MATRIX AND TENSOR COMPARISONS OF GENOMIC PROFILES TO PREDICT CANCER SURVIVAL AND DRUG TARGETS

by

Preethi Sankaranarayanan

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Bioengineering

The University of Utah

May 2015

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Preethi Sankaranarayanan**

has been approved by the following supervisory committee members:

Orly Alter	, Chair	**08-29-2014** Date Approved
Robert S. MacLeod	, Member	**08-29-2014** Date Approved
Karen Eilbeck	, Member	**08-29-2014** Date Approved
Andrea Bild	, Member	**08-29-2014** Date Approved
Tolga Tasdizen	, Member	**08-29-2014** Date Approved

and by **Patrick A. Tresco** , Chair/Dean of

the Department of **Bioengineering**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Despite recent large-scale profiling efforts, the best predictor of a glioblastoma (GBM) brain cancer patient's survival remains the patient's age at diagnosis. The best predictor of an ovarian serous cystadenocarcinoma (OV) patient's survival remains the tumor's stage, an assessment – numbering I to IV – of the spread of the cancer. To identify DNA copy-number alterations (CNAs) that might predict GBM or OV patients' survival, we comparatively modeled matched genomic profiles from The Cancer Genome Atlas (TCGA).

Generalized singular value decomposition (GSVD) of patient-matched but probe-independent GBM and normal profiles uncovered a previously unknown global pattern of tumor-exclusive co-occurring CNAs that is correlated, and possibly causally related to, GBM patients' survival and response to chemotherapy. This suggests that the GBM survival phenotype is an outcome of its global genotype. The GSVD, formulated as a framework for comparatively modeling two composite datasets, removes from the pattern variations that occur in the normal human genome (e.g., female-specific X chromosome amplification) and experimental variations, without a-priori knowledge of these variations. The pattern is independent of age, and combined with age, makes a better predictor than age alone. The pattern suggests previously unrecognized targets for personalized GBM drug therapy, the kinase *TLK2* and the methyltransferase *METTL2A*.

A novel tensor GSVD of patient- and platform-matched OV and normal genomic profiles revealed multiple chromosome arm-wide patterns of CNAs that are correlated with OV patients' survival. These indicate several, previously unrecognized, subtypes of OV. The tensor GSVD is an exact simultaneous decomposition of two high-dimensional datasets arranged in higher-order tensors. The tensor GSVD generalizes the GSVD, which is limited to two second-order tensors, i.e., matrices. The chromosome arm-wide patterns of CNAs are independent of the OV tumor stage. Combined with stage, each of the patterns makes a better predictor than stage alone.

We conclude that the GSVD and the novel tensor GSVD can uncover the relations, and possibly causal coordinations, between different recorded aspects of the same medical phenomenon. GSVD and tensor GSVD comparisons can be used to determine one patient's

medical status in relation to other patients in a set, and inform the patient's prognosis, and possibly also treatment.

CONTENTS

ABSTRACT . . . . . **iii**

LIST OF FIGURES . . . . . **vii**

LIST OF TABLES . . . . . **ix**

ACKNOWLEDGMENTS . . . . . **x**

CHAPTERS

1. INTRODUCTION **1**

1.1 Motivation 1
1.1.1 Personalized Cancer Prognosis and Therapy 3
1.1.2 Discover Mechanisms of Cancer 3
1.2 Existing Methods 4
1.2.1 Principal Component Analysis and Singular Value Decomposition 4
1.2.2 Limitations 6
1.2.3 Generalized Singular Value Decomposition (GSVD) 7
1.3 Dissertation Contributions 7
1.4 Overview of Chapters 9

2. GSVD COMPARISON OF HUMAN BRAIN TUMOR AND NORMAL GENOMIC PROFILES **11**

2.1 Glioblastoma Multiforme (GBM) 11
2.1.1 Tumor and Normal Datasets 12
2.2 Generalized Singular Value Decomposition (GSVD) 12
2.2.1 Construction of the GSVD 13
2.2.2 Mathematical Exclusivity and Significance 16
2.2.3 Biological Interpretation of the Mathematical Patterns 16
2.3 Results: GSVD of GBM and Normal Genomic Profiles 18
2.3.1 GSVD Removes Batch Effects 18
2.3.2 Discover Copy Number Changes Associated with GBM 23
2.3.3 Patient Prognosis and Drug Target Prediction 25

3. TENSOR GSVD (tGSVD) COMPARISONS OF MATCHED GENOMIC PROFILES **32**

3.1 Ovarian Serous Cystadenocarcinoma (OV) 32
3.1.1 Tumor and Normal Datasets 33
3.2 Tensor Generalized Singular Value Decomposition (tGSVD) 34
3.2.1 Introduction 34
3.2.2 tGSVD : Formulation and Construction 34

3.2.3 Existence, Uniqueness and Special Cases 45
3.2.4 Interpretation 47
3.2.5 Discovery and Validation of CNAs Predicting OV Survival 49
3.3 Biological Results 50
3.3.1 Independent Chromosome Arm-Wide Predictors of OV Survival 50
3.3.2 Novel Frequent Focal CNAs Indicating Survival 55
3.3.3 Possible Roles in OV Pathogenesis and Personalized Therapy 62

4. DISCUSSION 69

4.1 Summary 69
4.1.1 The GSVD Comparison of GBM and Normal Genomic Profiles 69
4.1.2 The Tensor GSVD Comparisons of Matched OV Genomic Profiles 70
4.2 Future Directions 72
4.2.1 Additional Applications in Personalized Medicine 72

APPENDICES

A. SUPPLEMENT I 73

B. SUPPLEMENT II 89

C. STATISTICAL METHODS 91

REFERENCES 99

LIST OF FIGURES

1.1 Structure of high-dimensional datasets with one or more common axes 2

1.2 Generalized singular value decomposition (GSVD) 8

2.1 Generalized singular value decomposition (GSVD) of the TCGA patient-matched tumor and normal aCGH profiles. 14

2.2 Most significant probelets in the tumor and normal datasets. 15

2.3 Significant probelets and corresponding tumor and normal arraylets uncovered by GSVD of the patient-matched GBM and normal aCGH profiles. 17

2.4 Differences in copy numbers among the TCGA annotations associated with the significant probelets. 20

2.5 The first most tumor-exclusive probelet and corresponding tumor arraylet uncovered by GSVD of the patient-matched GBM and normal aCGH profiles. 21

2.6 The first most normal-exclusive, i.e., 251st probelet and corresponding normal arraylet uncovered by GSVD. 22

2.7 Copy-number distributions of the 246th probelet and the corresponding 246th normal arraylet and 246th tumor arraylet. 23

2.8 Survival analyses of the three sets of patients classified by GSVD, age at diagnosis or both. 26

2.9 Kaplan-Meier (KM) survival analyses of the initial set of 251 patients classified by copy number changes in segments containing biochemically putative drug targets in GBM. 28

3.1 Tensor generalized singular value decomposition (tGSVD) of the patient- and platform-matched DNA copy-number profiles of the Xq chromosome arm. . . . 36

3.2 The tGSVD of the patient- and platform-matched DNA copy-number profiles of the chromosome arm combination 6p+12p. 39

3.3 The tGSVD of the patient- and platform-matched DNA copy-number profiles of chromosome arm 7p. 42

3.4 Most significant subtensors in the tumor and normal discovery datasets. 48

3.5 Survival analyses of the discovery set of patients classified by the standard OV indicators. 51

3.6 Survival analyses of the validation set of patients classified by the standard OV indicators. 52

3.7 Survival analyses of the discovery and validation sets of patients classified by tGSVD, or tGSVD and tumor stage at diagnosis. 54

3.8 Survival analyses of the discovery set of patients classified by tGSVD and standard OV indicators. . . . 57
3.9 Survival analyses of the validation set of patients classified by tGSVD and standard OV indicators. . . . 58
3.10 Tumor-exclusive and platform-consistent DNA copy-number alterations (CNAs) correlate with ovarian serous cystadenocarcinoma (OV) patients' survival. . . . 60
3.11 Survival analyses of the discovery and validation sets of patients classified by the novel frequent focal CNAs included in the tGSVD arraylets. . . . 61
3.12 Differential mRNA expression between the tGSVD classes is consistent with the CNAs. . . . 63
3.13 Differential microRNA expression between the tGSVD classes is consistent with the CNAs. . . . 65
3.14 Differential protein expression between the tGSVD classes is consistent with the CNAs. . . . 66
A.1 The 247th, normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD. . . . 74
A.2 The 248th, normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD. . . . 75
A.3 The 249th, normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD. . . . 76
A.4 The 250th, normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD. . . . 77
A.5 Kaplan-Meier (KM) survival analyses of only the chemotherapy patients from the three sets classified by GSVD. . . . 78
A.6 KM survival analysis of the initial set of 251 patients classified by a mutation in the gene *IDH1*. . . . 79
A.7 KM survival analysis of only the chemotherapy patients in the initial set classified by a mutation in *IDH1*. . . . 80
A.8 KM survival analyses of the initial set of 251 patients classified by GBM-associated chromosome number changes. . . . 81
A.9 KM survival analyses of the initial set of 251 patients classified by copy number changes in selected segments containing GBM-associated genes or genes previously unrecognized in GBM. . . . 83
A.10 KM survival analyses of only the chemotherapy patients in the initial set of 251 patients classified by copy number changes in selected segments. . . . 84
A.11 Survival analyses of the patients from the three sets classified by chemotherapy alone or GSVD and chemotherapy both. . . . 86
C.1 Box-whisker plot . . . 92

LIST OF TABLES

2.1 Enrichment of the significant probelets in TCGA annotations. 19

2.2 Cox proportional hazard models of the three sets of patients classified by GSVD, age at diagnosis or both. 27

2.3 Cox proportional hazard models of the three sets of patients classified by GSVD, chemotherapy or both. 27

3.1 Cox bivariate proportional hazard models of the patients in the discovery and validation sets classified by both tGSVD and the standard OV indicators. . . . 55

3.2 Cox univariate proportional hazard models of the discovery and validation sets of patients classified by either tGSVD or the standard OV indicators. 56

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Orly Alter, without whose guidance, support and encouragement this research would not have been possible. She was always there for me whenever I needed to be pointed in the right direction. Her strong drive for perfection is contagious and over the past years, I have been greatly influenced and inspired by it. I would also like to thank her for the academic opportunities that she has provided me as well as her support during difficult times both professionally and personally. My sincere thanks to my committee members, Dr. Rob MacLeod, Dr. Karen Eilbeck, Dr. Andrea Bild and Dr. Tolga Tasdizen, for their guidance, insightful comments, and very helpful suggestions.

I also owe a great deal of thanks to many others for helping me grow as a researcher during my Ph.D. training. I thank Ted Schomay and Katie Aiello for the impromptu conversations in the hallway and more importantly their friendship. I thank both of them for their valuable suggestions for improving my presentations and co-authoring my manuscript. My sincere thanks to my former lab-mates, Ben Alpert, Cheng Lee and Nicolas Bertagnolli, for the stimulating discussions and their willingness to help me anytime.

Many thanks to Christine Pickett for her help in editing this dissertation for style, syntax, and grammar. I would also like to thank the Scientific Computing and Imaging (SCI) Institute members, Brenda Peterson, Ed Cask, Nathan Galli, Tony Portillo, Deborah Zemek and Magali Coburn, for helping me out with all the administrative work in a timely manner and for making my stay at SCI so much fun. My sincere thanks to the SCI facility management team, especially Ali Moeinvaziri for patiently solving all the problems with Mathematica software installation and upgrade. Without the state-of-the-art coffee machines, many experiments in this dissertation would not have been possible. I thank Dr. Chris Johnson for the wonderful infrastructure at SCI.

I thank my friends, Avantika Vardhan, Prasanna Muralidharan, Joanita D'souza, Ksheeraja Yekkala, Gopal Veni, Karthik Raman, Umadevi Nagaraja and Amrish Kapoor, for making my time at Salt Lake City the most memorable, filled with joy and positive energy when I was going through difficult times; words alone cannot describe my gratitude.

My family deserves a very special thanks. I would like to thank my parents for their enthusiastic cheerleading and moral support. My sincere thanks to my brother Karthik Sankaranarayanan and his wife Meenakshi Narasimhamurthy, who also happens to be my childhood friend, for their tremendous support.

I thank my best friend and lovely husband, Nikhil Singh, for his unflinching faith in me. Whenever I had doubts in myself, he was there to offer kind and encouraging words. I thank my darling daughter, Iva Singh, for all the "I love you mamma"s and random kisses for no reason with her cute smile, especially on the days when the going gets tough. I also thank my younger daughter, Aria Singh, whose recent arrival has filled our family with joy.

CHAPTER 1

INTRODUCTION

1.1 Motivation

Biological sciences have been transformed over the past two decades by the development of technologies capable of performing large-scale measurements of cellular states. In particular, DNA sequencing instruments and microarrays have undergone an extraordinary increase in efficiency that has reduced the time and cost of experiments by several orders of magnitude. This breakthrough in high-throughput genomic data measurement is revolutionizing personalized medicine. It has resulted in tremendous growth in the number of large-scale multidimensional datasets recording different aspects (such as DNA copy number changes, mutations, gene expression, etc.) of a single disease such as cancer [1]. Often, these data take the structure or form as shown in Figure 1.1 with one or more common axes. However, along with the exponential growth of these large-scale datasets, the need for mathematical frameworks that can identify disease-specific changes by *simultaneously* comparing and contrasting the data, while still *retaining their original structure*, is also growing. A coherent model of these data that simultaneously finds the similarities and dissimilarities can enhance biological understanding of the disease such as cancer and inform a patient's diagnosis, prognosis and treatment.

This comparison is especially important to understand the spatiotemporal interactions among different cellular components such as genes, proteins, microRNAs and metabolites that are present only in the disease state but not in the normal disease-free state. For example, gene perturbation experiments (e.g., knockouts or RNA interference) reveal relationships between genes that may imply direct physical interactions or indirect logical interactions. In contrast, chromatin immunoprecipitation chip data can reveal direct protein-DNA interactions or cofactor associations with bound transcription factors. Combined, these technologies can provide a much more detailed view of a transcriptional regulatory network than either alone.

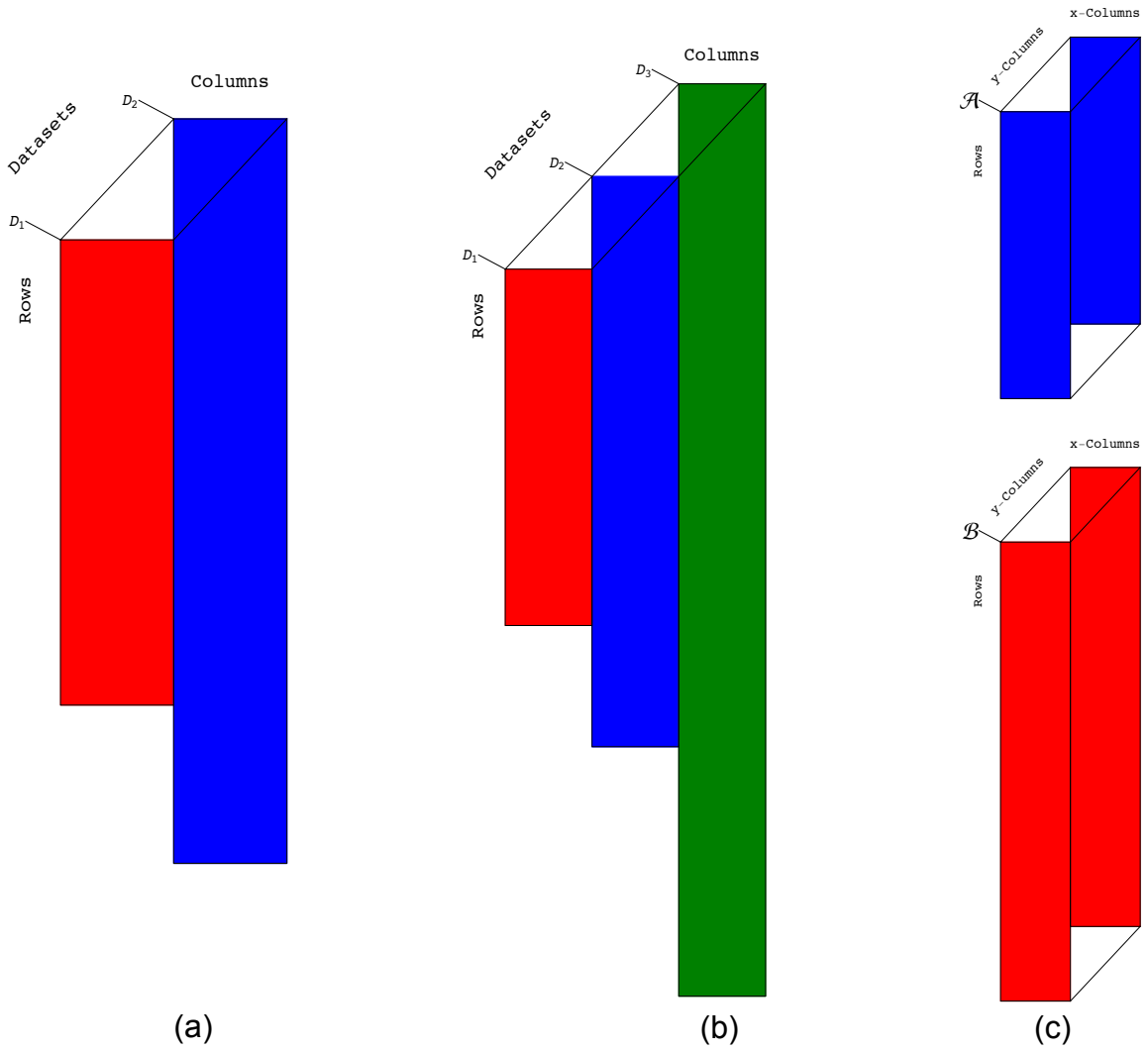


Figure 1.1: Structure of large-scale datasets recording different aspects of a single phenomenon with one or more common axes. Such high-dimensional datasets arise in many fields ranging from finance to biology.
(a) For example, in astronomy, we have the core temperature measurements of different stars (rows) in two galaxies measured over the same time period (columns).
(b) In finance, we have datasets recording stock prices (rows) in different markets over the same time period (columns).
(c) In biology, the tumor and normal genomic profiles (rows) measured on two platforms (y-columns) for the same set of patients (x-columns) can be represented as the two tensors $\mathcal{A}$ and $\mathcal{B}$.

1.1.1 Personalized Cancer Prognosis and Therapy

The complete genetic material of an organism makes up its genome. Genomics is the study of the structure and function of genomes. In cancer cells, some structural and functional changes associated with the genome result in the uncontrolled growth of cells. Cancer genomics is a field of genomics that focuses on acquiring a comprehensive overview of the cancer's formation, growth and development and thereby lays the foundation for the early detection, therapy and prevention of cancer.

The field of cancer genomics continues to advance at an extraordinarily rapid pace with the declining cost of next-generation sequencing and major international efforts, including the International Cancer Genome Consortium [2] and The Cancer Genome Atlas (TCGA). TCGA is a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing and microarrays. The cancers chosen by TCGA are often characterized by poor prognosis and have a high overall public health impact. This initiative aims to develop a comprehensive catalog of the genomic changes that occur in tumors and obtain an in-depth understanding about the relation of these changes to the biological processes in cancer. The end goal is to be able to diagnose, treat and prevent cancer.

Despite extensive studies with the latest tools, the best prognostic predictor of glioblastoma multiforme (GBM), the most common brain tumor, is the patient's age at diagnosis. The best prognostic predictor of the ovarian cystadenocarcinoma (OV), the most lethal gynecologic malignancy, remains the tumor's stage, an assessment – numbering I to IV – of the spread of the cancer. In this dissertation, we describe a global pattern of tumor-exclusive co-occurring copy-number alterations (CNAs) that is correlated and possibly coordinated with GBM patients' survival and response to chemotherapy [3]. The pattern is independent of age, and combined with age, makes a better predictor than age alone. We also describe chromosome arm-wide patterns of tumor-exclusive and platform-consistent DNA CNAs that correlate with OV patients' survival [4]. The patterns, across 6p+12p, 7p and Xq, are independent of the tumor's stage, the best predictor of OV survival to date, and include known as well as previously unreported, yet frequent, CNAs.

1.1.2 Discover Mechanisms of Cancer

The tumor-exclusive pattern of CNAs that we find in GBM data includes most known GBM-associated changes in chromosome numbers and focal CNAs, as well as several previously unreported CNAs in >3% of patients. These CNAs include the biochemically putative drug target, cell cycle-regulated serine/threonine kinase-encoding *TLK2*, the cy-

clin E1-encoding *CCNE1* and the Rb-binding histone demethylase-encoding *KDM5A*. The CNAs previously unrecognized in OV include a deletion of the p38-encoding *MAPK14* and p21-encoding *CDKN1A*; an amplification of *RAD51AP1*, which are drug-targeted in other cancers; a deletion of *TNF*, *RPA3* and *PABPC5*; and focal amplifications of *ASUN*, *ITPR2*, *POLD2*, *BCAP31* and the 5' ends of isoforms a and e and exons 5 and 6 of *SOX5*. These CNAs identified in GBM and OV genomic profiles act as a new link between the tumor's genome and a patient's prognosis, offering insights into the cancer's formation and growth and suggest promising drug targets.

1.2 Existing Methods

The existing standards for presenting and exchanging microarray data involve tabulation of the biological data in rows and columns where, for example, the rows denote the m genes and the columns denote the n arrays [5], thus taking the structure of a matrix of dimensions $m \times n$. Microarray data are high-dimensional and noisy. Therefore, analysis of such data requires mathematical tools that are adaptable to the large quantities of data, while reducing the complexity of the data to make them comprehensible. Some of the commonly used methods for this purpose are clustering methods [6–8] and matrix decomposition methods such as singular value decomposition (SVD) [9–11], nonnegative matrix factorization [12] and partial least squares [13].

1.2.1 Principal Component Analysis and Singular Value Decomposition

Principal Component Analysis (PCA) [14] is a method of statistical analysis for high-dimensional data. PCA is used to reduce the dimensionality of a dataset by decomposing it in a set of successive orthogonal components in such a way that the first principal component has the largest variance (i.e, accounts for the maximum variabiility in the data), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components [15]. PCA can equivalently be formulated as an eigenvalue problem. In particular, the principal components are essentially the same as the eigenvectors of the covariance matrix [16]. SVD is a method of matrix decomposition, which is commonly used to compute the principal components by solving this eigenvalue problem. SVD decomposes the matrix, D, such that,

$$D = U\Sigma V^T \tag{1.1}$$

where the D is the $m \times n$ data matrix; U is an $m \times n$ matrix with orthonormal columns that are called left singular vectors; Σ is a diagonal matrix containing the singular values

and V^T is an $n \times n$ matrix with orthonormal rows also known as right singular vectors. From Equation (1.1), it can be shown that,

$$D^T D = V \Sigma^2 V^T$$

Therefore, when the columns of D are treated as n samples of the m-dimensional data tabulated in matrix D, and D is column-centered, the right singular vectors are the same as the principal components of its covariance matrix, $D^T D$. The right singular vectors are also said to span the column space of D. If instead, the rows are treated as a m samples of the n-dimensional data represented by the matrix D, and the matrix is row centered, the left singular vectors are the principal components of the covariance matrix DD^T. From Equation (1.1), the left singular vectors are given by,

$$DD^T = U \Sigma^2 U^T$$

The left singular vectors thus span the row space of D.

Although the principal components are the same as the left and right singular vectors when the datasets are appropriately preprocessed, there are a few differences between PCA and SVD. PCA requires a specific standardization of the data matrix, i.e., centering of the columns (or the rows) whereas SVD has no such limitation. PCA usually reduces and classifies the data based upon the two or three patterns (components) of greatest weights, while the SVD retains the full data and the full set of patterns. The two methods also differ in their applications. PCA is commonly used for dimensionality reduction, whereas in our lab, SVD and its generalizations are used primarily to find patterns in the data. Since the most significant patterns or components often represent experimental rather than biological variation, PCA often leads to classification based upon, e.g., date of hybridization or other batch effects rather than the true biological difference, e.g., tumor subtype.

In the context of DNA microarray gene expression analysis, Alter et al. [11] showed that SVD is a linear transformation of the expression data D from the genes $\times$ arrays space to the reduced "eigengenes" $\times$ "eigenarrays" space where the eigengenes (or eigenarrays) are unique orthonormal superpositions of the genes (or arrays). The "eigengenes" obtained in this analysis can be interpreted as the principal components in the traditional PCA of the covariance matrix DD^T whereas the "eigenarrays" can be viewed as the principal components of the covariance matrix $D^T D$ if D is row and column centered, respectively. The "eigengenes" uncovered by SVD correlate with independent processes, biological or experimental, such as observed genome-wide effects of known regulators or transcription factors. The corresponding "eigenarrays" correlate with the corresponding cellular states,

such as measured samples in which these regulators or transcription factors are overactive or underactive [10].

SVD has a variety of applications in mathematics, biology and medicine. SVD can be used to filter out noise or remove experimental artifacts from the data to enable meaningful comparison of the expression of different genes across different arrays in different experiments [11]. Bertagnolli et al. [9] used SVD to identify the length distribution functions of sets and subsets of eukaryotic mRNA transcripts from DNA microarray data and reveal global relations among transcript length, cellular metabolism and tumor development. The global relations suggest a previously unrecognized physical mode for tumor and normal cells to differentially regulate metabolism in a transcript length-dependent manner. The identified distribution functions support a previous hypothesis from mathematical modeling of evolutionary forces that act upon transcript length in the manner of the restoring force of the harmonic oscillator. Alter et al. [17] showed that SVD of yeast genome-scale mRNA lengths distribution reveals asymmetry in RNA gel electrophoresis band broadening. SVD was used to uncover a global correlation, and predict causal coordination between eukaryotic DNA replication and mRNA transcription during the cell cycle in yeast [18, 19].

SVD provides a useful mathematical framework for processing and modeling genome-wide expression data, in which both the mathematical variables and operations may be assigned biological meaning [10, 17, 20].

1.2.2 Limitations

It is common in biology to compare two matrices of dimensions $m_1 \times n$ and $m_2 \times n$ simultaneously. For example, it is often necessary to compare the messenger RNA expression across arrays in two organisms at the same time points or to compare tumor and normal genomic profiles of the same set of patients. In both cases, the comparisons are row-independent, i.e., different organisms have a different number of genes and the tumor and normal genomic profiles measure different regions of DNA. Also, there is a one-to-one correspondence between the columns, i.e., the same time points or the same set of patients. One major limitation in all the above methods is that when comparing two or more matrices, these methods unfold or flatten the individual matrices to form a single matrix.

When comparing two (or more) matrices, the structure of the datasets is of an order higher than that of a single matrix. Unfolded into a single matrix, some of the degrees of freedom are lost and much of the information in the datasets might also be lost. Additionally, these methods are not capable of directly addressing the fundamental question of what is *similar* and *dissimilar* between the datasets of interest in a single comparison. For example,

PCA, the most commonly used method, cannot *simultaneously* compare two datasets of tumor and normal genomic profiles and identify the common biological phenomena shared between tumor and normal tissues (such as the X chromosome amplification in females that is evident in both tumor and normal tissues) and biological changes specific to tumor tissues alone. When the data are in the form of a flattened matrix, the biologically common and tumor-specific information gets mixed up, resulting in added biological "noise."

1.2.3 Generalized Singular Value Decomposition (GSVD)

The only two decompositions to date that can simultaneously compare and contrast two or more row-independent and column-matched matrices *preserving all the information* are the GSVD [21, 22] and higher-order GSVD (HO GSVD) [23]. The decompositions discussed in this dissertation (GSVD, HO GSVD and tensor GSVD) are exact, meaning no information is lost due to the process of decomposition itself.

Alter et al. [22, 24] showed that GSVD (Figure 1.2) provides a comparative mathematical framework for genome-scale expression datasets from two organisms tabulated as two matrices with different row dimensions but the same column dimension where the mathematical variables and operations represent the underlying biological reality. They also showed that mathematical similarity and dissimilarity between the two matrices correspond to the biological similarity and dissimilarity. This study inspired a new method to compute the GSVD [25] and also paved the way for several new applications of the GSVD in *biology* [26–29]. It also led to integrating large-scale biomedical data using pseudo-inverse projection [30] and tensor decompositions such as the higher-order SVD [19, 31–33]. The success of the GSVD is the main motivation behind this dissertation.

However, the GSVD framework is limited to the comparison of only two second-order tensors, i.e., matrices. In order to compare two higher-order tensors, there is a need for a new mathematical framework to be defined that enables the comparison of two row-independent but column matched higher-order (greater than two) tensors.

1.3 Dissertation Contributions

While the applications presented in this dissertation are in the field of biology, the methods and theory should be widely applicable to many fields, including finance, astronomy and medical imaging. For example, in astronomy, core temperature measurements of different stars in two galaxies measured in the same time period (Figure 1.1a) comprise a dataset of two matrices with matched columns (the time points). In finance, we have datasets recording stock prices in different markets over the same time period (Figure 1.1b).

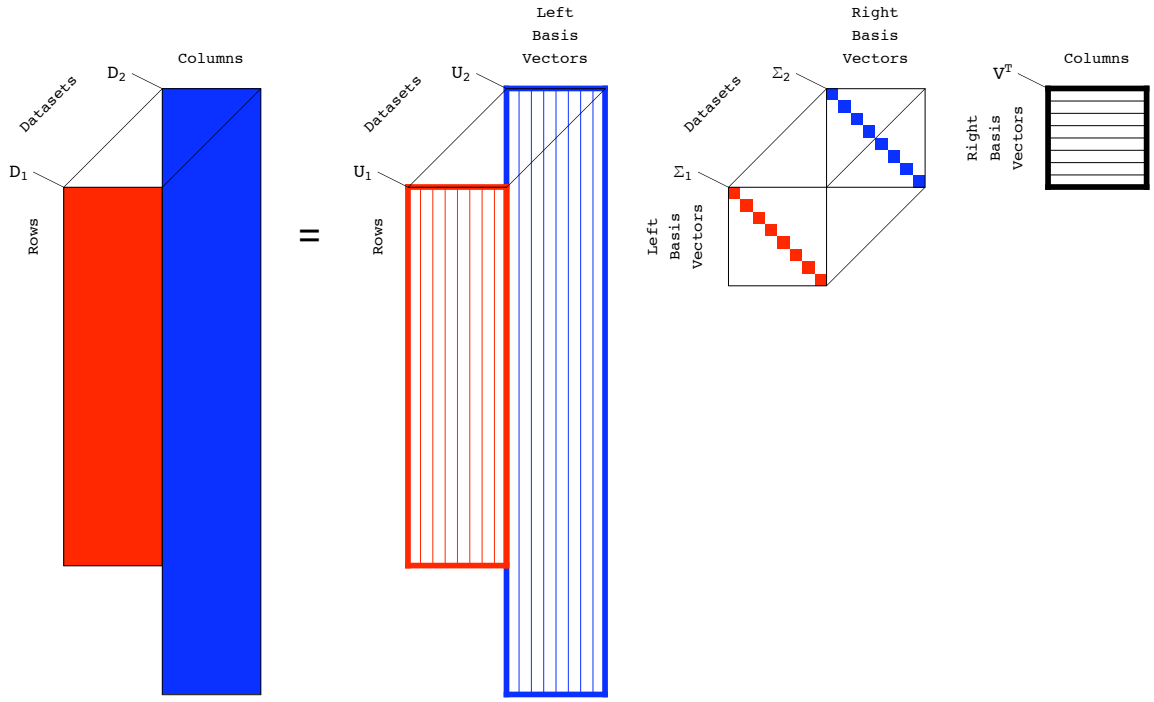


Figure 1.2: The GSVD of two matrices D_1 and D_2, that are row-independent but column-matched. The decomposition yields two sets of dataset-specific left basis vectors U_1 and U_2 that are orthonormal, one set of shared right basis vectors V^T and two sets of singular values in the form of diagonal matrices $\sum_1$ and $\sum_2$. Figure adapted from [23].

Comparative analyses of such datasets will simultaneously distinguish the similar from the dissimilar. The frameworks discussed in this dissertation can be used in understanding the common factors as well as the specific factors, for example, the common factors affecting the price of a stock worldwide and the factors specific to a particular stock market.

This dissertation describes the following contributions in tensor decompositions and genomic signal processing:

1. We show that the GSVD of patient-matched but probe-independent genomic profiles can be used to uncover a previously unknown global pattern of DNA aberrations that is correlated with, and possibly causally related to, brain cancer survival.

2. We also discover several previously unreported focal CNAs and most known GBM associated changes. Yet, we find that GBM survival phenotype is an outcome of its global genotype.

3. We define a novel mathematical framework called tensor GSVD (tGSVD) and prove that it extends the matrix GSVD. This simultaneous decomposition is by definition exact. We also mathematically derive several important properties of this decomposition.

4. We illustrate the tGSVD with comparisons of patient- and platform-matched but probe-independent genomic profiles of ovarian serous cystadenocarcinoma (OV) tumor and normal DNA copy-number profiles measured on two platforms. The tGSVD reveals chromosome arm-wide patterns of tumor-exclusive and platform-consistent CNAs, across 6p and 12p taken together, 7p and Xq, that are correlated with, and possibly causally related to, OV patients' survival.

5. We perform additional analyses using mRNA, microRNA and protein expression data between the tGSVD classes and find that these differential expressions consistently map to the DNA CNAs.

1.4 Overview of Chapters

The remainder of this dissertation is arranged in the following fashion:

Chapter 2 illustrates the GSVD with a comparison of genomic profiles from tumor and normal samples of the same set of GBM patients from TCGA and discusses the known and novel co-occurring CNAs predicting patient prognosis and potential drug targets.

Chapter 3 defines the tGSVD framework for the comparison of two column-matched but row-independent higher-order (greater than two) tensors and investigates its mathematical

properties and specific cases. It demonstrates the tGSVD in comparisons of patient- and platform-matched but probe-independent genomic profiles of OV tumor and normal samples to discover tumor-specific CNAs that correlate with patients' survival. It also shows that the respective differential microRNA, mRNA and protein expressions between the tGSVD classes are consistent with the DNA CNAs revealed by tGSVD.

Chapter 4 concludes with a discussion of contributions of this dissertation and possible future research on open questions.

CHAPTER 2

GSVD COMPARISON OF HUMAN BRAIN TUMOR AND NORMAL GENOMIC PROFILES

2.1 Glioblastoma Multiforme (GBM)

Glioblastoma multiforme (GBM), the most common brain tumor in adults, is characterized by poor prognosis [34]. GBM tumors exhibit a range of copy-number alterations (CNAs), many of which play roles in the cancer's pathogenesis [35–37]. Recent large-scale gene expression [38–40] and DNA methylation [41] profiling efforts identified GBM molecular subtypes, distinguished by small numbers of biomarkers. However, despite these efforts, GBM's best prognostic predictor remains the patient's age at diagnosis [42, 43].

To identify CNAs that might predict GBM patients' survival, we comparatively model patient-matched GBM and normal array CGH (aCGH) profiles from The Cancer Genome Atlas (TCGA) by using the generalized singular value decomposition (GSVD) [25].

We also find that, in probe-independent comparison of aCGH data from patient-matched tumor and normal samples, the mathematical variables of the GSVD, i.e., shared tumor and normal patterns of copy-number variation across the patients and the corresponding tumor- and normal-specific patterns of copy-number variation across the tumor and normal probes, represent experimental or biological reality. Patterns that are mathematically significant in both datasets represent copy-number variations (CNVs) in the normal human genome that are conserved in the tumor genome (e.g., female-specific X chromosome amplification). Patterns that are mathematically significant in the normal but not the tumor dataset represent experimental variations that exclusively affect the normal dataset. Similarly, some patterns that are mathematically significant in the tumor but not in the normal dataset represent experimental variations that exclusively affect the tumor dataset.

One pattern that is mathematically significant in the tumor, but not in the normal dataset, represents tumor-exclusive co-occurring CNAs, including most known GBM-associated changes in chromosome numbers and focal CNAs, as well as several previously

unreported CNAs in >3% of the patients [44]. This pattern is correlated, possibly coordinated with GBM patients' survival and response to therapy. We find that the pattern provides a prognostic predictor that is better than the chromosome numbers or any one focal CNA that it identifies, suggesting that the GBM survival phenotype is an outcome of its global genotype. The pattern is independent of age, and combined with age, makes a better predictor than age alone.

We confirm our results with GSVD comparison of matched profiles of a larger set of TCGA patients, inclusive of the initial set. We validate the prognostic contribution of the pattern with GSVD classification of the GBM profiles of a set of patients that is independent of both the initial set and the inclusive confirmation set [45].

2.1.1 Tumor and Normal Datasets

To compare TCGA patient-matched GBM and normal (mostly blood) aCGH profiles (Dataset S1 and Mathematica Notebooks S1 and S2), Agilent Human aCGH 244A-measured 365 tumor and 360 normal profiles were selected, corresponding to the same N=251 patients. Each profile lists $\log_2$ of the TCGA level 1 background-subtracted intensity in the sample relative to the Promega DNA reference, with signal to background >2.5 for both the sample and reference in more than 90% of the 223,603 autosomal probes on the microarray. The profiles are organized in one tumor and one normal dataset, of M_1=212,696 and M_2=211,227 autosomal and X chromosome probes, each probe with valid data in at least 99% of either the tumor or normal arrays, respectively. Each profile is centered at its autosomal median copy number. The <0.2% missing data entries in the tumor and normal datasets are estimated by using singular value decomposition (SVD) as described [22, 46]. Within each set, the medians of profiles of samples from the same patient are taken.

2.2 Generalized Singular Value Decomposition (GSVD)

Previously, we formulated the GSVD as a framework for comparatively modeling two composite datasets [22] (see also [23]), and illustrated its application in sequence-independent comparison of DNA microarray data from two organisms, where, as we showed, the mathematical variables and operations of the GSVD represent experimental or biological reality. The variables, subspaces of significant patterns that are uncovered in the simultaneous decomposition of the two datasets and are mathematically significant in either both (i.e., common to both) datasets or only one (i.e., exclusive to one) of the datasets, correlate with cellular programs that are either conserved in both or unique to only one of the organisms, respectively. The operation of reconstruction in the subspaces that are mathematically

common to both datasets outlines the biological similarity in the regulation of the cellular programs that are conserved across the species. Reconstruction in the common and exclusive subspaces of either dataset outlines the differential regulation of the conserved relative to the unique programs in the corresponding organism.

2.2.1 Construction of the GSVD

The structure of the patient-matched but probe-independent tumor and normal datasets D_1 and D_2, of N patients, i.e., N-arrays $\times$ M_1-tumor and M_2-normal probes, is of an order higher than that of a single matrix. The patients, the tumor and normal probes as well as the tissue types, each represent a degree of freedom. Unfolded into a single matrix, some of the degrees of freedom are lost and much of the information in the datasets might also be lost.

To compare the tumor and normal datasets, therefore, we use the GSVD, formulated to simultaneously separate the paired datasets into paired weighted sums of N outer products of two patterns each: One pattern of copy-number variation across the patients, i.e., a "probelet" v_n^T, which is identical for both the tumor and normal datasets, combined with either the corresponding tumor-specific pattern of copy-number variation across the tumor probes, i.e., the "tumor arraylet" $u_{1,n}$, or the corresponding normal-specific pattern across the normal probes, i.e., the "normal arraylet" $u_{2,n}$ (Figure 2.1),

$$\begin{aligned} D_1 &= U_1 \Sigma_1 V^T = \sum_{n=1}^{N} \sigma_{1,n} u_{1,n} \otimes v_n^T, \\ D_2 &= U_2 \Sigma_2 V^T = \sum_{n=1}^{N} \sigma_{2,n} u_{2,n} \otimes v_n^T. \end{aligned} \tag{2.1}$$

The probelets are, in general, non-orthonormal, but are normalized, such that $v_n^T v_n = 1$. The tumor and normal arraylets are orthonormal, such that $U_1^T U_1 = U_2^T U_2 = I$.

The significance of the probelet v_n^T in either the tumor or normal dataset, in terms of the overall information that it captures in this dataset, is proportional to either of the weights $\sigma_{1,n}$ or $\sigma_{2,n}$, respectively (Figure 2.2),

$$\begin{aligned} p_{1,n} &= \sigma_{1,n}^2 / \sum_{n=1}^{N} \sigma_{1,n}^2, \\ p_{2,n} &= \sigma_{2,n}^2 / \sum_{n=1}^{N} \sigma_{2,n}^2. \end{aligned} \tag{2.2}$$

The "generalized normalized Shannon entropy" of each dataset,

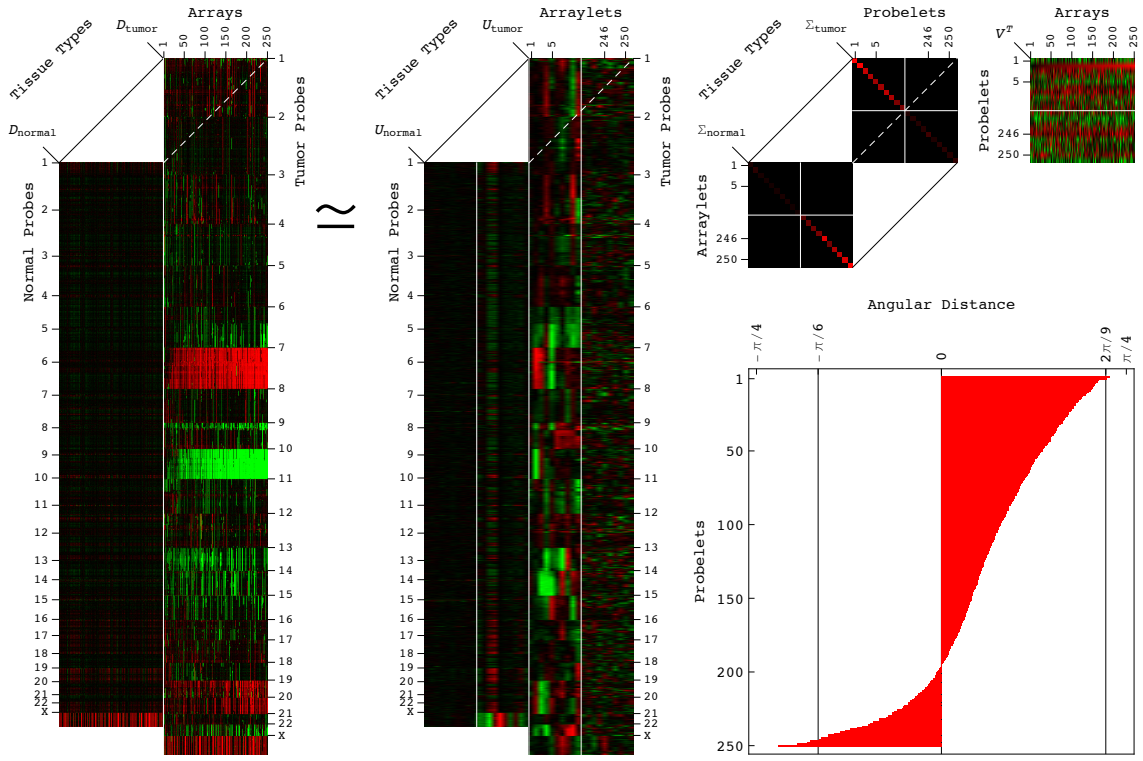


Figure 2.1: Generalized singular value decomposition (GSVD) of the TCGA patient-matched tumor and normal aCGH profiles.
The GSVD simultaneously separates the paired tumor and normal datasets into paired weighted sums of N outer products of two patterns each: One pattern of copy-number variation across the patients, i.e., a "probelet" v_n^T, which is identical for both the tumor and normal datasets, combined with either the corresponding tumor-specific pattern of copy-number variation across the tumor probes, i.e., the "tumor arraylet" $u_{1,n}$, or the corresponding normal-specific pattern across the normal probes, i.e., the "normal arraylet" $u_{2,n}$ (Equation (2.1)). This is depicted in a raster display, with relative copy-number gain (red), no change (black) and loss (green), explicitly showing only the first though the 10th and the 242nd through the 251st probelets and corresponding tumor and normal arraylets, which capture ∼52% and 71% of the information in the tumor and normal dataset, respectively. The significance of the probelet v_n^T in the tumor dataset relative to its significance in the normal dataset is defined in terms of an "angular distance" that is proportional to the ratio of these weights (Equation (2.4)). This is depicted in a bar chart display, showing that the first and second probelets are almost exclusive to the tumor dataset with angular distances $>2\pi/9$, the 247th to 251st probelets are approximately exclusive to the normal dataset with angular distances $\lesssim -\pi/6$, and the 246th probelet is relatively common to the normal and tumor datasets with an angular distance $> -\pi/6$.

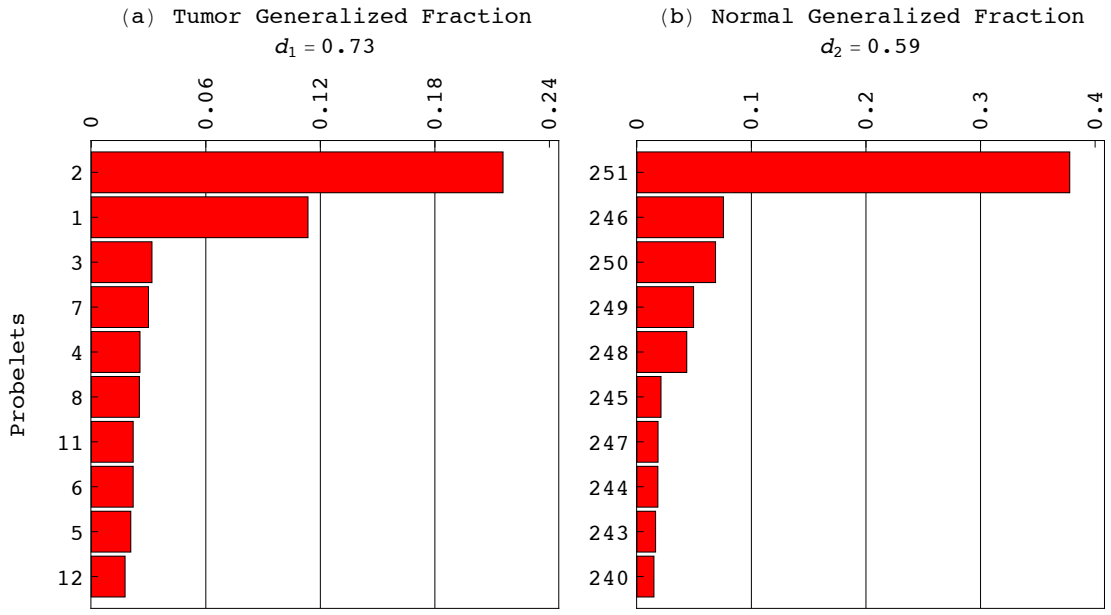


Figure 2.2: Most significant probelets in the tumor and normal datasets.
(a) Bar chart of the ten most significant probelets in the tumor dataset in terms of the generalized fraction that each probelet captures in this dataset (Equation (2.2)), showing that the two most tumor-exclusive probelets, i.e., the first probelet and the second probelet (Figure 2.3a–c), with angular distances $>2\pi/9$, are also the two most significant probelets in the tumor dataset, with $\sim$11% and 22% of the information in this dataset, respectively. The "generalized normalized Shannon entropy" (Equation (2.3)) of the tumor dataset is d_1=0.73. (b) Bar chart of the generalized fractions of the ten most significant probelets in the normal dataset, showing that the five most normal-exclusive probelets, the 247th to 251st probelets, with angular distances $\lesssim -\pi/6$, are among the seven most significant probelets in the normal dataset, capturing together $\sim$56% of the information in this dataset. The 246th probelet (Figure 2.1d–f), which is relatively common to the normal and tumor datasets with an angular distance $>-\pi/6$, is the second most significant probelet in the normal dataset with $\sim$8% of the information. The generalized entropy of the normal dataset, d_2=0.59, is smaller than that of the tumor dataset. This means that the normal dataset is more redundant and less complex than the tumor dataset.

$$\begin{aligned} 0 \leq d_1 &= (\log N)^{-1} \sum_{n=1}^{N} p_{1,n} \log p_{1,n} \leq 1, \\ 0 \leq d_2 &= (\log N)^{-1} \sum_{n=1}^{N} p_{2,n} \log p_{2,n} \leq 1, \end{aligned} \tag{2.3}$$

measures the complexity of the data from the distribution of the overall information among the different probelets and corresponding arraylets. An entropy of zero corresponds to an ordered and redundant dataset in which all the information is captured by a single probelet and its corresponding arraylet. An entropy of one corresponds to a disordered and random dataset in which all probelets and arraylets are of equal significance.

2.2.2 Mathematical Exclusivity and Significance

The significance of the probelet v_n^T in the tumor dataset relative to its significance in the normal dataset is defined in terms of an "angular distance" θ_n that is proportional to the ratio of these weights,

$$-\pi/4 \leq \theta_n = \arctan(\sigma_{1,n}/\sigma_{2,n}) - \pi/4 \leq \pi/4. \tag{2.4}$$

An angular distance of $\pm\pi/4$ indicates a probelet that is exclusive to either the tumor or normal dataset, respectively, whereas an angular distance of zero indicates a probelet that is common to both the tumor and normal datasets. The probelets are arranged in decreasing order of their angular distances, i.e., their significance in the tumor dataset relative to the normal dataset.

We find that the two most tumor-exclusive mathematical patterns of copy-number variation across the patients, i.e., the first probelet and the second probelet (Figure 2.3 *a*–*c*), with angular distances $> 2\pi/9$, are also the two most significant probelets in the tumor dataset, with ∼11% and 22% of the information in this dataset, respectively. Similarly, the five most normal-exclusive probelets, the 247th to 251st probelets, with angular distances $\lesssim -\pi/6$, are among the seven most significant probelets in the normal dataset, capturing together ∼56% of the information in this dataset. The 246th probelet (Figure 2.3 *d*–*f*), which is the second most significant probelet in the normal dataset with ∼8% of the information, is relatively common to the normal and tumor datasets with an angular distance $> -\pi/6$.

2.2.3 Biological Interpretation of the Mathematical Patterns

To biologically or experimentally interpret these significant probelets, we correlate or anticorrelate each probelet with relative copy-number gain or loss across a group of patients according to the TCGA annotations of the group of n patients with largest or smallest relative copy numbers in this probelet among all N patients, respectively. The P-value of a given association is calculated assuming hypergeometric probability distribution of the K annotations among the N patients, and of the subset of $k \subseteq K$ annotations among the subset of n patients, as described [8], $P(k; n, N, K) = \binom{N}{n}^{-1} \sum_{i=k}^{n} \binom{K}{i} \binom{N-K}{n-i}$. We visualize the copy-number distribution between the annotations that are associated with largest or smallest relative copy numbers in each probelet by using boxplots, and by calculating the corresponding Mann-Whitney-Wilcoxon P-value (please refer to Appendix C.2 for more information). To interpret the corresponding tumor and normal arraylets, we map the tumor and normal probes onto the National Center for Biotechnology Information (NCBI)

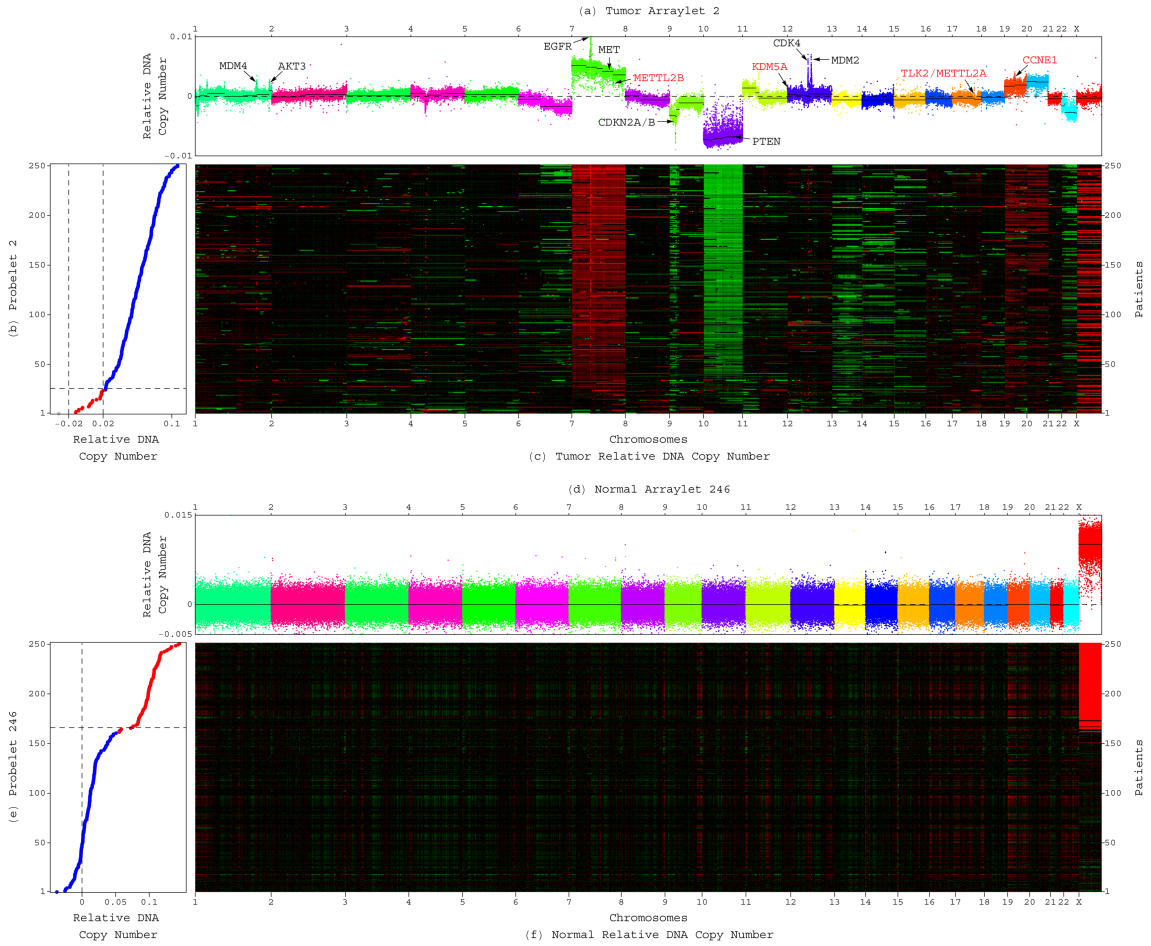


Figure 2.3: Significant probelets and corresponding tumor and normal arraylets uncovered by GSVD of the patient-matched GBM and normal aCGH profiles.
(a) Plot of the second tumor arraylet describes a global pattern of tumor-exclusive co-occurring CNAs across the tumor probes. The probes are ordered, and their copy numbers are colored, according to each probe's chromosomal location. Segments (black lines) identified by circular binary segmentation (CBS) include most known GBM-associated focal CNAs (black) and previously unrecognized CNAs (red). (b) Plot of the second most tumor-exclusive probelet, which is also the most significant probelet in the tumor dataset, describes the corresponding variation across the patients. The patients are ordered and classified according to each patient's relative copy number in this probelet. There are 227 patients (blue) with high (>0.02) and 23 patients (red) with low, approximately zero, numbers in the second probelet. One patient (gray) remains unclassified with a large negative (<-0.02) number. (c) Raster display of the tumor dataset, with relative gain (red), no change (black) and loss (green) of DNA copy numbers, shows the correspondence between the GBM profiles and the second probelet and tumor arraylet. (d) Plot of the 246th normal arraylet describes an X chromosome-exclusive amplification across the normal probes. (e) Plot of the 246th probelet, which is approximately common to both the normal and tumor datasets, and is the second most significant in the normal dataset, describes the corresponding copy-number amplification in the female (red) relative to the male (blue) patients. (f) Raster display of the normal dataset shows the correspondence between the normal profiles and the 246th probelet and normal arraylet.

human genome sequence build 36, by using the Agilent Technologies probe annotations posted at the University of California at Santa Cruz (UCSC) human genome browser [47,48]. We segment each arraylet and assign each segment a P-value by using the circular binary segmentation (CBS) algorithm as described [49, 50]. We find that the significant probelets and corresponding tumor and normal arraylets, as well as their interpretations, are robust to variations in the preprocessing of the data, e.g., in the data selection cutoffs.

2.3 Results: GSVD of GBM and Normal Genomic Profiles

2.3.1 GSVD Removes Batch Effects

We find that, first, the GSVD identifies significant experimental variations that exclusively affect either the tumor or the normal dataset, as well as CNVs that occur in the normal human genome and are common to both datasets, without a-priori knowledge of these variations (Table 2.1). The mathematically most tumor-exclusive probelet, i.e., the first probelet, correlates with tumor-exclusive experimental variation in the genomic center where the GBM samples were hybridized, with the P-values $< 10^{-5}$ (Table 2.1 and Figure 2.4a).

Similarly, the five most normal-exclusive probelets, i.e., the 247th to 251st probelets (Appendix A and Figure 2.6), correlate with experimental variations among the normal samples in genomic center, DNA microarray hybridization or scan date as well as the tissue batch and hybridization scanner, with P-values $< 10^{-3}$. Consistently, the corresponding arraylets, i.e., the first tumor arraylet (Figure 2.5a) and the 247th to 251st normal arraylets (Appendix A), describe copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths.

The 246th probelet (Figure 2.3e), which is mathematically approximately common to both the normal and tumor datasets, describes copy-number amplification in the female relative to the male patients that is biologically common to both the normal and tumor datasets. Consistently, both the 246th normal arraylet (Figure 2.3d) and 246th tumor arraylet describe an X chromosome-exclusive amplification. The P-values are $< 10^{-38}$ (Table 2.1 and Figure 2.7). To assign the patients gender, we calculate for each patient the standard deviation of the mean X chromosome number from the autosomal genomic mean in the patient's normal profile (Figure 2.3f).

Patients with X chromosome amplification greater than twice the standard deviation are assigned the female gender. For three of the patients, this copy-number gender assignment conflicts with the TCGA gender annotation. For three additional patients, the TCGA

Table 2.1: Enrichment of the significant probelets in TCGA annotations.

		Relative DNA Copy Number Gain					Relative DNA Copy Number Loss				
Probelet	Phenotype	Annotation	n	K	k	P-value	Annotation	n	K	k	P-value
1	Tumor Sample Center	HMS	183	34	34	8.5×10^{-6}	MSKCC	68	103	55	3.9×10^{-15}
246	Patient Gender	Female	86	86	84	8.0×10^{-62}	Male	165	165	163	8.0×10^{-62}
247	Normal Sample Scan Date	10.8.2009	51	6	6	5.5×10^{-5}	7.22.2009	38	11	10	2.0×10^{-8}
248	Normal Sample Batch/Scanner	HMS 8/2331	19	19	19	6.2×10^{-29}	–	–	–	–	–
249	Normal Sample Batch/Scanner	–	–	–	–	–	HMS 8/2331	22	19	19	9.6×10^{-26}
250	Normal Sample Scan Date	4.18.2007	26	9	9	3.3×10^{-10}	7.22.2009	25	11	9	1.1×10^{-8}
251	Normal Sample Center	HMS	139	46	46	2.8×10^{-14}	MSKCC	112	101	89	2.1×10^{-32}

Probabilistic significance of the enrichment of the n patients, with largest or smallest relative copy numbers in each significant probelet, in the respective TCGA annotations. The P-value of each enrichment is calculated assuming hypergeometric probability distribution of the K annotations among the N=251 patients of the initial set, and of the subset of $k \subseteq K$ annotations among the subset of n patients, as described [8], $P(k; n, N, K) = \binom{N}{n}^{-1} \sum_{i=k}^{n} \binom{K}{i} \binom{N-K}{n-i}$.

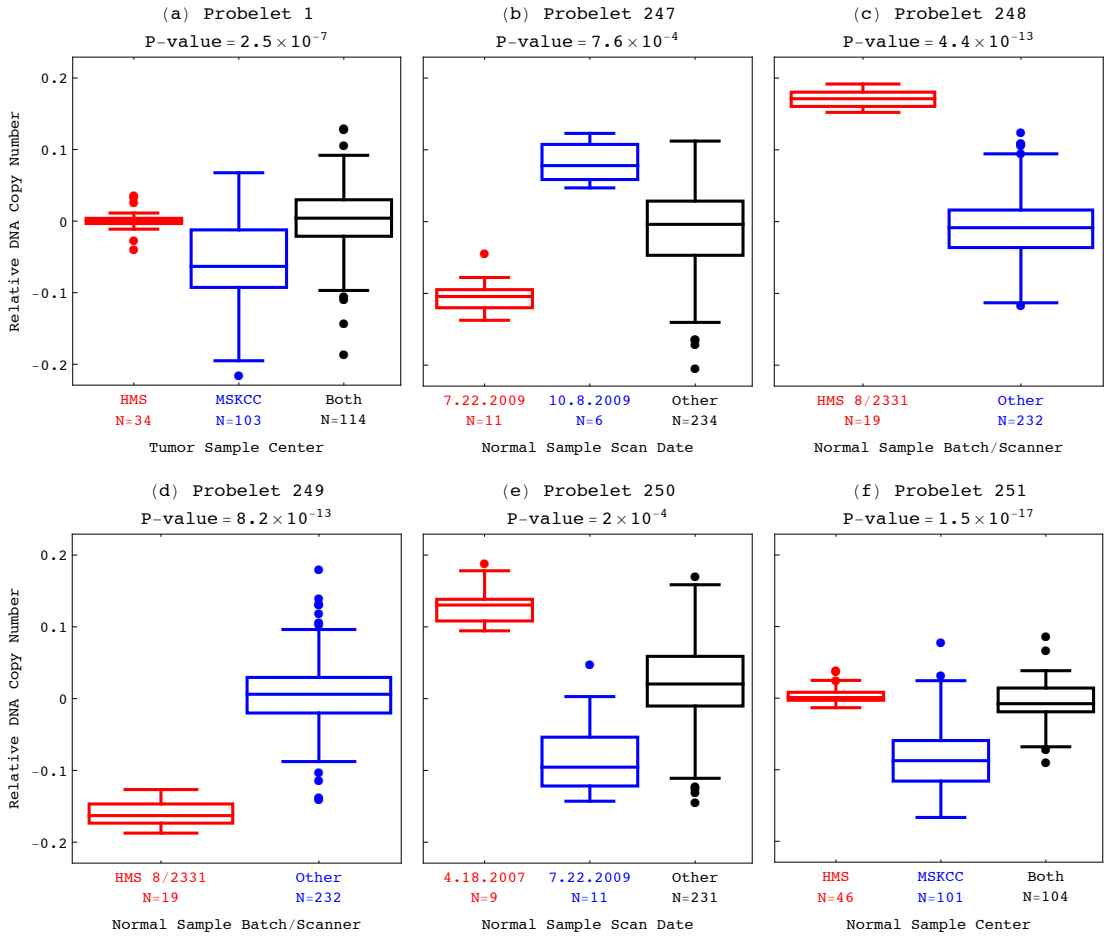


Figure 2.4: Differences in copy numbers among the TCGA annotations associated with the significant probelets.
Boxplot visualization of the distribution of copy numbers of the (*a*) first, most tumor-exclusive probelet among the associated genomic centers where the GBM samples were hybridized at (Table 2.1); (*b*) 247th, normal-exclusive probelet among the dates of hybridization of the normal samples; (*c*) 248th, normal-exclusive probelet between the associated tissue batches/hybridization scanners of the normal samples; (*d*) 249th, normal-exclusive probelet between the associated tissue batches/hybridization scanners of the normal samples; (*e*) 250th, normal-exclusive probelet among the dates of hybridization of the normal samples; (*f*) 251st, most normal-exclusive probelet among the associated genomic centers where the normal samples were hybridized. The Mann-Whitney-Wilcoxon *P*-values correspond to the two annotations that are associated with largest or smallest relative copy numbers in each probelet.

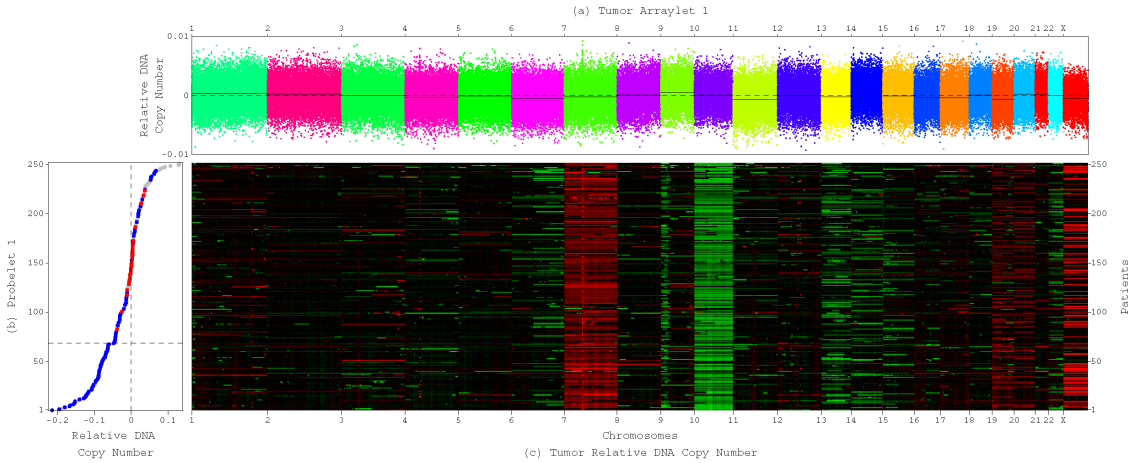


Figure 2.5: The first most tumor-exclusive probelet and corresponding tumor arraylet uncovered by GSVD of the patient-matched GBM and normal aCGH profiles.
(a) Plot of the first tumor arraylet describes unsegmented chromosomes (black lines), each with copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. The probes are ordered, and their copy numbers are colored, according to each probe's chromosomal location. (b) Plot of the first most tumor-exclusive probelet, which is also the second most significant probelet in the tumor dataset (Figure 2.2a), describes the corresponding variation across the patients. The patients are ordered according to each patient's relative copy number in this probelet. These copy numbers significantly correlate with the genomic center where the GBM samples were hybridized, HMS (red), MSKCC (blue) or multiple locations (gray), with the P-values $<10^{-5}$ (Table 2.1 and Figure 2.4a). (c) Raster display of the tumor dataset, with relative gain (red), no change (black) and loss (green) of DNA copy numbers, shows the correspondence between the GBM profiles and the first probelet and tumor arraylet.

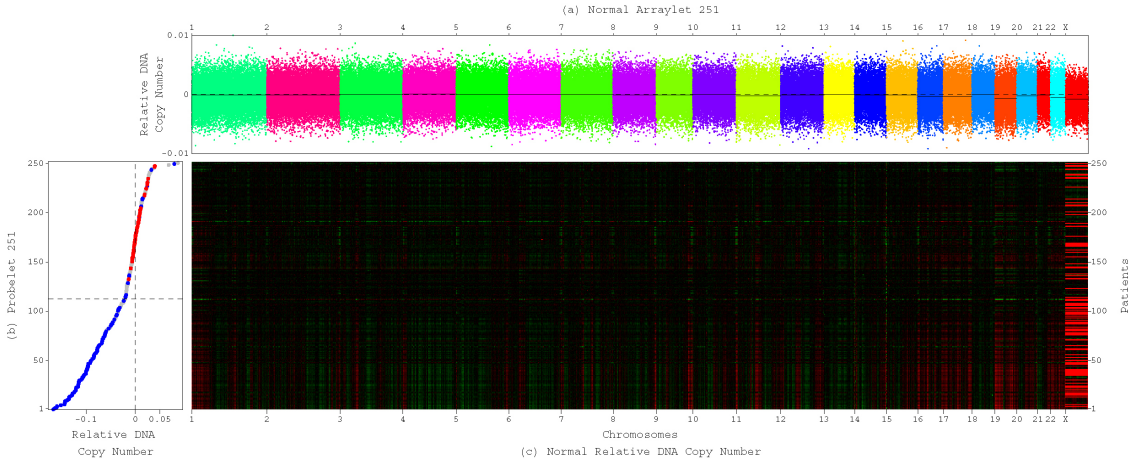


Figure 2.6: The first most normal-exclusive, i.e., 251st probelet and corresponding normal arraylet uncovered by GSVD.
(a) Plot of the 251st normal arraylet describes unsegmented [20,21] chromosomes (black lines), each with copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. (b) Plot of the first most normal-exclusive probelet, which is also the most significant probelet in the normal dataset (Figure 2.2b), describes the corresponding variation across the patients. Copy numbers in this probelet significantly correlate with the genomic center where the normal samples were hybridized, HMS (red), MSKCC (blue) or multiple locations (gray), with the P-values $<10^{-13}$ (Table 2.1 and Figure 2.4f). (c) Raster display of the normal dataset shows the correspondence between the normal profiles and the 251st probelet and normal arraylet.

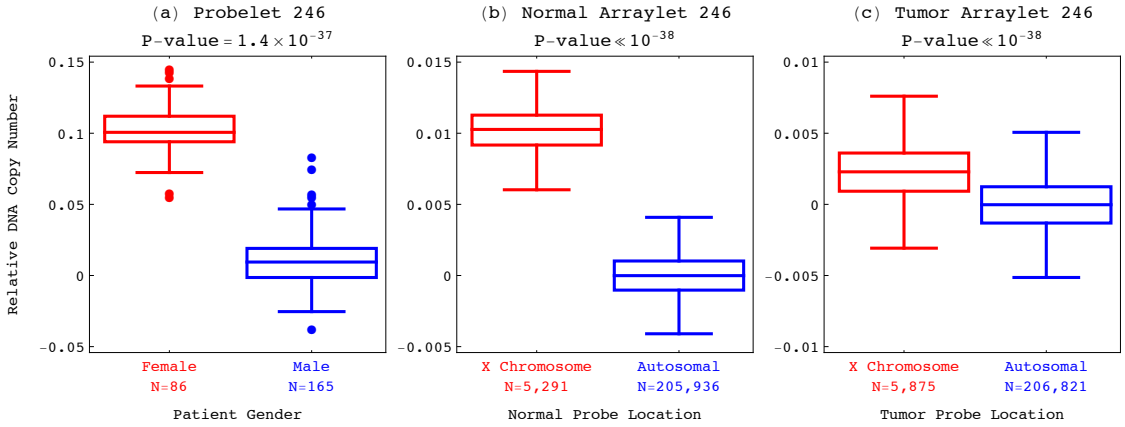


Figure 2.7: Copy-number distributions of the 246th probelet and the corresponding 246th normal arraylet and 246th tumor arraylet.
Boxplot visualization and Mann-Whitney-Wilcoxon P-values of the distribution of copy numbers of the (a) 246th probelet, which is approximately common to both the normal and tumor datasets, and is the second most significant in the normal dataset (Figure 2.2b), between the gender annotations (Table 2.1); (b) 246th normal arraylet between the autosomal and X chromosome normal probes; (c) 246th tumor arraylet between the autosomal and X chromosome tumor probes.

gender annotation is missing. In all these cases, the classification of the patients by the 246th probelet agrees with the copy-number assignment.

2.3.2 Discover Copy Number Changes Associated with GBM

Second, we find that the GSVD identifies a global pattern of tumor-exclusive co-occurring CNAs that includes most known GBM-associated changes in chromosome numbers and focal CNAs. This global pattern is described by the second tumor arraylet (Figure 2.3a and Dataset S3). The second most tumor-exclusive probelet (Figure 2.3b), which describes the corresponding copy-number variation across the patients, is the most significant probelet in the tumor dataset. Dominant in the global pattern, and frequently observed in GBM samples [35], is a co-occurrence of a gain of chromosome 7 and losses of chromosome 10 and the short arm of chromosome 9 (9p). To assign a chromosome gain or loss, we calculate for each tumor profile the standard deviation of the mean chromosome number from the autosomal genomic mean, excluding the outlying chromosomes 7, 9p and 10. The gain of chromosome 7 and the losses of chromosomes 10 and 9p are greater than twice the standard deviation in the global pattern as well as the tumor profiles of ∼20%, 41% and 12% of the patients, respectively.

Focal CNAs that are known to play roles in the origination and development of GBM and are described by the global pattern include amplifications of segments containing the

genes *MDM4* (1q32.1), *AKT3* (1q44), *EGFR* (7p11.2), *MET* (7q31.2), *CDK4* (12q14.1) and *MDM2* (12q15), and deletions of segments containing the genes *CDKN2A/B* (9p21.3) and *PTEN* (10q23.31), that occur in >3% of the patients. To assign a CNA in a segment, we calculate for each tumor profile the mean segment copy number. Profiles with segment amplification or deletion greater than twice the standard deviation from the autosomal genomic mean, excluding the outlying chromosomes 7, 9p and 10, or greater than one standard deviation from the chromosomal mean, when this deviation is consistent with the deviation from the genomic mean, are assigned a segment gain or loss, respectively. The frequencies of amplification or deletion we observe for these segments are similar to the reported frequencies of the corresponding focal CNAs [37].

Novel CNAs, previously unrecognized in GBM, are also revealed by the global pattern [44]. These include an amplification of a segment that contains *TLK2* (17q23.2) in ∼22% of the patients, with the corresponding CBS P-value$< 10^{-140}$. Copy-number amplification of *TLK2* has been correlated with overexpression in several other cancers [51, 52]. The human gene *TLK2*, with homologs in the plant *Arabidopsis thaliana* but not in the yeast *Saccharomyces cerevisiae*, encodes for a multicellular organisms-specific serine/threonine protein kinase, a biochemically putative drug target [53], which activity is directly dependent on ongoing DNA replication [54]. On the same segment with *TLK2*, we also find the gene *METTL2A*. Another amplified segment (CBS P-value$< 10^{-13}$) contains the homologous gene *METTL2B* (7q32.1). Overexpression of *METTL2A/B* was linked with prostate cancer metastasis [55], cAMP response element-binding (CREB) regulation in myeloid leukemia [56] and breast cancer patients' response to chemotherapy [57].

An amplification of a segment (CBS P-value$< 10^{-145}$) encompassing the cyclin E1-encoding *CCNE1* (19q12) is revealed in ∼4% of the patients. Cyclin E1 regulates entry into the DNA synthesis phase of the cell division cycle. Copy number increases of *CCNE1* have been linked with multiple cancers [58, 59], but not GBM. Amplicon-dependent expression of *CCNE1*, together with the genes *POP4*, *PLEKHF1*, *C19orf12* and *C19orf2* that flank *CCNE1* on this segment, was linked with primary treatment failure in ovarian cancer, possibly due to rapid repopulation of the tumor after chemotherapy [60].

Another rare amplification in ∼4% of the patients, of a segment (CBS P-value$< 10^{-28}$) that overlaps with the 5' end of *KDM5A* (12p13.33), is also revealed. The protein encoded by *KDM5A*, a retinoblastoma tumor suppressor (Rb)-binding lysine-specific histone demethylase [61], has been recently implicated in cancer drug tolerance [62]. The same amplified segment includes the solute carrier (SLC) sodium-neurotransmitter symporters

SLC6A12/13, biochemically putative carriers of drugs that might overcome the blood-brain barrier [63]. On the same segment, we also find *IQSEC3*, a mature neuron-specific guanine nucleotide exchange factor (GEF) for the ADP-ribosylation factor (ARF) *ARF1*, a key regulator of intracellular membrane traffic [64].

Note that although the tumor samples exhibit female-specific X chromosome amplification (Figure 2.3*c*), the second tumor arraylet exhibits an unsegmented X chromosome copy-number distribution, that is approximately centered at zero with a relatively small width. This illustrates the mathematical separation of the global pattern of tumor-exclusive co-occurring CNAs, that is described by the second tumor arraylet, from all other biological and experimental variations that compose either the tumor or the normal dataset, such as the gender variation that is common to both datasets, and is described by the 246th probelet and the corresponding 246th tumor and 246th normal arraylets.

2.3.3 Patient Prognosis and Drug Target Prediction

Third, we find that the GSVD classifies the patients into two groups of significantly different prognoses. The classification is according to the copy numbers listed in the second probelet, which correspond to the weights of the second tumor arraylet in the GBM aCGH profiles of the patients. A group of 227 patients, 224 of which with TCGA annotations, displays high (>0.02) relative copy numbers in the second probelet, and a Kaplan-Meier (KM) [65] (please refer to Appendix C.4 for more information) median survival time of ∼13 months (Figure 2.8*a*). A group of 23 patients, i.e., ∼10% of the patients, displays low, approximately zero, relative copy numbers in the second probelet, and a KM median survival time of ∼29 months, which is more than twice longer than that of the previous group. The corresponding log-rank test (please refer to Appendix C.5 for more information) P-value is $< 10^{-3}$. The univariate Cox [66] proportional hazard ratio (please refer to Appendix C.6 for more information) is 2.3, with a P-value $< 10^{-2}$ (Table 2.2), meaning that high relative copy numbers in the second probelet confer more than twice the hazard of low numbers.

Note that the cutoff of ±0.02 was selected to enable classification of as many of the patients as possible. Only one of the 251 patients has a negative copy number in the second probelet <-0.02, and remains unclassified. This patient is also missing the TCGA annotations. Survival analysis of only the chemotherapy patients classified by GSVD gives similar results (Table 2.3 and Appendix A). The P-values are calculated without adjusting for multiple comparisons [67]. We observe, therefore, that a negligible weight of the global pattern in a patient's GBM aCGH profile is indicative of a significantly longer survival time, as well as an improved response to treatment among chemotherapy patients.

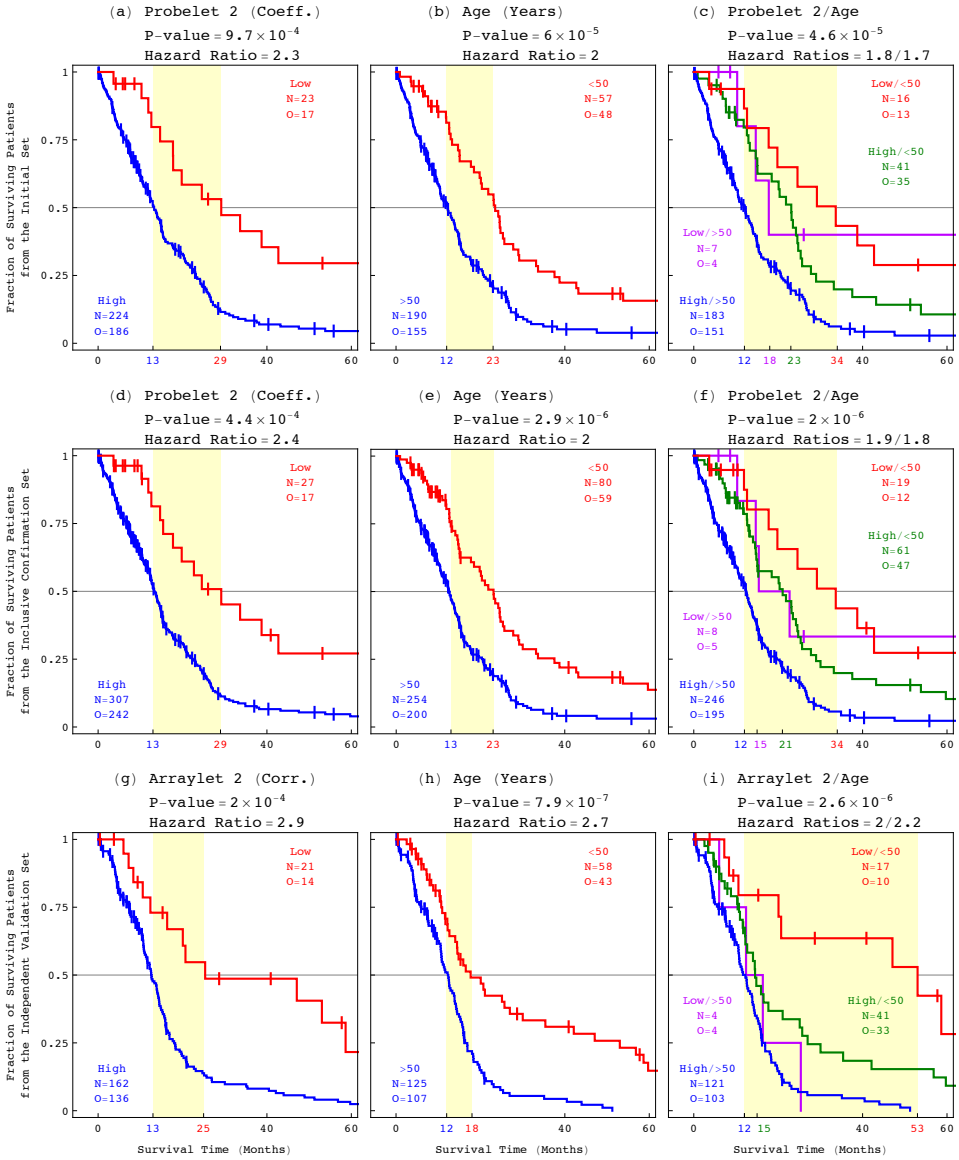


Figure 2.8: Survival analyses of the three sets of patients classified by GSVD, age at diagnosis or both.
(a) Kaplan-Meier (KM) curves for the 247 patients with TCGA annotations in the initial set of 251 patients, classified by copy numbers in the second probelet, which is computed by GSVD for the 251 patients. (b) Survival analyses of the 247 patients classified by age, i.e., >50 or <50 years old at diagnosis. (c) Survival analyses of the 247 patients classified by both GSVD and age. (d) Survival analyses of the 334 patients with TCGA annotations and a GSVD classification in the inclusive confirmation set of 344 patients, classified by copy numbers in the second probelet, which is computed by GSVD for the 344 patients. (e) Survival analyses of the 334 patients classified by age. (f) Survival analyses of the 334 patients classified by both GSVD and age. (g) Survival analyses of the 183 patients with a GSVD classification in the independent validation set of 184 patients, classified by correlations of each patient's GBM profile with the second tumor arraylet, which is computed by GSVD for the 251 patients. (h) Survival analyses of the 183 patients classified by age. (i) Survival analyses of the 183 patients classified by both GSVD and age.

Table 2.2: Cox proportional hazard models of the three sets of patients classified by GSVD, age at diagnosis or both.
In each set of patients, the multivariate Cox proportional hazard ratios [66] for GSVD and age are similar and do not differ significantly from the corresponding univariate hazard ratios. This means that GSVD and age are independent prognostic predictors.

Cox Proportional Hazard Model	Predictor	Initial Set		Inclusive Confirmation Set		Independent Validation Set	
		Hazard Ratio	*P*-value	Hazard Ratio	*P*-value	Hazard Ratio	*P*-value
Univariate	GSVD	2.3	1.3×10^{-3}	2.4	6.5×10^{-4}	2.9	3.6×10^{-4}
	Age	2.0	7.9×10^{-5}	2.0	4.3×10^{-6}	2.7	1.7×10^{-6}
Multivariate	GSVD	1.8	2.2×10^{-2}	1.9	1.2×10^{-2}	2.0	2.2×10^{-2}
	Age	1.7	2.0×10^{-3}	1.8	1.0×10^{-4}	2.2	2.0×10^{-4}

Table 2.3: Cox proportional hazard models of the three sets of patients classified by GSVD, chemotherapy or both.
In each set of patients, the multivariate Cox proportional hazard ratios for GSVD and chemotherapy are similar and do not differ significantly from the corresponding univariate hazard ratios. This means that GSVD and chemotherapy are independent prognostic predictors. The *P*-values are calculated without adjusting for multiple comparisons [67].

Cox Proportional Hazard Model	Predictor	Initial Set		Inclusive Confirmation Set		Independent Validation Set	
		Hazard Ratio	*P*-value	Hazard Ratio	*P*-value	Hazard Ratio	*P*-value
Univariate	GSVD	2.4	1.2×10^{-3}	2.4	6.4×10^{-4}	2.8	1.3×10^{-3}
	Chemotherapy	2.6	1.5×10^{-8}	2.7	6.3×10^{-11}	2.2	7.3×10^{-4}
Multivariate	GSVD	3.0	5.2×10^{-5}	3.1	2.5×10^{-5}	3.3	2.3×10^{-4}
	Chemotherapy	3.1	7.9×10^{-11}	3.2	1.9×10^{-13}	2.7	3.0×10^{-5}

A mutation in the gene *IDH1* was recently linked with improved GBM prognosis [34,39] and associated with a CpG island methylator phenotype [41]. We find, however, only seven patients (six chmeotherapy patients), i.e., <3%, with *IDH1* mutation. This is less than a third of the 23 patients in the long-term survival group defined by the global pattern. The corresponding survival analyses are, therefore, statistically insignificant (Appendix A).

Chromosome 10 loss, chromosome 7 gain and even loss of 9p, which are dominant in the global pattern, have been suggested as indicators of poorer GBM prognoses for over two decades [35, 36]. However, the KM survival curves for the groups of patients with either one of these chromosome number changes almost overlap the curves for the patients with no changes (Appendix A). The log-rank test *P*-values for all three classifications are $\gtrsim 10^{-1}$, with the median survival time differences $\lesssim$3 months. Similarly, in the KM survival analyses of the groups of patients with either a CNA or no CNA in either one of the 130 segments identified by the global pattern (Appendix A), log-rank test *P*-values $< 5 \times 10^{-2}$ are calculated for only 12 of the classifications. Of these, only six correspond to a KM

median survival time difference that is $\gtrsim$5 months, approximately a third of the $\sim$16 months difference observed for the GSVD classification.

One of these segments contains the genes *TLK2* and *METTL2A* and another segment contains the homologous gene *METTL2B* (Figure 2.9), previously unrecognized in GBM. The KM median survival times we calculate for the 56 patients with *TLK2/METTL2A* amplification and, separately, for the 19 patients with *METTL2B* amplifications are $\sim$5 and 8 months longer than that for the remaining patients in each case.

Similarly, the KM median survival times we calculate for the 43 chemotherapy patients with *TLK2/METTL2A* amplification and, separately, for the 15 chemotherapy patients with *METTL2B* amplification are both $\sim$7 months longer than that for the remaining chemotherapy patients in each case (Appendix A). This suggests that drug-targeting the kinase that *TLK2* encodes and/or the methyltransferase-like proteins that *METTL2A/B* encode may affect not only the pathogenesis but also the prognosis of GBM as well as the patient's response to chemotherapy.

Taken together, we find that the global pattern provides a better prognostic predictor than the chromosome numbers or any one focal CNA that it identifies. This suggests that the GBM survival phenotype is an outcome of its global genotype.

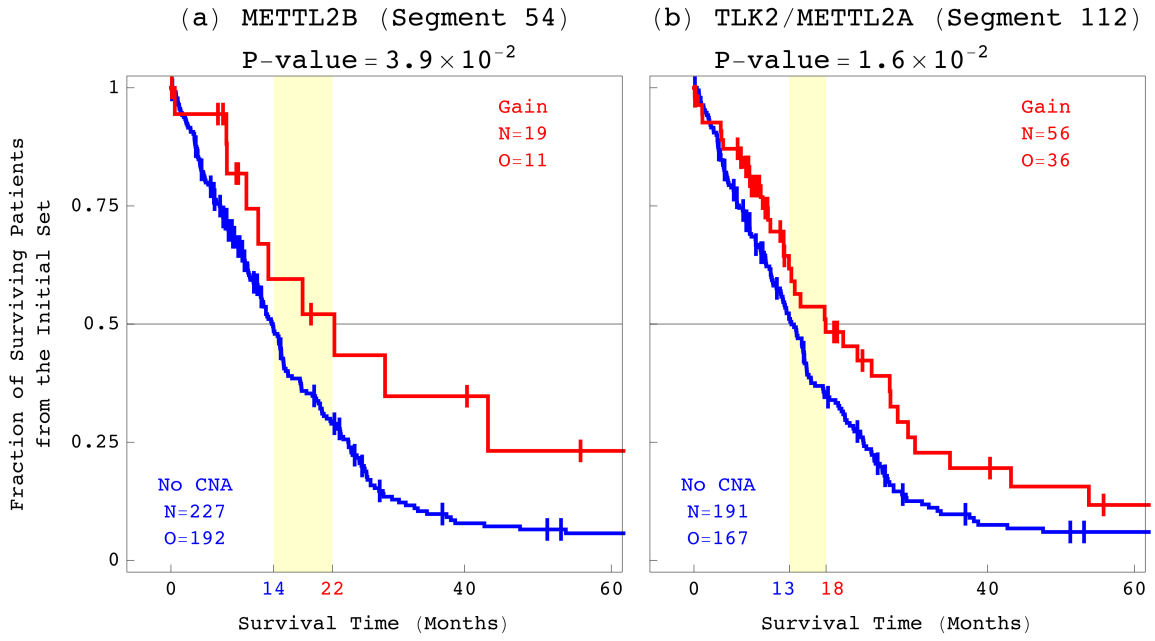


Figure 2.9: Kaplan-Meier (KM) survival analyses of the initial set of 251 patients classified by copy number changes in segments containing biochemically putative drug targets in GBM.

Despite the recent genome-scale molecular profiling efforts, age at diagnosis remains the best prognostic predictor for GBM in clinical use. The KM median survival time difference between the patients >50 or <50 years old at diagnosis is ∼11 months, approximately two thirds of the ∼16 months difference observed for the global pattern, with the log-rank test P-value $< 10^{-4}$ (Figure 2.8*b*). The univariate Cox proportional hazard ratio we calculate for age is 2, i.e., similar to that for the global pattern. Taken together, the prognostic contribution of the global pattern is comparable to that of age. Similarly, we find that the prognostic contribution of the global pattern is comparable to that of chemotherapy (Appendix A).

To examine whether the weight of the global pattern in a patient's GBM aCGH profile is correlated with the patient's age at diagnosis, we classify the patients into four groups, with prognosis of longer-term survival according to both, only one or neither of the classifications (Figure 2.8*c*). The KM curves for these four groups are significantly different, with the log-rank test P-value $< 10^{-4}$. Within each age group, the subgroup of patients with low relative copy numbers in the second probelet consistently exhibits longer survival than the remaining patients. The median survival time of the 16 patients <50 years old at diagnosis with low copy numbers in the second probelet is ∼34 months, almost three times longer than the ∼12 months median survival time of the patients >50 years old at diagnosis with high numbers in the second probelet. The multivariate Cox proportional hazard ratios for the global pattern and age are 1.8 and 1.7, respectively, with both corresponding P-values $< 3\times10^{-2}$. These ratios are similar, meaning that both a high weight of the global pattern in a patient's GBM aCGH profile and an age >50 years old at diagnosis confer similar relative hazard. These ratios also do not differ significantly from the univariate ratios of 2.3 and 2 for the global pattern and age, respectively. Taken together, the prognostic contribution of the global pattern is not only comparable to that of age, but is also independent of age. Combined with age, the global pattern makes a better predictor than age alone. Similarly, we find that the global pattern is independent of chemotherapy (Appendix A).

To confirm the global pattern, we use GSVD to compare matched profiles of a larger, more recent, set of 344 TCGA patients, that is inclusive of the initial set of 251 patients [45]. Agilent Human aCGH 244A-measured 458 tumor and 459 normal profiles were selected, corresponding to the inclusive confirmation set of N=344 patients (Dataset S4). The profiles, centered at their autosomal median copy numbers, are organized in one tumor and one normal dataset, of M_1=200,139 and M_2=198,342 probes, respectively. Within each set, the medians of profiles of samples from the same patient are taken after estimating missing

data by using SVD. We find that the significant probelets and corresponding tumor and normal arraylets, as well as their interpretations, are robust to the increase from 251 patients in the initial set to 344 patients in the inclusive confirmation set, and the accompanying decreases in tumor and normal probes, respectively.

The second tumor arraylet computed by GSVD for the 344 patients of the inclusive confirmation set correlates with that of the initial set, with the correlation ∼0.99. To classify the patients according to the copy numbers listed in the corresponding second probelet of the inclusive confirmation set, the classification cutoff ±0.02 of the initial set of 251 patients is scaled by the norm of the copy numbers listed for these patients, resulting in the cutoff ±0.017. Only four of the 251 patients in the initial set, i.e., ∼1.5%, with copy numbers that are near the classification cutoffs of both sets, change classification. Of the 344 patients, we find that 315 patients, 309 with TCGA annotations, display high (>0.017) and 27, i.e., ∼8%, display low, approximately zero, relative copy numbers in the second probelet. Only two patients, one missing TCGA annotations, remain unclassified with large negative (<-0.017) copy numbers in the second probelet. Survival analyses of the inclusive confirmation set of 344 patients give qualitatively the same results as these of the initial set of 251 patients. These analyses confirm that a negligible weight of the global pattern, which is described by the second tumor arraylet, i.e., a low copy number in the second probelet, is indicative of a significantly longer survival time (Figure 2.8d). Survival analysis of only the chemotherapy patients in the inclusive confirmation set classified by GSVD gives similar results (Appendix A). These analyses confirm that the prognostic contribution of the global pattern is comparable to that of age (Figure 2.8e) and is independent of age (Figure 2.8f). Similarly, we confirm that the global pattern is independent of chemotherapy (Appendix A).

To validate the prognostic contribution of the global pattern, we classify GBM profiles of an independent set of 184 TCGA patients, that is mutually exclusive of the initial set of 251 patients. Agilent Human aCGH 244A-measured 280 tumor profiles were selected, corresponding to the independent validation set of 184 patients with available TCGA status annotations (Dataset S5). Each profile lists relative copy numbers in more than 97.5% of the 206,820 autosomal probes among the M_1=212,696 probes that define the second tumor arraylet computed by GSVD for the 251 patients of the initial set. Medians of profiles of samples from the same patient are taken. To classify the 184 patients according to the correlations of their GBM profiles with the second tumor arraylet of the initial set, the classification cutoff of the initial set of 251 patients is scaled by the norm of the correlations calculated for these patients, resulting in the cutoff ±0.15. For the profiles of 162 patients,

we calculate high (>0.15) and for 21, i.e., ∼11%, low, approximately zero, correlation with the second tumor arraylet. One patient remains unclassified with a large negative (<-0.15) correlation.

We find that survival analyses of the independent validation set of 184 patients give qualitatively the same results as these of the initial set of 251 patients and the inclusive confirmation set of 344 patients (Figures 2.3*g*–*i* and Appendix A). These analyses validate the prognostic contribution of the global pattern, which is computed by GSVD of patient-matched tumor and normal aCGH profiles, also for patients with measured GBM aCGH profiles in the absence of matched normal profiles.

CHAPTER 3

TENSOR GSVD (tGSVD) COMPARISONS OF MATCHED GENOMIC PROFILES

3.1 Ovarian Serous Cystadenocarcinoma (OV)

Ovarian serous cystadenocarcinoma (OV) accounts for about 90% of all ovarian cancers. Despite recent large-scale profiling efforts [68], very few DNA copy-number alterations (CNAs) that frequently occur in OV have been identified so far. The best predictor of OV survival to date has remained the tumor's stage at diagnosis, a pathological assessment of the spread of the cancer numbering I to IV [69]. To identify CNAs that might predict OV patients' survival, we comparatively model patient- and platform-matched but probe-independent genomic profiles of ovarian serous cystadenocarcinoma (OV) tumor and normal samples from the Cancer Genome Atlas (TCGA) using a novel tensor GSVD (tGSVD).

The tGSVD reveals chromosome arm-wide patterns of tumor-exclusive and platform-consistent CNAs, across 6p+12p, 7p and Xq, that are correlated with, and possibly causally related to, OV patients' survival. We find that, first, the patterns are independent of the tumor's stage. Therefore, combined with stage, each pattern makes a better predictor than stage alone. Second, the amplified and deleted segments identified by the patterns [47,49] include most known OV-associated CNAs [70], and several previously unreported yet frequent CNAs [71–74]. Third, differential mRNA expression between the tGSVD classes is enriched in ontologies that include genes, which consistently map to the DNA CNAs [75,76]. Differential microRNA and protein expression also consistently map to the CNAs [77].

A coherent picture emerges for each of the tGSVD patterns, across 6p+12p, 7p and Xq, suggesting roles for the DNA CNAs in OV pathogenesis as well as personalized therapy. The 6p+12p pattern describes previously unrecognized co-occurring alterations, including loss of the p21-encoding *CDKN1A* and the p38-encoding *MAPK14* on 6p, and gain of *KRAS* on 12p, which together, but not separately, can lead to transformation of human normal to tumor cells [78,79]. These alterations, together with deletion of *TNF* on 6p, and amplification of *RAD51AP1* and *ITPR2* on 12p, which are also included in the 6p+12p

pattern, correlate with a suppression of cell cycle arrest, senescence and apoptosis in the OV tumor cell, and an OV patient's shorter survival time [80–91]. Since drugs interacting with *CDKN1A*, *MAPK14*, *RAD51AP1* and *KRAS* exist [91], the 6p+12p tGSVD may prove useful in OV personalized therapy. In 7p, *RPA3* deletion and *POLD2* amplification correlate with DNA double-strand break (DSB) repair via homologous recombination (HR) during replication, reduced genomic instability and a longer survival time [92, 93]. In Xq, *PABPC5* deletion and *BCAP31* amplification correlate with an active cellular immune response and a longer survival time [94].

3.1.1 Tumor and Normal Datasets

To identify links between an OV tumor's genome and a patient's prognosis, which might offer insights into the cancer's pathogenesis and suggest targets for drug therapy, we modeled patient- and platform-matched OV and normal genomic, i.e., array CGH (aCGH) profiles from TCGA by using a novel tGSVD. We selected a discovery set of 249 TCGA patients with both primary OV and normal aCGH profiles, each measured by two DNA microarray platforms, and a validation set of 148 patients, mutually exclusive of the discovery set, with primary OV aCGH profiles measured by either one of the two platforms or both [68] (Datasets S1 and S2).

Each profile in the discovery datasets lists $\log_2$ of the TCGA level 1 background-subtracted intensity in the sample relative to the male Promega DNA reference, with signal to background $\geq$2.5 for both the sample and reference in $\geq$90% of the 391,190 autosomal probes and $\geq$65% of the 10,911 X chromosome probes that match between the two Agilent Human aCGH DNA microarray platforms, G4447A and G4124A. Tumor and normal probes were selected with valid data in $\geq$99% of the tumor or normal arrays of each platform, respectively. For each chromosome arm or combination of two chromosome arms, and for each platform, the <0.5% missing data entries in the tumor and normal profiles were estimated by using the SVD, as previously described [22]. Each profile was then centered at its copy-number median and normalized by its copy-number sMAD.

For the validation dataset, we selected 131 and 41 stage III-IV OV aCGH profiles measured by the Agilent Human aCGH G4447A and G4124A microarray platforms, respectively. Each profile lists $\log_2$ of the TCGA level 1 background-subtracted intensity in the sample relative to the male Promega DNA reference, with signal to background $\geq$2.5 for both the sample and reference in $\geq$99.5% of the 391,190 autosomal probes and $\geq$96.5% of the 10,911 X chromosome probes that match between the platforms. Medians of the profiles of samples from the same patient were then taken.

For each chromosome arm or combination of two chromosome arms, the structure of the patient- and platform-matched but probe-independent tumor and normal discovery datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, of K_1-tumor and K_2-normal probes $\times$ L-patients, i.e., arrays $\times$ M-platforms, is that of two third-order tensors with a one-to-one mapping between the column dimensions but different row dimensions, where $K_1, K_2 \geq LM$.

3.2 Tensor Generalized Singular Value Decomposition (tGSVD)

3.2.1 Introduction

We define a novel tGSVD, an exact simultaneous decomposition of two such datasets, arranged in two higher-than-second-order tensors of the same column dimensions but different row dimensions. We prove that the tGSVD extends the matrix GSVD [3, 21, 22, 25, 95–98] and the tensor higher-order singular value decomposition (HOSVD) [31, 32, 99] from two matrices and one tensor, respectively, to two tensors [100]. The tGSVD identifies patterns of varying mathematical significance in one dataset relative to the other. We show that the mathematical properties of the tGSVD allow interpreting the patterns in terms of the biomedical similarities and dissimilarities between the two datasets. We demonstrate the tGSVD in comparisons of patient- and platform-matched but probe-independent genomic profiles of ovarian serous cystadenocarcinoma (OV) tumor and normal samples.

3.2.2 tGSVD : Formulation and Construction

To compare the OV tumor and normal datasets described above, that are each of the form of a third-order tensor, we define a novel tGSVD that simultaneously separates the paired datasets into weighted sums of LM paired "subtensors," i.e., combinations or outer products of three patterns each: either one tumor-specific pattern of copy-number variation across the tumor probes, i.e., a "tumor arraylet" $u_{1,a}$, or the corresponding normal-specific pattern across the normal probes, i.e., the "normal arraylet" $u_{2,a}$, combined with one pattern of copy-number variation across the patients, i.e., an "x-probelet" $v^T_{x,b}$ and one pattern across the platforms, i.e., a "y-probelet" $v^T_{y,c}$, which are identical for both the tumor and normal datasets (Figures 3.1, 3.2 and 3.3),

Figure 3.1. Tensor generalized singular value decomposition (tGSVD) of the patient- and platform-matched DNA copy-number profiles of the Xq chromosome arm. For each chromosome arm or combination of two chromosome arms, the structure of the tumor and normal discovery datasets ($\mathcal{D}_1$ and $\mathcal{D}_2$) is that of two third-order tensors with a one-to-one mapping between the column dimensions but different row dimensions. The patients, platforms, tissue types and probes, each represents a degree of freedom. Unfolded into a single matrix, some of the degrees of freedom are lost and much of the information in the datasets might also be lost. We define a tGSVD that simultaneously separates the paired datasets into weighted sums of paired subtensors, i.e., combinations or outer products of three patterns each: either one tumor-specific pattern of copy-number variation across the tumor probes, i.e., a tumor arraylet (a column basis vector of U_1), or the corresponding normal-specific arraylet (a column basis vector of U_2), combined with one pattern of variation across the patients, i.e., an x-probelet (a row basis vector of V_x^T), and one pattern across the platforms, i.e., a y-probelet (a row basis vector of V_y^T), which are identical for both the tumor and normal datasets (Equation (3.1)). The tGSVD is depicted in a raster display, with relative copy-number gain (red), no change (black) and loss (green), explicitly showing the first through the 5th, and the 245th through the 249th Xq x-probelets, both Xq y-probelets, and the first through the 10th, and the 489th through the 498th Xq tumor and normal arraylets. We prove that the significance of a subtensor in the tumor dataset relative to that of the corresponding subtensor in the normal dataset, i.e., the tGSVD angular distance, equals the row mode GSVD angular distance, i.e., the significance of the corresponding tumor arraylet in the tumor dataset relative to that of the normal arraylet in the normal dataset. The tGSVD angular distances for the 498 pairs of Xq arraylets are depicted in a bar chart display, where the angular distance corresponding to the first pair of arraylets is $\sim\pi/4$. For the Xq chromosome arm, we find that the most significant subtensor in the tumor dataset (which corresponds to the coefficient of largest magnitude in $\mathcal{R}_1$) is a combination of (*i*) the first y-probelet, which is approximately invariant across the platforms, (*ii*) the first x-probelet, which classifies the discovery set of patients into two groups of high and low coefficients, of significantly and robustly different prognoses, and (*iii*) the first, most tumor-exclusive tumor arraylet, which classifies the validation set of patients into two groups of high and low correlations of significantly different prognoses consistent with the x-probelet's classification of the discovery set.

Platforms
V_y^T
y-Probelets
Arrays
V_x^T
x-Probelets
$\mathcal{R}_{\text{tumor}}$
x-Probelets
y-Probelets
Arraylets
U_{tumor}
Tumor Probes
Arraylets
Tumor Probes
$\mathcal{D}_{\text{tumor}}$
Arrays
Platforms
Tumor Probes

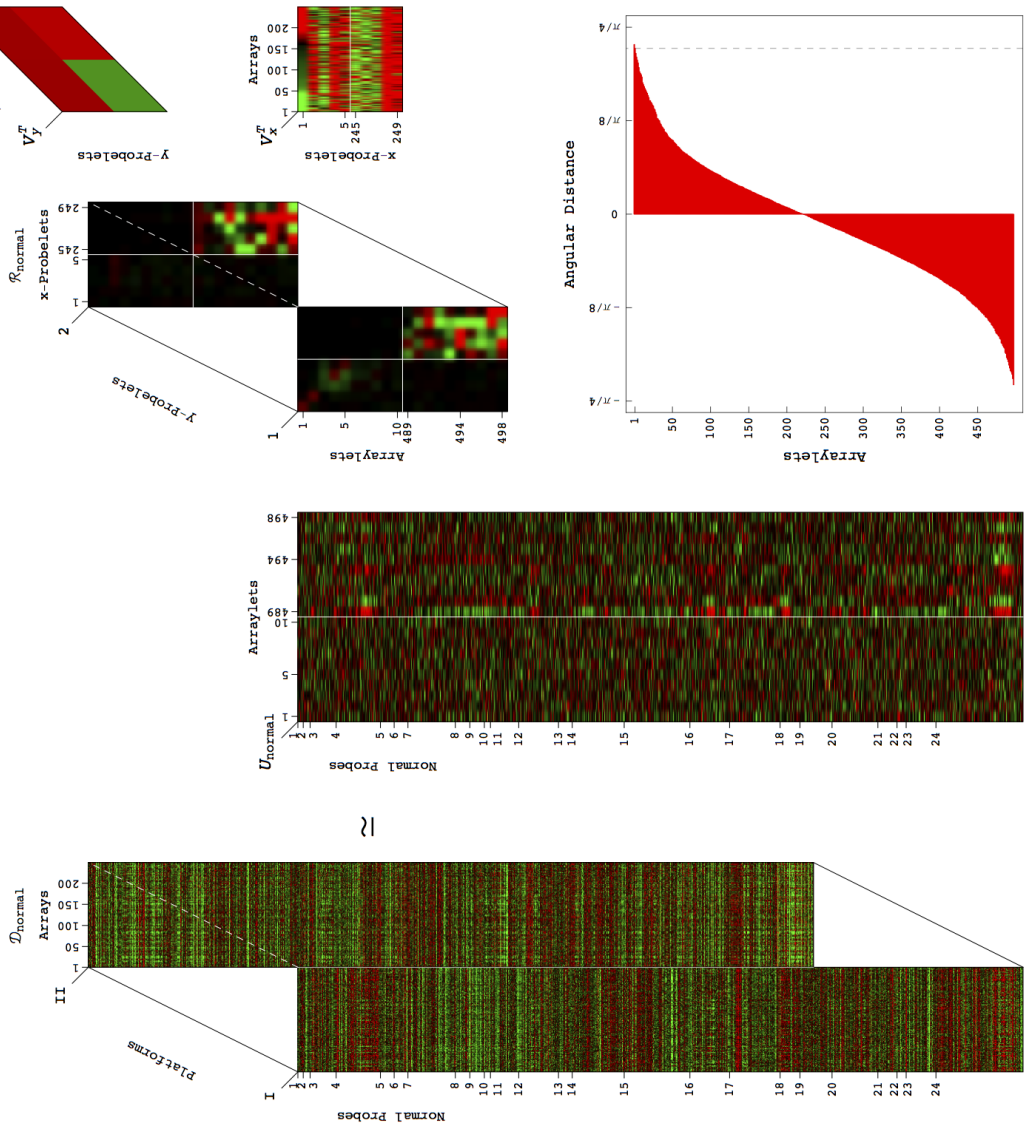
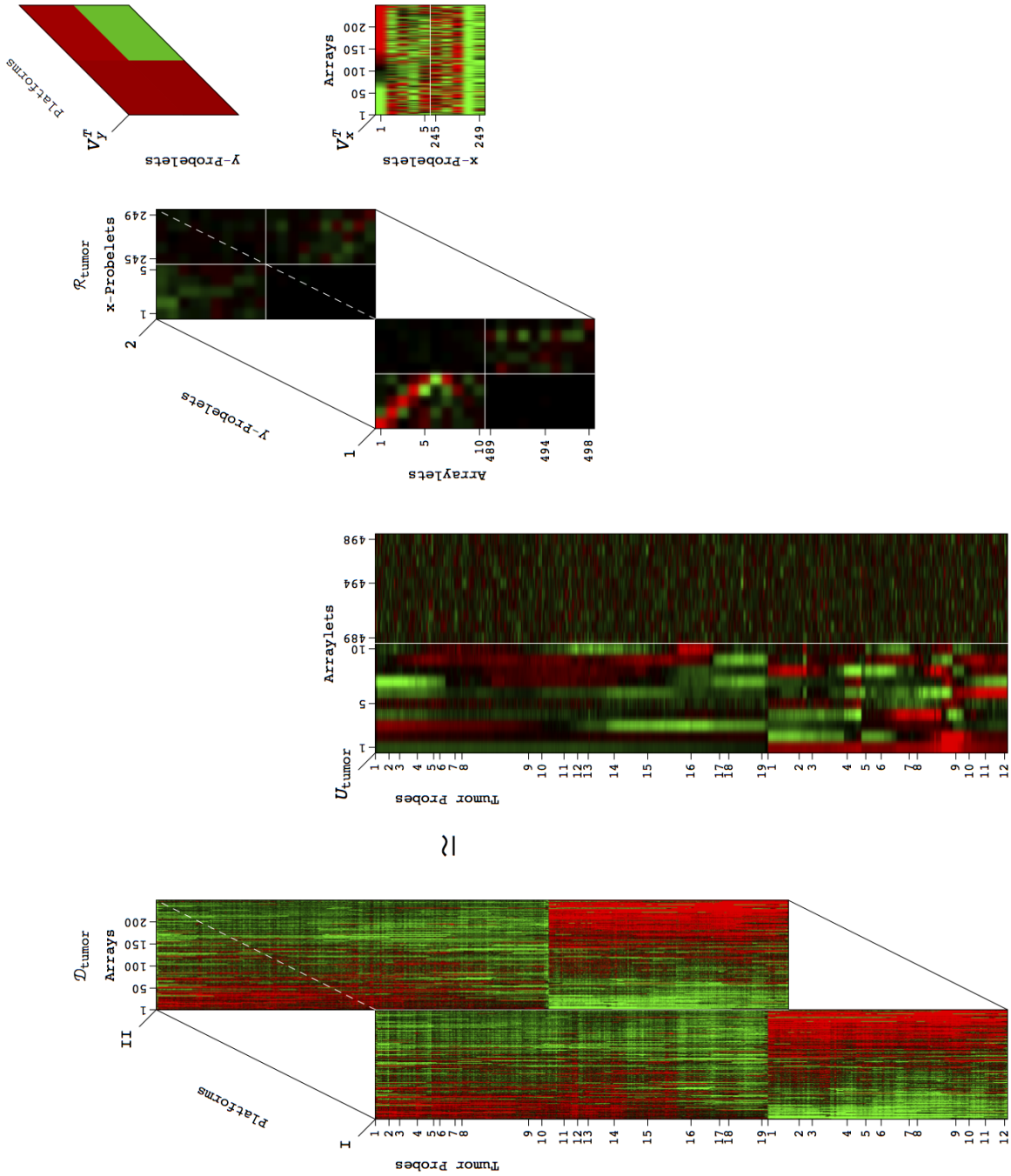


Figure 3.1: Continued.

Figure 3.2. The tGSVD is depicted in a raster display, with relative copy-number gain (red), no change (black) and loss (green), explicitly showing the first through the 5th, and the 245th through the 249th 6p+12p x-probelets, both 6p+12p y-probelets, and the first through the 10th, and the 489th through the 498th 6p+12p tumor and normal arraylets. We prove that the significance of a subtensor in the tumor dataset relative to that of the corresponding subtensor in the normal dataset, i.e., the tGSVD angular distance, equals the row mode GSVD angular distance, i.e., the significance of the corresponding tumor arraylet in the tumor dataset relative to that of the normal arraylet in the normal dataset. The tGSVD angular distances for the 498 pairs of 6p+12p arraylets are depicted in a bar chart display, where the angular distance corresponding to the first pair of arraylets is $\sim\pi/4$. For the 6p+12p chromosome arm combination, we find that the most significant subtensor in the tumor dataset (which corresponds to the coefficient of largest magnitude in $\mathcal{R}_1$) is a combination of (*i*) the first y-probelet, which is approximately invariant across the platforms; (*ii*) the first x-probelet, which classifies the discovery set of patients into two groups of high and low coefficients, of significantly and robustly different prognoses; and (*iii*) the first, most tumor-exclusive tumor arraylet, which classifies the validation set of patients into two groups of high and low correlations of significantly different prognoses consistent with the x-probelet's classification of the discovery set.


Platforms
V_y^T
y-Probelets
V_x^T
Arrays
x-Probelets
$\mathcal{R}_{\text{tumor}}$
x-Probelets
y-Probelets
Arraylets
U_{tumor}
Arraylets
Tumor Probes
$\mathcal{D}_{\text{tumor}}$
Arrays
Platforms
Tumor Probes

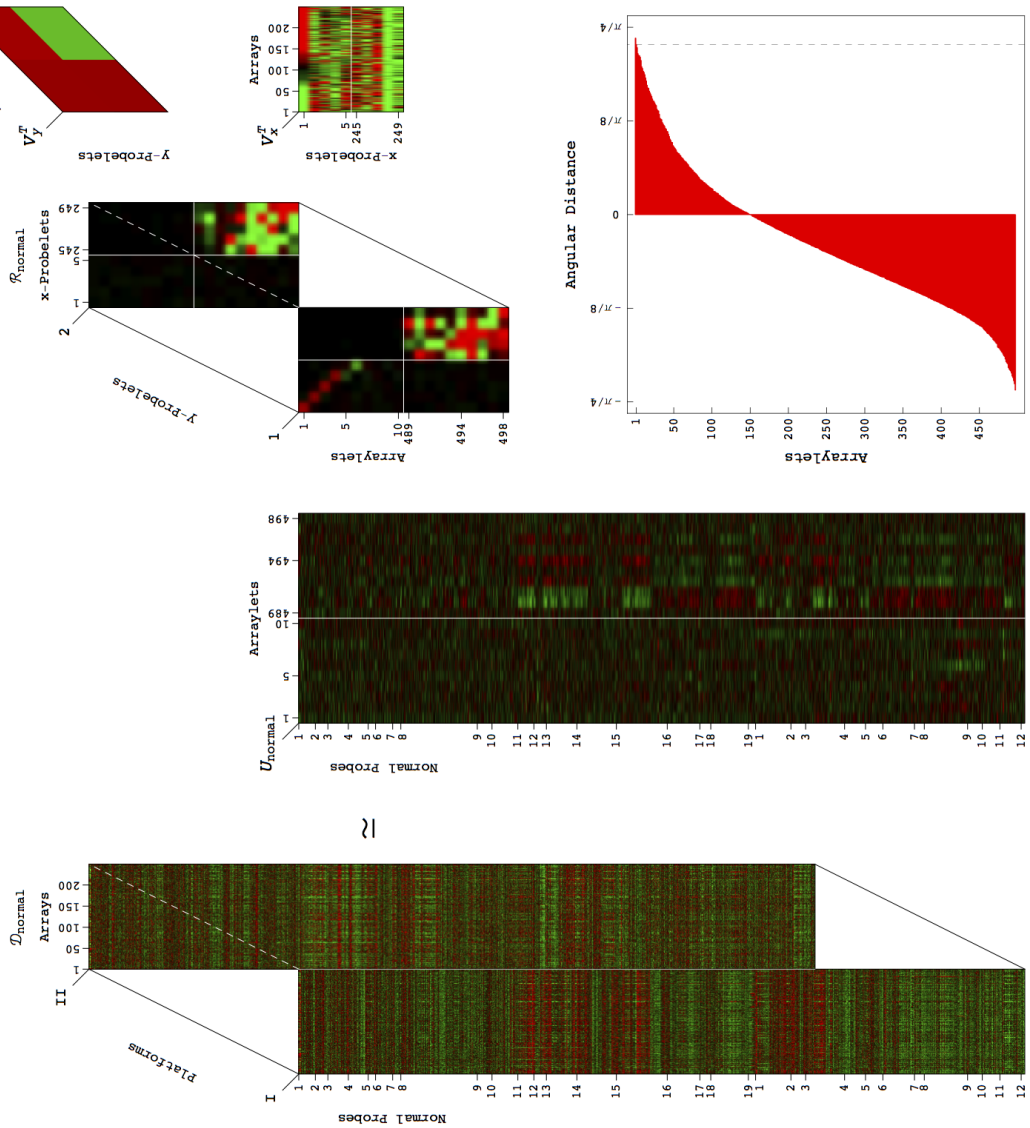
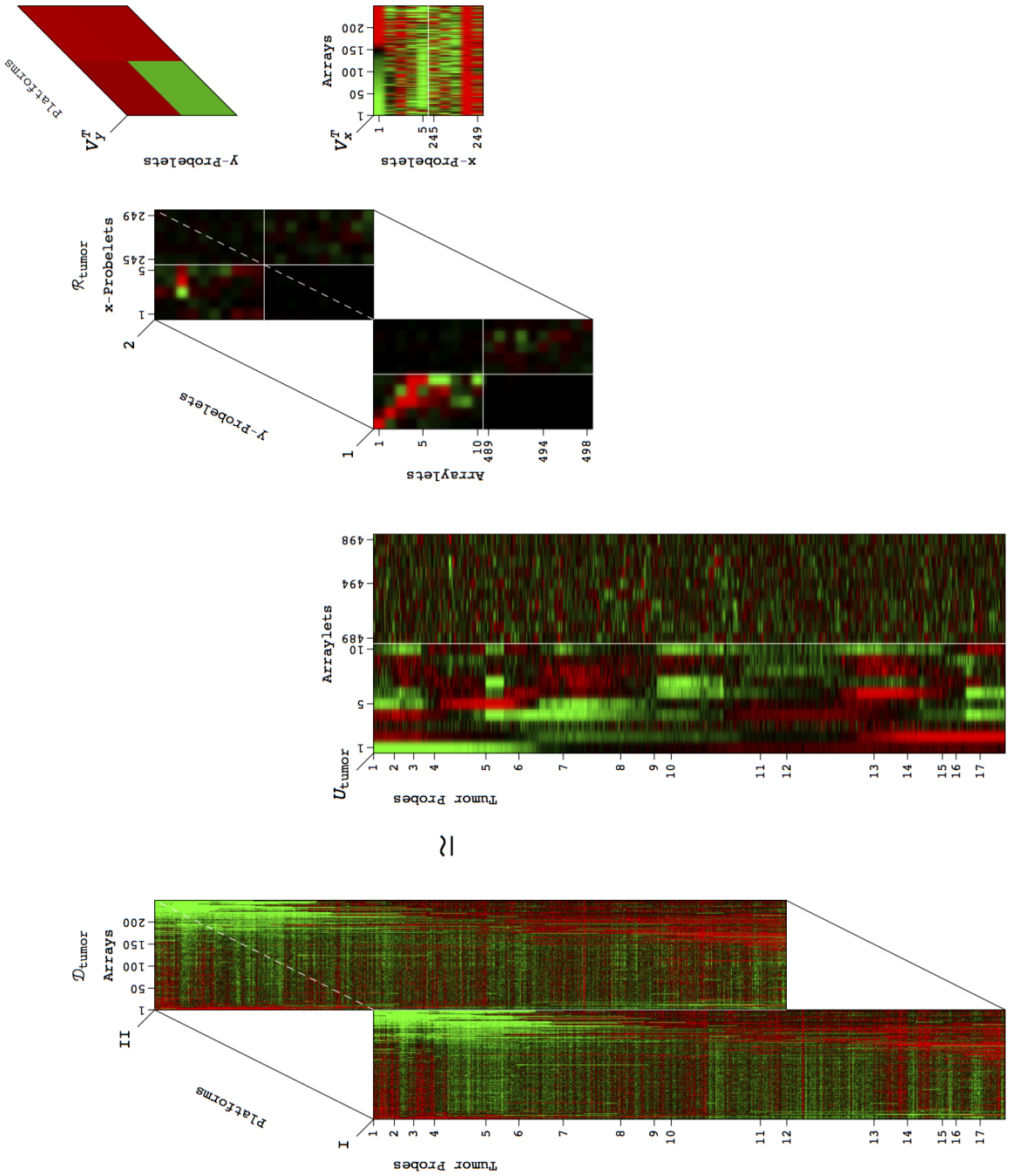


Figure 3.2: Continued.

Figure 3.3. The tGSVD is depicted in a raster display, with relative copy-number gain (red), no change (black) and loss (green), explicitly showing the first through the 5th, and the 245th through the 249th 7p x-probelets, both 7p y-probelets, and the first through the 10th, and the 489th through the 498th 7p tumor and normal arraylets. The tGSVD angular distances for the 498 pairs of 7p arraylets are depicted in a bar chart display, where the angular distance corresponding to the first pair of arraylets is $\sim\pi/4$.


Platforms
V_y^T
y-Probelets
Arrays
V_x^T
x-Probelets
$\mathcal{R}_{\text{tumor}}$
x-Probelets
y-Probelets
Arraylets
U_{tumor}
Tumor Probes
$\mathcal{D}_{\text{tumor}}$
Arrays
Platforms
Tumor Probes

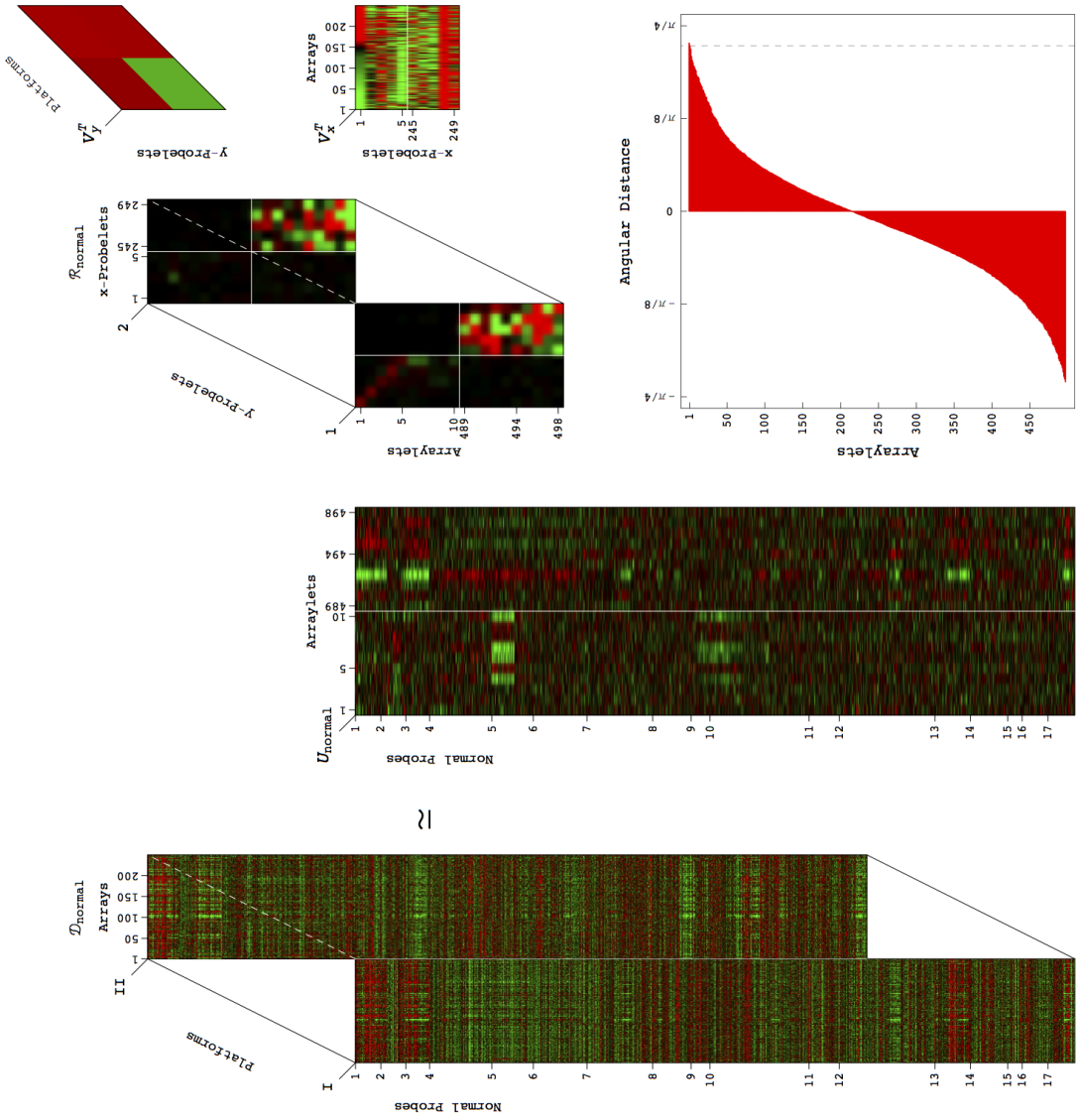


Figure 3.3: Continued.

$$\begin{aligned}\mathcal{D}_i &= \mathcal{R}_i \times_a U_i \times_b V_x \times_c V_y \\ &= \sum_{a=1}^{LM}\sum_{b=1}^{L}\sum_{c=1}^{M} \mathcal{R}_{i,abc}\mathcal{S}_i(a,b,c), \\ \mathcal{S}_i(a,b,c) &= u_{i,a} \otimes v_{x,b}^T \otimes v_{y,c}^T, \quad i = 1,2, \end{aligned} \tag{3.1}$$

where $\times_a U_i$, $\times_b V_x$ and $\times_c V_y$ denote contractions of the LM-arraylet, L-x-probelet and M-y-probelet dimensions of the "core tensor" $\mathcal{R}_i$ with those of U_i, V_x and V_y, respectively, and where $\otimes$ denotes an outer product.

Suppose that unfolding both tensors $\mathcal{D}_i$ into matrices, each preserving the K_i-row dimension, e.g., by appending the LM columns $\mathcal{D}_{i,:lm}$ of the corresponding tensor, gives two full column-rank matrices $D_i \in \mathbb{R}^{K_i \times LM}$. We obtain the column bases vectors U_i from the GSVD of D_i [21, 22, 25, 95–98], i.e., the "row mode GSVD"

$$D_i = (\ldots, \mathcal{D}_{i,:lm}, \ldots) = U_i \Sigma_i V^T, \quad i = 1,2. \tag{3.2}$$

Suppose, similarly, that unfolding both tensors $\mathcal{D}_i$ into matrices, each preserving the L-x- (or M-y-) column dimension, e.g., by appending the K_iM rows $\mathcal{D}^T_{i,K_i:M}$ (or the K_iL rows $\mathcal{D}^T_{i,K_iL:}$) of the corresponding tensor, gives two full column-rank matrices $D_{ix} \in \mathbb{R}^{K_iM \times L}$ (or $D_{iy} \in \mathbb{R}^{K_iL \times M}$). We obtain the x- (or y-) row basis vectors V_x^T (or V_y^T), from the GSVD of D_{ix} (or D_{iy}), i.e., the x- (or y-) column mode GSVD,

$$\begin{aligned} D_{ix} = &\quad (\ldots, \mathcal{D}^T_{i,k:m}, \ldots) &&= U_{ix}\Sigma_{ix}V_x^T, \\ D_{iy} = &\quad (\ldots, \mathcal{D}^T_{i,kl:}, \ldots) &&= U_{iy}\Sigma_{iy}V_y^T, \quad i = 1,2. \end{aligned} \tag{3.3}$$

Note that the x- and y-row bases vectors are, in general, nonorthogonal but normalized, and V_x and V_y are invertible. The column bases vectors are orthonormal, such that $U_i^T U_i = I$.

The generalized singular values are arranged in Σ_i, Σ_{ix} and Σ_{iy} in decreasing orders of the corresponding "GSVD angular distances," i.e., decreasing orders of the ratios $\sigma_{1,a}/\sigma_{2,a}$, $\sigma_{1x,b}/\sigma_{2x,b}$ and $\sigma_{1y,c}/\sigma_{2y,c}$, respectively. We then compute the core tensors $\mathcal{R}_i$ by contracting the row-, x- and y-column dimensions of the tensors $\mathcal{D}_i$ with those of the matrices U_i, V_x^{-1} and V_y^{-1}, respectively. For real tensors, the "tensor generalized singular values" $\mathcal{R}_{i,abc}$ tabulated in the core tensors are real but not necessarily nonnegative.

Our tGSVD construction generalizes the GSVD to higher orders in analogy with the generalization of the singular value decomposition (SVD) by HOSVD [31, 32, 99], and is different from other approaches to the decomposition of two tensors [100].

3.2.3 Existence, Uniqueness and Special Cases

We prove that our tGSVD exists for two tensors of any order because it is constructed from the GSVDs of the tensors unfolded into full column-rank matrices (Lemma 3.1). The tGSVD has the same uniqueness properties as the GSVD, where the column bases vectors $u_{i,a}$ and the row bases vectors $v^T_{x,b}$ and $v^T_{y,c}$ are unique, except in degenerate subspaces, defined by subsets of equal generalized singular values σ_i, σ_{ix} and σ_{iy}, respectively, and up to phase factors of ± 1, such that each vector captures both parallel and antiparallel patterns (Lemma 3.2). The tGSVD of two second-order tensors reduces to the GSVD of the corresponding matrices (Corollary 3.1). The tGSVD of the tensor $\mathcal{D}_1 \in \mathbb{R}^{LM\times L\times M}$, which row mode unfolding gives the identity matrix $D_1 = I \in \mathbb{R}^{LM\times LM}$, and a tensor $\mathcal{D}_2$ of the same column dimensions reduces to the HOSVD of $\mathcal{D}_2$ (Theorem 3.1).

Lemma 3.1 *The tGSVD exists for any two, e.g., third-order tensors $\mathcal{D}_i \in \mathbb{R}^{K_i\times L\times M}$ of the same column dimensions L and M but different row dimensions K_i, where $K_i \geq LM$ for $i = 1, 2$, if the tensors unfold into full column-rank matrices, $D_i \in \mathbb{R}^{K_i\times LM}$, $D_{ix} \in \mathbb{R}^{K_iM\times L}$ and $D_{iy} \in \mathbb{R}^{K_iL\times M}$, each preserving either the K_i-row dimension, L-x-, or M-y- column dimension, respectively.*

Proof: The tGSVD of Equation (3.1), of the pair of third-order tensors $\mathcal{D}_i$, is constructed from the GSVDs of Equations (3.2) and (3.3), of the pairs of full column-rank matrices D_i, D_{ix} and D_{iy}, where $i = 1, 2$. From the existence of the GSVDs of Equations (3.2) and (3.3) [21, 25, 95–98], the orthonormal column bases vectors of U_i, as well as the normalized x- and y-row bases vectors of the invertible V^T_x or V^T_y, exist, and, therefore, the tGSVD of Equation (3.1) also exists. Note that the proof holds for tensors of higher-than-third order.
□

Lemma 3.2 *The tGSVD has the same uniqueness properties as the GSVD.*

Proof: From the uniqueness properties of the GSVDs of Equations (3.2) and (3.3), the orthonormal column bases vectors $u_{i,a}$, and the normalized row bases vectors $v^T_{x,b}$, and $v^T_{y,c}$ of the tGSVD of Equation (3.1) are unique, except in degenerate subspaces, defined by subsets of equal generalized singular values σ_i, σ_{ix} and σ_{iy}, respectively, and up to phase factors of ± 1. The tGSVD, therefore, has the same uniqueness properties as the GSVD. Note that the proof holds for tensors of higher-than-third order.
□

Corollary 3.1 *For two second-order tensors, the tGSVD reduces to the GSVD of the corresponding matrices.*

Proof: For two second-order tensors, e.g., the matrices $D_i \in \mathbb{R}^{K_i \times L}$, the tGSVD of Equation (3.1) is

$$\begin{aligned} D_i &= R_i \times_a U_i \times_b V_x \\ &= U_i R_i V_x^T, \qquad i = 1, 2. \end{aligned} \tag{3.4}$$

The row- and x-column mode GSVDs of Equations (3.2) and (3.3) are identical, because unfolding each matrix D_i while preserving either its K_i-row dimension, or L-x-column dimension results in D_i, up to permutations of either its rows or columns, respectively,

$$D_i = U_i \Sigma_i V_x^T = D_{ix}, \quad i = 1, 2. \tag{3.5}$$

From the uniqueness properties of the tGSVD of Equation (3.4), and the GSVDs of Equation (3.5) it follows that $R_i = \Sigma_i$, and that for two second-order tensors, i.e., matrices, the tGSVD is equivalent to the GSVD.
□

Theorem 3.1 *The tGSVD of the tensor $\mathcal{D}_1 \in \mathbb{R}^{LM \times L \times M}$, which row mode unfolding gives the identity matrix $D_1 = I \in \mathbb{R}^{LM \times LM}$, and a tensor $\mathcal{D}_2$ of the same column dimensions reduces to the HOSVD of $\mathcal{D}_2$.*

Proof: Consider the GSVD of Equation (3.2), of the matrices $D_1 = I$ and D_2, as computed by using the QR decomposition of the appended D_1 and D_2, and the SVD of the block of the resulting column-wise orthonormal Q that corresponds to D_2, i.e., $Q_2 = U_{Q_2} \Sigma_{Q_2} V_{Q_2}^T$ [25],

$$\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} = \begin{bmatrix} I \\ D_2 \end{bmatrix} = QR = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} R = \begin{bmatrix} R^{-1} \\ U_{Q_2} \Sigma_{Q_2} V_{Q_2}^T \end{bmatrix} R, \tag{3.6}$$

where R is upper triangular and, therefore, invertible. Since Q is column-wise orthonormal, $V_{Q_2}^T$ is orthonormal and Σ_{Q_2} is positive diagonal, it follows that

$$\begin{aligned} I &= Q_1^T Q_1 + Q_2^T Q_2 \\ &= R^{-T} R^{-1} + V_{Q_2} \Sigma_{Q_2}^2 V_{Q_2}^T \\ &= (V_{Q_2}^T R)^{-T} (V_{Q_2}^T R)^{-1} + \Sigma_{Q_2}^2, \\ (I - \Sigma_{Q_2}^2)^{-1} &= (V_{Q_2}^T R)(V_{Q_2}^T R)^T, \end{aligned} \tag{3.7}$$

and that $(I - \Sigma_{Q_2}^2)^{\frac{1}{2}} V_{Q_2}^T R$ is orthonormal. The GSVD of Equation (3.2) factors the matrix D_2 into a column-wise orthonormal U_{Q_2}, a positive diagonal $\Sigma_{Q_2}(I - \Sigma_{Q_2}^2)^{-\frac{1}{2}}$ and an orthonormal $(I - \Sigma_{Q_2}^2)^{\frac{1}{2}} V_{Q_2}^T R$, and is, therefore, reduced to the SVD of D_2.

Note that this proof holds for the GSVDs of Equation (3.3). This is because the x- and y-column unfoldings of the tensor $\mathcal{D}_1 \in \mathbb{R}^{LM \times L \times M}$, which row mode unfolding gives the identity matrix $D_1 = I \in \mathbb{R}^{LM \times LM}$, give

$$D_{1x} = \begin{bmatrix} I \\ \vdots \\ I \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} M \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} M(M-1) \end{matrix}, \qquad D_{1y} = \begin{bmatrix} I \\ \vdots \\ I \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} L \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} L(L-1) \end{matrix}. \tag{3.8}$$

The GSVDs of Equations (3.2) and (3.3), of either one of the matrices D_1, D_{1x} or D_{1y} with the corresponding full column-rank matrices D_2, D_{2x} or D_{2y}, are, therefore, reduced to the SVDs of D_2, D_{2x} or D_{2y}, respectively.

The tGSVD of Equation (3.1), where the orthonormal column bases vectors $u_{2,a}$, and the normalized row bases vectors $v_{x,b}^T$, and $v_{y,c}^T$ in the factorization of the tensor $\mathcal{D}_2$ are computed via the SVDs of the unfolded tensor is, therefore, reduced to the HOSVD of $\mathcal{D}_2$ [31, 32, 99]. Note that the proof holds for tensors of higher-than-third order.

□

3.2.4 Interpretation

The significance of the subtensor $\mathcal{S}_i(a, b, c)$ in the tensor $\mathcal{D}_i$ is proportional to the magnitude of the corresponding $\mathcal{R}_{i,abc}$ (Figure 3.4),

$$\mathcal{P}_{i,abc} = \mathcal{R}_{i,abc}^2 / \sum_{a=1}^{LM} \sum_{b=1}^{L} \sum_{c=1}^{M} \mathcal{R}_{i,abc}^2, \quad i = 1, 2. \tag{3.9}$$

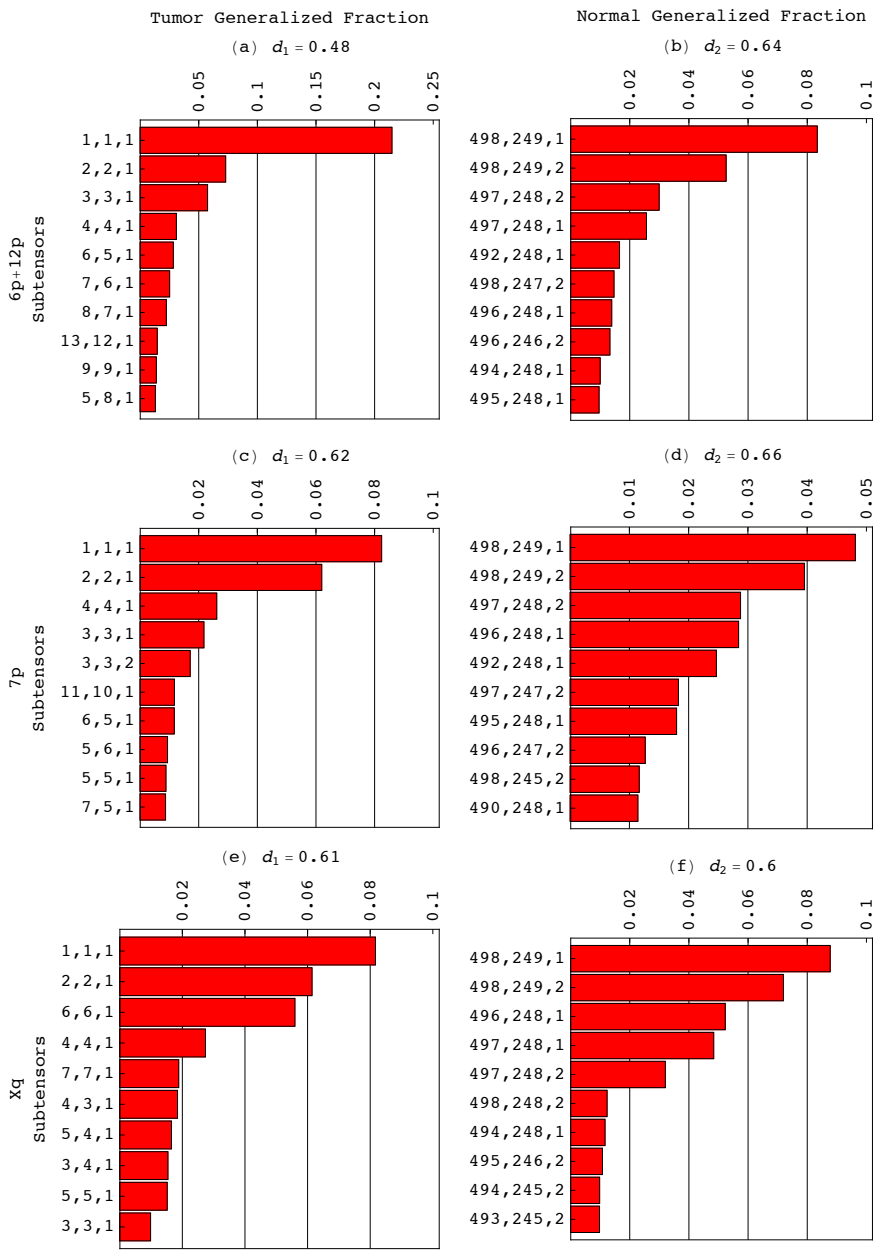


Figure 3.4: Most significant subtensors in the tumor and normal discovery datasets. Bar charts of the ten subtensors $\mathcal{S}_i(a, b, c)$ of Equation (3.1) that are most significant in the 6p+12p (a) tumor, and (b) normal, 7p (c) tumor, and (d) normal, and Xq (e) tumor, and (f) normal datasets, in terms of the fractions $\mathcal{P}_{i,abc}$ of Equation (3.9), i.e., the subtensors which correspond to the coefficients $\mathcal{R}_{i,abc}$ of largest magnitudes. The most significant subtensor in each of the tumor datasets, e.g., is $\mathcal{S}_1(1, 1, 1)$, which is a combination or outer product of the first, most tumor-exclusive arraylet, and the first x- and y-probelets. The most significant subtensor in each of the normal datasets is $\mathcal{S}_2(498, 249, 1)$, which is a combination or outer product of the 498th, most normal-exclusive arraylet, the 249th x-probelet and the first y-probelet. The tensor generalized Shannon entropy d_i of Equation (3.12) of each dataset is also noted.

The significance of $\mathcal{S}_1(a,b,c)$ in $\mathcal{D}_1$ relative to that of $\mathcal{S}_2(a,b,c)$ in $\mathcal{D}_2$ is defined by the "tGSVD angular distance" Θ_{abc}, in analogy with, e.g., the row mode GSVD angular distance θ_a, which defines the significance of the column basis vector $u_{1,a}$ in the matrix D_1 of Equation (3.2) relative to that of $u_{2,a}$ in D_2

$$\begin{aligned} \Theta_{abc} &= \arctan(\mathcal{R}_{1,abc}/\mathcal{R}_{2,abc}) - \pi/4, \\ \theta_a &= \arctan(\sigma_{1,a}/\sigma_{2,a}) - \pi/4. \end{aligned} \tag{3.10}$$

Note that $|\Theta_{abc}| \leq \pi/4$, where $\Theta_{abc} = \pm\pi/4$ indicates a subtensor exclusive to either $\mathcal{D}_1$ or $\mathcal{D}_2$, respectively, and $\Theta_{abc} = 0$ indicates a subtensor common to both.

Theorem 3.2 *The tGSVD angular distance equals the row mode GSVD angular distance, i.e.,* $\Theta_{abc} = \theta_a$.

Proof: The unfolding of $\mathcal{D}_i$ of Equation (3.1) into D_i of Equation (3.2) unfolds the core tensors $\mathcal{R}_i$ of Equation (3.1) into matrices R_i, which preserve the row dimensions, i.e., the LM-column bases dimensions of $\mathcal{R}_i$, and gives

$$\begin{aligned} D_i &= U_i R_i (V_x^T \otimes V_y^T), \\ R_i &= \Sigma_i V^T (V_x^{-T} \otimes V_y^{-T}), \quad i = 1,2, \end{aligned} \tag{3.11}$$

where $\otimes$ denotes a Kronecker product. Because Σ_i are positive diagonal matrices, it follows that $\mathcal{R}_{1,abc}/\mathcal{R}_{2,abc} = R_{1,a}/R_{2,a} = \sigma_{1,a}/\sigma_{2,a}$. Substituting this in Equation (3.10) gives $\Theta_{abc} = \theta_a$. Note that the proof holds for tensors of higher-than-third order.

Also, the "tensor generalized Shannon entropy" of each dataset,

$$\begin{aligned} 0 \leq d_i &= -(2\log LM)^{-1} \sum_{a=1}^{LM} \sum_{b=1}^{L} \sum_{c=1}^{M} \mathcal{P}_{i,abc} \log \mathcal{P}_{i,abc} \leq 1, \\ i &= 1,2, \end{aligned} \tag{3.12}$$

measures the complexity of each dataset from the distribution of the overall information among the different subtensors. An entropy of zero corresponds to an ordered and redundant dataset in which all the information is captured by a single subtensor. An entropy of one corresponds to a disordered and random dataset in which all subtensors are of equal significance (Figure 3.4).

3.2.5 Discovery and Validation of CNAs Predicting OV Survival

We computed the tGSVD for the discovery datasets of each chromosome arm or combination of two chromosome arms (Mathematica Notebook S1 in Appendix B), and selected

those for which the most significant subtensor in the tumor dataset is a combination of (*i*) a *y*-probelet of consistent, i.e., approximately equal copy numbers in both platforms, (*ii*) an *x*-probelet that classifies the discovery set of patients into two groups of high (>0.5 standardized median absolute deviation, i.e., sMAD, from the median) and low coefficients, of significantly (log-rank test *P*-value <0.05; please refer to Appendix C.5 for more information on log-rank test) and robustly (throughout the range of ±0.1 sMAD around the cutoff) different prognoses (please refer to Appendix C.4 for more information on survival analysis), and (*iii*) the most tumor-exclusive tumor arraylet, if that arraylet classifies the validation set of patients into two groups of high and low Spearman's rank correlation coefficients of significantly different prognoses consistent with the *x*-probelet's classification of the discovery set.The arraylet correlation cutoff is the *x*-probelet coefficient cutoff scaled by the norm/$\sqrt{2}$ of the Spearman's rank correlation coefficients of the 498 tumor profiles of the discovery set of patients, as previously described [3].

3.3 Biological Results

3.3.1 Independent Chromosome Arm-Wide Predictors of OV Survival

To date, the best predictor of OV survival has remained the tumor's stage at diagnosis [69]. Additional indicators, such as the residual disease after surgery, the outcome of subsequent therapy and the neoplasm status, which is the last known status of the disease, are determined during treatment (Figures 3.5 and 3.6).

The tGSVD revealed patterns of tumor-exclusive and platform-consistent CNAs across the chromosome arms 6p+12p, 7p, and Xq that correlate with survival in the discovery and, separately, validation sets (Figure 3.7).

To interpret the 6p+12p, 7p and Xq tumor arraylets, we mapped the tumor probes onto the National Center for Biotechnology Information (NCBI) human genome sequence build 37, by using the Agilent Technologies probe annotations posted at the University of California at Santa Cruz (UCSC) human genome browser [47]. We segmented each of the arraylets and assigned each segment a *P*-value by using the circular binary segmentation (CBS) algorithm as described [49]. To assign a CNA in a segment, we calculated the segment's median copy number, and sMAD from the median in the corresponding arraylet. If the segment's median is at least one sMAD greater (or lesser) than the arraylet's median, then the arraylet is assigned a gain (or a loss) in the segment. Similarly, we calculated the segment's median copy number, and sMAD from the median in each tumor profile. If the segment's median is at least one sMAD greater (or lesser) than the profile's median, then

the patient is assigned a gain (or a loss) in the segment.

We find that each of the patterns is independent of each of the standard indicators (Tables 3.1 and 3.2). Survival analyses of the discovery set classified, e.g., by the 6p+12p tGSVD into high and low x-probelet coefficients, and by pathology at diagnosis into tumor stages I-II and III-IV, give the bivariate Cox hazard ratios (please refer to Appendix C.6 for more information) of 1.7 and 4.4, which are similar to the corresponding univariate ratios 1.7 and 3.7, respectively. Therefore, combined with either one of the standard indicators, each of the three tGSVD patterns makes a better predictor than the standard indicator alone (Figures 3.8 and 3.9). The Kaplan-Meier (KM) median survival time difference of 61 months among the discovery set groups classified, e.g., by both the 6p+12p tGSVD and stage, is more than 50% and almost two years greater than the 39 month difference between the patient groups classified by stage alone.

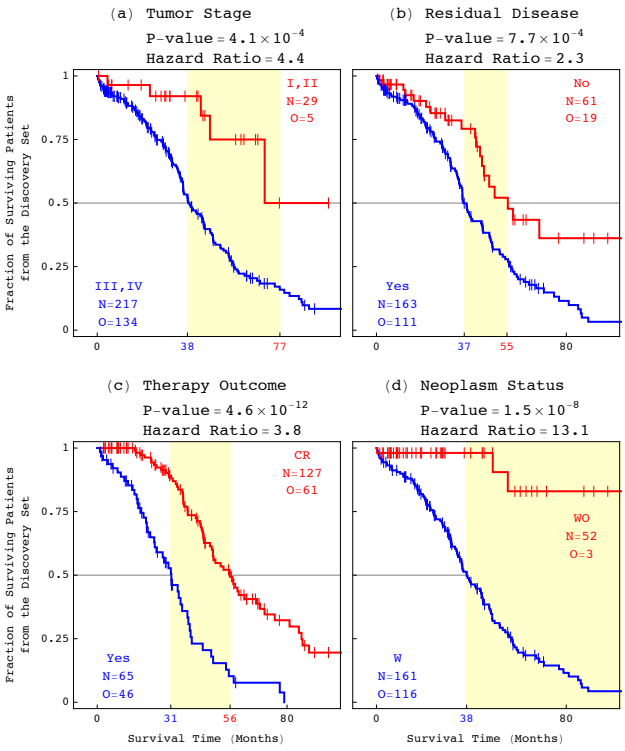


Figure 3.5: Survival analyses of the discovery set of patients classified by the standard OV indicators.
KM curves for the discovery set of 249 patients classified by (a) tumor stage at diagnosis, the best predictor of OV survival to date, (b) residual disease after surgery, (c) outcome of subsequent therapy, and (d) neoplasm status, which is the last known status of the disease.

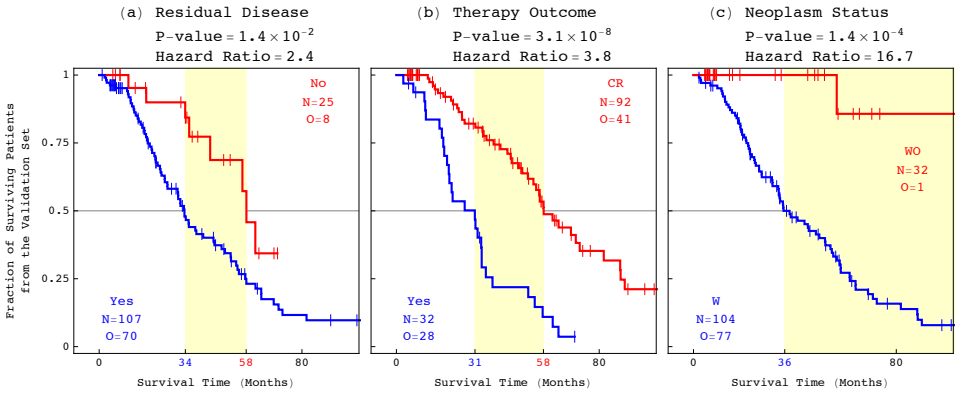


Figure 3.6: Survival analyses of the validation set of patients classified by the standard OV indicators.
KM curves for the validation set of 148 stage III-IV patients classified by (a) residual disease after surgery, (b) outcome of subsequent therapy, and (c) neoplasm status.

Figure 3.7: Survival analyses of the discovery and validation sets of patients classified by tGSVD, or tGSVD and tumor stage at diagnosis. (a) Kaplan-Meier (KM) curves for the discovery set of 249 patients classified by the 6p+12p x-probelet coefficient, show a median survival time difference of 11 months, with the corresponding log-rank test P-value $< 10^{-2}$. The univariate Cox proportional hazard ratio is 1.7, with a P-value $< 10^{-2}$. (b) Survival analyses of the 249 patients classified by the 7p x-probelet coefficient. (c) The 249 patients classified by the Xq x-probelet coefficient. (d) The 249 patients classified by both the 6p+12p tGSVD and tumor stage at diagnosis, show the bivariate Cox hazard ratios of 1.5 and 4, which do not differ significantly from the corresponding univariate hazard ratios, of 1.7 and 4.4, respectively. This means that the 6p+12p tGSVD is independent of stage, the best predictor of OV survival to date. The 61 months KM median survival time difference is more than 50% and almost two years greater than the 39 month difference of the 249 patients classified by stage alone. This means that tGSVD and stage combined make a better predictor than stage alone. (e) The 249 patients classified by both the 7p tGSVD and stage. (f) The 249 patients classified by both the Xq tGSVD and stage. (g) KM curves for the validation set of 148 stage III-IV patients classified by the 6p+12p arraylet correlation, show a median survival time difference of 22 months, with the corresponding log-rank test P-value $< 10^{-2}$, and the univariate Cox proportional hazard ratio 1.9. This validates the survival analyses of the discovery set of 249 patients. (h) Survival analyses of the 148 patients classified by the 7p arraylet correlation. (i) The 148 patients classified by the Xq arraylet correlation.

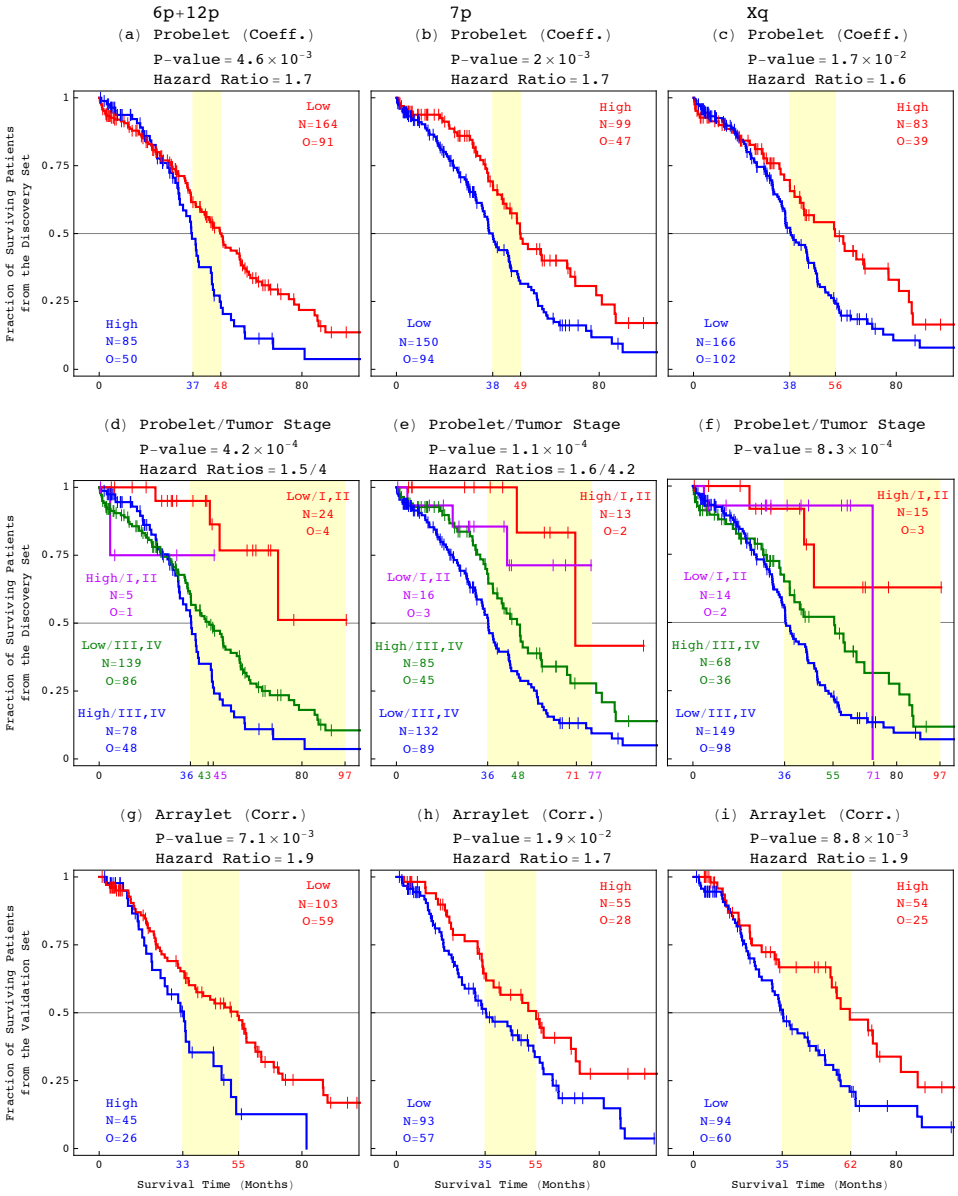

6p+12p
(a) Probelet (Coeff.)
P-value = 4.6 × 10^-3
Hazard Ratio = 1.7
7p
(b) Probelet (Coeff.)
P-value = 2 × 10^-3
Hazard Ratio = 1.7
Xq
(c) Probelet (Coeff.)
P-value = 1.7 × 10^-2
Hazard Ratio = 1.6
(d) Probelet/Tumor Stage
P-value = 4.2 × 10^-4
Hazard Ratios = 1.5/4
(e) Probelet/Tumor Stage
P-value = 1.1 × 10^-4
Hazard Ratios = 1.6/4.2
(f) Probelet/Tumor Stage
P-value = 8.3 × 10^-4
(g) Arraylet (Corr.)
P-value = 7.1 × 10^-3
Hazard Ratio = 1.9
(h) Arraylet (Corr.)
P-value = 1.9 × 10^-2
Hazard Ratio = 1.7
(i) Arraylet (Corr.)
P-value = 8.8 × 10^-3
Hazard Ratio = 1.9
Fraction of Surviving Patients from the Discovery Set
Fraction of Surviving Patients from the Validation Set
Survival Time (Months)
Survival Time (Months)
Survival Time (Months)
Low N=164 O=91
High N=85 O=50
High N=99 O=47
Low N=150 O=94
High N=83 O=39
Low N=166 O=102
Low/I,II N=24 O=4
High/I,II N=5 O=1
Low/III,IV N=139 O=86
High/III,IV N=78 O=48
High/I,II N=13 O=2
Low/I,II N=16 O=3
High/III,IV N=85 O=45
Low/III,IV N=132 O=89
High/I,II N=15 O=3
Low/I,II N=14 O=2
High/III,IV N=68 O=36
Low/III,IV N=149 O=98
Low N=103 O=59
High N=45 O=26
High N=55 O=28
Low N=93 O=57
High N=54 O=25
Low N=94 O=60

Table 3.1: Cox bivariate proportional hazard models of the patients in the discovery and validation sets classified by both tGSVD and the standard OV indicators.

Chromosome Arm	Predictor	Discovery and Validation Sets	
		Hazard Ratio	P-value
6p+12p	tGSVD	1.7	4.4×10^{-4}
	Tumor Stage	3.7	3.9×10^{-3}
	tGSVD	1.6	2.5×10^{-3}
	Residual Disease	2.2	1.2×10^{-4}
	tGSVD	1.7	1.2×10^{-3}
	Therapy Outcome	3.7	1.9×10^{-15}
	tGSVD	1.6	1.2×10^{-3}
	Neoplasm Status	13.0	3.9×10^{-7}
7p	tGSVD	1.7	4.2×10^{-4}
	Tumor Stage	3.9	2.4×10^{-3}
	tGSVD	1.6	1.3×10^{-3}
	Residual Disease	2.2	1.1×10^{-4}
	tGSVD	1.5	1.6×10^{-2}
	Therapy Outcome	3.5	2.4×10^{-14}
	tGSVD	1.7	6.0×10^{-4}
	Neoplasm Status	13.3	3.0×10^{-7}
Xq	tGSVD	1.6	1.7×10^{-3}
	Tumor Stage	3.8	3.2×10^{-3}
	tGSVD	1.9	1.1×10^{-4}
	Residual Disease	2.2	9.3×10^{-5}
	tGSVD	1.8	8.5×10^{-4}
	Therapy Outcome	3.8	1.1×10^{-16}
	tGSVD	1.7	6.7×10^{-4}
	Neoplasm Status	14.5	1.3×10^{-7}

3.3.2 Novel Frequent Focal CNAs Indicating Survival

OV tumors exhibit significant CNA variation among them, much more so than, e.g., glioblastoma brain tumors [3]. Very few frequently occurring OV CNAs have been identified to date. We find that the three tGSVD patterns include most known OV-associated CNAs that map to the corresponding chromosome arms [47], and several previously unreported yet frequent CNAs in >23% of the patients (Figure 3.10).

The 6p+12p arraylet, for example, includes two segments [49] corresponding to the only known OV focal CNAs that map to 6p+12p, 7p or Xq. One, a deletion (6p11.2), overlaps the 3' end unique to isoform a of the DNA primase polypeptide 2-encoding *PRIM2* [68].

Table 3.2: Cox univariate proportional hazard models of the discovery and validation sets of patients classified by either tGSVD or the standard OV indicators.

Predictor		Discovery and Validation Sets	
		Hazard Ratio	P-value
tGSVD	6p+12p	1.8	1.0×10^{-4}
	7p	1.7	1.7×10^{-4}
	Xq	1.7	4.8×10^{-4}
Tumor Stage		4.1	1.8×10^{-3}
Residual Disease		2.3	8.4×10^{-5}
Therapy Outcome		3.8	8.3×10^{-17}
Neoplasm Status		14.0	1.8×10^{-7}

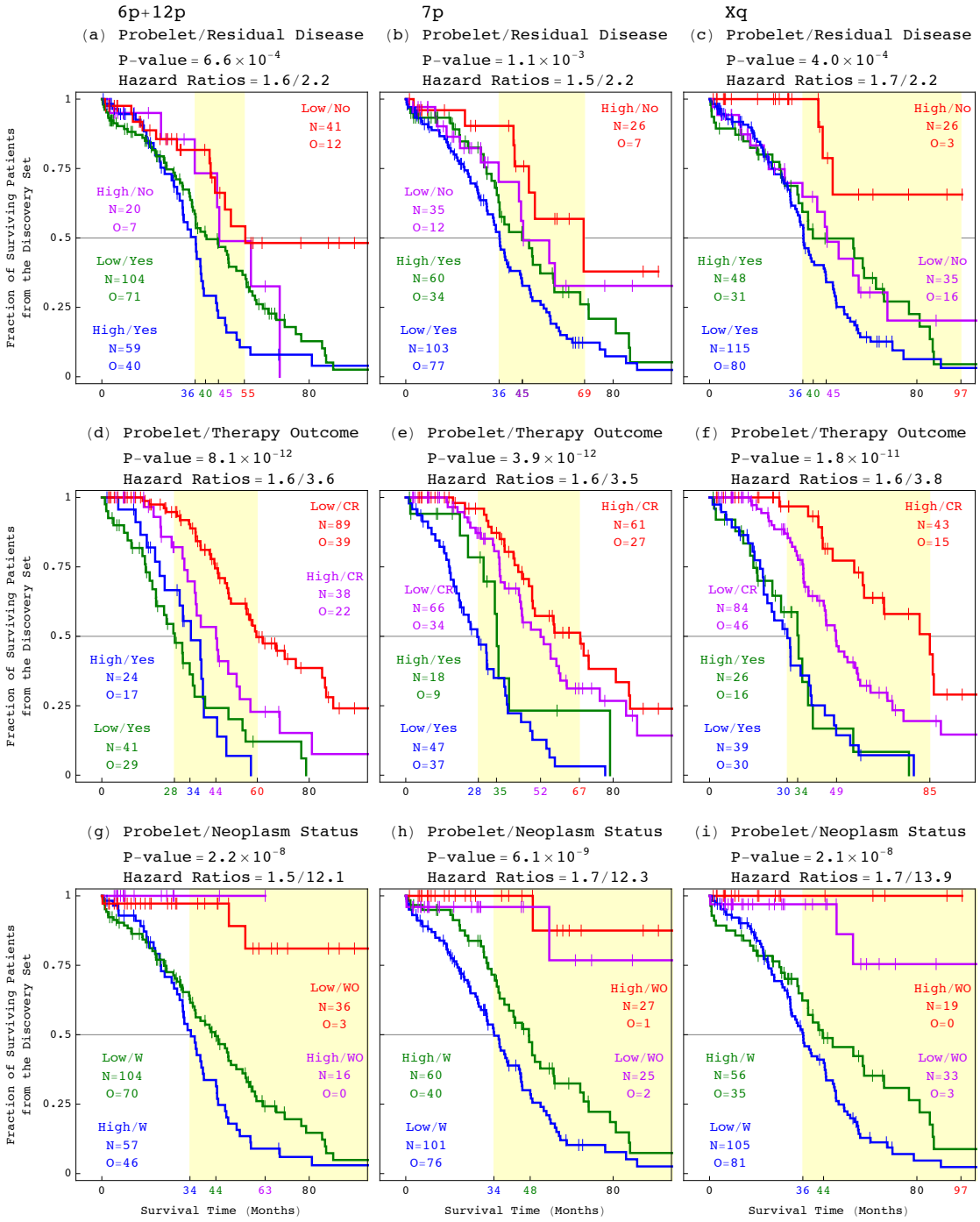


Figure 3.8: Survival analyses of the discovery set of patients classified by tGSVD and standard OV indicators.
KM curves for the discovery set of 249 patients classified by both the (*a*) 6p+12p, (*b*) 7p or (*c*) Xq tGSVD, and residual disease after surgery, the (*d*) 6p+12p, (*e*) 7p or (*f*) Xq tGSVD, and outcome of subsequent therapy, and (*g*) 6p+12p, (*h*) 7p or (*i*) Xq tGSVD, and neoplasm status.

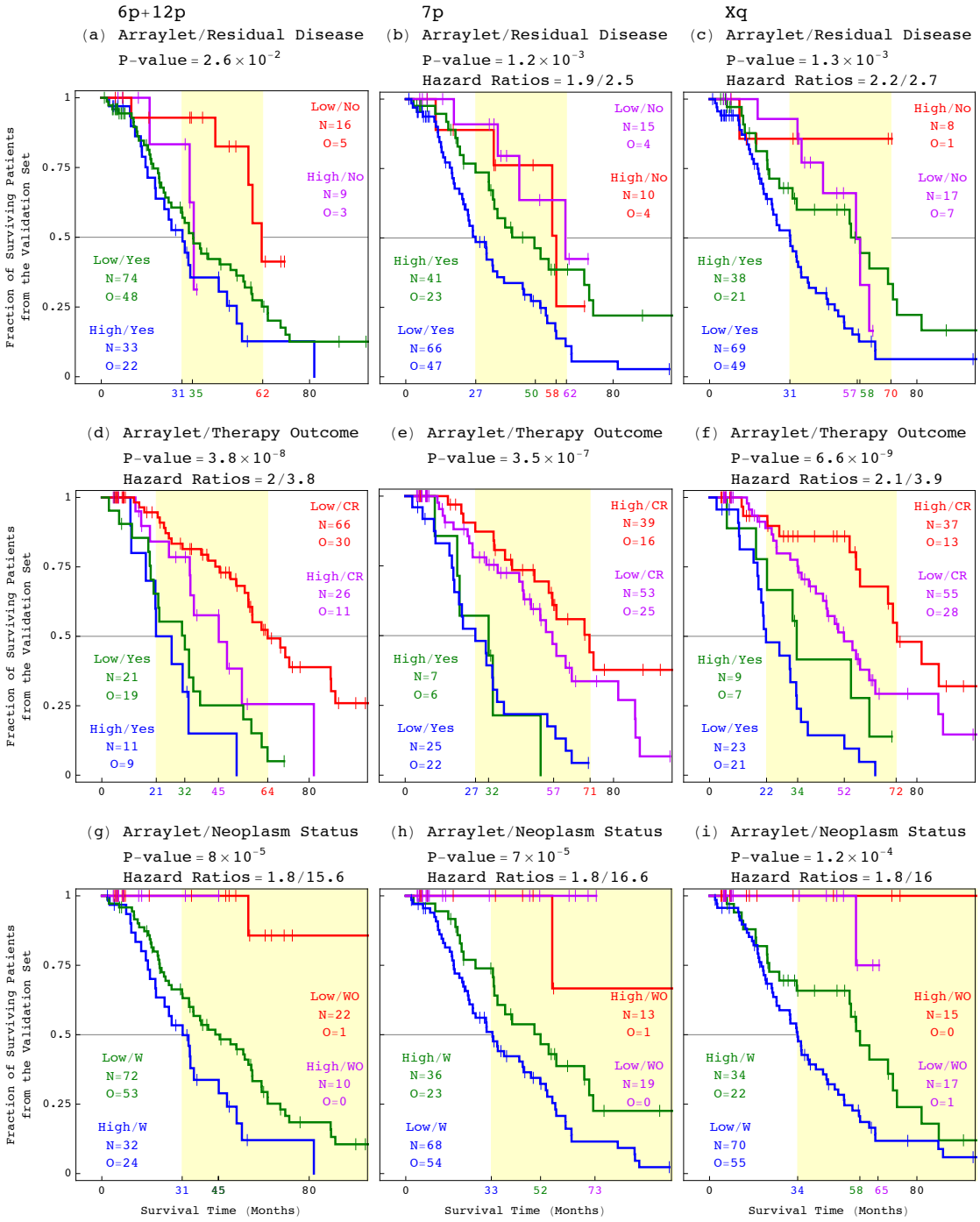
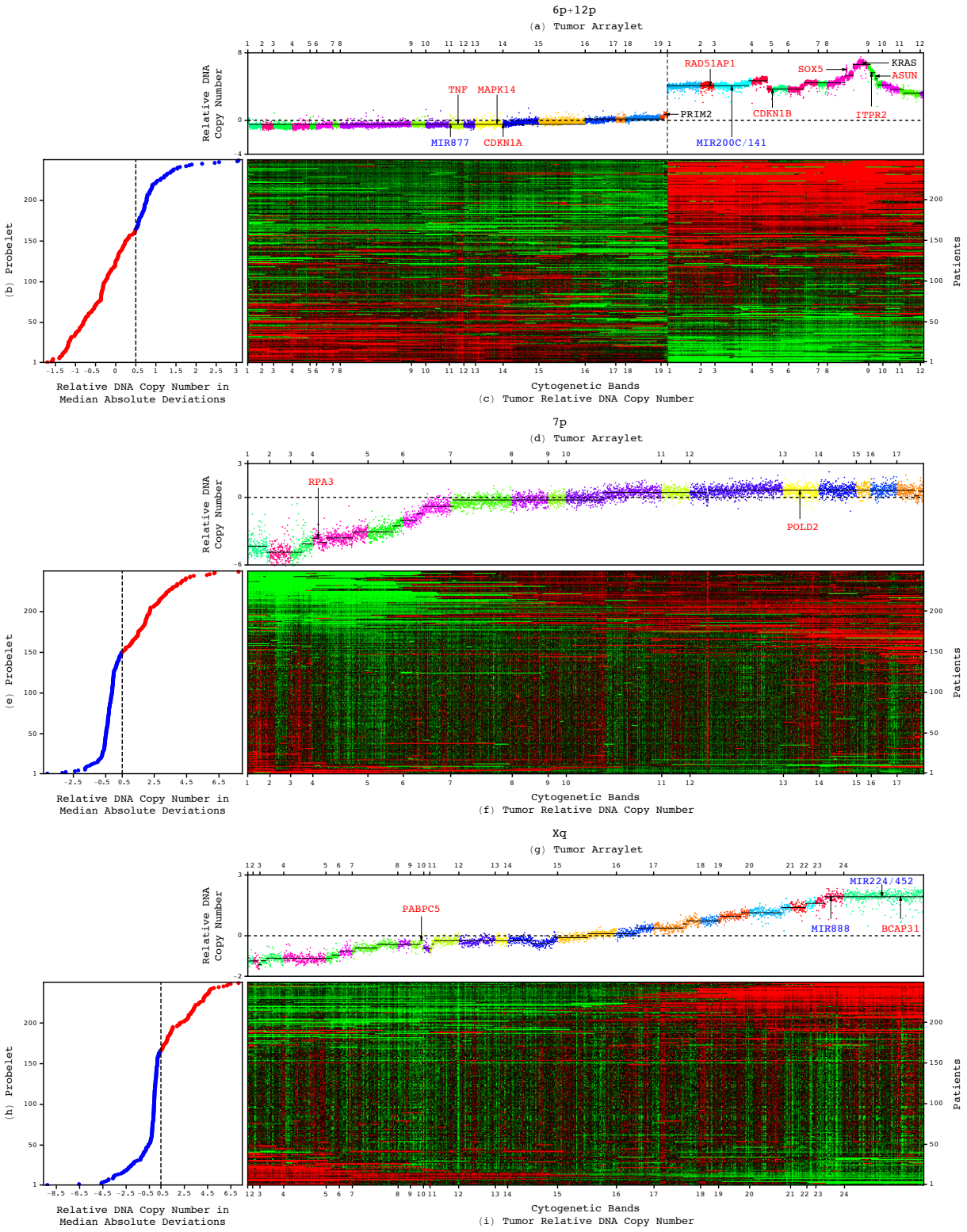


Figure 3.9: Survival analyses of the validation set of patients classified by tGSVD and standard OV indicators.
KM curves for the validation set of 148 stage III-IV patients classified by both the (*a*) 6p+12p, (*b*) 7p or (*c*) Xq tGSVD, and residual disease after surgery, the (*d*) 6p+12p, (*e*) 7p or (*f*) Xq tGSVD, and outcome of subsequent therapy, and (*g*) 6p+12p, (*h*) 7p or (*i*) Xq tGSVD, and neoplasm status.

Figure 3.10: Tumor-exclusive and platform-consistent DNA copy-number alterations (CNAs) correlate with ovarian serous cystadenocarcinoma (OV) patients' survival. (a) Plot of the first 6p+12p tumor arraylet describes a pattern of tumor-exclusive and platform-consistent co-occurring CNAs across the chromosome arm combination of 6p+12p. The probes are ordered, and their copy numbers are colored according to each probe's chromosomal band location. Segments (black lines) amplified and deleted include most known OV-associated CNAs that map to 6p+12p (black), including an amplification of *KRAS* and a deletion of *PRIM2*. CNAs previously unrecognized in OV (red) include a deletion of the p38-encoding *MAPK14*, and p21-encoding *CDKN1A*, and an amplification of *RAD51AP1*, which are drug-targeted in other cancers, a deletion of *TNF*, and focal amplifications of *ASUN*, *ITPR2*, and the 5' ends of isoforms a and e, and exons 5 and 6 of *SOX5*. A high 6p+12p arraylet correlation significantly correlates with a patient's shorter survival time. (b) Plot of the first x-probelet describes the classification of the discovery set of patients into two groups of high (blue) and low (red) coefficients. A high 6p+12p x-probelet coefficient significantly and robustly correlates with a patient's shorter survival time. (c) Raster display of the 6p+12p tumor profiles, where medians of the profiles of the same patient measured by the two platforms were taken, with relative gain (red), no change (black) and loss (green) of DNA copy numbers. (d) Plot of the first 7p tumor arraylet describes a pattern of CNAs across the chromosome arm 7p. CNAs previously unrecognized in OV (red) include a focal deletion of *RPA3* and an amplification of *POLD2*. A high 7p arraylet correlation significantly correlates with a patient's longer survival time. (e) Plot of the first x-probelet describes the classification of the discovery set of patients into two groups of high (red) and low (blue) coefficients. A high 7p x-probelet coefficient significantly and robustly correlates with a patient's longer survival time. (f) Raster display of the 7p tumor profiles. (g) Plot of the first Xq tumor arraylet. CNAs previously unrecognized in OV (red) include a focal deletion of *PABPC5* and an amplification of *BCAP31*. A high Xq arraylet correlation significantly correlates with a patient's longer survival time. (h) Plot of the first x-probelet describes the classification of the discovery set of patients into two groups of high (red) and low (blue) coefficients. A high Xq x-probelet coefficient significantly and robustly correlates with a patient's longer survival time. (i) Raster display of the Xq tumor profiles.


6p+12p
(a) Tumor Arraylet
Relative DNA Copy Number
RAD51AP1
SOX5
KRAS
ASUN
TNF
MAPK14
PRIM2
CDKN1B
ITPR2
MIR877
CDKN1A
MIR200C/141
(b) Probelet
Patients
Relative DNA Copy Number in Median Absolute Deviations
Cytogenetic Bands
(c) Tumor Relative DNA Copy Number
7p
(d) Tumor Arraylet
RPA3
POLD2
(e) Probelet
(f) Tumor Relative DNA Copy Number
Xq
(g) Tumor Arraylet
PABPC5
MIR224/452
MIR888
BCAP31
(h) Probelet
(i) Tumor Relative DNA Copy Number

The other, an amplification (12p12.1-p11.23), contains several genes, including the Kirsten rat sarcoma viral oncogene homolog *KRAS*, one of three human Ras genes and the 5' ends of isoforms b and d of the SRY (sex determining region Y)-box 5-encoding *SOX5* [70], and is significantly (log-rank test P-value <0.05, and KM median survival time difference ≥ 12 months) correlated with OV survival (Dataset S3 in Appendix B).

Novel frequent focal CNAs (segments <125 probes) include four amplifications and two deletions that are significantly correlated with OV survival (Figure 3.11). The amplifications

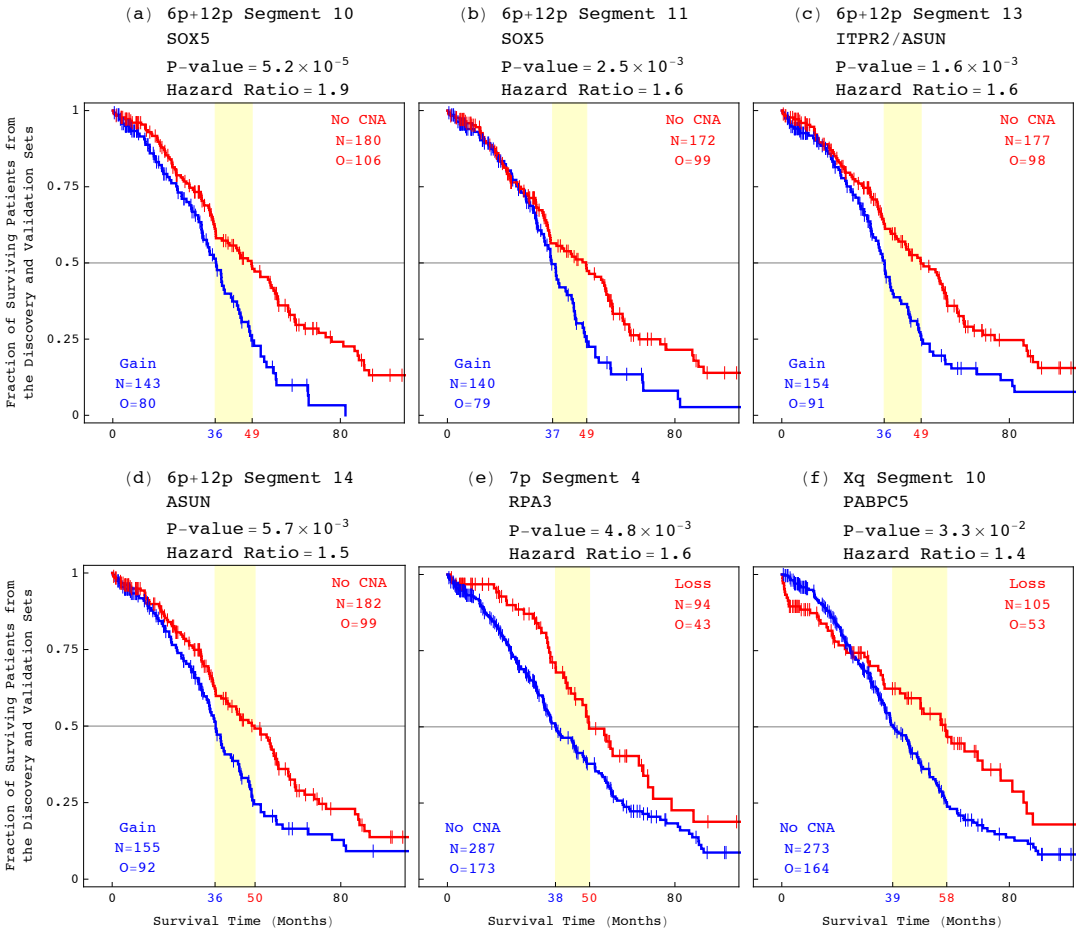


Figure 3.11: Survival analyses of the discovery and validation sets of patients classified by the novel frequent focal CNAs included in the tGSVD arraylets.
Six novel frequent focal CNAs that are included in the tGSVD arraylets are significantly correlated with OV survival. Two amplified consecutive segments (12p12.1) contain (*a*) the 5' ends of isoforms a and e of *SOX5*, and (*b*) exons 5 and 6, the first exons that are common to isoforms a, b, d and e of *SOX5*. Two other amplified consecutive segments (12p11.23) contain (*c*) *ITPR2* and (*d*) *ASUN*. One deletion (7p22.1-p21.3) contains (*e*) *RPA3*. Another deletion (Xq21.31) contains (*f*) *PABPC5*, and the sequence tag site DXS241 adjacent to translocation breakpoints observed in premature ovarian failure.

flank the segment that contains *KRAS*. Two consecutive segments (12p12.1) contain the 5' ends of isoforms a and e of *SOX5*, and exons 5 and 6, the first exons that are common to isoforms a, b, d and e of *SOX5* [71]. Two other consecutive segments (12p11.23) contain the inositol 1,4,5-trisphosphate receptor type 2-encoding *ITPR2*, and the asunder spermatogenesis regulator-encoding *ASUN*. *ASUN* was discovered in a screen of expressed sequence tags on 12p11-p12, which DNA amplification correlated with mRNA overexpression in four human testicular seminomas and one ovarian papillary serous adenocarcinoma cell line, exemplifying human germ cell tumors [72]. *ASUN* and its homologs are essential for nuclear division after DNA replication in the HeLa human cervical cancer cell line, the frog and the fly [73]. One deletion (7p22.1-p21.3) contains the replication protein A3-encoding *RPA3*. The other (Xq21.31) contains the cytoplasmic poly(A)-binding protein 5-encoding *PABPC5*, and the sequence tag site DXS241 adjacent to translocation breakpoints observed in premature ovarian failure [74].

3.3.3 Possible Roles in OV Pathogenesis and Personalized Therapy

To compare the variation in DNA copy numbers with that in gene expression, we used mRNA expression profiles that were available for 394 of the 397 TCGA patients in the discovery and validation sets. Each profile lists the TCGA level 3 mRNA expression for 11,457 autosomal and X chromosome genes on the Affymetrix Human Genome U133A Array platform with UCSC coordinates [47] and GO annotations [75]. Medians of the profiles of samples from the same patient were taken. To examine the possible relations between a tGSVD class and the OV pathogenesis, we assessed the enrichment of the subsets of genes that are differentially expressed between the tGSVD classes in any one of the multiple GO annotations [76]. The P-value of a given enrichment was calculated assuming hypergeometric probability distribution of the annotations among the genes in the global set, and of the subset of annotations among the subset of genes, as previously described [3].

We find that differential mRNA expression between the tGSVD classes is enriched in ontologies that include genes, which consistently map to the CNAs [75,76] (Figure 3.12 and Dataset S4 in Appendix B).

To compare with the variation in microRNA expression, we used microRNA expression profiles that were available for 395 of the 397 patients. Each profile lists the TCGA level 3 microRNA expression for 639 autosomal and X chromosome microRNAs on the Agilent Human microRNA Array 8x15K platform with UCSC coordinates. Medians of the profiles of samples from the same patient were taken. To compare with the variation in protein expression, we used protein expression profiles that were available for 282 of the 397 patients.

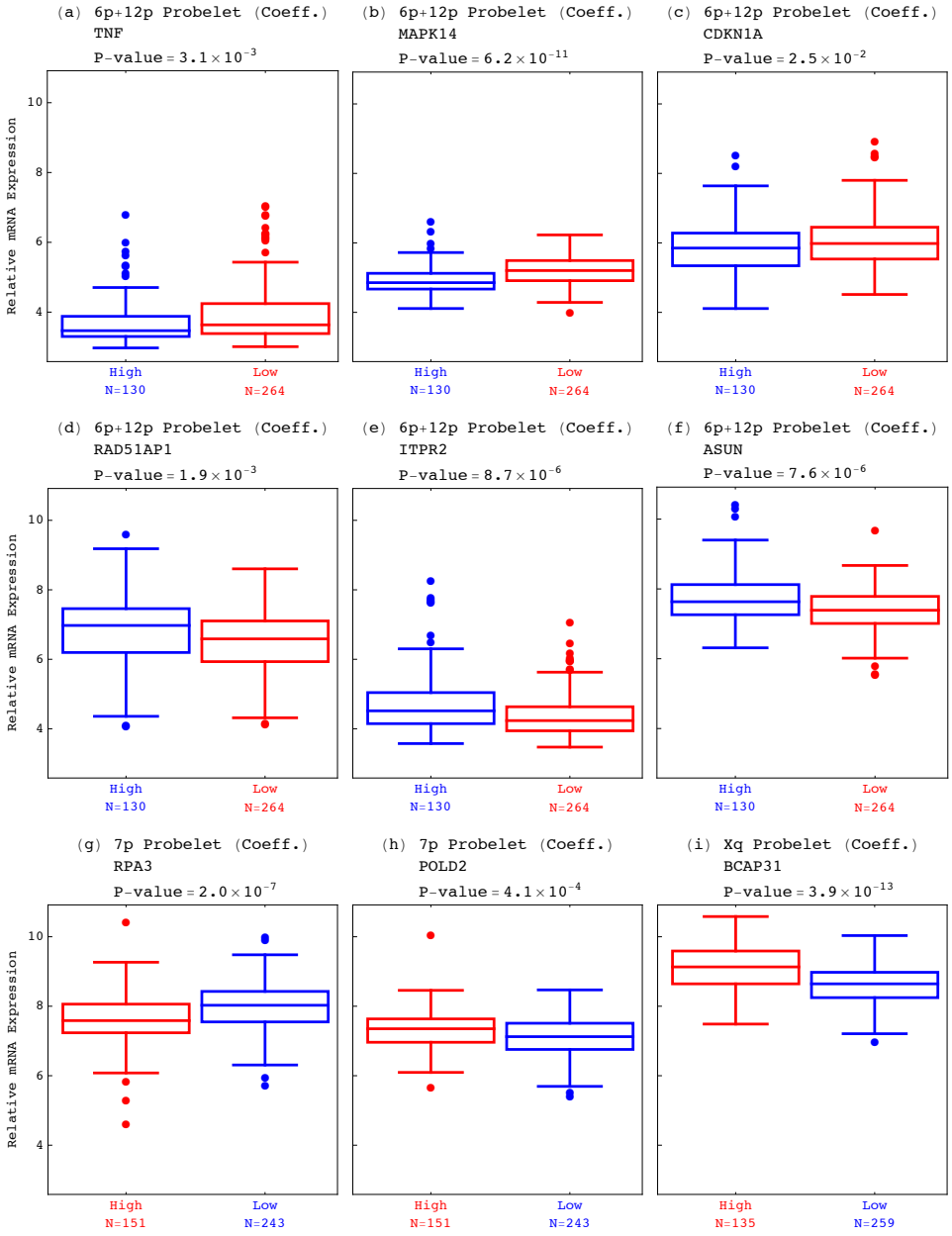


Figure 3.12: Differential mRNA expression between the tGSVD classes is consistent with the CNAs.
(*a*) *TNF*, (*b*) *MAPK14* and (*c*) *CDKN1A*, which are deleted in the 6p+12p arraylet, are significantly (Mann-Whitney P-value <0.05) underexpressed in the tGSVD class of a high 6p+12p x-probelet coefficient, or arraylet correlation relative to the tGSVD class of a low 6p+12p x-probelet coefficient, or arraylet correlation. (*d*) *RAD51AP1*, (*e*) *ITPR2* and (*f*) *ASUN*, which are amplified in the 6p+12p arraylet, are significantly underexpressed in the tGSVD class of a high 6p+12p x-probelet coefficient, or arraylet correlation. (*g*) *RPA3* and (*h*) *POLD2*, which are deleted, and amplified in the 7p arraylet, are significantly underexpressed and overexpressed, respectively, in the tGSVD class of a high 7p x-probelet coefficient, or arraylet correlation. (*i*) *BCAP31*, which is amplified in the Xq arraylet, is significantly overexpressed in the tGSVD class of a high Xq x-probelet coefficient, or arraylet correlation.

Each profile lists the TCGA level 3 protein expression for the 165 antibodies on the MD Anderson Reverse Phase Protein Array (RPPA), which probe for 136 proteins encoded by autosomal and X chromosome genes.

We find that the CNAs are consistent with differential mRNA, microRNA and protein expression between the tGSVD classes (Figures 3.12, 3.13 and 3.14).

The mRNA and protein encoded by, for example, *MAPK14*, which is deleted in the 6p+12p arraylet, are both significantly (Mann-Whitney P-values $<10^{-5}$; please refer to Appendix C.2 for more information) underexpressed in the tGSVD class of a high 6p+12p x-probelet coefficient, or arraylet correlation relative to the tGSVD class of a low 6p+12p x-probelet coefficient, or arraylet correlation. The microRNA mir-877* that maps to the same deletion as *MAPK14* is also significantly (Mann-Whitney P-value <0.05) underexpressed.

A coherent picture emerges for each chromosome arm or combination of two chromosome arms that suggests roles for the DNA CNAs in OV pathogenesis and personalized therapy.

3.3.3.1 *6p+12p*

The genes, which are significantly (Mann-Whitney P-values <0.05) differentially expressed in the patient group of high 6p+12p x-probelet coefficient or arraylet correlation, relative to the group of low coefficient or correlation, are enriched (hypergeometric P-values $<10^{-3}$) in the ontologies of cellular response to ionizing radiation (GO:0071479), and major histocompatibility (MHC) protein complex (GO:0042611). Most of the GO:0071479 genes are underexpressed, including the p21 cyclin-dependent kinase inhibitor-encoding *CDKN1A*, and the p38 mitogen-activated protein kinase-encoding *MAPK14*, which map to a deletion >45 Mbp on the telomeric part of 6p (6p25.3-p21.1). Also underexpressed is p38, the protein encoded by *MAPK14*. All GO:0042611 genes, including the tumor necrosis factor-encoding *TNF*, are underexpressed, and map to the same deletion. The one microRNA that is significantly differentially expressed in the 6p+12p tGSVD classes, and maps to the same deletion, is the splicing-dependent microRNA miR-877*, which is encoded by the 13th intron of the ATP-binding cassette subfamily F member 1-encoding gene *ABCF1* [77]. Both miR-877* and *ABCF1* are consistently underexpressed.

One of only two GO:0071479 overexpressed genes is the RAD51-associated protein 1-encoding *RAD51AP1*, which maps to an amplification >9 Mbp on the telomeric part of 12p (12p13.33-p13.31) that is significantly correlated with OV survival. All four microRNAs that are differentially expressed in the 6p+12p tGSVD classes, and map to the same amplification, miR-200c, miR-200c*, miR-141 and miR-141*, are consistently overexpressed. The second protein that is significantly differentially expressed in the 6p+12p tGSVD

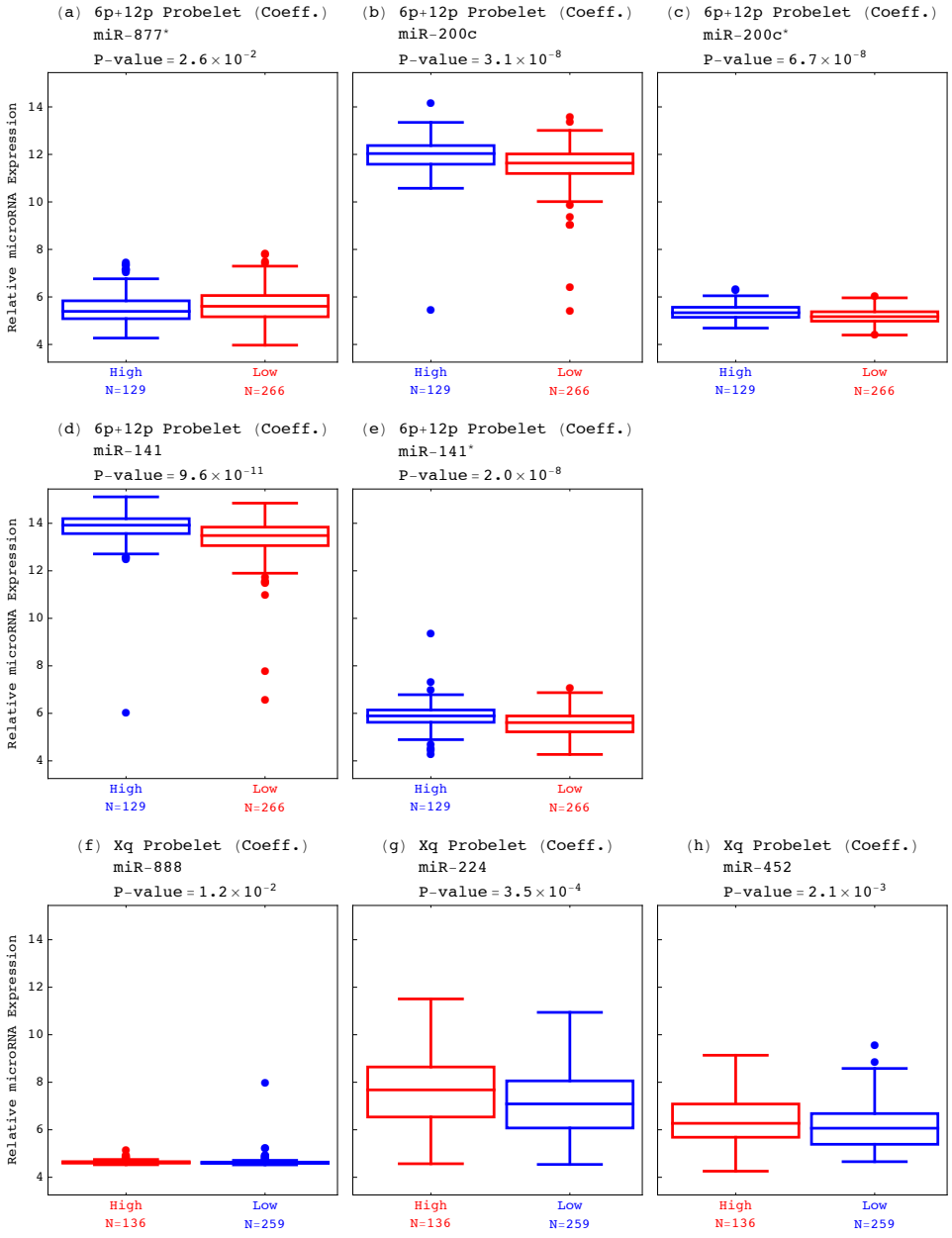


Figure 3.13: Differential microRNA expression between the tGSVD classes is consistent with the CNAs.
(a) mir-877*, which is deleted, and (b) mir-200c, (c) mir-200c*, (d) mir-141 and (e) mir-141*, which are amplified in the 6p+12p arraylet, are significantly (Mann-Whitney P-value <0.05) overexpressed and underexpressed, respectively, in the tGSVD class of a high 6p+12p x-probelet coefficient, or arraylet correlation relative to the tGSVD class of a low 6p+12p x-probelet coefficient, or arraylet correlation. (f) mir-888, (g) mir-224 and (h) mir-452, which are amplified in the Xq arraylet, are significantly overexpressed in the tGSVD class of a high Xq x-probelet coefficient, or arraylet correlation.

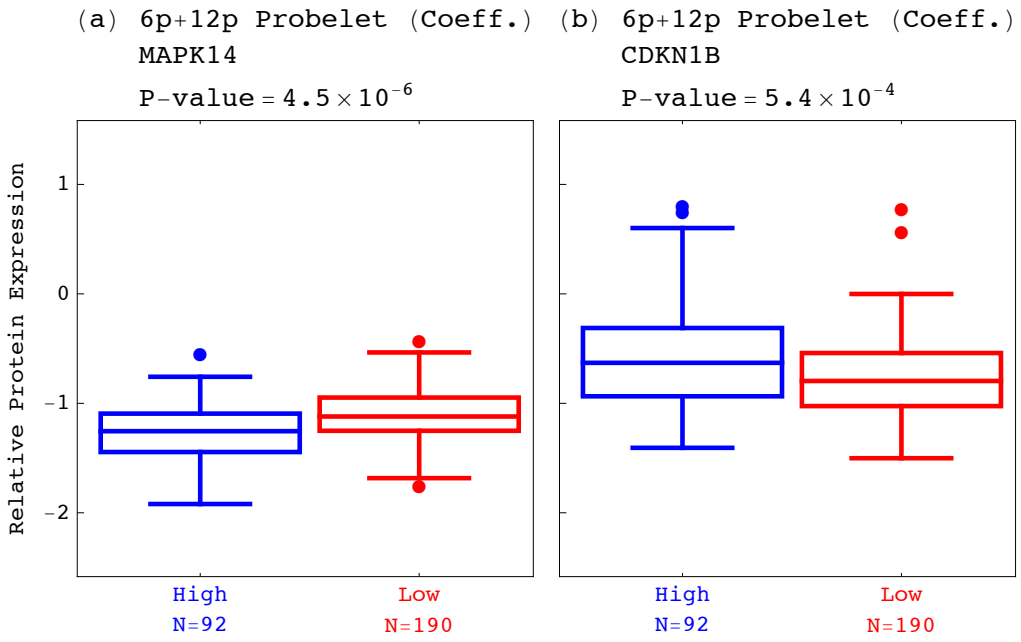


Figure 3.14: Differential protein expression between the tGSVD classes is consistent with the CNAs.
(*a*) *MAPK14*, which is deleted, and (*b*) *CDKN1B*, which is amplified in the 6p+12p arraylet, are significantly (Mann-Whitney P-value <0.05) overexpressed and underexpressed, respectively, in the tGSVD class of a high 6p+12p x-probelet coefficient, or arraylet correlation relative to the tGSVD class of a low 6p+12p x-probelet coefficient, or arraylet correlation.

classes is p27. Consistently, the cyclin-dependent kinase inhibitor *CDKN1B*, which encodes p27, maps to a 4.5 Mbp amplification (12p13.2-p12.3) that is significantly correlated with OV survival, and its mRNA is overexpressed. The mRNA encoded by *KRAS* is also overexpressed.

Note that while the tGSVD 6p+12p pattern of CNAs correlates with survival in the discovery and, separately, validation sets, neither the 6p nor the 12p pattern alone correlates with survival. Indeed, experiments studying the conditions for the transformation of human normal to tumor cells indicate that cells, where both p21 and p38 are inactive, are susceptible to Ras-mediated transformation [78, 79]. However, the activation of Ras alone induces tumor-suppressing cellular senescence via the activities of either p21 or p38. The 6p+12p pattern, therefore, which includes the co-occurring loss of the p21-encoding *CDKN1A* and the p38-encoding *MAPK14* on 6p, and gain of *KRAS* on 12p, encodes for cellular conditions that together, but not separately, can lead to transformation.

In addition, p21 and p38 are necessary for p53-mediated cell cycle arrest [80] and apoptosis [81], respectively, in response to DNA damage. Overexpression of the p21-encoding *CDKN1A* is correlated with a low malignant potential of an ovarian tumor [82]. *RAD51AP1* overexpression disrupts cell cycle arrest and apoptosis, can lead to cellular resistance to DNA-damaging cancer therapies and may increase genomic instability [83]. *TNF*-induced apoptosis is correlated with downregulation of *ITPR2* [84]. Overexpression of miR-200c, and miR-141, both of which putatively target the *BRCA1* associated protein-1 oncosuppressor-encoding *BAP1*, correlates with OV tumor growth, dedifferentiation and invasiveness [85, 86]. Overexpression of the *CDKN1B*-encoded p27, which can promote cellular migration [87] and even proliferation [88], was found to correlate with poorer OV patients' prognosis [89, 90].

Overall, the 6p+12p pattern encodes for cellular conditions that together, but not separately, can induce the transformation of human normal to tumor cells, i.e., deletion of *CDKN1A* and *MAPK14* on 6p, and amplification of *KRAS* on 12p, together with deletion of *TNF* on 6p, and amplification of *RAD51AP1* and *ITPR2* on 12p. These previously unrecognized co-occurring CNAs correlate with a suppression of cell cycle arrest, senescence, and apoptosis in the OV tumor cell, and an OV patient's shorter survival time. Since drugs interacting with *CDKN1A*, *MAPK14*, *RAD51AP1* and *KRAS* exist [91], the 6p+12p tGSVD may prove useful in OV personalized therapy.

3.3.3.2 *7p*

The significantly differentially expressed genes in the 7p tGSVD classes are enriched (hypergeometric P-value $<10^{-10}$) in the ontology of DNA strand elongation involved in DNA replication (GO:0006271). Most of these genes are overexpressed, including the DNA polymerase delta subunit 2-encoding *POLD2* that is essential for DNA replication and repair, which maps to an amplification >17 Mbp on the centromeric part of 7p (7p14.1-p11.2). Only two genes are underexpressed: *RPA3* on 7p and the DNA ligase IV-encoding *LIG4* on 13q. The interaction of p53 with the *RPA3*-encoded protein mediates suppression of HR, the preferred cellular mechanism for DNA DSB repair during replication [92]. *LIG4* is essential for DSB repair via the more error-prone nonhomologous end joining pathway [93]. HR defects are thought to facilitate the genomic disarray among OV tumors [68].

Taken together, previously unrecognized co-occurring deletion and underexpression of *RPA3*, and amplification and overexpression of *POLD2* on 7p correlate with DNA DSB repair via HR during replication, reduced genomic instability and a longer survival time.

3.3.3.3 *Xq*

The differentially expressed genes in the Xq tGSVD classes are enriched (hypergeometric P-value $<10^{-6}$) in the ontology of antigen processing and presentation of peptide antigen (GO:0048002). Most of these genes are overexpressed, including the B-cell receptor-associated protein 31-encoding *BCAP31*, which maps to an amplification >11 Mbp on the telomeric part of Xq (Xq27.3-q28). All three microRNAs that are differentially expressed in the Xq tGSVD classes, and map to the same amplification, miR-888, miR-224 and miR-452, together with the gamma-aminobutyric acid (GABA) A receptor epsilon-encoding *GABRE*, which hosts mir-224 and mir-452 in its introns, are consistently overexpressed. Underexpression of miR-224 was implicated in OV pathogenesis [85]. *PABPC5*, which maps to a focal deletion on Xq, is suppressed upon viral infection [94].

Taken together, previously unrecognized co-occurring deletion of *PABPC5*, and amplification and overexpression of *BCAP31* on Xq correlate with increased cellular immune activity, and longer survival.

Our tGSVD comparisons of patient- and platform-matched OV and normal genomic profiles revealed previously unrecognized links between a tumor's genome and a patient's prognosis, which offer insights into ovarian cancer formation and growth, and suggest targets for personalized drug therapy. Previously, the best prognostic indicator of OV was the tumor's stage at diagnosis.

CHAPTER 4

DISCUSSION

4.1 Summary

4.1.1 The GSVD Comparison of GBM and Normal Genomic Profiles

Previously, Alter et al. [22, 24] showed that the GSVD provides a mathematical framework for sequence-independent comparative modeling of DNA microarray data from two organisms, where the mathematical variables and operations represent experimental or biological reality. The variables, subspaces of significant patterns that are common to both or exclusive to either of the datasets, correlate with cellular programs that are conserved in both or unique to either of the organisms. The operation of reconstruction in the subspaces common to both datasets outlines the biological similarity in the regulation of the cellular programs that are conserved across the species. Reconstruction in the common and exclusive subspaces of either dataset outlines the differential regulation of the conserved relative to the unique programs in the corresponding organism. Recent experimental results [32] verify a computationally predicted genome-wide mode of regulation [19, 31], and demonstrate that GSVD modeling of DNA microarray data can be used to correctly predict previously unknown cellular mechanisms.

Recently, Ponnapalli et al. [23] mathematically defined a higher-order GSVD (HO GSVD) for more than two large-scale matrices with different row dimensions and the same column dimensions. They proved that this novel HO GSVD extends almost all the mathematical properties of the GSVD to higher orders. They showed, comparing global mRNA expression from the three disparate organisms, *S. pombe*, *S. cerevisiae* and human, that the HO GSVD provides a sequence-independent comparative framework for more than two genomic datasets, where the variables and operations represent experimental or biological reality. The approximately common HO GSVD subspace represents biological similarity among the organisms. Simultaneous reconstruction in the common subspace removes the experimental artifacts, which are dissimilar, from the datasets.

We now also show that in a probe-independent comparison of aCGH data from patient-matched tumor and normal samples, the mathematical variables of the GSVD, i.e., shared probelets and the corresponding tumor- and normal-specific arraylets, represent experimental or biological reality. Probelets that are mathematically significant in both datasets correspond to normal arraylets representing copy-number variations (CNVs) in the normal human genome that are conserved in the tumor genome (e.g., female-specific X chromosome amplification) and are represented by the corresponding tumor arraylets. Probelets that are mathematically significant in the normal but not in the tumor dataset represent experimental variations that exclusively affect the normal dataset. Similarly, some probelets that are mathematically significant in the tumor but not in the normal dataset represent experimental variations that exclusively affect the tumor dataset.

We find that the mathematically second most tumor-exclusive probelet, which is also the mathematically most significant probelet in the tumor dataset, is statistically correlated, possibly biologically coordinated with GBM patients' survival and response to chemotherapy. The corresponding tumor arraylet describes a global pattern of tumor-exclusive co-occurring CNAs, including most known GBM-associated changes in chromosome numbers and focal CNAs, as well as several previously unreported CNAs, including the biochemically putative drug target-encoding *TLK2* [44]. We find that a negligible weight of the second tumor arraylet in a patient's GBM aCGH profile, mathematically defined by either the corresponding copy number in the second probelet, or by the correlation of the GBM profile with the second arraylet, is indicative of a significantly longer GBM survival time. This GSVD comparative modeling of aCGH data from patient-matched tumor and normal samples, therefore, draws a mathematical analogy between the prediction of cellular modes of regulation and the prognosis of cancers.

We confirm our results with GSVD comparison of matched profiles of a larger set of TCGA patients, inclusive of the initial set. We validate the prognostic contribution of the pattern with GSVD classification of the GBM profiles of a set of patients that is independent of both the initial set and the inclusive confirmation set [45].

4.1.2 The Tensor GSVD Comparisons of Matched OV Genomic Profiles

In personalized medicine, the growing numbers of large-scale multidimensional datasets recording different aspects of a single disease, e.g., in TCGA [68], promise to enhance basic biological understanding of the disease, lead to the development of new therapies and inform a patient's diagnosis, prognosis and treatment. This rapid growth in biomedical datasets is accompanied by a fundamental need for mathematical frameworks that can create one

coherent model from multiple datasets arranged in multiple tensors of matched columns, e.g., patients, platforms and tissues, but independent rows, e.g., probes. The recent HO GSVD is the only simultaneous decomposition to date of more than two such datasets that is by definition exact, and which mathematical properties allow interpreting its variables and operations in terms of, e.g., biomedical reality [23, 101]. This decomposition, however, is limited to datasets arranged in second-order tensors, i.e., matrices.

We define a novel tensor GSVD (tGSVD), an exact simultaneous decomposition of two such datasets, arranged in two higher-than-second-order tensors of the same column dimensions but different row dimensions. We prove that the tGSVD extends the matrix GSVD [3, 21, 22, 25, 95–98] and the tensor higher-order singular value decomposition (HOSVD) [31,32,99] from two matrices and one tensor, respectively, to two tensors [100].We show that the tGSVD can simultaneously find the similarities and dissimilarities, i.e., patterns of varying relative significance, in one dataset relative to the other. The mathematical properties of the tGSVD allow interpreting the patterns in terms of the biomedical similarities and dissimilarities between the two datasets.

We demonstrate the tGSVD in comparisons of patient- and platform-matched but probe-independent genomic profiles of ovarian serous cystadenocarcinoma (OV) tumor and normal samples from TCGA. The tGSVD reveals chromosome arm-wide patterns of tumor-exclusive and platform-consistent DNA copy-number alterations (CNAs) that correlate with OV patients' survival. The patterns, across 6p+12p, 7p and Xq, are independent of the tumor's stage, the best predictor of OV survival to date, and include known as well as previously unreported, yet frequent CNAs. Differential mRNA expression between the tGSVD classes is enriched in ontologies that include genes, that consistently map to the DNA CNAs [75, 76]. Differential microRNA and protein expression also consistently map to the CNAs [77]. Taken together, these patterns revealed by tGSVD suggest roles for the CNAs in OV pathogenesis and therapy.

Unlike previous analyses, notably by TCGA Research Network [37,68], our tGSVD and GSVD analyses were not limited to the 22 human autosomal chromosomes and included the X chromosome because the tGSVD or the matrix GSVD does not make any apriori assumptions about the data. Our analyses, therefore, can be used to create a single coherent mathematical model that simultaneously finds the similarities and dissimilarities between any two datasets arranged in (*i*) two higher-than-second-order tensors (tGSVD) or (*ii*) two second-order tensors or matrices (GSVD) of the same column dimensions but different row dimensions.

4.2 Future Directions

4.2.1 Additional Applications in Personalized Medicine

Our mathematical frameworks, GSVD and tGSVD, are capable of identifying the *similarities as well as the dissimilarities* between datasets, which is very important in personalized medicine because such identification allows us the flexibility to address different types of biological questions. For example, when comparing tumor and normal genomic profiles using GSVD, one might be interested in what is *exclusive* to tumor and contributing to patient prognosis. However, when one is comparing tumor genomic profiles from two different measuring platforms using GSVD, the same mathematical framework as mentioned in the previous example, the question will change to what is *common* between the two profiling platforms that is attributed to "true biological signal" and contributing to patient prognosis.

Additional possible applications of the GSVD, HO GSVD and tGSVD in personalized medicine include comparisons of multiple patient-matched datasets (for GSVD and HO GSVD comparisons) or two patient- and platform-matched datasets (for tGSVD comparisons), each corresponding to (*i*) a set of large-scale molecular biological profiles (such as DNA copy numbers) acquired by a high-throughput technology, e.g., DNA microarrays from the same tissue type (such as tumor or normal); (*ii*) a set of biomedical images or signals; or (*iii*) a set of anatomical or clinical pathology test results or phenotypical observations (such as age or tumor stage). For example, tGSVD comparisons of tumor and normal profiles of patient- and platform-matched Single Nucleotide Polymorphism (SNP) data measured using microarray technology and DNA sequencing technology from the same set of TCGA breast invasive carcinoma patients can reveal the biological phenomena common and exclusive to tumor and normal datasets.

GSVD and tGSVD comparisons can uncover the relations and possibly even causal coordinations among these different recorded aspects of the same medical phenomenon. GSVD and tGSVD comparisons can be used to determine a single patient's medical status in relation to all other patients in the set and inform the patient's diagnosis, prognosis and treatment.

APPENDIX A

SUPPLEMENT I

A.1 Figures

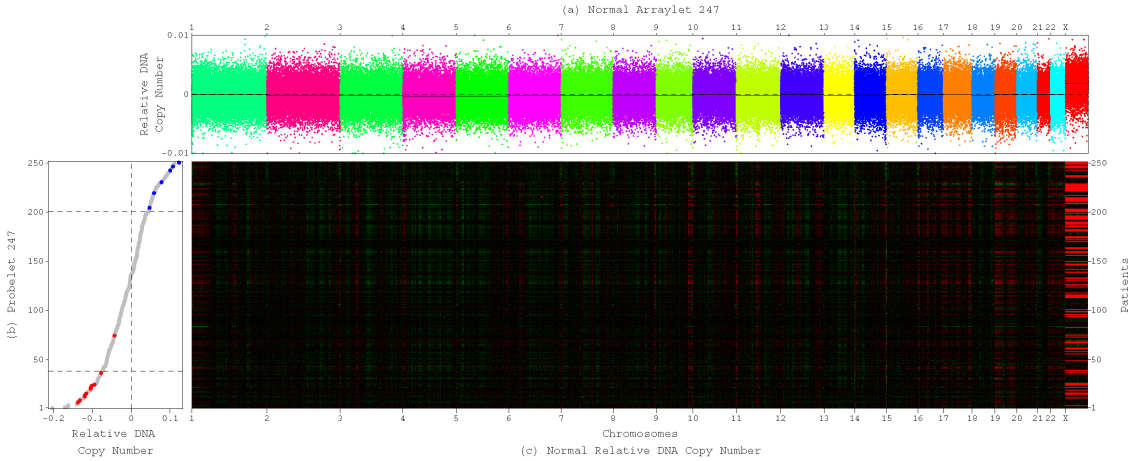


Figure A.1: The 247th, normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD.
(a) Plot of the 247th normal arraylet describes copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. The normal probes are ordered, and their copy numbers are colored, according to each probe's chromosomal location. (b) Plot of the 247th probelet describes the corresponding variation across the patients. Copy numbers in this probelet correlate with the date of hybridization of the normal samples, 7.22.2009 (red), 10.8.2009 (blue) or other (gray), with the P-values $<10^{-3}$ (Table 2.1 and Figure 2.4b). (c) Raster display of the normal dataset shows the correspondence between the normal profiles and the 247th probelet and normal arraylet.

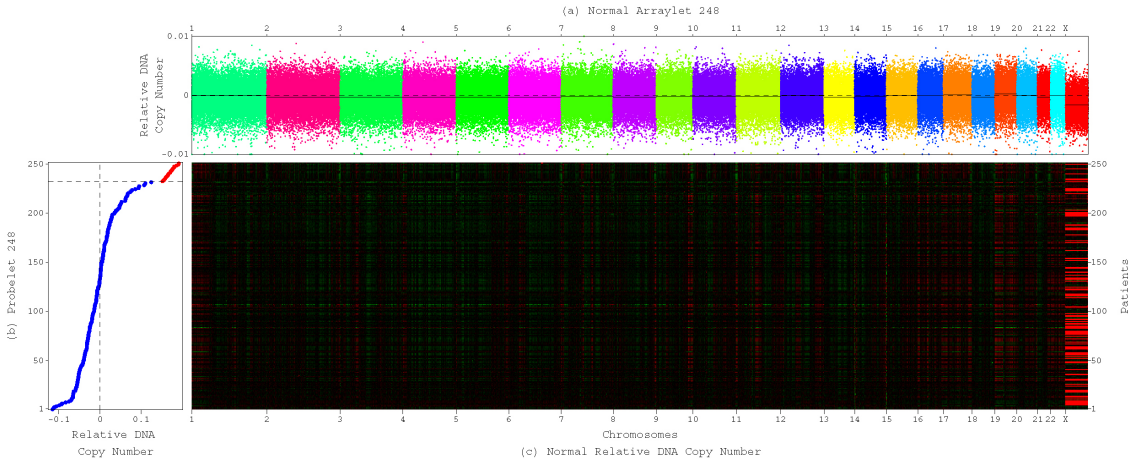


Figure A.2: The 248th, normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD.
(a) Plot of the 248th normal arraylet describes copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. (b) Plot of the 248th probelet describes the corresponding variation across the patients. Copy numbers in this probelet significantly correlate with the tissue batch/hybridization scanner of the normal samples, HMS 8/2331 (red) and other (gray), with the P-values $<10^{-12}$ (Table 2.1 and Figure 2.4c). (c) Raster display of the normal dataset shows the correspondence between the normal profiles and the 248th probelet and normal arraylet.

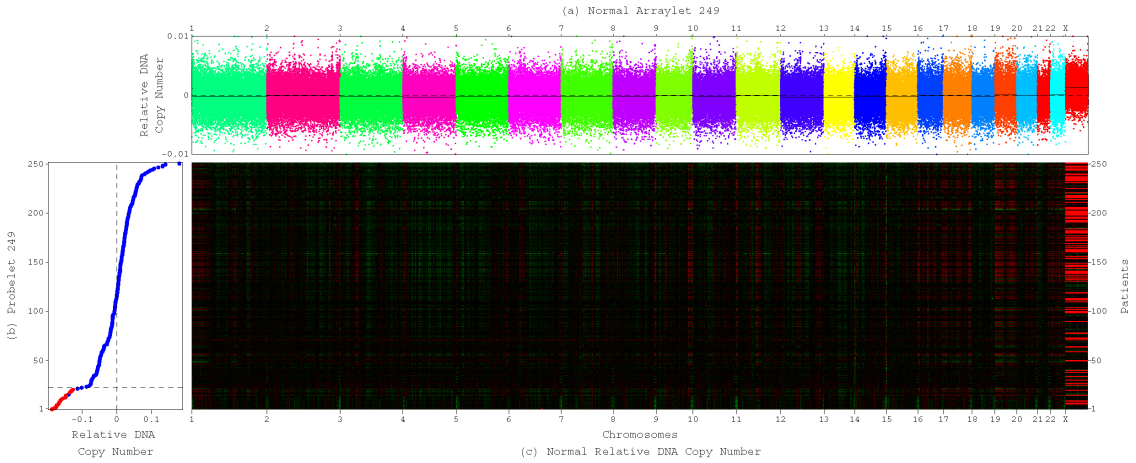


Figure A.3: The 249th, normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD.
(a) Plot of the 249th normal arraylet describes copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. (b) Plot of the 249th probelet describes the corresponding variation across the patients. Copy numbers in this probelet significantly correlate with the tissue batch/hybridization scanner of the normal samples, HMS 8/2331 (red) and other (gray), with the P-values $<10^{-12}$ (Table 2.1 and Figure 2.4d). (c) Raster display of the normal dataset shows the correspondence between the normal profiles and the 249th probelet and normal arraylet.

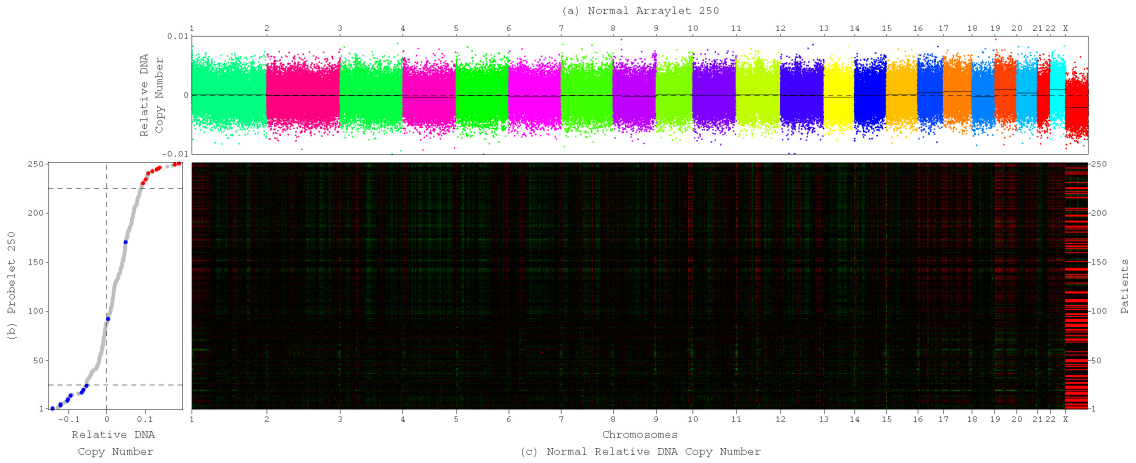


Figure A.4: The 250th, normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD.
(a) Plot of the 250th normal arraylet describes copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. (b) Plot of the 250th probelet describes the corresponding variation across the patients. Copy numbers in this probelet correlate with the date of hybridization of the normal samples, 4.18.2007 (red), 7.22.2009 (blue) or other (gray), with the P-values $<10^{-3}$ (Table 2.1 and Figure 2.4e). (c) Raster display of the normal dataset shows the correspondence between the normal profiles and the 250th probelet and normal arraylet.

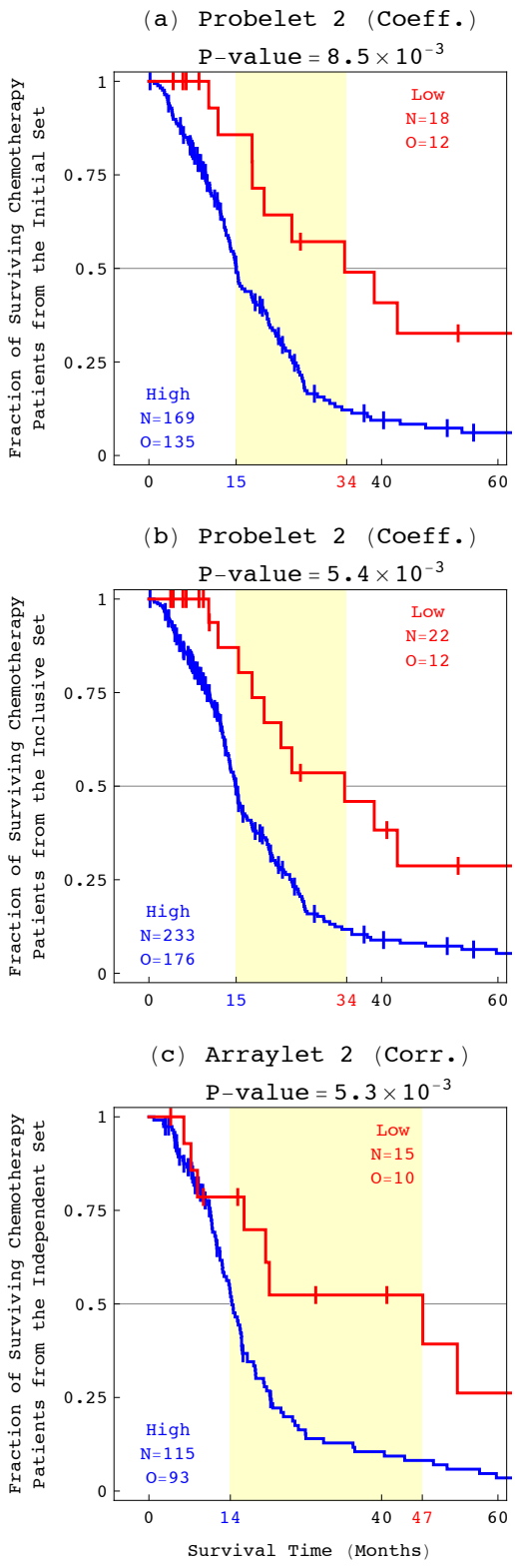


Figure A.5: Kaplan-Meier (KM) survival analyses of only the chemotherapy patients from the three sets classified by GSVD.

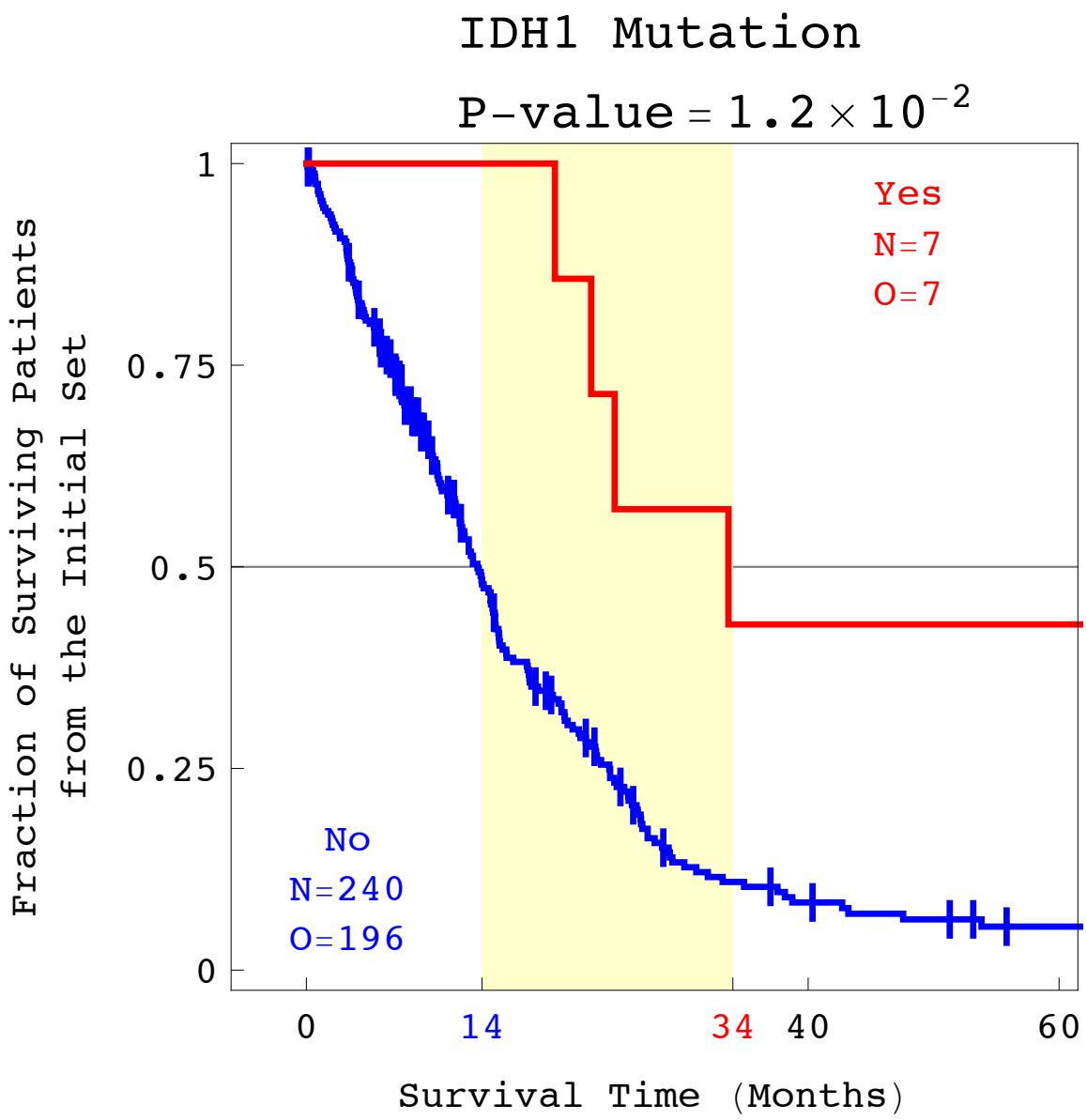


Figure A.6: KM survival analysis of the initial set of 251 patients classified by a mutation in the gene *IDH1*.

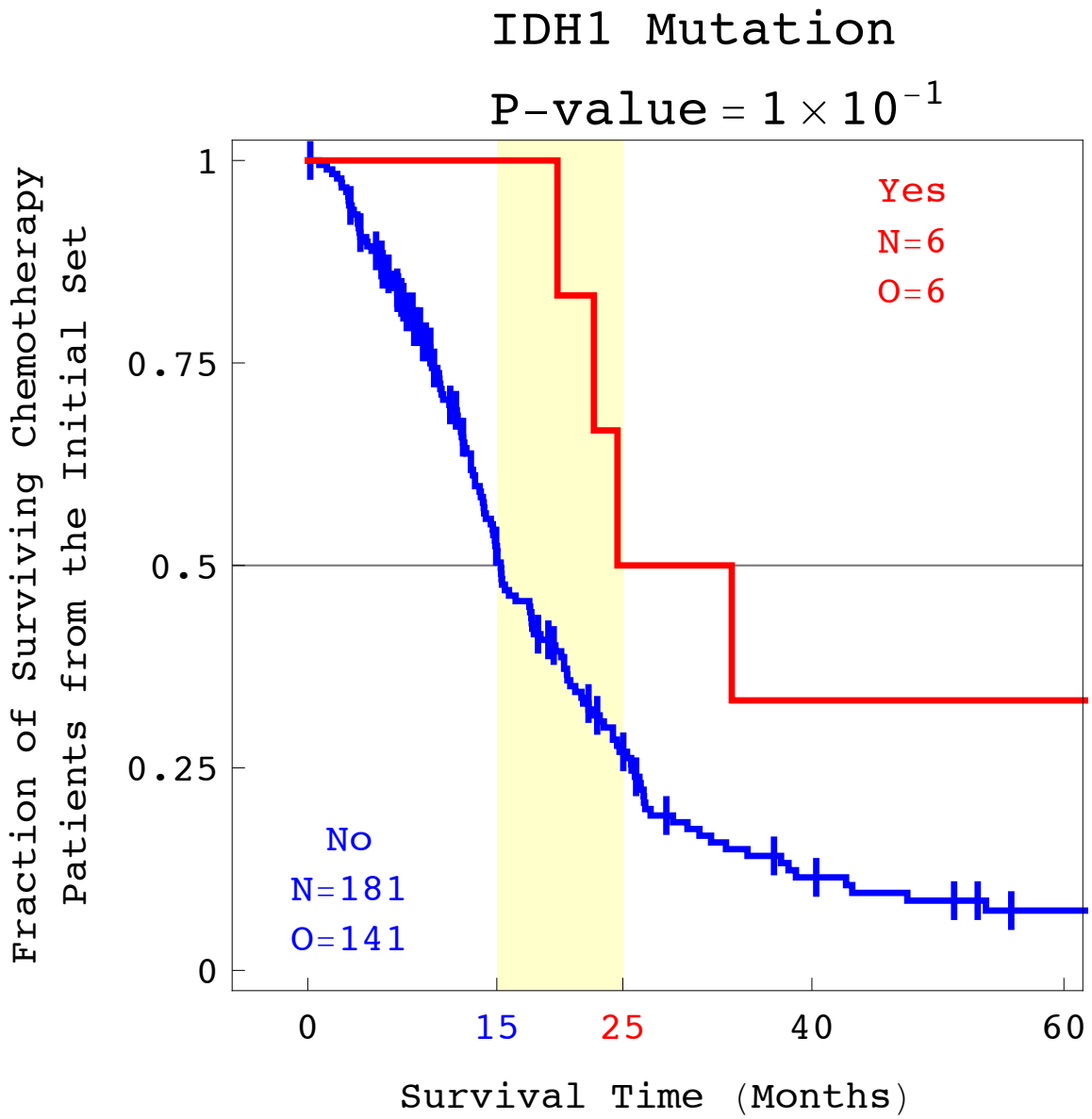


Figure A.7: KM survival analysis of only the chemotherapy patients in the initial set classified by a mutation in *IDH1*.

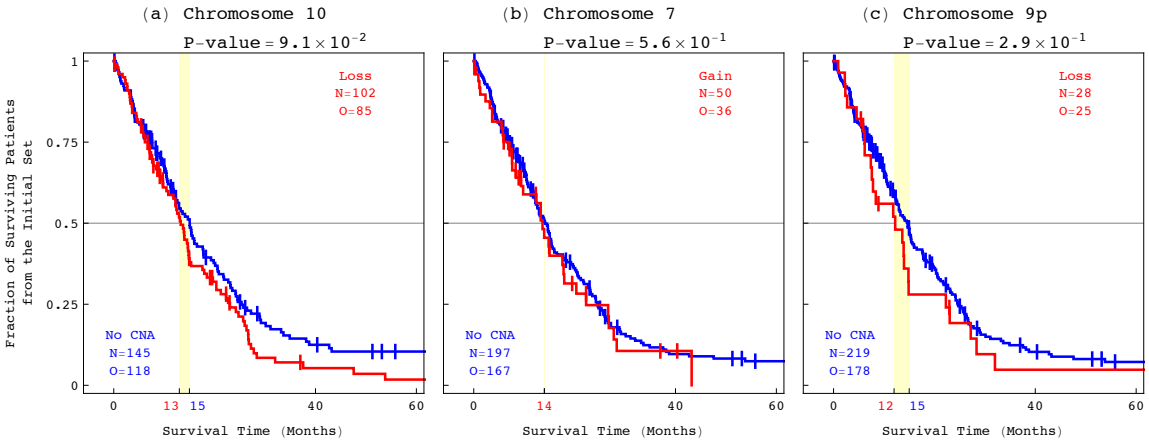
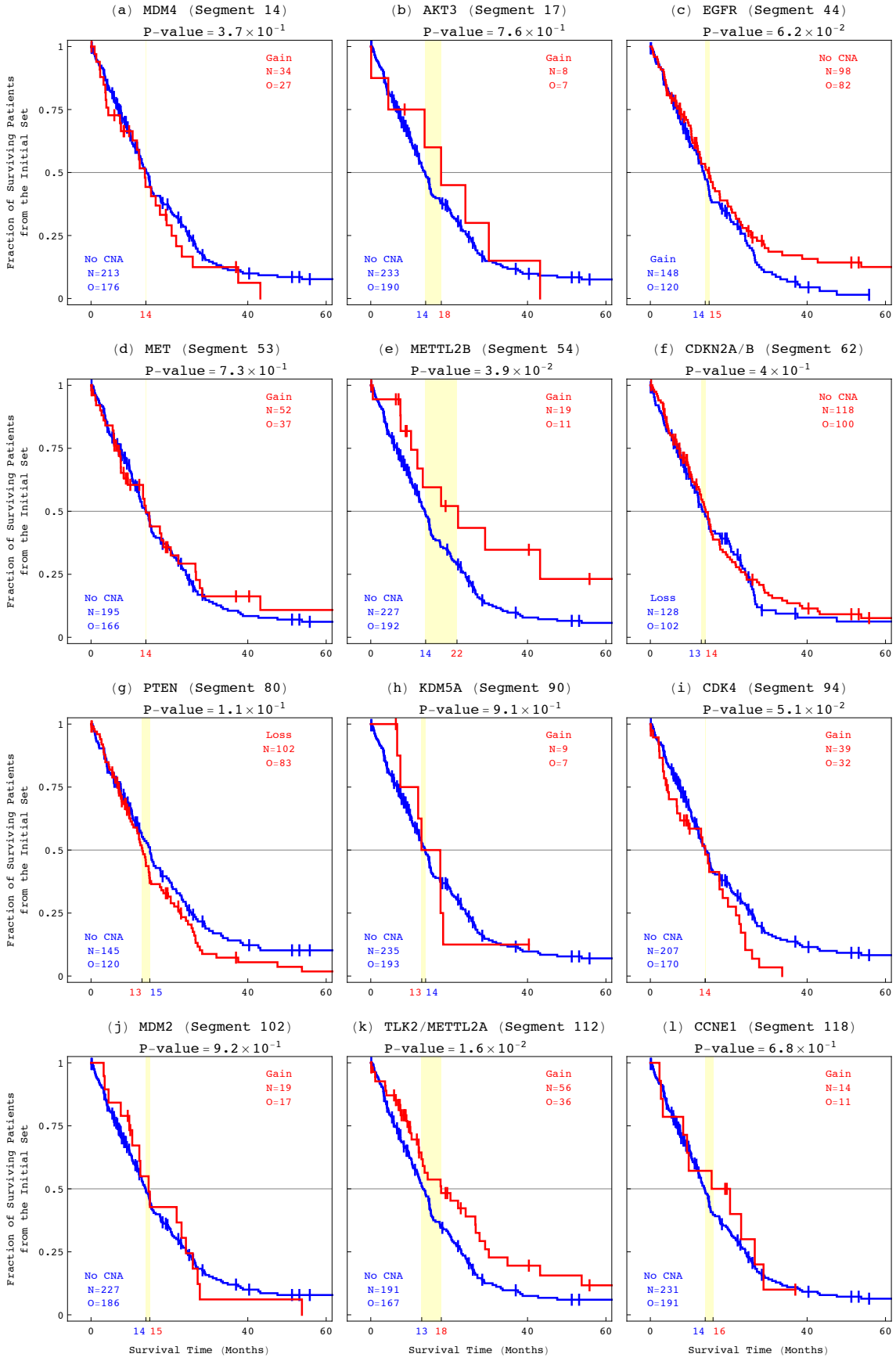


Figure A.8: KM survival analyses of the initial set of 251 patients classified by GBM-associated chromosome number changes.
(a) Analysis of the 247 patients with TCGA annotations in the initial set of 251 patients, classified by number changes in chromosome 10, shows almost overlapping Kaplan-Meier (KM) curves with a KM median survival time difference of $\sim$2 months, and a corresponding log-rank test P-value $\sim 10^{-1}$, meaning that chromosome 10 loss, frequently observed in GBM, is a poor predictor of GBM patients' survival. (b) KM survival analysis of the 247 patients classified by number changes in chromosome 7 shows almost overlapping KM curves with a KM median survival time difference of <one month, and a corresponding log-rank test P-value $>5\times 10^{-1}$, meaning that chromosome 7 gain is a poor predictor of GBM survival. (c) KM survival analysis of the 247 patients classified by number changes in chromosome 9p shows a KM median survival time difference of $\sim$3 months, and a log-rank test P-value $>10^{-1}$, meaning that chromosome 9p loss is a poor predictor of GBM survival.

Figure A.9: KM survival analyses of the initial set of 251 patients classified by copy number changes in selected segments containing GBM-associated genes or genes previously unrecognized in GBM. In the KM survival analyses of the groups of patients with either a CNA or no CNA in either one of the 130 segments identified by the global pattern, i.e., the second tumor-exclusive arraylet (Dataset S3), log-rank test P-values $<5\times10^{-2}$ are calculated for only 12 of the classifications. Of these, only six correspond to a KM median survival time difference that is $\gtrsim$5 months, approximately a third of the $\sim$16 months difference observed for the GSVD classification. One of these segments contains the genes *TLK2* and *METTL2A*, previously unrecognized in GBM. The KM median survival time we calculate for the 56 patients with *TLK2* amplification is $\sim$5 months longer than that for the remaining patients. This suggests that drug-targeting the kinase and/or the methyltransferase-like protein that *TLK2* and *METTL2A* encode, respectively, may affect not only the pathogenesis but also the prognosis of GBM.


(a) MDM4 (Segment 14)
P-value = 3.7 × 10^-1
Gain N=34 O=27
No CNA N=213 O=176
(b) AKT3 (Segment 17)
P-value = 7.6 × 10^-1
Gain N=8 O=7
No CNA N=233 O=190
(c) EGFR (Segment 44)
P-value = 6.2 × 10^-2
No CNA N=98 O=82
Gain N=148 O=120
(d) MET (Segment 53)
P-value = 7.3 × 10^-1
Gain N=52 O=37
No CNA N=195 O=166
(e) METTL2B (Segment 54)
P-value = 3.9 × 10^-2
Gain N=19 O=11
No CNA N=227 O=192
(f) CDKN2A/B (Segment 62)
P-value = 4 × 10^-1
No CNA N=118 O=100
Loss N=128 O=102
(g) PTEN (Segment 80)
P-value = 1.1 × 10^-1
Loss N=102 O=83
No CNA N=145 O=120
(h) KDM5A (Segment 90)
P-value = 9.1 × 10^-1
Gain N=9 O=7
No CNA N=235 O=193
(i) CDK4 (Segment 94)
P-value = 5.1 × 10^-2
Gain N=39 O=32
No CNA N=207 O=170
(j) MDM2 (Segment 102)
P-value = 9.2 × 10^-1
Gain N=19 O=17
No CNA N=227 O=186
(k) TLK2/METTL2A (Segment 112)
P-value = 1.6 × 10^-2
Gain N=56 O=36
No CNA N=191 O=167
(l) CCNE1 (Segment 118)
P-value = 6.8 × 10^-1
Gain N=14 O=11
No CNA N=231 O=191
Fraction of Surviving Patients from the Initial Set
Survival Time (Months)

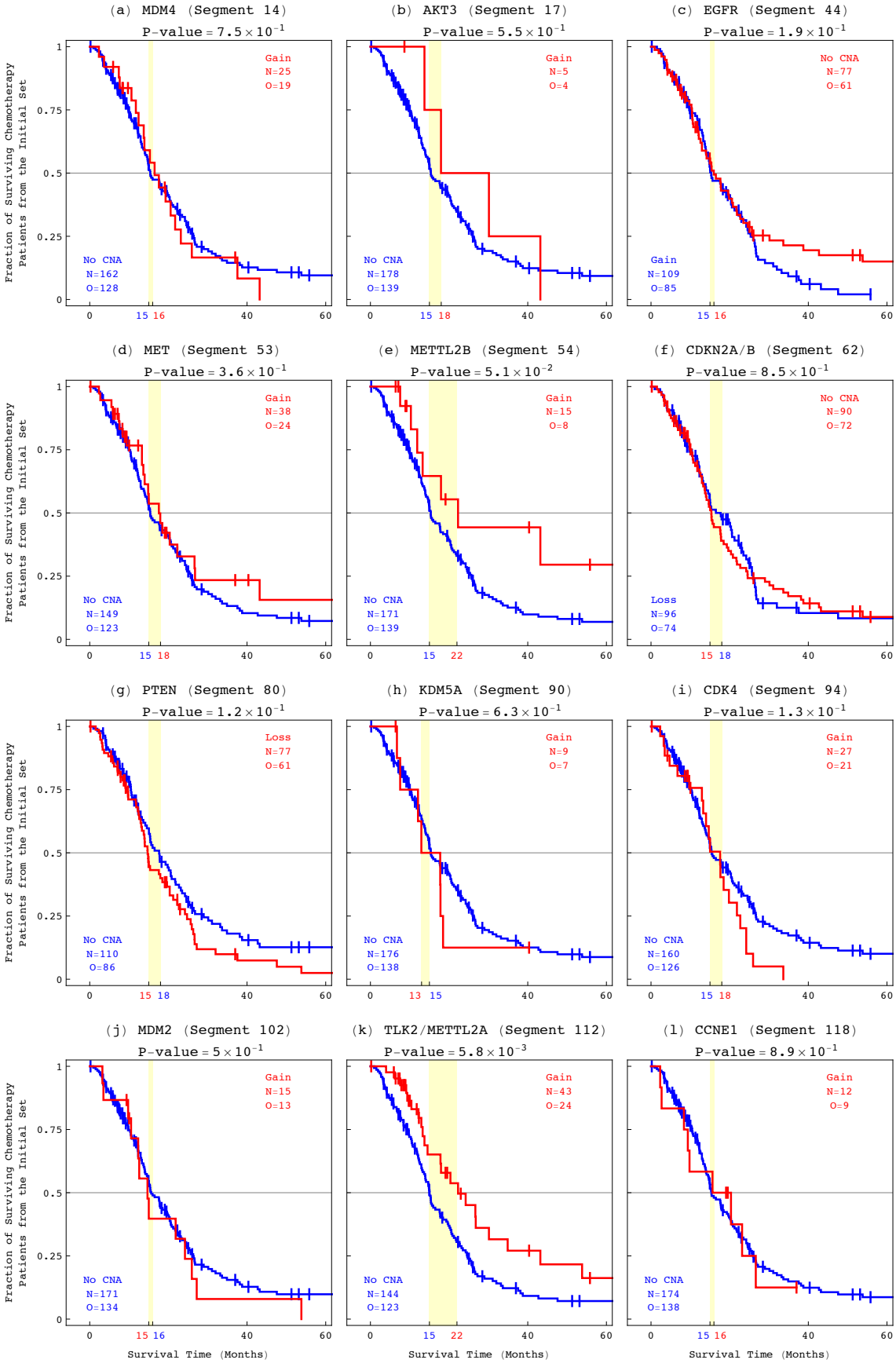
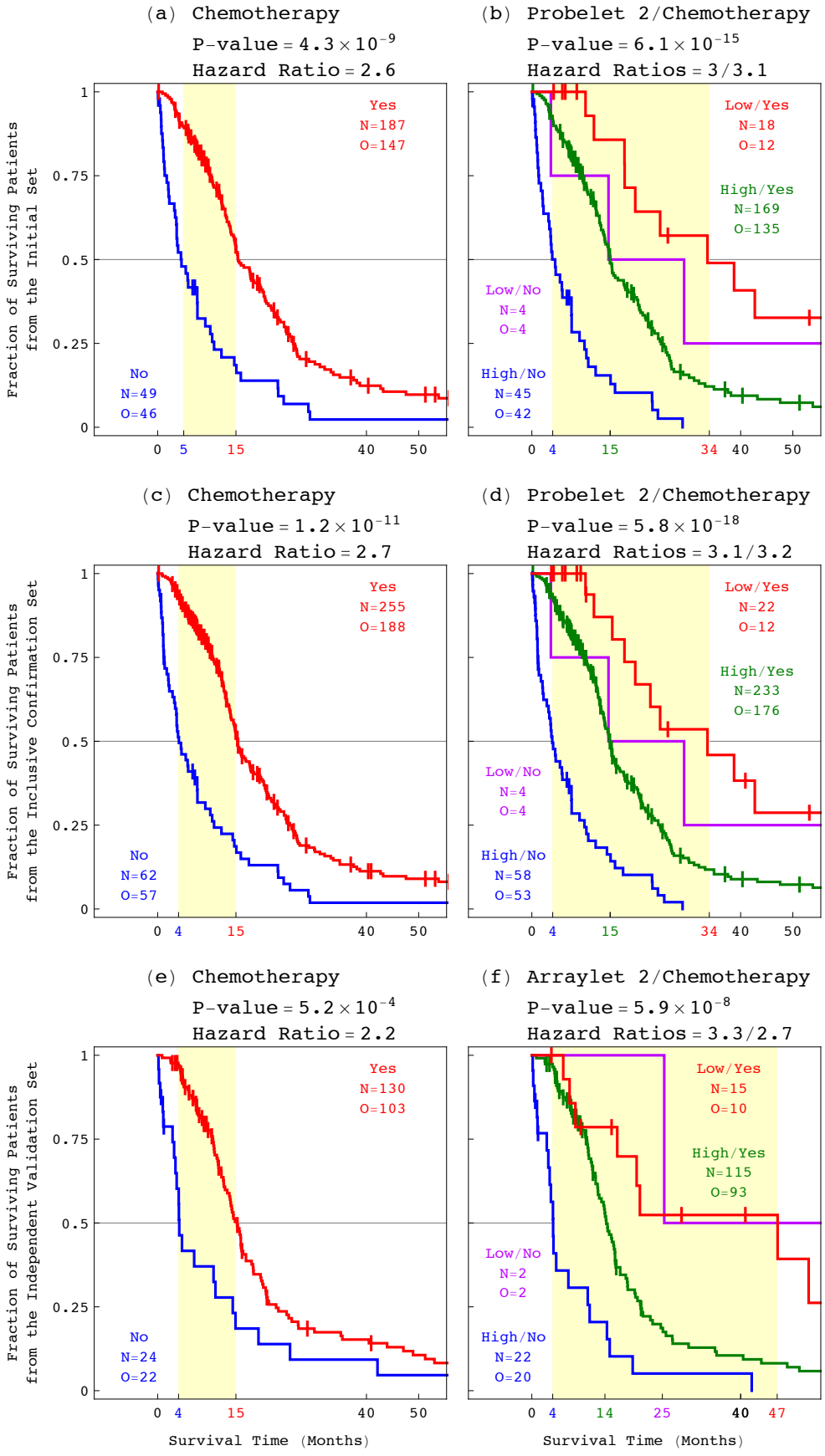


Figure A.10: KM survival analyses of only the chemotherapy patients in the initial set of 251 patients classified by copy number changes in selected segments.

Figure A.11: Survival analyses of the patients from the three sets classified by chemotherapy alone or GSVD and chemotherapy both. (a) KM and Cox survival analyses of the 236 patients with TCGA chemotherapy annotations in the initial set of 251 patients, classified by chemotherapy, show that lack of chemotherapy, with a KM median survival time difference of ∼10 months and a univariate hazard ratio of 2.6 (Table 2.3), confers more than twice the hazard of chemotherapy. (b) Survival analyses of the 236 patients classified by both GSVD and chemotherapy show similar multivariate Cox hazard ratios, of 3 and 3.1, respectively. This means that GSVD and chemotherapy are independent prognostic predictors. With a KM median survival time difference of ∼30 months, GSVD and chemotherapy combined make a better predictor than chemotherapy alone. (c) Survival analyses of the 317 patients with TCGA chemotherapy annotations in the inclusive confirmation set of 344 patients, classified by chemotherapy, show a KM median survival time difference of ∼11 months and a univariate hazard ratio of 2.7, and confirm the survival analyses of the initial set of 251 patients. (d) Survival analyses of the 317 patients classified by both GSVD and chemotherapy show similar multivariate Cox hazard ratios, of 3.1 and 3.2, and a KM median survival time difference of ∼30 months, with the corresponding log-rank test P-value $<10^{-17}$. This confirms that the prognostic contribution of GSVD is independent of chemotherapy, and that combined with chemotherapy, GSVD makes a better predictor than chemotherapy alone. (e) Survival analyses of the 154 patients with TCGA chemotherapy annotations in the independent validation set of 184 patients, classified by chemotherapy, show a KM median survival time difference of ∼11 months and a univariate hazard ratio of 2.2, and validate the survival analyses of the initial set of 251 patients. (f) Survival analyses of the 154 patients classified by both GSVD and chemotherapy, show similar multivariate Cox hazard ratios, of 3.3 and 2.7, and a KM median survival time difference of ∼43 months. This validates that the prognostic contribution of GSVD is independent of chemotherapy, and that combined with chemotherapy, GSVD makes a better predictor than chemotherapy alone, also for patients with measured GBM aCGH profiles in the absence of matched normal profiles.


(a) Chemotherapy
P-value = 4.3 × 10^-9
Hazard Ratio = 2.6
Yes
N=187
O=147
No
N=49
O=46
(b) Probelet 2/Chemotherapy
P-value = 6.1 × 10^-15
Hazard Ratios = 3/3.1
Low/Yes
N=18
O=12
High/Yes
N=169
O=135
Low/No
N=4
O=4
High/No
N=45
O=42
(c) Chemotherapy
P-value = 1.2 × 10^-11
Hazard Ratio = 2.7
Yes
N=255
O=188
No
N=62
O=57
(d) Probelet 2/Chemotherapy
P-value = 5.8 × 10^-18
Hazard Ratios = 3.1/3.2
Low/Yes
N=22
O=12
High/Yes
N=233
O=176
Low/No
N=4
O=4
High/No
N=58
O=53
(e) Chemotherapy
P-value = 5.2 × 10^-4
Hazard Ratio = 2.2
Yes
N=130
O=103
No
N=24
O=22
(f) Arraylet 2/Chemotherapy
P-value = 5.9 × 10^-8
Hazard Ratios = 3.3/2.7
Low/Yes
N=15
O=10
High/Yes
N=115
O=93
Low/No
N=2
O=2
High/No
N=22
O=20
Fraction of Surviving Patients from the Initial Set
Fraction of Surviving Patients from the Inclusive Confirmation Set
Fraction of Surviving Patients from the Independent Validation Set
Survival Time (Months)
Survival Time (Months)

A.2 Supplementary Files

Mathematica Notebook S1. Generalized singular value decomposition (GSVD) of the TCGA patient-matched tumor and normal aCGH profiles. A Mathematica 8.0.1 code file, executable by Mathematica 8.0.1 and readable by Mathematica Player, freely available at http://www.wolfram.com/products/player/.
doi:10.1371/journal.pone.0030098.s002.

Mathematica Notebook S2. Generalized singular value decomposition (GSVD) of the TCGA patient-matched tumor and normal aCGH profiles. A PDF format file, readable by Adobe Acrobat Reader.
doi:10.1371/journal.pone.0030098.s003.

Dataset S1. Initial set of 251 patients. A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing The Cancer Genome Atlas (TCGA) [4] annotations of the initial set of 251 patients and the corresponding normal and tumor samples. The tumor and normal profiles of the initial set of 251 patients, in tab-delimited text format files, tabulating relative copy number variation across 212,696 and 211,227 tumor and normal probes, respectively, are available at http://www.alterlab.org/GBM_prognosis/.
doi:10.1371/journal.pone.0030098.s004.

Dataset S2. Segments of the significant tumor and normal arraylets, computed by GSVD for the initial set of 251 patients. A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, tabulating segments identified by circular binary segmentation (CBS) [49]Venkatraman2007
doi:10.1371/journal.pone.0030098.s005.

Dataset S3. Segments of the second tumor arraylet, computed by GSVD for the initial set of 251 patients. A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, tabulating, for each of the 130 CBS segments of the second tumor arraylet, the segment's coordinates, the CBS P-value, and the log-rank test P-value corresponding to the Kaplan-Meier (KM) survival analysis of the initial set of 251 patients classified by either a gain or a loss of this segment.
doi:10.1371/journal.pone.0030098.s006.

Dataset S4. Inclusive confirmation set of 344 patients. A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing the TCGA annotations of the inclusive confirmation set of 344 patients. The tumor and normal profiles of the inclusive confirmation set of 344 patients, in tab-delimited text format files, tabulating relative copy number variation across 200,139 and 198,342 tumor and normal probes, respectively, are available at http://www.alterlab.org/GBM_prognosis/.
doi:10.1371/journal.pone.0030098.s007

Dataset S5. Independent validation set of 184 patients. A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing the TCGA annotations of the independent validation set of 184 patients. The tumor profiles of the independent validation set of 184 patients, in a tab-delimited text format file, tabulating relative copy number variation across 212,696 autosomal and X chromosome probes, are available at http://www.alterlab.org/GBM_prognosis/.
doi:10.1371/journal.pone.0030098.s008.

APPENDIX B

SUPPLEMENT II

B.1 Supplementary Files

Mathematica Notebook S1. Tensor generalized singular value decomposition (tGSVD) of patient- and platform-matched tumor and normal genomic profiles. A PDF format file, readable by Adobe Acrobat Reader. The corresponding Mathematica 9.0.1 code file, executable by Mathematica and readable by Mathematica Player, is available at http://www.alterlab.org/OV_prognosis/.

Dataset S1. Discovery Set of Patients. A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing TCGA annotations of the discovery set of 249 patients. The tumor and normal profiles of the discovery set of patients measured by each of the two DNA microarray platforms, tabulating relative copy-number variation across the 6p+12p, 7p and Xq tumor and normal probes, are available in tab-delimited text format files at http://www.alterlab.org/OV_prognosis/.

Dataset S2. Validation Set of Patients. A tab-delimited text format file reproducing TCGA annotations of the validation set of 148 patients. The tumor profiles of the validation set of patients, tabulating relative copy-number variation across the 6p+12p, 7p and Xq tumor probes, are available in tab-delimited text format files at http://www.alterlab.org/OV_prognosis/.

Dataset S3. Most Tumor-Exclusive Tumor Arraylets. A tab-delimited text format file tabulating the segments of the first, most tumor-exclusive tumor arraylets computed by tGSVD for the discovery set of patients across 6p+12p, 7p and Xq.

Dataset S4. Differential mRNA Expression. A tab-delimited text format file tabulating differential expression of 11,457 autosomal and X chromosome mRNAs in the 6p+12p,

7p and Xq tGSVD classes. The mRNA expression profiles of 394 of the 397 patients in the discovery and validation sets are available in tab-delimited text format files at http://www.alterlab.org/OV_prognosis/.

Dataset S5. Differential microRNA Expression. A tab-delimited text format file tabulating differential expression of 639 autosomal and X chromosome microRNAs in the 6p+12p, 7p and Xq tGSVD classes. The microRNA expression profiles of 395 patients are available in tab-delimited text format files at http://www.alterlab.org/OV_prognosis/.

Dataset S6. Differential Protein Expression. A tab-delimited text format file tabulating differential expression of 165 antibodies that probe for 136 autosomal and X chromosome proteins in the 6p+12p, 7p and Xq tGSVD classes. The protein expression profiles of 282 patients are available in tab-delimited text format files at http://www.alterlab.org/OV_prognosis/.

APPENDIX C

STATISTICAL METHODS

C.1 Box-Whisker Plot

In *statistics*, it is common to use box plots or box-and-whiskers plots as a tool for exploratory data analysis. In this dissertation, we use box plots to visualize batch effects in the GBM patients (Figure 2.4). We also visualize the differential mRNA, microRNA and protein expression of OV patients in two tGSVD classes using box-and-whisker plots (Figures 3.12, 3.13 and 3.14).

A box plot is nonparametric and does not make any assumption about the underlying statistical distribution of the data. It is used to graphically depict the five-number summary (minimum, first quartile, median, third quartile and maximum) of the data using a *box* with a band within it. Consider the following data:

4.3 , 5.1 , 3.6 , 4.5 , 4.4 , 4.9 , 5.5 , 4.7 , 4.1 , 4.6 , 4.4 , 4.3 , 4.8 , 4.4 , 4.2 , 4.5 , 4.4

To obtain the five-number summary, we first arrange the data in ascending order.

3.6 , 4.1 , 4.2 , 4.3 , 4.3 , 4.4 , 4.4 , 4.4 , 4.4 , 4.5 , 4.5 , 4.6 , 4.7 , 4.8 , 4.9 , 5.1 , 5.5

The minimum value is 3.6 and the maximum value is 5.5. The median is the middle value of the data. After sorting the data, out of the 17 data points, the middle value is the 9^{th} data point. Therefore, 4.4 is the median of this dataset. The median divides the dataset into two groups of eight data points each. The medians of these two groups are 4.3 and 4.7. These are the first and third quartiles, respectively (Figure C.1). The bottom and top lines of the box are the first ($Q1$) and third ($Q3$) quartiles and the band inside the box represents the median ($Q2$) of the data. The spacings between the different parts of the box indicate the degree of spread and skewness in the data. In this dissertation, the lower and upper ends of the whiskers (lines connecting the box to the fences) represent the lowest and the highest data points within 1.5 inter quartile range, i.e., $1.5(Q3 - Q1)$ of the lower and upper quartiles, respectively [102]. The data points lying outside of this range are

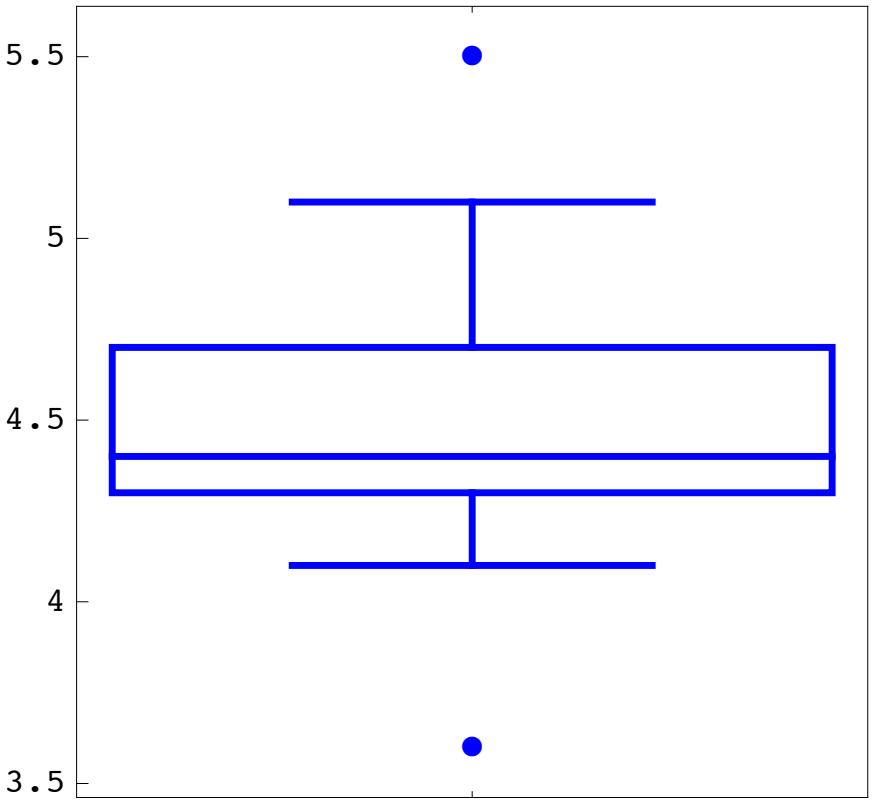


Figure C.1: Box-whisker plot of sample data

considered outliers and are marked as dots. Unlike mean and standard deviation, median and quartiles are more robust to skewed or heavy-tailed distributions such as microarray data. For example, if we replace the maximum value with 10 and the minimum value with 2, the median and quartiles remain unchanged whereas the mean and standard deviation are sensitive to these changes.

C.2 Mann-Whitney Test

Mann-Whitney test is a nonparametric test used to test whether the medians of two populations are equal. This test is also known as Wilcoxon Rank Sum or Mann-Whitney Wilcoxon test [103, 104]. In many applications, this test is used in place of the two-sample t-test when the normality assumption of the underlying population distributions is questionable. Another advantage of this test is that it depends only on the ranks of the observations and not on the actual values. The following assumptions are made about the samples:

1. There is a symmetry between populations with respect to the probability of a random drawing of a larger observation.
2. The two samples are independent.

The statistical hypothesis is formulated as follows:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

where μ_1 and μ_2 are the true population medians. A small P-value $< \alpha$ suggests that the evidence from the samples is statistically significant to reject the null hypothesis. The α value used in this dissertation is 0.05 unless otherwise specified. For a small sample size of two independent samples X and Y, the U statistic is calculated directly by ranking all the observations into a single ranked list without regard to the sample in which they are found.

$$U_x = n_x n_y + n_x \frac{(n_x + 1)}{2} - R_x$$
$$U_y = n_x n_y + n_y \frac{(n_y + 1)}{2} - R_y$$
$$U = \min(U_x, U_y)$$

where n_x and n_y are the sample sizes of X and Y, respectively, and R_x and R_y are the sum of the ranks in X and Y, respectively. U is the test statistic. For large samples, U is approximately normally distributed. The z statistic in this case would be

$$z = \frac{U - \mu}{\sigma}$$

where μ and σ are the mean and standard deviation of U. Ties in ranking the observations should first be resolved before calculating the U statistic in cases of both large and small sample sizes. After the test statistic is obtained, the P-value is calculated as the probability of obtaining a test statistic $< -|U|$ or $> |U|$.

In this dissertation, the Mann-Whitney test was used to test whether two groups of measurements of DNA copy number, mRNA, microRNA or protein expression classified by the probelets differ significantly.

C.3 Hypergeometric Probability Distribution

The hypergeometric distribution is a discrete probability distribution that gives the distribution of successes in n draws from a population of size N containing K successes in the category of interest. The sampling is done *without* replacement. This distribution assumes independence of K categories. The probability mass function (pmf) of the random variable X is given by

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

which is the probability to observe the subset of k among the total K items in the category of interest in a subset of n items selected from the total N items without repetitions. The probability to observe *at least k items* among the total K items in the category of interest in a subset of n items selected from the total N items without repetitions is given by

$$\begin{aligned} P(k; n, N, K) &= \frac{\sum_{i=k}^{n}\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \\ &= 1 - \frac{\sum_{i=0}^{k-1}\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \end{aligned} \tag{C.1}$$

since

$$\frac{\sum_{i=0}^{n}\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} = 1$$

In this dissertation, the hypergeometric probability distribution is assumed while calculating the P-values for enrichments annotation or gene ontology enrichment of TCGA annotations in Chapter 2 and enrichment of gene annotations in Chapter 3. In these cases, we consider the population to be the total number of patients or the total number of genes in the array

(N), the sample size is the number of genes or patients in that particular group or category (K) and a success is any gene or patient labeled to be of the category of interest by the annotation that is being tested for enrichment. The probability [8] that the enrichment is due to random chance is given by (C.1).

A P-value $< \alpha$ suggests that the null hypothesis that the enrichment is due to random chance can be rejected. The α value chosen in this dissertation for the hypergeometric test is 0.05.

C.4 Kaplan-Meier Survival Analysis

In medical research, nonparametric methods such as Kaplan-Meier survival analysis [65] and the log-rank test are used to determine the distributions of patient survival times and compare them even when only partial observations are available. An important advantage of the Kaplan-Meier (KM) estimation is that this method can take into account censored data, especially right censoring. For example, suppose a patient drops out of a study before the study ends and the partial observation is that the patient was alive *at least up until that point*. This type of partial measurement is known as right-censored data where we know that a data point is above a certain value but the exact measurement is unknown [105].

In this dissertation, we use KM curves to analyze the survival times of two or more groups of patients classified either by probelet of interest or other prognostic indicators or both. We use the log-rank test to statistically test if the difference that we observe between the groups is significant.

The KM estimator is an estimator $\hat{S}(t)$, for the true survival function $S(t)$. It can be computed from the sample data as a product of all the conditional probabilities that the probability of a patient surviving *past* an event time $t_{(i)}$ given a patient survives *up until* that particular event time $t_{(i)}$.

$$\hat{S}(t_{(j)}) = \prod_{i=0}^{t_{(j)}} \frac{n_{(i)} - (c_{(i)} + d(i))}{n_{(i)}} \tag{C.2}$$

where $n_{(i)}$ is the number of patients surviving *up until* the event time $t_{(i)}$, $c_{(i)}$ is the number of censored patients *exactly* at that particular event time $t_{(i)}$ and $d_{(i)}$ is the number of deaths that occurred at the event time $t_{(i)}$. $\hat{S}(t_{(j)})$ calculated from equation C.2 for various event times $t_{(j)}$ are then plotted in the form of a stepwise curve with time in the x axis and fraction of patients surviving in the y axis. This plot is called a KM curve. With a sufficiently large sample size, $\hat{S}(t)$ approaches the true survival function $S(t)$. A KM curve

can be used to qualitatively verify, for example, if patients in the treatment group live longer than the patients in the control group of a clinical trial.

C.5 Log-Rank Test

When comparing the survival functions between two or more groups, an observed difference can be a true difference observed or may be due to sampling error. Therefore, it is essential to perform significance tests to determine if the difference is true. In a survival analysis of two groups, typically the log-rank test is used with censored data whereas the Wilcoxon rank sum test is used if the data are not censored [106]. The log-rank test is a large-sample χ^2 test where the null hypothesis H_o is that the two populations have identical survival distributions, and the alternate hypothesis H_a is that the two populations do not have identical survival distributions. The test statistic is called the log-rank test statistic. Like any other type of χ^2 test, the computation of the test statistic involves observed and expected cell counts and is computed as

$$Z = \frac{\sum_{i=0}^{t} (O_{ji} - E_{ji})^2}{\sum_{i=0}^{t} V_i}$$

where O_{ji} is the number of observed events in the group j and E_{ji} is the number of expected events (here a death) in the group j if there were in reality no difference between the two groups. t is usually the time at which the last event in the pooled groups occurs. V_i is the variance at the event time t, which is calculated as

$$V_i = \frac{O_i(n_i - O_i)(\frac{n_{ji}}{n_i})(1 - \frac{n_{ji}}{n_i})}{n_i - 1}$$

where O_i is the observed number of events at time i across both the groups; n_i is the number of patients at risk at the time point i, i.e., the number of patients alive at the time point i across both groups; and n_{ji} is the number of patients alive in the group j at the time point i. From the Z statistic, the P-value is calculated assuming that the test statistic follows the χ^2 distribution with one degree of freedom (number of groups -1).

C.6 Cox Proportional Hazards Model

The log-rank test cannot be used to explore (and adjust for) the several variables, such as age, tumor stage, residual disease, therapy outcome, neoplasm status and various other predictors, known to affect patient survival. The Cox proportional hazards regression analysis [66, 105] is used to investigate several variables at a time. This model gives the hazard functions for the explanatory variables. The hazard function is the probability that

an event will occur within a small time interval (for the scope of this dissertation, the event is death) given that the individual has survived up to the beginning of that time point t, i.e., it is the risk of dying at time t. The hazard function at a given time t is given by

$$h(t, \mathbf{X}) = h_0(t) \times e^{\sum_{i=1}^{p} \beta_i X_i}$$

where p is the number of explanatory or confounding variables, $\mathbf{X}$ is $\langle X_1 X_2 .. X_p \rangle$ and $h_0(t)$ is the baseline hazard function and it denotes the probability of dying when all the explanatory variables X_is are zero. The baseline hazard function is analogous to the intercept in ordinary regression. β_i s are the parameters and their estimates $\hat{\beta}_i$ s are calculated by maximizing a partial likelihood function L. L is called a partial likelihood function because it considers probabilities only for patients who fail (death occurs) and does not consider censored data.

$$L = \prod_{j=1}^{k} L_j$$

where k is the number of failure times. At the jth failure time, L_j denotes the likelihood of failing at this time, given survival up to this time. The set of individuals at risk at the jth failure time is called the "risk set" and this number decreases as the failure time increases. In order to maximize L, we solve for each

$$\frac{\partial ln L}{\partial \beta_i} = 0$$
$$i = 1, 2, .., p$$

One important assumption of this model is that the hazard functions for any two individuals at any time point are proportional. A simple test to verify this assumption is to check if the KM survival curves of the two groups cross each other. If they do not, then the assumption holds good. Cox model is a robust semiparametric model because no assumption is made about the baseline hazard function $h_0(t)$. In fact, this model is very popular because it can be shown that only the estimates $\hat{\beta}_i$ s and their standard error are required to obtain a point estimate or test for the significance of the effect of explanatory variable (in this dissertation, this variable is either the probelet value or the arraylet correlation of the patients) adjusted for the other confounding variables.

To test for the significance, the Wald statistic, which is a z statistic, is calculated as:

$$Z_i = \hat{\beta}_i / Standard\ Error(\hat{\beta}_i)\ where\ i = 1, 2, ..p$$

The two tailed p-value is calculated assuming a standard normal distribution.

In order to obtain the point estimate of the effect of the variable of interest adjusted for other variables, hazard ratios are calculated for each of the regression coefficients $\hat{\beta}_i$ as

$$Hazard\ Ratio\ (HR_i) = e^{\hat{\beta}_i}\ where\ i = 1, 2, ..p$$

For example, if the estimated hazard ratio is 2 for the effect of probelet adjusted for age of the patient as a confounding variable, then we see that the hazard for the patients with high (or low) probelet value is twice as much as the hazard for the patients with low (or high) probelet value. In this dissertation, hazard ratios and the p-values associated with the Wald test statistic are used in determining if the probelet values or arraylet correlations of patients are confounded by other predictors such as age, tumor stage, neoplasm status etc.

REFERENCES

[1] D. Hwang, A. G. Rust, S. Ramsey, J. J. Smith, D. M. Leslie, A. D. Weston, P. De Atauri, J. D. Aitchison, L. Hood, A. F. Siegel *et al.*, "A data integration methodology for systems biology," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 48, pp. 17 296–17 301, 2005.

[2] T. J. Hudson, W. Anderson, A. Aretz, A. D. Barker, C. Bell, R. R. Bernabé, M. Bhan, F. Calvo, I. Eerola, D. S. Gerhard *et al.*, "International network of cancer genome projects," *Nature*, vol. 464, no. 7291, pp. 993–998, 2010.

[3] C. H. Lee, B. O. Alpert, P. Sankaranarayanan, and O. Alter, "GSVD comparison of patient-matched normal and tumor aCGH profiles reveals global copy-number alterations predicting glioblastoma multiforme survival," *PLoS One*, vol. 7, no. 1, p. e30098, 2012.

[4] P. Sankaranarayanan, T. E. Schomay, K. A. Aiello, and O. Alter, "Tensor GSVD of patient- and platform-matched tumor and normal DNA copy-number profiles uncovers chromosome arm-wide patterns of tumor-exclusive platform-consistent alterations encoding for cell transformation and predicting ovarian cancer survival," *PLoS One*, p. e121396, 2015.

[5] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton *et al.*, "Minimum information about a microarray experiment (MIAME) - toward standards for microarray data," *Nature genetics*, vol. 29, no. 4, pp. 365–371, 2001.

[6] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church, "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation," *Nature biotechnology*, vol. 16, no. 10, pp. 939–945, 1998.

[7] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14 863–14 868, 1998.

[8] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nat Genet*, vol. 22, pp. 281–285, 1999.

[9] N. M. Bertagnolli, J. A. Drake, J. M. Tennessen, and O. Alter, "SVD identifies transcript length distribution functions from DNA microarray data and reveals evolutionary forces globally affecting GBM metabolism," *PLoS One*, vol. 8, no. 11, p. e78913, 2013.

[10] O. Alter, "Genomic signal processing: from matrix algebra to genetic networks," in *Microarray Data Analysis*. Springer, 2007, pp. 17–59.

[11] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proceedings of the National Academy of Sciences*, vol. 97, no. 18, pp. 10 101–10 106, 2000.

[12] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the national academy of sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

[13] D. V. Nguyen and D. M. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. 39–50, 2002.

[14] T. W. Anderson, *An introduction to multivariate statistical analysis*. Wiley, 1958.

[15] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in *A practical approach to microarray data analysis*. Springer, 2003, pp. 91–109.

[16] I. Jolliffe, *Principal component analysis*. Springer-Verlag, 1986.

[17] O. Alter and G. H. Golub, "Singular value decomposition of genome-scale mRNA lengths distribution reveals asymmetry in RNA gel electrophoresis band broadening," *Proceedings of the National Academy of Sciences*, vol. 103, no. 32, pp. 11 828–11 833, 2006.

[18] O. Alter, G. Golub, P. Brown, and D. Botstein, "Novel genome-scale correlation between DNA replication and RNA transcription during the cell cycle in yeast is predicted by data-driven models," in *2004 Miami Nature Winter Symposium, Jan*, 2004.

[19] O. Alter and G. H. Golub, "Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription," *Proc Natl Acad Sci USA*, vol. 101, pp. 16 577–16 582, 2004.

[20] O. Alter, P. O. Brown, and D. Botstein, "Processing and modeling genome-wide expression data using singular value decomposition," in *BiOS 2001 The International Symposium on Biomedical Optics*. International Society for Optics and Photonics, 2001, pp. 171–186.

[21] C. F. Van Loan, "Generalizing the singular value decomposition," *SIAM Journal on Numerical Analysis*, vol. 13, no. 1, pp. 76–83, 1976.

[22] O. Alter, P. O. Brown, and D. Botstein, "Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms," *Proceedings of the National Academy of Sciences*, vol. 100, no. 6, pp. 3351–3356, 2003.

[23] S. P. Ponnapalli, M. A. Saunders, C. F. Van Loan, and O. Alter, "A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms," *PLoS One*, vol. 6, no. 12, p. e28072, 2011.

[24] O. Alter, "Discovery of principles of nature from mathematical modeling of DNA microarray data," *Proc Natl Acad Sci USA*, vol. 103, pp. 16 063–16 064, 2006.

[25] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore: Johns Hopkins University Press, third edition, 694 p., 1996.

[26] R. A. Burton and G. B. Fincher, "(1, 3; 1, 4)-β-d-glucans in cell walls of the poaceae, lower plants, and fungi: a tale of two linkages," *Molecular Plant*, vol. 2, no. 5, pp. 873–882, 2009.

[27] A. W. Schreiber, N. J. Shirley, R. A. Burton, and G. B. Fincher, "Combining transcriptional datasets using the generalized singular value decomposition," *BMC bioinformatics*, vol. 9, no. 1, p. 335, 2008.

[28] J. A. Berger, S. Hautaniemi, S. K. Mitra, and J. Astola, "Jointly analyzing gene expression and copy number data in breast cancer using data reduction models," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 3, no. 1, p. 2, 2006.

[29] J. A. Berger, S. Hautaniemi, and S. K. Mitra, "Comparative analysis of gene expression and DNA copy number data for pancreatic and breast cancers using an orthogonal decomposition," in *Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE.* IEEE, 2004, pp. 584–585.

[30] O. Alter and G. H. Golub, "Reconstructing the pathways of a cellular system from genome-scale signals by using matrix and tensor computations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 49, pp. 17 559–17 564, 2005.

[31] L. Omberg, G. H. Golub, and O. Alter, "A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies," *Proc Natl Acad Sci USA*, vol. 104, pp. 18 371–18 376, 2007.

[32] L. Omberg, J. R. Meyerson, K. Kobayashi, L. S. Drury, J. F. Diffley, and O. Alter, "Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression," *Mol Syst Biol*, vol. 5, p. 312, 2009.

[33] C. Muralidhara, A. M. Gross, R. R. Gutell, and O. Alter, "Tensor decomposition reveals concurrent evolutionary convergences and divergences and correlations with structural motifs in ribosomal RNA," *PLoS One*, vol. 6, no. 4, p. e18768, 2011.

[34] B. Purow and D. Schiff, "Advances in the genetics of glioblastoma: are we reaching critical mass?" *Nat Rev Neurol*, vol. 5, pp. 419–426, 2009.

[35] R. N. Wiltshire, B. A. Rasheed, H. S. Friedman, A. H. Friedman, and S. H. Bigner, "Comparative genetic patterns of glioblastoma multiforme: potential diagnostic tool for tumor classification," *Neuro Oncol*, vol. 2, pp. 164–173, 2000.

[36] J. M. Nigro, A. Misra, L. Zhang, I. Smirnov, H. Colman, C. Griffin, N. Ozburn, M. Chen, E. Pan, D. Koul *et al.*, "Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma," *Cancer Res*, vol. 65, pp. 1678–1686, 2005.

[37] R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, G. M. Mastrogianakis, J. J. Olson, T. Mikkelsen, N. Lehman, K. Aldape *et al.*, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061–1068, 2008.

[38] P. S. Mischel, R. Shai, T. Shi, S. Horvath, K. V. Lu, G. Choe, D. Seligson, T. J. Kremen, A. Palotie, L. M. Liau *et al.*, "Identification of molecular subtypes of glioblastoma by gene expression profiling," *Oncogene*, vol. 22, pp. 2361–2673, 2003.

[39] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov *et al.*, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, pp. 98–110, 2010.

[40] H. Colman, L. Zhang, E. P. Sulman, J. M. McDonald, N. L. Shooshtari, A. Rivera, S. Popoff, C. L. Nutt, D. N. Louis, J. G. Cairncross *et al.*, "A multigene predictor of outcome in glioblastoma," *Neuro Oncol*, vol. 12, pp. 49–57, 2010.

[41] H. Noushmehr, D. J. Weisenberger, K. Diefes, H. S. Phillips, K. Pujara, B. P. Berman, F. Pan, C. E. Pelloski, E. P. Sulman, K. P. Bhat *et al.*, "Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma," *Cancer Cell*, vol. 17, pp. 510–522, 2010.

[42] W. J. Curran, C. B. Scott, J. Horton, J. S. Nelson, A. S. Weinstein, A. J. Fischbach, C. H. Chang, M. Rotman, S. O. Asbell, R. E. Krisch *et al.*, "Recursive partitioning analysis of prognostic factors in three Radiation Therapy Oncology Group malignant glioma trials," *J Natl Cancer Inst*, vol. 85, pp. 704–710, 1993.

[43] T. Gorlia, M. J. van den Bent, M. E. Hegi, R. O. Mirimanoff, M. Weller, J. G. Cairncross, E. Eisenhauer, K. Belanger, A. A. Brandes, A. Allgeier *et al.*, "Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of EORTC and NCIC trial 26981-22981/CE.3," *Lancet Oncol*, vol. 9, pp. 29–38, 2008.

[44] C. H. Lee and O. Alter, "Known and novel copy number alterations in GBM and their patterns of co-occurrence are revealed by GSVD comparison of array CGH data from patient-matched normal and tumor TCGA samples," in *60th Annual American Society of Human Genetics (ASHG) Meeting (November 2–6, 2010, Washington, DC)*, 2010.

[45] B. O. Alpert, P. Sankaranarayanan, C. H. Lee, and O. Alter, "Glioblastoma multiforme prognosis by using a patient's array CGH tumor profile and a generalized SVD-computed global pattern of copy-number alterations," in *2nd DNA and Genome World Day (April 25–29, 2011, Dalian, China)*, 2011.

[46] T. O. Nielsen, R. B. West, S. C. Linn, O. Alter, M. A. Knowling, J. X. O'Connell, S. Zhu, M. Fero, G. Sherlock, J. R. Pollack *et al.*, "Molecular characterisation of soft tissue tumours: a gene expression study," *Lancet*, vol. 359, pp. 1301–1307, 2002.

[47] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, "The human genome browser at UCSC," *Genome Res*, vol. 12, pp. 996–1006, 2002.

[48] P. A. Fujita, B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik, M. S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho *et al.*, "The UCSC Genome Browser database: update 2011," *Nucleic Acids Res*, vol. 39, pp. D876–D882, 2011.

[49] A. B. Olshen, E. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics*, vol. 5, pp. 557–572, 2004.

[50] E. Venkatraman and A. B. Olshen, "A faster circular binary segmentation algorithm for the analysis of array CGH data," *Bioinformatics*, vol. 23, pp. 657–663, 2007.

[51] M. Heidenblad, D. Lindgren, J. A. Veltman, T. Jonson, E. H. Mahlamäki, L. Gorunova, A. G. van Kessel, E. F. Schoenmakers, and M. Höglund, "Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications," *Oncogene*, vol. 24, pp. 1794–1801, 2005.

[52] Q. Wang, S. Diskin, E. Rappaport, E. Attiyeh, Y. Mosse, D. Shue, E. Seiser, J. Jagannathan, S. Shusterman, M. Bansal *et al.*, "Integrative genomics identifies distinct molecular classes of neuroblastoma and shows that multiple genes are targeted by regional alterations in DNA copy number," *Cancer Res*, vol. 66, pp. 6050–6062, 2006.

[53] A. L. Hopkins and C. R. Groom, "The druggable genome," *Nat Rev Drug Discov*, vol. 1, pp. 727–730, 2002.

[54] H. Silljé, K. Takahashi, K. Tanaka, G. Van Houwe, and E. Nigg, "Mammalian homologues of the plant Tousled gene code for cell-cycle-regulated kinases with maximal activities linked to ongoing DNA replication," *EMBO J*, vol. 18, pp. 5691–5702, 1999.

[55] U. R. Chandran, C. Ma, R. Dhir, M. Bisceglia, M. Lyons-Weiler, W. Liang, G. Michalopoulos, M. Becich, and F. A. Monzon, "Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process," *BMC Cancer*, vol. 7, p. 64, 2007.

[56] M. Pellegrini, J. C. Cheng, J. Voutila, D. Judelson, J. Taylor, S. F. Nelson, and K. M. Sakamoto, "Expression profile of CREB knockdown in myeloid leukemia cells," *BMC Cancer*, vol. 8, p. 264, 2008.

[57] M. Millour, C. Charbonnel, F. Magrangeas, S. Minvielle, C. Loic *et al.*, "Gene expression profiles discriminate between pathological complete response and resistance to neoadjuvant FEC100 in breast cancer," *Cancer Genomics Proteomics*, vol. 3, pp. 89–95, 2006.

[58] A. M. Snijders, M. E. Nowee, J. Fridlyand, J. M. Piek, J. C. Dorsman, A. N. Jain, D. Pinkel, P. J. van Diest, R. H. Verheijen, and D. G. Albertson, "Genome-wide-array-based comparative genomic hybridization reveals genetic homogeneity and frequent copy number increases encompassing CCNE1 in fallopian tube carcinoma," *Oncogene*, vol. 22, pp. 4281–4286, 2003.

[59] P. J. Campbell, S. Yachida, L. J. Mudie, P. J. Stephens, E. D. Pleasance, L. A. Stebbings, L. A. Morsberger, C. Latimer, S. McLaren, M.-L. Lin *et al.*, "The patterns and dynamics of genomic instability in metastatic pancreatic cancer," *Nature*, vol. 467, pp. 1109–1113, 2010.

[60] D. Etemadmoghadam, J. George, P. A. Cowin, C. Cullinane, M. Kansara, K. L. Gorringe, G. K. Smyth, D. D. Bowtell, A. O. C. S. Group *et al.*, "Amplicon-dependent CCNE1 expression is critical for clonogenic survival after cisplatin treatment and is correlated with 20q11 gain in ovarian cancer," *PLoS One*, vol. 5, p. e15498, 2010.

[61] D. Defeo-Jones, P. S. Huang, R. E. Jones, K. M. Haskell, G. A. Vuocolo, M. G. Hanobik, H. E. Huber, and A. Oliff, "Cloning of cDNAs for cellular proteins that bind to the retinoblastoma gene product," *Nature*, vol. 352, pp. 251–254, 1991.

[62] S. V. Sharma, D. Y. Lee, B. Li, M. P. Quinlan, F. Takahashi, S. Maheswaran, U. McDermott, N. Azizian, L. Zou, M. A. Fischbach *et al.*, "A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations," *Cell*, vol. 141, pp. 69–80, 2010.

[63] W. M. Pardridge, "The blood-brain barrier: bottleneck in brain drug development," *NeuroRx*, vol. 2, pp. 3–14, 2005.

[64] Y. Hattori, S. Ohta, K. Hamada, H. Yamada-Okabe, Y. Kanemura, Y. Matsuzaki, H. Okano, Y. Kawakami, and M. Toda, "Identification of a neuron-specific human gene, KIAA1110, that is a guanine nucleotide exchange factor for ARF1," *Biochem Biophys Res Commun*, vol. 364, pp. 737–742, 2007.

[65] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *J Amer Statist Assn*, vol. 53, pp. 457–481, 1958.

[66] D. R. Cox, "Regression models and life-tables," *J Roy Statist Soc B*, vol. 34, pp. 187–220, 1972.

[67] K. J. Rothman, "No adjustments are needed for multiple comparisons," *Epidemiology*, vol. 1, pp. 43–46, 1990.

[68] Cancer Genome Atlas Research Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, pp. 609–615, 2011.

[69] M. G. Prisco, G. F. Zannoni, I. De Stefano, V. G. Vellone, L. Tortorella, A. Fagotti, L. Mereu, G. Scambia, and D. Gallo, "Prognostic role of metastasis tumor antigen 1 in patients with ovarian cancer: a clinical study," *Human pathology*, vol. 43, no. 2, pp. 282–288, 2012.

[70] D. A. Engler, S. Gupta, W. B. Growdon, R. I. Drapkin, M. Nitta, P. A. Sergent, S. F. Allred, J. Gross, M. T. Deavers, W.-L. Kuo *et al.*, "Genome wide DNA copy number analysis of serous type ovarian carcinomas identifies genetic markers predictive of clinical outcome," *PLoS One*, vol. 7, no. 2, p. e30996, 2012.

[71] T. Ikeda, J. Zhang, T. Chano, A. Mabuchi, A. Fukuda, H. Kawaguchi, K. Nakamura, and S. Ikegawa, "Identification and characterization of the human long form of Sox5 (*L-SOX5*) gene," *Gene*, vol. 298, no. 1, pp. 59–68, 2002.

[72] V. Bourdon, F. Naef, P. H. Rao, V. Reuter, S. C. Mok, G. J. Bosl, S. Koul, V. V. Murty, R. S. Kucherlapati, and R. Chaganti, "Genomic and expression analysis of the 12p11-p12 amplicon using EST arrays identifies two novel amplified and overexpressed genes," *Cancer research*, vol. 62, no. 21, pp. 6218–6223, 2002.

[73] L. A. Lee, E. Lee, M. A. Anderson, L. Vardy, E. Tahinci, S. M. Ali, H. Kashevsky, M. Benasutti, M. W. Kirschner, and T. L. Orr-Weaver, "*Drosophila* genome-scale screen for PAN GU kinase substrates identifies Mat89Bb as a cell cycle regulator," *Developmental cell*, vol. 8, no. 3, pp. 435–442, 2005.

[74] P. Blanco, C. A. Sargent, C. A. Boucher, G. Howell, M. Ross, and N. A. Affara, "A novel poly(A)-binding protein gene (*PABPC5*) maps to an X-specific subinterval in the Xq21.3/Yp11.2 homology block of the human sex chromosomes," *Genomics*, vol. 74, no. 1, pp. 1–11, 2001.

[75] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[76] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists," *BMC bioinformatics*, vol. 10, no. 1, p. 48, 2009.

[77] C. R. Sibley, Y. Seow, S. Saayman, K. K. Dijkstra, S. El Andaloussi, M. S. Weinberg, and M. J. Wood, "The biogenesis and characterization of mammalian microRNAs of mirtron origin," *Nucleic acids research*, p. gkr722, 2011.

[78] A. E. Karnoub and R. A. Weinberg, "Ras oncogenes: split personalities," *Nature reviews Molecular cell biology*, vol. 9, no. 7, pp. 517–531, 2008.

[79] W. C. Hahn, C. M. Counter, A. S. Lundberg, R. L. Beijersbergen, M. W. Brooks, and R. A. Weinberg, "Creation of human tumour cells with defined genetic elements," *Nature*, vol. 400, no. 6743, pp. 464–468, 1999.

[80] T. Waldman, K. W. Kinzler, and B. Vogelstein, "p21 is necessary for the p53-mediated G_1 arrest in human cancer cells," *Cancer research*, vol. 55, no. 22, pp. 5187–5190, 1995.

[81] D. V. Bulavin, S. Saito, M. C. Hollander, K. Sakaguchi, C. W. Anderson, E. Appella, and A. J. Fornace Jr, "Phosphorylation of human p53 by p38 kinase coordinates N-terminal phosphorylation and apoptosis in response to UV radiation," *The EMBO journal*, vol. 18, no. 23, pp. 6845–6854, 1999.

[82] M. S. Anglesio, J. M. Arnold, J. George, A. V. Tinker, R. Tothill, N. Waddell, L. Simms, B. Locandro, S. Fereday, N. Traficante *et al.*, "Mutation of ERBB2 provides a novel alternative mechanism for the ubiquitous activation of RAS-MAPK in ovarian serous low malignant potential tumors," *Molecular Cancer Research*, vol. 6, no. 11, pp. 1678–1690, 2008.

[83] H. L. Klein, "The consequences of Rad51 overexpression for normal and tumor cells," *DNA repair*, vol. 7, no. 5, pp. 686–693, 2008.

[84] F. Diaz and L. Bourguignon, "Selective down-regulation of IP_3 receptor subtypes by caspases and calpain during TNF α-induced apoptosis of human T-lymphoma cells," *Cell Calcium*, vol. 27, no. 6, pp. 315–328, 2000.

[85] M. V. Iorio, R. Visone, G. Di Leva, V. Donati, F. Petrocca, P. Casalini, C. Taccioli, S. Volinia, C.-G. Liu, H. Alder *et al.*, "MicroRNA signatures in human ovarian cancer," *Cancer research*, vol. 67, no. 18, pp. 8699–8707, 2007.

[86] D. Yang, Y. Sun, L. Hu, H. Zheng, P. Ji, C. V. Pecot, Y. Zhao, S. Reynolds, H. Cheng, R. Rupaimoole *et al.*, "Integrated analyses identify a master microRNA regulatory network for the mesenchymal subtype in serous ovarian cancer," *Cancer cell*, vol. 23, no. 2, pp. 186–199, 2013.

[87] H. Nagahara, A. M. Vocero-Akbani, E. L. Snyder, A. Ho, D. G. Latham, N. A. Lissy, M. Becker-Hapak, S. A. Ezhevsky, and S. F. Dowdy, "Transduction of full-length TAT fusion proteins into mammalian cells: TAT-$p27^{Kip1}$ induces cell migration," *Nature medicine*, vol. 4, no. 12, pp. 1449–1452, 1998.

[88] Y. H. Kwon, A. Jovanovic, M. S. Serfas, and A. L. Tyner, "The cdk inhibitor p21 is required for necrosis, but it inhibits apoptosis following toxin-induced liver injury," *Journal of Biological Chemistry*, vol. 278, no. 32, pp. 30 348–30 355, 2003.

[89] I. M. Chu, L. Hengst, and J. M. Slingerland, "The cdk inhibitor p27 in human cancer: prognostic potential and relevance to anticancer therapy," *Nature Reviews Cancer*, vol. 8, no. 4, pp. 253–267, 2008.

[90] T. J. Duncan, A. Al-Attar, P. Rolland, S. Harper, I. Spendlove, and L. G. Durrant, "Cytoplasmic p27 expression is an independent prognostic factor in ovarian cancer," *International Journal of Gynecologic Pathology*, vol. 29, no. 1, pp. 8–18, 2010.

[91] J. Ahmed, T. Meinel, M. Dunkel, M. S. Murgueitio, R. Adams, C. Blasse, A. Eckert, S. Preissner, and R. Preissner, "CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D960–D967, 2011.

[92] L. Y. Romanova, H. Willers, M. V. Blagosklonny, and S. N. Powell, "The interaction of p53 with replication protein a mediates suppression of homologous recombination," *Oncogene*, vol. 23, no. 56, pp. 9025–9033, 2004.

[93] M. E. Moynahan and M. Jasin, "Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis," *Nature reviews Molecular cell biology*, vol. 11, no. 3, pp. 196–207, 2010.

[94] G. R. Kumar, L. Shum, and B. A. Glaunsinger, "Importin α-mediated nuclear import of cytoplasmic poly(A) binding protein occurs as a direct consequence of cytoplasmic mRNA depletion," *Mol Cell Biol*, vol. 31, no. 15, pp. 3113–3125, 2011.

[95] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge university press, 2012.

[96] C. C. Paige and M. A. Saunders, "Towards a generalized singular value decomposition," *SIAM Journal on Numerical Analysis*, vol. 18, no. 3, pp. 398–405, 1981.

[97] Z. Bai and J. W. Demmel, "Computing the generalized singular value decomposition," *SIAM Journal on Scientific Computing*, vol. 14, no. 6, pp. 1464–1486, 1993.

[98] S. Friedland, "A new approach to generalized singular value decomposition," *SIAM journal on matrix analysis and applications*, vol. 27, no. 2, pp. 434–444, 2005.

[99] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[100] J. Vandewalle, L. De Lathauwer, and P. Comon, "The generalized higher order singular value decomposition and the oriented signal-to-signal ratios of pairs of signal tensors and their use in signal processing," in *Proc ECCTD03-European Conf on Circuit Theory and Design. pp I-389–I-392*, 2003.

[101] S. Ponnapalli, G. Golub, and O. Alter, "A novel higher-order generalized singular value decomposition for comparative analysis of multiple genome-scale datasets." *In: Stanford University and Yahoo! Research Workshop on Algorithms for Modern Massive Datasets (MMDS) (June 21–24, 2006, Stanford, CA)*, 2006.

[102] J. Tukey, *Exploratory Data Analysis*. Addison-Wesley Publishing Company, 1977.

[103] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[104] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, pp. 80–83, 1945.

[105] D. G. Kleinbaum and M. Klein, *Survival analysis*. Springer, 1996.

[106] X. Liu, *Survival analysis: models and applications*. John Wiley & Sons, 2012.