pablo Tamayo, Ji-Yeon Yang, Diana Hubbard, Hailei Zhang, Chad J Creighton, Sian Fereday, Michael Lawrence, Scott L Carter, Craig H Mermel, et al. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *The Journal of clinical investigation*, 123(1), 2012. 3.3.2

[103] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1073–1080. ACM, 2009. 3.2.2

[104] Alan Walker and Julian Parkhill. Single-cell genomics. *Nature Reviews Microbiology*, 6 (3):176–177, 2008. 1.1

[105] Yong Wang and Nicholas E Navin. Advances and applications of single-cell sequencing

technologies. *Molecular cell*, 58(4):598–609, 2015. 1.1

[106] Larry Wasserman. All of nonparametric statistics, 2006. 2.1.3

[107] Andreas Weingessel and Kurt Hornik. Local pca algorithms. *IEEE Transactions on neural Networks*, 11(6):1242–1250, 2000. 2.1.3

[108] Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W Laird, Douglas A Levine, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications*, 4, 2013. 2.2.2

[109] Habil Zare, Junfeng Wang, Alex Hu, Kris Weber, Josh Smith, Debbie Nickerson, ChaoZhong Song, Daniela Witten, C Anthony Blau, and William Stafford Noble. Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput Biol*, 10(7):e1003703, 2014. 4.1

[110] Wenyu Zhang, Jiajia Chen, Yang Yang, Yifei Tang, Jing Shang, and Bairong Shen. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PloS one*, 6(3):e17915, 2011. 1.1