TUNABLE BICLUSTERING ALGORITHM FOR ANALYZING LARGE GENE EXPRESSION DATA SETS

By

AMARTYA SINGH

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Physics and Astronomy

Written under the direction of

Hossein Khiabanian and Gyan Bhanot

And approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2019

ABSTRACT OF THE DISSERTATION

Tunable Biclustering Algorithm for Analyzing Large Gene Expression Data Sets

By AMARTYA SINGH

Dissertation Director:

Hossein Khiabanian and Gyan Bhanot

Traditional clustering approaches for gene expression data are not well adapted to address the complexity and heterogeneity of tumors, where small sets of genes may be aberrantly co-expressed in specific subsets of tumors. Biclustering algorithms that perform local clustering on subsets of genes and conditions help address this problem. We have proposed a graph-based Tunable Biclustering Algorithm (TuBA) (Chapter 2) based on a novel pairwise proximity measure that leverages the size of the data sets to identify subsets of tumor samples that co-express subsets of genes at their highest or lowest levels relative to other samples.

We applied TuBA to three large gene expression datasets encompassing a total of 3,940 breast invasive carcinoma (BRCA) patients (Chapter 3). We demonstrated that there was significant agreement between the results obtained for each data set, and discovered that about 50% of the altered co-expression signatures were associated with a subtype of the disease that exhibits low levels of expression of the estrogen hormone receptor 1 (ER) and the human epidermal growth factor receptor 2 (HER2) genes. Tumors belonging to this subtype are labelled as ER−/HER2−. Since only 15% of all BRCA patients are estimated to have tumors that belong to this subtype, our algorithm was able to highlight the tremendous heterogeneity in alterations within tumors of this subtype. Quite significantly, more than 50% of these signatures were associated with alterations in the DNA that results in amplification (or deletion) of genes copies, which subsequently result in higher (or lower) level of gene expression. Thus, TuBA was especially effective in identifying transcriptionally active

copy number variations in tumor samples. Finally, TuBA identified biclusters that were associated with the tumor microenvironment, which included biclusters associated with infiltrating immune and stromal cells. These can improve our understanding about the role played by the microenvironment in modulating tumor progression.

We showed that TuBA outperforms other algorithms in identification of co-expressed genes located in transcriptionally active copy number altered sites (Chapter 4). Moreover, from a differential co-expression perspective, TuBA offers an advantage over other methods since no prior specification of subsets of samples (conditions) is necessary; the nature of our proximity measure ensures that such differential co-expression signatures are preferentially identified.

In summary, our method identified a multitude of altered transcriptional profiles associated with the tremendous heterogeneity of diseased states in breast cancer. Exploring the diversity of these aberrant signatures can help identify potential biomarkers of clinical relevance that can further improve treatment outcomes, especially for ER$-$/HER2$-$ breast cancers. Although transcriptomic alterations are not the ultimate determinants of progression of disease, our algorithm holds the promise to improve therapeutic selection and design by identifying significantly altered transcriptional patterns associated with tumors.

Acknowledgements

Published work

The work described in Chapters 2, 3, and 4 is published in the *GigaScience* journal [1] (referenced below). Work described in Chapter 5 are currently being prepared for publication.

Chapters 2, 3, and 4: Amartya Singh, Gyan Bhanot, and Hossein Khiabanian.

TuBA: Tunable biclustering algorithm reveals clinically relevant tumor transcriptional profiles in breast cancer. *GigaScience*, 8(6), 06 2019. ISSN 2047-217X. doi:10.1093/gigascience/giz064.

Work of others

The work on ribosomal genes described section 5.2 in Chapter 5 is part of a collaborative study (of which I am a co-author) for which the manuscript is currently under preparation. The joint first authors of this study are Anshuman Panda and Anupama Yadav. Other authors of this study are Huwate Yeerna, Michael Biehl, Markus Lux, Alexander Schulz, Tyler Klecha, Sebastian Doniach, Hossein Khiabanian, Shridar Ganesan, Pablo Tamayo, and Gyan Bhanot. The data analysis for the results described in section 5.2 was performed by me.

Source of data

I relied on publicly available data sets for the analysis described in this thesis. The results discussed in this thesis are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The processed TCGA data sets were downloaded from the UCSC Xena portal (https://xena.ucsc.edu/). I also relied on the cBio portal (https://www.cbioportal.org) and the Gene Expression Omnibus (GEO) portal (https://www.ncbi.nlm.nih.gov/gds) for other cancer data sets.

Gratitudes

I am deeply grateful to my advisor Dr. Hossein Khiabanian for giving me the opportunity to work with him and develop this project from scratch. Without his constant support and feedback this

project would not have been possible. As my advisor, it must have been exhausting to witness the countless acts of foolishness I committed, and the innumerable times I floundered even with the simplest of problems. Perhaps, he found it all quite amusing. All the same, he was always extremely generous with his time and attention, even for those not so infrequent occasions when I would be upset about everything that is wrong with the world. I am also greatly indebted to my co-advisor Prof. Gyan Bhanot. Four summers ago, I walked into his office completely oblivious to the fact that that meeting would alter the course of my life. A long list of books to read, and some extremely kind words of support and encouragement were what I walked out with that day. The words of encouragement never ceased over the next four years, which speaks more about how he is as a human being than about how I was as a student. I would also like to take this opportunity to thank some very special people with whom I have lived the better part of the past 5 years. They put up with all my tantrums and outbursts, and still cared for me selflessly. These people truly are my family away from home: Suryateja Gavva, Deepti Jain, Aditya Ballal, Ruturaj Apte, Jay Vora, Aditya Potukucchi, and Abhishek Bhrushundi.

Finally, a special note of thanks for the truly amazing Berklee Indian Ensemble. During moments of despair and despondency, their beautiful songs soothed and healed me by offering the warmest of embraces. I'm especially grateful to Vasundhara Gupta for 'One'. Her heartfelt compositions overwhelm me with feelings of pure elation and gratitude every single time I listen to them. To her, and the entire Berklee Indian Ensemble, I will forever owe a great debt of gratitude.

Dedication

To my family, who could not have been more supportive and caring

Table of Contents

List of Tables

List of Figures

Chapter 1
Background

"Answer. That you are here - that life exists and identity. That the powerful play goes on, and you may contribute a verse."

– Walt Whitman, *Leaves of Grass*

"..vices are sometimes only virtues carried to an excess!"

– Charles Dickens, *Dombey and Son*

1.1 A cell biology primer

Cells are the building blocks of all of life on Earth [2]. The building blocks themselves are made of a diverse array of molecules that serve a variety of functions. Most of these molecules belong to one of three classes of molecules - deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins. The entire set of DNA within the nucleus of the cell is called the genome. According to Matt Ridley [3], if we imagine the human genome as a book, it consists of 23 chapters, called chromosomes. Each of these chapters contain thousands of stories, called genes. The paragraphs that convey parts of the stories are called exons. Interspersed between such paragraphs are advertisements called introns. All the words, called codons, are three letters long and use only 4 letters - A, T, C, G - called bases. The bases are A - adenine, T - thymine, C - cytosine, and G - guanine. While the book analogy provides a good perspective about the form of the genome, it does not suffice to explain how the information manifests itself to enable life's processes.

Described as the central dogma of molecular biology by Francis Crick in 1970 [4], the flow of information between DNA, RNA, and proteins holds the key to the sustenance and perpetuation of life in all its forms. Figure 1.1 shows the flow of information from DNA to RNA to protein as per the central dogma. The solid arrows indicate the information flows that are ubiquitous and essential in all cells. The dashed arrow indicates a flow that takes place occasionally. The steps numbered in the figure refer to the following processes respectively:

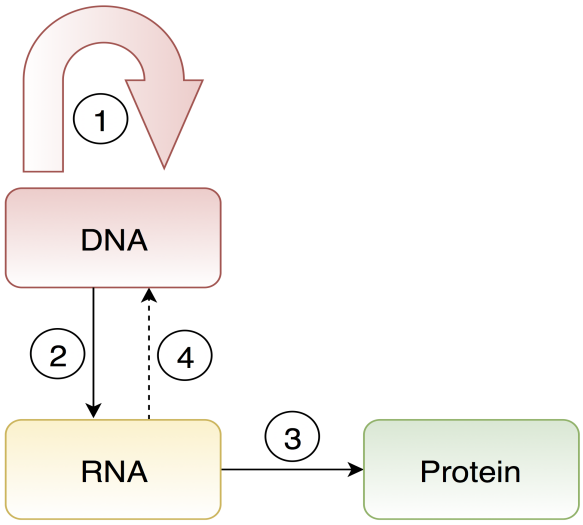

Figure 1.1: **Central dogma of molecular biology.**

1. *Replication*: Duplication of DNA during cell cycle of a parent cell while giving birth to two daughter cells using an enzyme called the DNA polymerase

2. *Transcription*: Production of single-stranded RNA copies of DNA templates using an enzyme called the RNA polymerase

3. *Translation*: Production of proteins based on RNA templates using a molecular machine made up of RNAs and proteins called the ribosome

4. *Reverse Transcription*: Production of complementary DNA (cDNA) from RNA templates using an enzyme known as the reverse transcriptase

All the cells in our body contain identical copies of DNA (some differences do exist, but the number of such differences are small compared to differences between DNA of cells from different individuals). This is because all of them are descendants of a single zygote (fertilized egg). Despite possessing the same DNA, there is an enormous difference in the identities and functions of cells that make up distinct tissues in our bodies (compare cells that make up your skin, to the cells that make up the cornea in your eyes). This is a truly remarkable fact. How are these differences realized? Put in the simplest possible way, these differences arise from choices about which stories get told when, where, and how often. In slightly more formal terms, every cell assumes its identity by one way or another arriving at answers to the following questions: (i) which genes should get transcribed, (ii) which of the transcripts should get translated, and (iii) how many transcripts are produced for each gene? Gene expression in cells is regulated through a number of mechanisms that operate at different levels of the flow of information according to the central dogma. At the transcription level, *transcription*

factors play the dominant role. Transcription factors (TF) are proteins that regulate transcription of their target genes by binding to specific sequences in the DNA. Their presence increases or decreases the likelihood of the binding of RNA polymerase to the DNA promoter regions near the target gene; when a TF binding to its target sites leads to an increase in the level of the transcripts, it is said to act as an *activator*, conversely, when it decreases the number of transcripts of the target gene produced in the cell, it is said to act as a *repressor*.

Another level of gene regulation is through what are known as *epigenetic* factors. These refer to heritable factors which influence gene expression, but are not associated with alterations in the sequence of the DNA itself [5, 6]. These changes include:

1. **DNA Methylation** - A heritable chemical modification of the DNA that is also very dynamic and can be created and/or modified due to external environmental stimuli. The most common form of methylation occurs at cytosine (C) sites that are immediately followed by a guanine (G). These are represented as CpGs. The regulatory regions of quite a few genes are enriched in CpGs. These enriched regions are popularly known as *CpG islands*. Methylation of Cs in these islands are frequently associated with a suppression of the expression levels of the gene downstream [7].

2. **Histone modifications** - Histones are protein complexes that are responsible for efficient packing of the DNA (chromatin) in the nucleus. The histones can get modified due to methylation or acetylation, such that certain portions of the DNA are no longer accessible by the transcriptional machinery. This leads to a suppression in the expression levels of the transcripts of the affected genes [8, 9].

After a gene is transcribed into an RNA, there are further controls that a cell relies on to regulate the expression levels of the final protein. These controls are refered to as *post-transcriptional* mechanisms and include:

1. **Alternative splicing** - The exons in the transcribed RNA, called the pre-RNA, can be rearranged and put together to yield different mRNAs, and as a result different proteins. This explains in part the diversity of phenotypes observed with such a limited number of protein coding genes in the genome [10, 11].

2. **Regulation by non-coding RNA (ncRNA)** - Based on research over the last few decades, it is now clear that the majority of the genomes of most complex organisms do not code for proteins. Are these regions never transcribed? Far from it. Several such regions in the genome are in fact transcribed into RNA, but these RNA do not undergo translation to make

corresponding proteins. These RNA molecules are called *noncoding* RNA (ncRNA). In fact, the RNAs that code for proteins are called messenger RNA (mRNA) to distinguish them from ncRNAs. The ncRNAs play a number of infrastructural roles (ribosomal RNAs, transfer RNAs etc), as well as regulatory roles [12]. RNA interference mechanisms involving interactions between ncRNAs and mRNAs have been described that have an inhibitory effect on gene expression at the translation stage [13]. The small ncRNAs that play a role in regulating gene expression are called micro RNAs (miRNAs). miRNAs have been predicted to have regulatory influence on almost half of the protein-coding genes [14]. Note, this mechanism of regulation is not captured within the paradigm of the central dogma.

Finally, even after the RNAs get translated into proteins, some of these proteins remain inactive and do not perform any function until they are activated by what are called *post-translational modifications* (PTMs). The most common PTM is phosphorylation, which is simply the addition of a phosphate group to a protein molecule. External stimuli (could also include internal stimuli such as DNA damage, osmotic pressure etc) undergo transduction by activating inactive proteins in the cell through a series of phosphorylations mediated by proteins called *kinases*. For different kinds of stimuli, there exist a large number of pathways related to particular cellular functions that determine either the equilibrium or the dynamic state of the cell.

In multicellular organisms, all of the mechanisms described above are involved in a complex interplay to ensure that each cell functions cooperatively, and fulfills its specific role. It is hardly a linear sequential flow of information as the central dogma may appear to suggest. It is through changes to these mechanisms that certain cells assume identities that are detrimental to the overall health of the organism. One of these identities, perhaps the most well-known and the one most dreaded, is that of cancer.

1.2 Cancer - A genetic disease

Cancer is not a singular disease. It is a constellation of diverse and evolving disorders manifested by uncontrolled proliferation of cells which can eventually lead to the death of the host [15]. Notwithstanding the diversity, as Vogelstein and Kinzler pithily observed [16], "*Cancer is, in essence, a genetic disease.*" While the genetic nature of cancer is accepted as the gospel truth today, it was only through the painstaking efforts of early geneticists in the 1970s that the first "cancer genes" (proto-oncogenes) were discovered. Over the years with the help of screening assays, many new oncogenes, as well as genes that fulfill tumor-suppressive roles have been discovered. At present, there are 723 genes listed in the Cancer Gene Census [17] of the Catalogue of Somatic Mutations in

Cancer (COSMIC) database (v88, 19-March-2019) that have been causally implicated in cancers. Additional evidence in support of the genetic nature of cancer came through studies that examined predisposition to diseases within families. The most compelling evidence for such familial predisposition are the associations between germ line mutations of the *BRCA1* and *BRCA2* genes and risk of breast cancers, and the germ line mutations of the *RB1* gene and risk of development of retinoblastoma, respectively. Individuals with mutated copies of *BRCA1* and *BRCA2* genes have a 10-fold lifetime risk of developing breast cancer, while individuals with mutated copies of the *RB1* gene have more than 90% risk of developing a retinoblastoma.

Accumulation of such mutations by cancer cells leads to a change in the transcriptional and translational regulatory programs described in the previous section. For example, in their review article [16], Vogel and Kinzler consider the *TP53* gene which is a transcription factor that plays a crucial role in normal cells as a tumor suppressor. Due to accumulation of mutations, a mutated version of TP53 protein can no longer bind successfully to its target genes, and as a result fails to fulfill its role as a tumor suppressor. Hypermethylation of CpG sites in promoter regions of tumor suppressor genes has also been frequently observed, which leads to suppression in the expression levels of those genes [18]. Histone methylation over large genomic regions leading to suppression in expression of genes located within those regions has been reported for multiple cancer types [6, 19, 20]. Moreover, alternative splice forms have been found in cancer cells that are not found in normal cells of the same tissue type [21, 22]. These are just a few of the alterations that may be present in cancer cells. From the point of view of the cancer cells, these are essential to enable them to survive and evolve successfully in adverse environments against great odds. Despite the heterogeneity in their alterations and identities, they nevertheless share some common traits highlighted by Hanahan and Weinberg [23, 24], which we describe briefly in the next section.

1.3 Hallmarks of Cancer

In their paper on the hallmarks of cancer [23], Hanahan and Weinberg proposed 6 fundamental traits that all cancers must exhibit for tumor growth and metastasis (spread of tumor cells to other tissues in the body). They defined the hallmarks as acquired functional capabilities that allow cancer cells to survive, proliferate, and disseminate:

1. **Sustaining proliferative signalling:** According to Hanahan and Weinberg, the most fundamental trait of cancer cells involves their ability to sustain proliferation. Cancer cells deregulate the signals that ensure homeostasis of cell number. They do not remain dependent on growth factors from external sources.

Figure 1.2: **Hallmarks of Cancer.** Hallmarks shown in blue circles are the enabling characteristics, and the ones in green are the emerging hallmarks [24]. Figure adapted from [23, 24].

2. **Evading growth suppressors:** Cancer cells gain the ability to circumvent programs that negatively regulate cell proliferation, including cell-to-cell contact inhibition present in normal cells.

3. **Resisting cell death:** The natural process of programmed cell death (apoptosis) is attenuated in cancer cells. The most common way in which cancer cells achieve this is by losing the tumor suppressive function of TP53.

4. **Enabling replicative immortality:** Cancer cells overcome the barriers that limit the number of times cells can go through the cell growth-and-division cycles. They exhibit unlimited replicative potential, and are therefore immortal.

5. **Inducing angiogenesis:** Angiogenesis is the term used to describe the process by which new blood vessels are formed. In tumors, angiogenesis is always activated and *remains on*, causing the surrounding blood vessels to continually sprout new vessels that help sustain expanding new growth.

6. **Activating invasion and metastasis:** Cancer cells develop alterations in their shape as well as in their attachment to other cells and to the extracellular matrix (ECM) (for example, by

losing E-cadherin, a key cell-to-cell adhesion molecule). They also up-regulate the expression of molecules associated with embryogenesis and inflammation which promote cell migration.

7. **Reprogramming energy metabolism (Emerging Hallmark [24]):** The chronic and often uncontrolled cell proliferation involves not only deregulated control of cell proliferation but also corresponding adjustments of energy metabolism in order to fuel cell growth and division. Cancer cells can reprogram their glucose metabolism, and thus their energy production, by limiting their energy metabolism largely to glycolysis (normally favored under anaerobic conditions) leading to a state called *aerobic glycolysis*.

8. **Evading immune destruction (Emerging Hallmark [24])**: Evidence suggests that the immune system operates as a significant barrier to tumor formation and progression, at least in some forms of non-virus-induced cancer. Therefore, cancer cells may be evading immune destruction by avoiding detection or suppressing normal immune response.

In a follow up paper [24], they elaborated on these hallmarks and asserted that the acquisition of the hallmarks is made possible by the following 2 *enabling characteristics*:

1. **Genome instability and mutation:** The mutations and chromosomal alterations enabled by loss of integrity and instability of the genome confer selection advantage to subsets of cancer cells. This characteristic is causally associated with the acquisition of hallmark capabilities.

2. **Tumor-promoting inflammation:** Inflammation can contribute to multiple hallmark capabilities by supplying bioactive molecules to the tumor microenvironment. These may include growth factors that sustain proliferative signaling, ECM-modifying enzymes that facilitate angiogenesis etc.

Note, the hallmarks of cancers are by no means independent. All of them are associated with gene networks and signalling pathways that are intimately connected with each other. Thus, cancer as a disease of genes is quintessentially a consequence of deregulation of these intimately linked gene networks or pathways that modulate growth and dissemination. In the following section, we explore the enabling characteristic of genome instability and mutation in a little more detail.

1.4 Chromosomal alterations in cancer

Cells with normal chromosomal configuration are said to be in the *euploid* karyotype state, where karyotype refers to the number and visual appearance of chromosomes within the nuclei of a cell. *Aneuploidy* (deviation from euploidy) is extremely common in cancers [25, 26]. Beyond a certain

number of mutations, it is no longer possible to carry out error-free duplication due to the fact that the defects induced by the mutations impair the DNA repair functions and also deregulate the cell cycle checkpoints.

Theodor Boveri first proposed the idea that irreparable chromosomal defects may be responsible for turning normal cells to cancer cells in 1902 [27, 28]. However, it's only now after several decades that the idea has been brought to the forefront of cancer research based on recent findings on chromosomal rearrangements [29, 30]. The following chromosomal alterations are frequently observed in cancers:

1. **Aneuploidy:** refers to the state of cells with additional or missing copies of one or more chromosomes.
2. **Polyploidy:** refers to the state of cells with more than 2 sets of chromosomes.
3. **Translocation:** refers to the movement of chromosomal segments either within the same chromosome, or to another chromosome.
4. **Inversion:** refers to the reversal of the order of the genes in subsequences in the chromosome relative to the neighbouring sequences.
5. **Point mutations:** refers to the substitution of bases in the DNA sequence with other bases due to errors during replication or repair.
6. **Amplification:** refers to the presence of more than 2 copies of contiguous genes in some regions within the chromosomes.
7. **Deletion:** refers to the loss of one or both copies of genes in some regions within the chromosomes.

Over the last few decades, technologies have evolved at a rapid pace and have enabled us to accurately characterize each tumor sample based on these chromosomal alterations and other molecular features. Techniques that enable an exhaustive characterization of the genomic profiles of samples are called *genome-wide* techniques. In particular, techniques that investigate changes at the DNA sequence level (such as point mutations, amplifications, and deletions) are called *genomic* techniques; techniques that investigate the changes at the level of proteins are called *proteomics* techniques. In this work, we have relied on data generated from high-throughput *transcriptomic* techniques that investigate the relative abundances of transcripts in the cells. In the following section, we describe how the data for gene expression profiling is generated using technologies that simultaneously measure the abundance levels of thousands of transcripts.

1.5 Measuring gene expression

1.5.1 Why do it?

Because we can. This trivial response to the question posed above masks a number of complex factors at play that make some biological measurements much more challenging compared to others. Since it is the proteins that perform most of the functions, finding out which of them are expressed and functionally active in the cells, and in what amounts, would appear to be the obvious course of action. Except that these measurements are difficult. Rapid advances in proteomics techniques have been made over the past decade but many challenges still remain [31]. High-throughput transcriptomic techniques on the other hand are significantly simpler (for one, segregation and purification of RNA is much easier compared to proteins). Because of studies based on these high-throughput transcriptomic techniques, our understanding of the regulation of gene expression at the level of the transcriptome has improved dramatically. However, there is a caveat. While associating gene expression levels with phenotypes, we might be tempted to assume that gene expression levels correlate perfectly with protein expression levels. In a landmark study of the correlation between mRNA levels and protein expression levels in Yeast, Gygi *et al.* [32] showed that for some genes, while the mRNA levels barely varied, the protein levels varied by more than 20-fold. Conversely, they observed that while the levels for certain proteins were consistent, their respective mRNA transcript levels varied by as much as 30-fold. Thus, one needs to be careful when associating transcript abundance levels with phenotypes. Notwithstanding the caveat, understanding the regulation of gene expression at the level of the transcriptome, especially for cancers, does help us build valuable bridges between the genotypes and the tumor phenotypes. In the following subsections, we describe the general principles behind the two widely adopted techniques used to measure gene expression.

1.5.2 DNA Microarrays

The two most commonly used types of DNA microarrays are: oligonucleotide microarrays, and complementary DNA (cDNA) microarrays. While the two differ in their manufacturing, labelling, and analysis approaches, the broad principles are similar:

1. Single-stranded DNA sequences complementary to the genes whose abundances we wish to quantify, are fixed on some solid support (glass, silicon etc.). These are called *probes*. Several copies of the probes are fixed very close to each other to form *spots*.

2. RNA is extracted from the sample.

3. The mRNA is copied into cDNA using reverse transcriptase and labelled fluorescent nucleotides. These labelled cDNA are the *targets*.

4. The targets are allowed to hybridize to the probes. Targets only hybridize to probes with complementary sequences.

5. The array is washed after hybridization and scanned with a fluorescence microscope.

6. The spots with higher intensity of fluorescence indicate higher relative abundance of the corresponding transcripts.

Limitations of microarrays

There are some key limitations of gene expression measurements using microarrays:

- Limited to known transcripts and organisms with sequenced genomes
- Non-specific hybridization or cross-hybridization of closely related gene family members
- Limited dynamic range (signal range) - for example, it is difficult to distinguish *no* expression from low levels of expression due to background noise

1.5.3 RNA Sequencing (RNA-seq)

RNA-seq refers to the application of high-throughput sequencing technologies to quantify mRNA abundances. These high-throughput technologies rely on diverse chemical and physical means to perform these measurements. Despite differences in details of the particular means adopted, RNA-seq broadly involves the following key steps:

1. *Extraction*: Total RNA is extracted from the sample and purified.

2. *RNA Fragmentation*: The mRNAs are fragmented into shorter fragments by random shearing.

3. *Reverse transcription*: The sheared fragments of the mRNAs are reverse transcribed to cDNA using *primers*. Primers are DNA (or RNA) molecules whose $3'$ end is the initiation point of DNA synthesis by DNA polymerase.

4. *Adapter ligation*: The $5'$ and $3'$ ends of the cDNAs are repaired, and adapter sequences which allow them to hybridize are added.

5. *Amplification*: Correctly ligated cDNA fragments are amplified by polymerase chain reaction (PCR) to ensure there is sufficient signal.

6. *Quantification*: Transcript abundances are quantified based on the signal assessed by image analysis.

Biases

While RNA-seq overcomes all the limitations of microarray analysis (quite importantly, RNA-seq can quantify low abundance reads reliably). However, there are a few biases that one needs to be mindful of when dealing with RNA-seq data:

- Bias can be introduced by PCR due to non-linear amplification of genomic regions with low nucleotide complexity such as sequential GpCs [33].
- Total number of reads per transcript is proportional not just to the abundance of the transcript but also to the transcript length. This is usually accounted for during the normalization step prior to any form of clustering analysis (more on this below).

1.6 Normalization

1.6.1 Microarray

The idea behind normalization is to account for systematic variation in gene expression that can arise as a consequence of:

- Biases introduced during experimental setup
- Variability in conditions (when the assumption is that the conditions are unchanged)
- Differences in sample collection and preparation

There are two kinds of normalizations that one may be required to perform depending on the question of interest:

1. *Within array normalization* - Required to account for systematic differences in the intensities (of spots) and location dependent biases of the fluorescent labels. This ensures an unbiased comparison between genes within a sample.
2. *Between array normalization* - Necessary for comparing genes across different conditions (for example, whether a given gene is differentially expresssed between two groups of samples representing two different conditions). Usual approaches include scaling each array by a constant factor to make sure that the median intensities of all the arrays are the same (global normalization), or by matching the percentiles of each array (quantile normalization)

1.6.2 RNA-seq

Even for RNA-seq gene expression data, depending on the question of interest one would need to perform:

1. *Within sample normalization*: As pointed out earlier, the number of reads corresponding to a transcript depends on its length (this is because longer transcripts will produce more fragments during the random shearing process). Thus, in order to compare genes of different lengths within a sample, the counts of the transcripts should be normalized by their lengths. The most popular method is to convert the raw counts into a measure called transcripts per million (TPM) [34]. TPM can be interpreted as follows: suppose we sequenced 1 million *full length* transcripts, TPM_i would be the number of reads we would see of type i, given the abundances of the other transcripts in our sample.

2. *Between sample normalization*: As noted earlier for microarrays, between sample normalization is required to enable us to compare expression features (genes, isoforms) across conditions. This is necessary for standard differential expression analysis, as well as clustering analysis of gene expression data [35]. Between sample normalization addresses the differences in the sequencing depth (total number of reads) between the samples of an experiment. Most popular methods (such as DESeq [36] and TMM [37]) broadly revolve around the same idea, which is that most of the genes are not differentially expressed, thus one could find genes that are similarly expressed across conditions and rely on these to determine the scaling factors for each sample.

1.7 Analyzing gene expression data

1.7.1 Clustering

Over the last decade and a half, comprehensive large-scale genomic studies have resulted in an unprecedented increase in both the depth (number of molecular and clinical aspects explored) and breadth (number of samples investigated) of well-curated data for the human genome and transcriptome. Collaborative efforts such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), have undertaken large-scale studies to explore various aspects of tumors at the molecular level. Many of these studies have generated large gene expression data sets for multiple cancer types. Fig. 1.3 shows a typical gene expression data set with genes along the rows and samples along the columns. The small es in the matrix can be any real number (raw count data from RNA-seq measurements is in integer form).

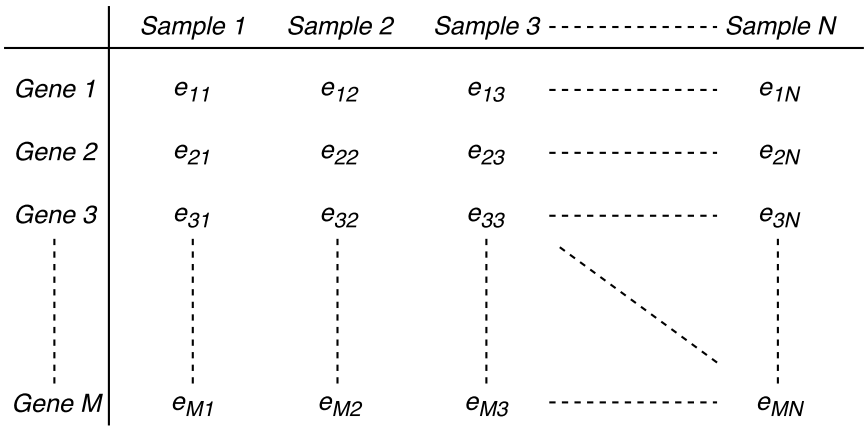

Figure 1.3: **Schematic illustration of a typical gene expression data set.**

.

A simple yet meaningful way to organize and comprehend large gene expression data sets is to group together (cluster) similar variables (genes) or conditions (samples) based on some mathematical measure of proximity between the entities of interest. The most commonly used measure of proximity or similarity is the *Pearson's correlation coefficient*, defined as:

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$

where $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is the mean of the values of vector x, and $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ is the mean of the values of vector y. The two vectors x and y could correspond to expression values of a pair of genes across the samples, or they could be expression values for a pair of samples across the genes. Based on whether we group rows or columns, the results could respectively yield information about:

- Which sets of genes exhibit similar patterns of expression levels across all the different conditions? Based on this, one could also classify a new gene by determining its proxmimity to a particular cluster.
- Which conditions are similar to each other across all the genes? Once again, we could rely on this information to classify a new sample.

Since very little is known about the variables as well as the conditions a priori, clustering is usually performed in an unsupervised manner whereby the genes are clustered based on pairwise (dis)similarity assessed over all the conditions and vice versa [38]. This approach to clustering is also known as global clustering. Some measure of success in analyzing gene expression data sets of cancers was achieved by relying on partially supervised approaches based on histological and/or clinical information [39, 40, 41]. In the following section, we introduce a form of unsupervised clustering that groups together similar subsets of genes and similar subsets of samples simultaneously.

1.7.2 Biclustering

Clustering approaches based on global proximity measures are not optimal for analyzing large gene expression data sets of complex diseases such as cancers, because of the following reasons:

1. Generally, and this is true in normal cells as well, only subsets of genes participate in any given cellular process.
2. A cellular process may be active only under a subset of conditions.
3. Cancers are heterogeneous, with distinct disease subtypes driven by deregulation of diverse gene networks, even within a single tissue type.

However, biclustering algorithms that perform local clustering of gene expression data sets are particularly well adapted to address these concerns. Such algorithms are subject to the following constraints [42]:

1. Clusters of genes should be defined with respect to only a subset of conditions, and vice versa.
2. The clusters should not be exclusive and/or exhaustive a gene/condition may belong to more than one cluster, or to none at all.

As described earlier, gene expression data sets are usually in the form of real valued matrices with genes along the rows and the samples along the columns. Biclusters within these matrices can broadly be defined as subsets of rows (genes) that exhibit similar or homogeneous behaviour across a subset of columns (samples). Depending on how this similarity or homogeneity is defined, the type of biclusters can be classified into 4 kinds that are briefly described below [42, 43].

1.7.3 Biclusters with constant values

Such a bicluster would correspond to a sub-matrix (I, J) where all the values in the sub-matrix are equal, i.e.,

$$a_{ij} = \mu \qquad \forall\, (i, j) \in (I, J)$$

The most intuitive metric to find such biclusters is the *variance*. The variance for a bicluster is defined as:

$$VAR(I, J) = \sum_{i \in I,\, j \in J} (a_{ij} - a_{IJ})^2$$

where a_{IJ} is the mean of all the elements of the bicluster.

The ideal bicluster would be one for which the variance is 0. Of course, this would trivially be true for sub-matrices of size 1 (single element). Thus, methods that seek constant-valued biclusters put in constraints for the minimum number of rows that should be included in the biclusters [44].

1.7.4 Biclusters with constant values on rows or columns

Biclusters with coherent variations across rows or columns of the matrix are of more practical interest for gene expression data, since rows with constant values across a subset of columns would correspond to genes that have the same expression value within a subset of conditions. Similarly, columns with constant values along a subset of rows would correspond to conditions within which a subset of genes have similar expression values.

Thus, the expression values within the bicluster (I, J) with constant rows can be given by either,

$$a_{ij} = \mu + \alpha_i \qquad \forall\, (i,j) \in (I,J)$$

or,

$$a_{ij} = \mu \times \alpha_i \qquad \forall\, (i,j) \in (I,J)$$

where μ is a base value within the bicluster (I, J) and α_i is the additive or multiplicative scaling factor for row i. Similarly, for a bicluster with constant columns, the expression values can be given by either,

$$a_{ij} = \mu + \beta_j \qquad \forall\, (i,j) \in (I,J)$$

or,

$$a_{ij} = \mu \times \beta_j \qquad \forall\, (i,j) \in (I,J)$$

where β_j is the additive or multiplicative scaling factor for column j. The usual way to find such biclusters is to first normalize the rows or columns of the candidate sub-matrix by the row mean or column mean, respectively [45]. This transforms the sub-matrix to a form similar to the one with constant values described in the previous subsection. One can then employ biclustering methods that can identify constant biclusters. We have provided simple illustrations of these two kind of biclusters with constant values along rows and columns in Fig. 1.4.

1.7.5 Biclusters with coherent values

A more general and practical extension of value based biclusters is to consider biclusters with coherent values on *both* rows and columns (Fig. 1.4). For gene expression data, this would correspond to subsets of genes that have coherent values across subsets of conditions, and vice versa. Two types of models describe the values within biclusters that belong to this type, (i) additive, and (ii) multiplicative.

The values in an ideal bicluster (I, J) based on the additive model are given by,

$$a_{ij} = \mu + \alpha_i + \beta_j \qquad \forall\, (i,j) \in (I,J)$$

Constant rows

50	50	50	50	50	50
80	80	80	80	80	80
36	36	36	36	36	36
55	55	55	55	55	55
41	41	41	41	41	41
64	64	64	64	64	64

Constant columns

50	80	36	55	41	64
50	80	36	55	41	64
50	80	36	55	41	64
50	80	36	55	41	64
50	80	36	55	41	64
50	80	36	55	41	64

Coherent values - additive

50	20	40	70	80	60
80	50	70	100	110	90
90	60	80	110	120	100
40	10	30	60	70	50
45	15	35	65	75	55
80	50	70	100	110	90

Coherent values - multiplicative

30	60	15	45	36	18
40	80	20	60	48	24
20	40	10	30	24	12
10	20	5	15	12	6
50	100	25	75	60	30
60	120	30	90	72	36

Figure 1.4: **Illustration of biclusters with constant or coherent values.** The upper two biclusters are constant valued along rows and columns, respectively. The bottom two biclusters have coherent values along their rows and columns. Figure adapted from [42].

.

where once again, μ is a base value for the bicluster, α_i is the additive scaling factor for row i, and β_j is the additive scaling factor for column j.

For the multiplicative model, the values in the bicluster (I, J) with coherent values would be given by,

$$a_{ij} = \mu' \times \alpha'_i \times \beta'_j \qquad \forall\, (i,j) \in (I,J)$$

where μ' is a base value for the bicluster, α'_i is the multiplicative scaling factor for row i, and β'_j is the multiplicative scaling factor for column j.

1.7.6 Biclusters with coherent evolution

The biclusters described in the previous subsections were all based on models for the actual expression values. Another approach to detect meaningful patterns in the data matrix is to seek coherent *evolution* across subsets of rows and/or columns. For gene expression data in particular, the question of interest may be to seek genes that are differentially regulated (up or down) within a subset of conditions. Co-evolution patterns are found without relying on models for the actual expression values in the data sets. This can be done by taking into account the ordinality (relative order) of the genes and conditions based on the expression values. The magnitudes and/or uniformity of the expression values is irrelevant. The only aspect of any relevance is their relative ranking. Quite often, some kind of data discretization step is employed to transform the numerical data into a form that is essentially categorical in nature [46, 47].

Broadly, these biclusters are of 4 types (see Fig. 1.5):

1. *Overall coherent evolution* - All the elements of the bicluster belong to the same category. For example, imagine finding biclusters that only contains 1s in a binary matrix obtained from a gene expression data set that was binarized using some threshold.

2. *Coherent evolution on rows* - The values in the rows of the bicluster fall within the same category across its columns. If these categories correspond to ranges of expression values, then lighter shades of red in Fig. 1.5 could be associated with lower expression values, and darker shades of red associated with higher expression values.

3. *Coherent evolution on columns* - The values in the columns of the bicluster fall within the same category across its rows. Once again, if these categories correspond to ranges of expression values, then lighter shades of red could be associated with lower expression values, while darker shades of red would be associated with higher expression values. A more general case that does not correspond to such a situation is illustrated in the bottom right panel of Fig. 1.5.

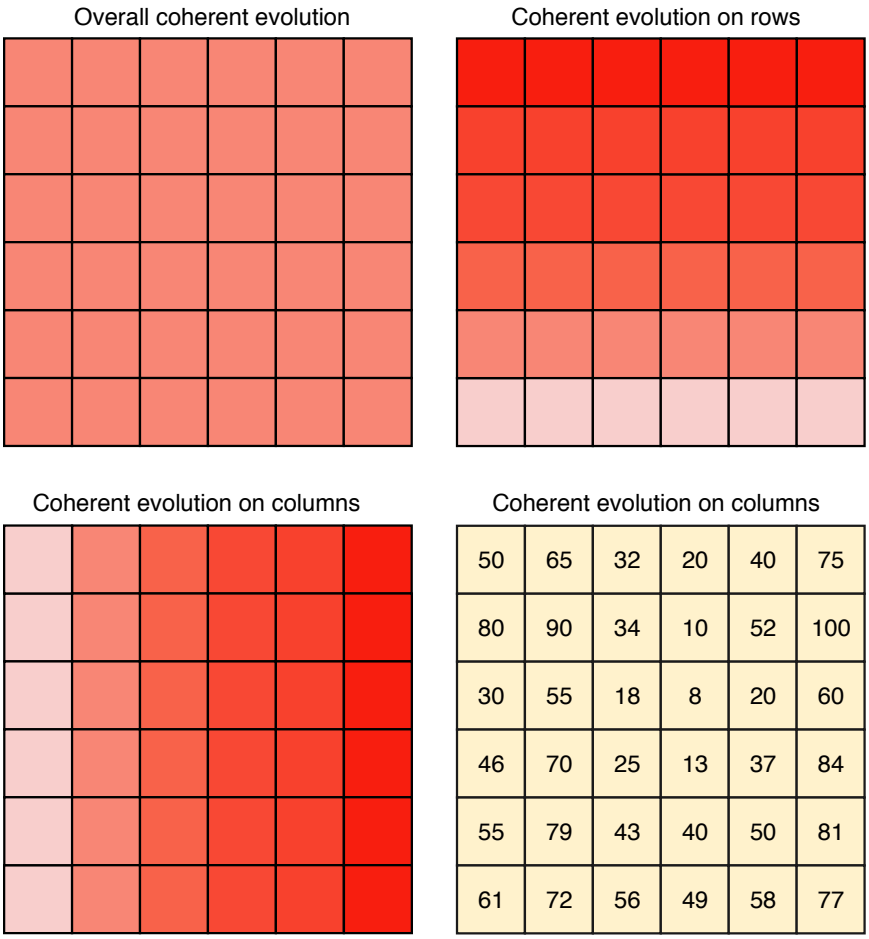

Figure 1.5: **Illustration of biclusters with coherent evolution along rows and/or columns.** Lighter shades of red may correspond to lower expression values for gene expression data, while higher shades indicate higher expression values. Figure adapted from [42].

.

The columns in the illustrated matrix can be permuted such that the values in all the rows are strictly increasing.

The algorithm proposed by us finds biclusters that seeks gene expression signatures associated with up-regulation or down-regulation without relying on models of the actual gene expression values. In that regard, our algorithm is similar to algorithms that find coherent evolution patterns in gene expression data. However, the biclusters identified by our algorithm are markedly different in nature from the ones described in this subsection. We describe our algorithm and the properties of its biclusters in the next chapter.

Chapter 2
The Algorithm

"Normals teach us rules; outliers teach us laws."

– Siddhartha Mukherjee, *The Laws of Medicine*

"Okay you guys, pair up in threes!"

– Yogi Berra

2.1 Motivation

A significant amount of research on cancers revolves around the following questions:

- What makes tumor cells different from normal cells belonging to the same tissue of origin?
- What makes tumor cells of one tumor arising from a given tissue of origin, different from tumor cells belonging to another tumor that also arises from the same tissue of origin? (Intertumoral heterogeneity)
- What makes tumor cells different from each other even within the same tumor? (Intratumoral heterogeneity)

In terms of gene expression, if these differences exist at the level of transcripts, they would be evident in terms of differences in the real numbers that represent the relative abundances of these transcripts in the gene expression matrix. In order to analyze these differences in the patterns of gene expression that may be associated with altered mechanisms in tumors, one could pursue two courses of action:

1. Assume that the expression level of genes across conditions can be approximated by a parametric distribution and determine whether some subsets of conditions show up as outliers.
2. Without assuming any underlying distribution, employ a non-parametric statistical test to determine differences in gene expression levels between subsets of conditions.

There is a significant difference between the two approaches described above. While the first approach does not explicitly require prior knowledge about the conditions in order to ascertain the subset of conditions that would qualify as outliers, implicitly it requires assumptions about the expected levels of expression across conditions. Given the heterogeneity of alterations exhibited by tumors, it is extremely challenging to propose models of gene expression patterns that capture the essential properties of all of the underlying alterations. The second approach on the other hand is limited by the requirement of prior knowledge about the conditions.

In this chapter, we describe a novel biclustering method, called the Tunable Biclustering Algorithm (TuBA), based on a measure of proximity that enables it to preferentially identify aberrantly co-expressed genes and associated pathways in subsets of cancer patients. The proposed proximity measure does not presume any parametric distribution for the expression levels of genes across conditions, nor does it require prespecification of prior information about the conditions. As described in greater detail in the following section, our proximity measure simply leverages the size of the datasets (specifically, the number of samples in the datasets) to infer gene co-expression signatures exhibited aberrantly within subsets of conditions.

2.2 Proximity measure underlying TuBA

TuBA's proximity measure is a pairwise measure based on the hypothesis that if a cellular mechanism is affected in a subset of tumors, genes relevant to the mechanism should co-exhibit similar up- or down-regulation in a significant fraction of those tumors. To illustrate this more clearly, we focus on the case corresponding to high expression. Assume there exists an underlying mechanism responsible for an increase in the expression levels of genes A and B in a given tumor sample, such that when the expression level of gene A goes up, so does the expression level of gene B, and vice versa. This suggests that for a gene expression data set with a large number of samples, if we do relative comparisons of the gene expression levels across tumors, the subset of tumors that harbor the given altered/aberrant mechanism would rank higher than the rest of the tumors for both genes A and B. Our pairwise proximity measure is essentially based on this idea and poses the following question: which pairs of genes exhibit higher (or lower) expression levels in similar subsets of samples (conditions) relative to the rest of the samples in the gene expression dataset? We can make this more concrete. If we do pairwise comparisons between the top (or bottom) percentile sets - say, top 5% samples - of all the genes, we should be able to identify the pairs that share a large number of samples in their percentile sets. The number of samples shared between percentile sets of any given gene pair is expected to follow the hypergeometric distribution. Based on this expectation we can

calculate the probability of observing the overlap of given number of samples between the percentile sets with the null hypothesis that the two sets are independent, against the alternative hypothesis that there is dependence/association between the two sets and the observed number of samples is greater than what would be expected by chance alone. The details about the calculation of these probabilities are provided in the subsection below.

2.2.1 Computation of the significance values (p-values) of overlaps

In order to succinctly describe the details of the computation that underlie TuBA's proximity measure, we focus on the case corresponding to high expression. Let us pick a pair of genes from a gene expression data set - say, gene 1 (red in Fig. 2.1) and gene 2 (blue in Fig. 2.1). For each gene, we first rearrange the samples such that when we plot the normalized expression levels along the vertical axis with the samples placed along the horizontal axis, then the samples with the highest expression levels would lie at the right most end of the horizontal axis (see panel A in Fig. 2.1). We then choose a percentile cutoff, and prepare the list of samples that make up the top percentile sets for each gene, respectively. To illustrate this, we take the help of the Venn diagram in panel B of Fig. 2.1. In the Venn diagram: (i) the red circle in the diagram represents the set of top (5%, 10%, etc.) percentile samples for gene 1, or the set of all non-zero samples for gene 1, whichever is smaller, (ii) the blue circle represents the set of top percentile samples for gene 2, or the set of all non-zero samples for gene 2, whichever is smaller, and (iii) the grey rectangular box (containing both the circles) represents the set of all samples that have non-zero expression values for both gene 1 and gene 2.

Thus, the regions labeled by **a**, **b**, **c** and **d** in the Venn diagram represent:

- **a**: the set of samples that are found in the top percentile sets (or the subsets of samples with non-zero values) of both gene 1 and gene 2
- **b**: the set of samples that are found only in the top percentile set (or the subset of samples with non-zero values) of gene 1
- **c**: the set of samples that are found only in the top percentile (or the subset of samples with non-zero values) of gene 2
- **d**: the set of samples that are neither found in the top percentile set (or the subset of samples with non-zero values) of gene 1, nor in the top percentile set (or the subset of samples with non-zero values) of gene 2

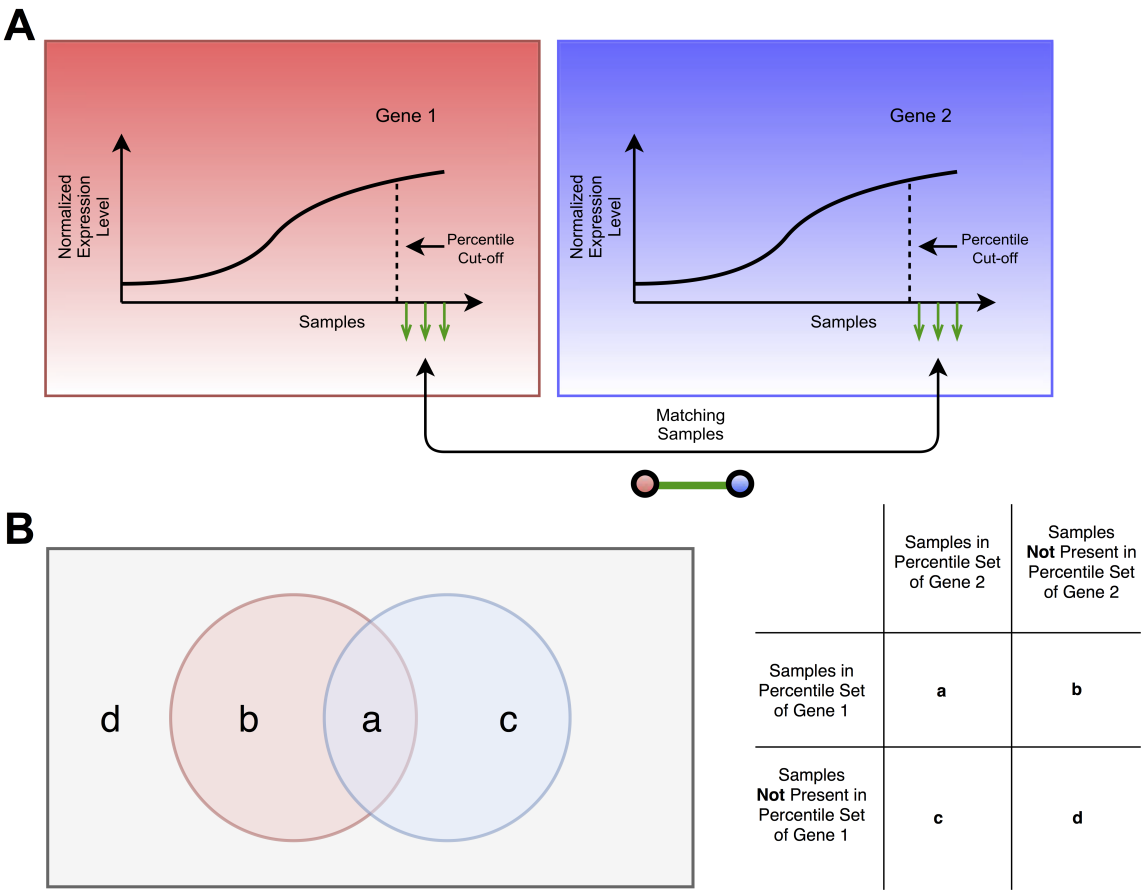

Figure 2.1: **Illustration of the idea underlying TuBA's proximity measure.** Panel A shows a schematic representation of plots of expression levels of a gene pair, where the samples are arranged along the horizontal axis such that ones with the highest expression levels are towards the right. For a given percentile cutoff (dashed vertical lines), we compare the top samples for the gene pair. Gene pairs with significant number of samples shared between their percentile sets are represented as a pair of nodes linked by an edge. Panel B shows a Venn diagram (*left*) that illustrates the setup of the contingency tables (*right*) for determining the significance of overlaps between the percentile sets of gene pairs. Figure reproduced from [1].

.

We use these four quantities - **a**, **b**, **c** and **d** - to set up 2×2 contingency tables, and then employ the one-sided Fisher's exact test to calculate the significance of the data in the table assuming that the null hypothesis is true, namely that the rows and columns of the contingency table are independent. Fisher showed that the probability of observing the set of values in any given 2×2 contingency table is given by the hypergeometric distribution [48]. The hypergeometric probability mass function is given by:

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

where for our case, $N = a + b + c + d =$ total number of samples in dataset , $K = a + b =$ maximum number of samples that can match , $n = a + c = a + b =$ number of samples in percentile set(s), and $k = a =$ number of matching samples.

Thus, for the contingency table,

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}}$$

Note, the equation above only gives the hypergeometric probability of observing an overlap or match of exactly **a** samples. The calculation of the significance value ($p-$value) requires us to find not just the probability of the number of matches observed but also for all possible number of matches that exceed the observed number up till the maximum number of matches possible. The sum of all these probabilities gives us the significance of overlap for each gene pair. However, even these $p-$values are not the final significance values of overlaps used by TuBA. We need to keep in mind that because we are computing significance values for every possible gene pair in the gene expression data set, these values need to be corrected for multiple hypothesis testing. Therefore, TuBA corrects these $p-$values for false discovery using the Benjamini-Hochberg method [49], and relies on the adjusted $p-$values or false discovery rates (FDR) for subsequent steps.

2.2.2 Salient features of TuBA's pairwise proximity measure and its relevance to biological systems

Some of the salient features of TuBA's proximity measure that distinguishes it from other pairwise measures of proximity are:

- It does not model the distributions of the measured expression levels of genes across samples. Thus, the inherent uncertainties imposed by biological noise and technical noise that undermine reliable modeling of actual expression values do not pose any problem.

- It does not quantify the differences between the expression levels of genes in samples that make up the top (or bottom) percentile sets and the rest of the samples in the data set. This enables

the identification of biologically relevant gene co-expression signatures without restricting the analysis exclusively to genes that exhibit differential expression across subsets of samples. In case of tumor data sets, this increases the likelihood for identification of gene co-expression signatures associated with the microenvironment.

- It does not impose a penalty for relative changes in ranks of samples in the respective percentile sets. This is a very important property that derives its value from the simple observation that despite differences in the ranks of matching samples in the two percentile sets of any given gene pair, there is still valuable information to be gleaned by virtue of the fact that these subsets of matching samples exhibit higher (or lower) expression levels for a given gene-pair compared to all the other samples. This feature of our proximity measure makes it less sensitive to noise compared to other proximity measures such as the Spearman's rank correlation.

TuBA's proximity measure is especially relevant for analyzing gene expression data corresponding to biological real systems. In these systems, we expect the following two scenarios to be prevalent:

1. Subsets of genes associated with particular biological processes/pathways are co-expressed across all the samples.

2. Subsets of genes may be deregulated in subsets of samples because of the same underlying mechanism(s), such that their expression levels are higher (or lower) compared to the rest of the samples not influenced by that mechanism(s).

For the first scenario, one would reasonably expect significant matches between the sets of samples that exhibit higher (or lower) expression levels of the involved genes. Thus, our proximity measure would reliably capture these co-expression signatures. The second scenario is of particular interest for data sets associated with diseased states, especially cancers, since these gene co-expression signatures and their underlying mechanisms could help us identify potential biomarkers with prognostic and/or predictive value. This is the basic motivation behind standard differential co-expression analyses as well [50]. However, unlike usual differential co-expression analyses, our proximity measure does not rely on any prior knowledge or specification of subtypes. Our proximity measure is especially effective in capturing the co-expression signatures associated with alterations in the expression profiles of the transcripts in disease states.

2.3 The bare minimum essentials about graphs

As TuBA is a graph-based method, in this section we introduce a few of the most basic concepts about graphs that will be helpful to understand the design of the algorithm.

All graphs consist of two basic elements -

1. Vertices (red circles in the graph in Fig. 2.2)

2. Edges (black lines connecting the vertices in the graph in Fig. 2.2)

Vertices are also frequently refered to as nodes, as we do in this chapter. The edges represent the associations between the nodes that they link together. Depending on the nature of associations they represent, the edges can be (i) directed or undirected, and/or (ii) weighted or unweighted. The weights, denoted by real numbers, could represent the strength of the connections between the respective nodes. We only consider the simplest graphs that have undirected, unweighted edges between every pair of vertices. The graph in Fig. 2.2 is an undirected and unweighted graph, which is the only kind of graph that TuBA prepares and examines for discovering biclusters.

A key concept in the analysis of graphs is that of the *measure of centrality*. A question that is commonly asked is - which is the most important vertex or node in the graph? Of course, the answer to this question depends on the notion of importance pertinent to the context of the data the graph represents. Nevertheless, the simplest measure of centrality for an undirected and unweighted graph is called the *degree*. The degree of a vertex or node in a graph is simply the number of edges connected to it. For instance, in the graph in Fig. 2.2 the degree of node **4** is 5, while that of node **9** is 1.

Another important concept for analyzing graphs is that of *complete subgraphs*, or *cliques*. Cliques are subgraphs within graphs that are composed of subsets of vertices that are *all* mutually connected to each other by edges. For example in Fig. 2.2, nodes **11** and **12** constitute a clique of size 2, nodes **5**, **6**, and **7** constitute a clique of size 3 and so on (the size of the cliques is given by the number of vertices or nodes it contains). One kind of clique is of particular interest to us, namely the *maximal clique*. Maximal cliques are cliques that cannot be extended further by addition of other vertices in the graph. A special kind of maximal clique is the *maximum clique* or *largest clique*, which as the name suggests is the largest possible maximal clique within a graph. For example, the clique made up of nodes **1**, **2**, **3**, and **4** is the largest clique within the graph in Fig. 2.2.

2.4 The Bron-Kerbosch (BK) algorithm for finding maximal cliques in undirected graphs

The task of identifying the maximal cliques within a graph is known to be computationally hard [51]. One could try brute force approaches by testing every possible subset of vertices in the graph to see if they are all mutually connected. However, for large graphs such an approach would be extremely

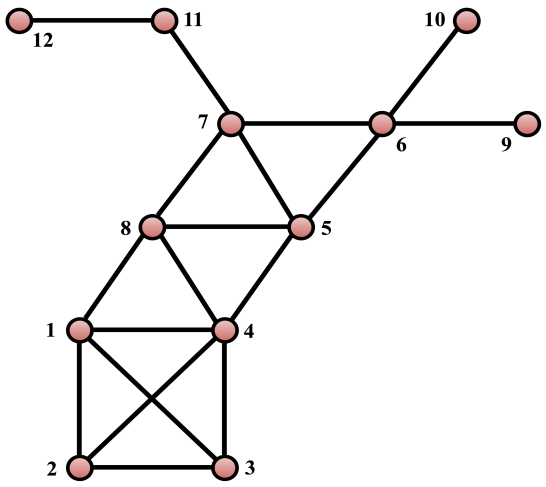

Figure 2.2: **Example of an undirected graph**. In this graph, the edges do not have any weights as well.

inefficient. In 1973, Bron and Kerbosch [52] proposed a recursive backtracking algorithm to find maximal cliques in undirected graphs. Improvements and variants of the BK algorithm are reported to be more efficient than the alternatives [53]. Here, we briefly describe the original Bron-Kerbosch algorithm along with a simple example to illustrate how it works in practice.

The BK algorithm requires three disjoint sets of vertices - R, S and X - as its argument. At the beginning, both R and X are kept empty, while P contains the entire set of vertices present in the graph. The subsequent steps of the algorithm are listed below:

1. Choose a vertex v from the set P
2. Add v to the set R
3. Remove non-neighbours of v from P and X (of course, if X is an empty set, it will remain empty)
4. Pick another vertex from the new P, and repeat steps 2 and 3.
5. Repeat till P becomes empty. If X is also empty then R is a new maximal clique.
6. Restore R,P, and X to how they were before the choice of vertex v
7. Remove v from P, and add it to X
8. Repeat the steps above till there are no more vertices left in P.

The pseudocode of the BK algorithm is provided below:

```
BronKerbosch(R, P, X)
    if P and X are both empty then
        report R as a maximal clique
    end if
for every vertex v in P apply
    BronKerbosch(R union v, P intersect N(v), X intersect N(v))
    P = P excluding v
    X = X union v
end for
```

N(v) stands for the set of vertices that are neighbours of vertex v.

We illustrate how the BK algorithm works by applying it to the graph in Fig. 2.2. Let us choose node **4** as our initial vertex v. Then, using the BK algorithm we get:

- $R = \{\mathbf{4}\}$, $P = \{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{5}, \mathbf{8}\}$, and $X = \phi$
- If we now pick node **3** as v, we get $R = \{\mathbf{3}, \mathbf{4}\}$, $P = \{\mathbf{1}, \mathbf{2}\}$, and $X = \phi$
- Repeating with node **2** as v this time, we get $R = \{\mathbf{2}, \mathbf{3}, \mathbf{4}\}$, $P = \{\mathbf{1}\}$, and $X = \phi$
- Since there is only one node left in P, we add it to R to get $R = \{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}\}$, while $P = X = \phi$.
- Since P and X are both empty, $R = \{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}\}$ must be a maximal clique (in this case, it is also a maximum clique)

Suppose, instead of node **3** we had chosen node **5** in step 2 above, then the sequence of steps from the beginning (with node **4** as the initial v) would be:

- $R = \{\mathbf{4}\}$, $P = \{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{5}, \mathbf{8}\}$, and $X = \phi$
- Picking node **5** as v, we get $R = \{\mathbf{4}, \mathbf{5}\}$, $P = \{\mathbf{8}\}$, and $X = \phi$
- With only one node left in P, we add it to R to get $R = \{\mathbf{4}, \mathbf{5}, \mathbf{8}\}$, while $P = X = \phi$.
- Since P and X are both empty, $R = \{\mathbf{4}, \mathbf{5}, \mathbf{8}\}$ must be a maximal clique

Similarly, with the choice of node **8** instead of **3** and **5**, the sequence of steps from the beginning (with node **4** as the initial v) would be:

- $R = \{\mathbf{4}\}$, $P = \{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{5}, \mathbf{8}\}$, and $X = \phi$
- Picking node **8** as v, we get $R = \{\mathbf{4}, \mathbf{8}\}$, $P = \{\mathbf{1}, \mathbf{5}\}$, and $X = \phi$

- Repeating with node **1** as v this time, we get $P = \phi$, and $X = \phi$
- Since P and X are both empty, $R = \{\mathbf{1}, \mathbf{4}, \mathbf{8}\}$ must be a maximal clique

The entire process is repeated for other vertices after excluding node **4** from P, and adding it to X ($X = \{\mathbf{4}\}$).

2.5 TuBA's graph-based iterative approach to identify biclusters

TuBA's proximity measure offers an elegant way to organize the information of the identified genes and their associations - all the gene pairs that share significant number of samples between their percentile sets can be represented as pairs of nodes linked by edges. The nodes represent the genes, and the edge linking the pair of nodes is representative of the set of samples that are common between the percentile sets of the two genes. After plotting the complete set of such pairwise associations we obtain large graphs that are then analyzed by the iterative approach described below to identify the most robust co-expression signatures (also see Fig. 2.3):

1. Prune the graph such that it is only made up of complete subgraphs (cliques) of size 3 (triangles). Thus, the elementary units of our graphs are triangles.
2. Identify the largest clique(s) (using a modified version of the Bron-Kerbosch algorithm [54]) and list the nodes belonging to the largest clique(s). If there are several cliques that qualify as largest, take the union of the sets of nodes of the ones that have a non-zero intersection. We call this subgraph a seed. The seeds represent the most robust gene co-expression signatures within the graphs.
3. Reduce the graph by removing the nodes of the seed identified in Step 2 *and* all the edges that contain any of the nodes contained in the seed. This step significantly reduces the computation time required to identify all the maximum cliques in the graph.
4. Repeat steps 2 and 3 on the reduced graph such that there are no more cliques of size 3 left.
5. Reintroduce the seeds in the original pruned graph and proceed sequentially to identify and add the nodes that share edges with at least two nodes in the seed. Add these edges and nodes to the seeds to obtain the final biclusters.

Note that the gene sets in the seeds identified by the recursive application of steps 2 and 3 to the pruned graph obtained after step 1 are mutually exclusive, i.e., they do not share any gene between them. There may be additional associations between subsets of genes in the seeds and other genes

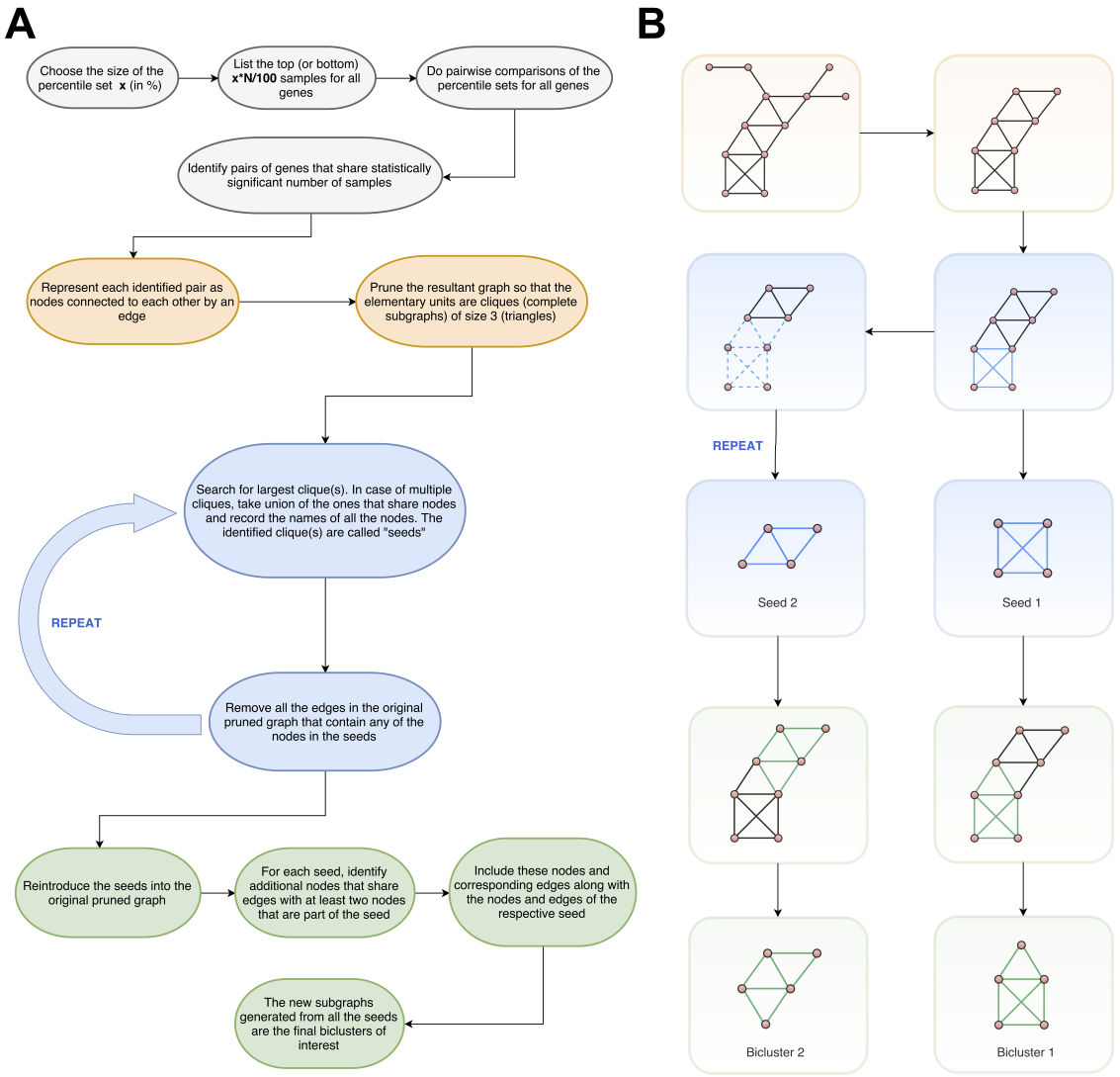

Figure 2.3: **TuBA's iterative graph-based pipeline.** Panel A shows the flowchart describing the steps of TuBA's iterative process, panel B shows an illustration of how TuBA discovers biclusters for a simple schematic graph. Figure reproduced from [1].

in the graph that were excluded simply because those genes were not members of the maximum cliques. It is to mitigate this limitation imposed by the criteria for a subgraph to qualify as a seed, we introduced step 5. In panel A of Fig. 2.3, we show the flowchart detailing all the steps that TuBA goes through to discover biclusters in any graph; panel B in Fig. 2.3 actually illustrates how the iterative process of TuBA works for the simple schematic graph of Fig. 2.2. Note, that TuBA's graph-based iterative process is only applicable for undirected, unweighted graphs.

2.6 Nature of TuBA's biclusters

One of the key steps in our algorithm is the identification of mutually exclusive largest cliques as the seeds of our biclusters. This enables the identification of shared altered mechanisms in subsets of samples that exhibit relatively higher (or lower) expression levels of genes co-expressed due to the altered mechanism. Of course, the co-expressed genes may also be associated with functionally related pathways that may or may not have been altered in the given subset of samples.

A crucial assumption implicit in the requirement of largest cliques as seeds is that the sets of genes comprising the seeds are co-expressed within subsets of samples that make up the edges. It is extremely important to note that this assumption is not the same as requiring all gene-pairs comprising the seed to share identical sets of samples, nor is it the same as assuming that all the samples comprising the final biclusters co-express all the genes present in the bicluster at the highest (or lowest) levels. Instead, our expectation is that the samples present in the final biclusters are *enriched* in the top (or bottom) samples for each gene comprising the biclusters. Another way to put it is that the samples in the bicluster exhibit higher (lower) expression levels of the genes in the bicluster relative to most samples that are not members of the bicluster. We illustrate and expand on this expectation in the following subsection below.

2.6.1 What do TuBA's biclusters look like?

In Chapter 1, we described four major classes of biclusters. TuBA's biclusters belong to the fourth class - biclusters with coherent evolution - since biclustering algorithms that look at coherent evolution essentially seek subsets of genes that are up-regulated or down-regulated across subsets of conditions *without relying on the actual expression values.*

However, TuBA's biclusters are quite distinct from the kinds of biclusters the other biclustering algorithms that belong to this class identify. We illustrate this with the help of Fig. 2.4. In panel A, we show a bicluster identified by TuBA that contains 6 genes (rows) and 6 samples (columns). We identified the gene within this bicluster that had the highest average expression across the samples,

and placed it as row 1 of the matrix. We then resorted the samples in increasing order of expression levels such that the sample with the lowest expression (light red) for the given gene was placed at the top left corner, and the sample with the highest expression was placed at the top right corner of the matrix. The order of the samples was kept the same as the one for the gene placed on row 1. We can observe that there is no explicit coherence in the expression values (higher expression values represented by deeper shades of red) across the columns or the rows of the matrix; contrast this with the biclusters associated with coherent evolution discovered by other biclustering algorithms that seek biclusters with coherent evolution (see Fig. 1.5).

This might appear to be a drawback of our algorithm, however for real gene expression data sets it is in fact of tremendous value. TuBA not only identifies patterns of coherent evolution identified by other methods (provided these patterns exist among the subset of samples that make up the top (or bottom) sample sets for the genes in the bicluster), but it actually accommodates a much more diverse array of expression patterns that nevertheless fulfill the basic expectation that the samples present in the final biclusters are enriched in the top (or bottom) samples for each gene comprising the biclusters. To make it easier to understand this expectation we have shown a schematic representation of TuBA's bicluster (panel A in Fig. 2.4) in comparison to a 6×6 submatrix made up of samples picked randomly from the gene expression matrix for the same set of genes as the ones in the bicluster (panel B in Fig. 2.4). There is a significant difference between the two matrices - for any given gene, the samples in TuBA's bicluster predominantly exhibit higher expression levels for than the samples in the random submatrix. This is the quintessential property of the biclusters discovered by TuBA.

The ability of TuBA to accommodate a diverse array of expression patterns stems from a salient feature of the proximity measure described earlier - it does not penalize differences in ranks of samples in the percentile sets. Instead, it relies on the simple fact that disease states are highly likely to exhibit aberrant co-expression of genes due to alterations in the underlying transcriptional programs. If the genes involved with the alteration exhibit higher (or lower) expression levels within the same subsets of samples, our proximity measure would be able to capture their co-expression signature.

2.6.2 Enrichment of TuBA's bicluster in top (or bottom) sample sets

We make the expectation that the samples present in the final biclusters are enriched in the top (or bottom) samples for each gene comprising the biclusters more concrete with the help of the following example: Suppose we have a data set that consists of 1000 samples. We apply TuBA to this data set (assume we are analyzing high expression), and discover a bicluster made up of 100 genes and 200 samples. For each of the 100 genes in the bicluster, we then:

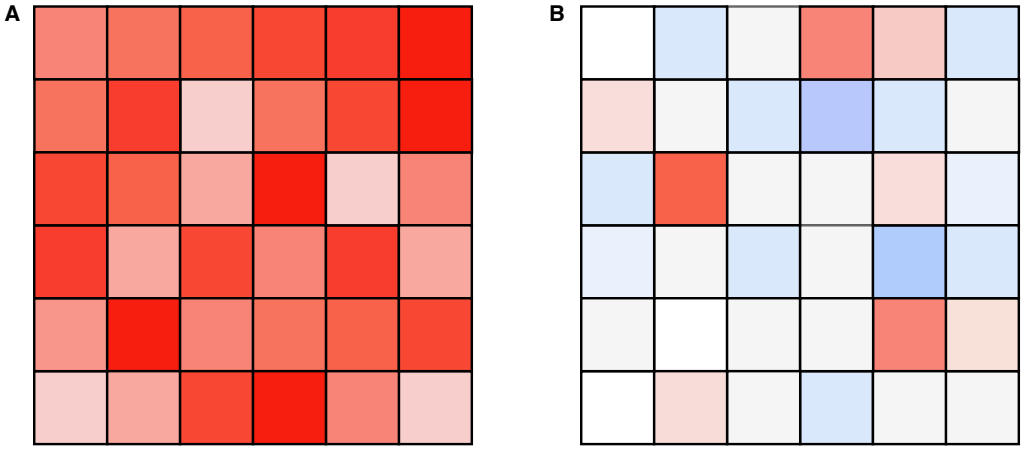

Figure 2.4: **TuBA's bicluster in comparison to a randomly chosen submatrix from a gene expression data set.** Panel A shows a schematic illustration of a bicluster with 6 genes and 6 samples found by TuBA, panel B shows a schematic illustration of a submatrix from the same data with 6 samples chosen randomly for the same set of genes as the bicluster.

1. Identify the top 200 samples

2. Test whether these 200 samples are enriched in the 200 samples present in the bicluster.

In order to calculate the significance values for the enrichment of the top samples in the samples in the bicluster, we can use the one-sided Fisher's exact test. Since we are performing the tests for each gene in the bicluster, we need to correct the $p-$values to obtain FDRs using the Benjamini-Hochberg method. A similar analysis can be performed for the samples in the bicluster. For the same bicluster with 100 genes and 200 samples, this time for each of the 200 samples:

1. Identify the genes in the bicluster that have the given sample present within their respective top 200 samples.

2. Test whether this subset of genes have a significant overlap with the complete set of genes that make up the bicluster.

Once again, we can use the one-sided Fisher's exact test to determine whether the overlaps between the two gene sets is significant. These need to be corrected for multiple hypothesis testing as well, since we are performing the tests for each sample in the bicluster.

2.6.3 Quality of TuBA's biclusters

The tests described above generate FDRs for each and every gene and sample in all the biclusters discovered by TuBA for any given data set. In a given bicluster, the FDR value for a gene (sample) can be viewed as a measure of its relevance to the respective bicluster the closer the value of the

FDR is to 0, the stronger is the association of the gene (sample) to the bicluster. We can use the FDR values for the genes and samples within any bicluster i to evaluate its overall quality, $Q(B_i)$. We define $Q(B_i)$ as the minimum of the proportion of genes in bicluster i with $FDR < 0.05$ or the proportion of samples in bicluster i with $FDR < 0.05$. $Q(B_i)$ takes values between 0 and 1; values close to 0 indicate weak associations between the constituent genes and samples within the bicluster while values close to 1 would indicate strong associations.

2.7 Tuning TuBA

TuBA has two adjustable parameters that determine the nature of the undirected graph, and consequently, the final biclusters:

1. **Percentile cutoff**: Dictates the number of samples considered for comparison between genes
2. **Overlap significance cutoff**: Dictates the minimum number of samples that must be shared between the percentile sets of a pair of genes for them to be represented on the graph

To illustrate how the choice of the percentile cutoff influences the graph, we consider a hypothetical data set consisting of 200 samples. Consider an ideal case wherein a gene pair is up-regulated in exactly the same 5% of tumors within this cohort, i.e., both the genes in the gene pair have the same top 10 samples. In Fig. 2.5, we show the $p-$values for overlaps as a function of the fraction of samples that overlap. We can see that the overlap significance value for this hypothetical case will be $p < 10^{-15}$ (see the dark blue curve at Fraction of Overlap = 1). If instead we had picked a top $10-$percentile cutoff, we would have 20 samples in the top percentile sets of each gene. Out of these 20 only 10 would overlap, resulting in a drop in overlap significance to $10^{-10} < p < 10^{-5}$ (see green curve at Fraction of Overlap = 0.5).

Thus, an increase in the size of the percentile set may result in loss of significance for aberrant gene-pair signatures found only in subsets of samples. This does not automatically imply that the best policy is to reduce the size of the percentile set, because a reduction in the size of the percentile set increases the likelihood that a number of samples match purely by chance. We can see this by comparing the $p-$values for Fraction of Overlap = 1 for the three cases in Fig. 2.5.

In reality, there is no optimal choice for the size of the percentile set. This is because of both the differences in the prevalence of aberrant gene expression signatures within and among tumors, and the differences in frequency of occurrence of disease subtypes in the population. The choice of the percentile cutoff should be viewed as a *knob* that determines the level of heterogeneity in the population that we aim to capture.

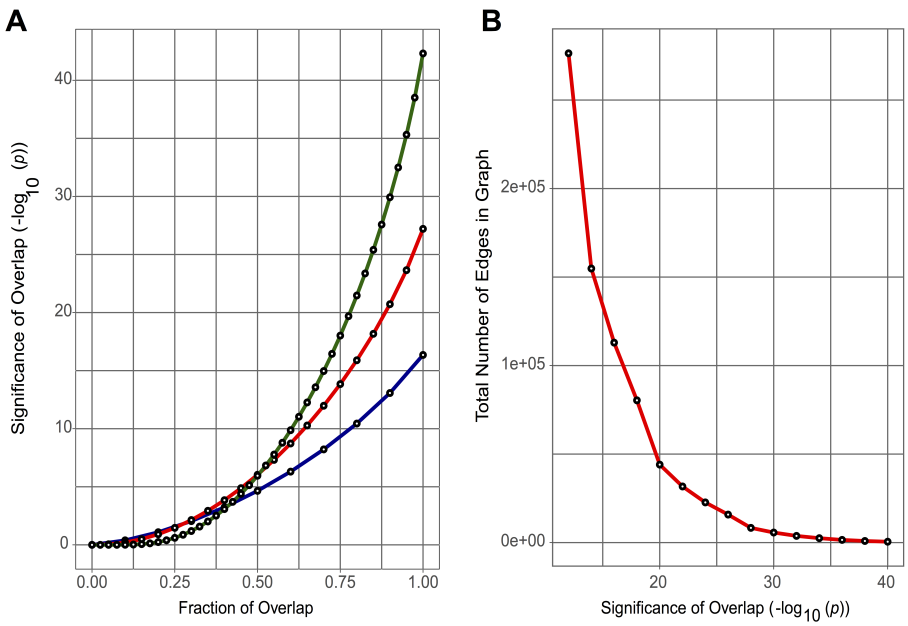

Figure 2.5: **TuBA's tunable parameters and their influence on the graphs.** Panel A shows the significance of overlap corresponding to fractions of overlap ranging between 0 and 1 for the top percentile sets of sizes: (i) 20% (dark green), (ii) 10% (red), and (iii) 5% for a hypothetical data set with 200 samples. Panel B shows the divergence of the total number of edges in the graphs as the overlap significance cutoff is lowered. Figure reproduced from [1].

The choice of the second parameter, the extent of patient/sample overlap between percentile sets, is primarily dictated by its impact on the size of the graph of connected gene-pairs. When the overlap significance cutoff $p-$value is raised (lowering of significance), new genes and samples appear, resulting in an increase in the number of edges in the graph. We can best understand this effect with the help of Fig. 2.6 that depicts the impact of the choice of overlap significance cutoffs on the respective graphs corresponding to 3 real gene expression data sets (examined in greater detail in the following chapter). Assume that in these data sets there is an aberrant tumor-specific gene co-expression signature in a subset of samples that is frequently (but not always) accompanied by an enhanced immune response/infiltration. Thus, we would expect that genes associated with the immune cells would be found to be expressed at relatively higher levels in these subset of samples compared to the rest of the samples that do not have immune cells infiltrating into the tumor. Since the subsets of samples that have the aberrant signature and higher immune infiltration have similar samples, the association between the aberrant signature and the immune response/infiltration would appear in the form of edges that link the two in the graph. As we increase the overlap cutoff further (lowering the level of significance), more such edges would get added to the graph that would indicate an association between various signatures identified in the graph (panels). However, as far as the biclusters are concerned this increase in the number of edges in the graphs is accompanied by only

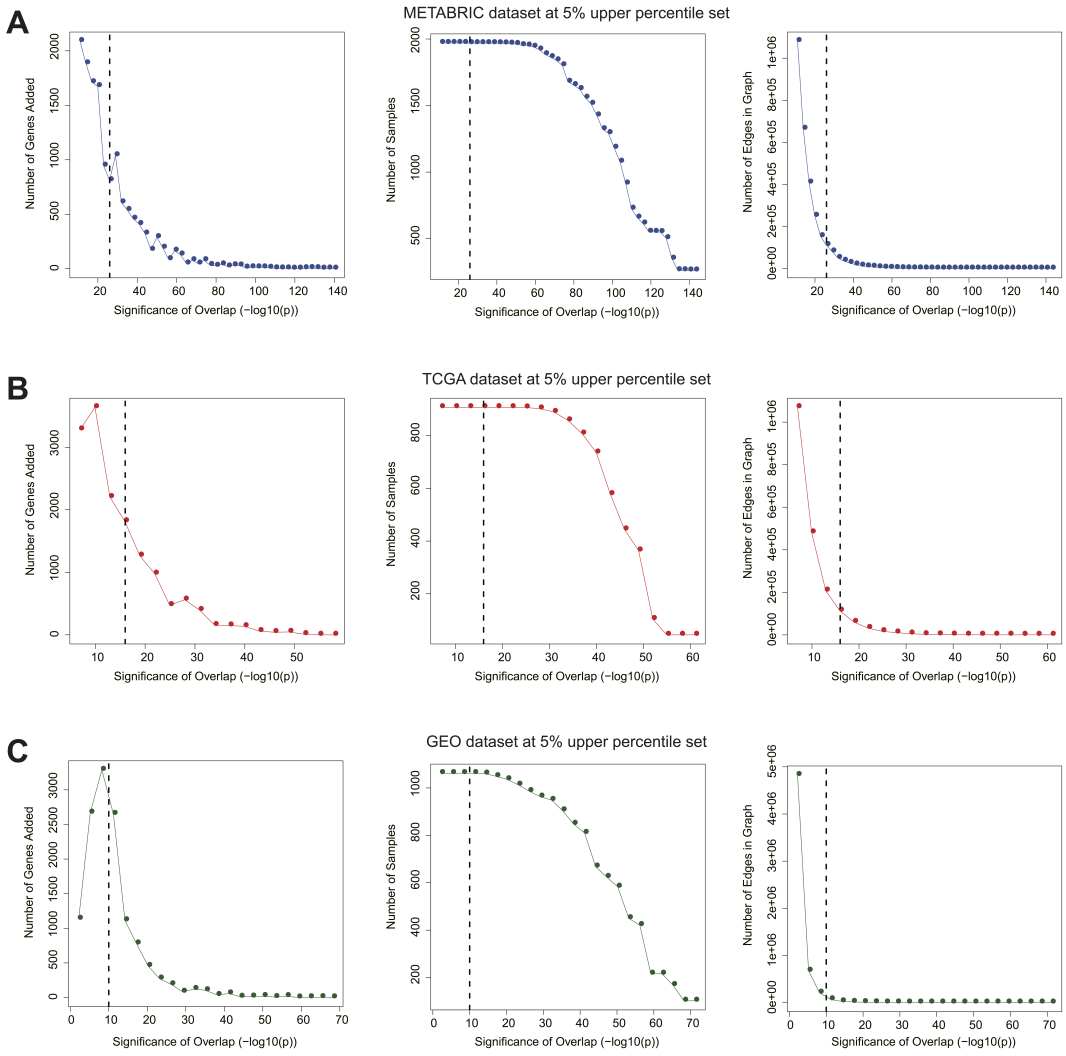

Figure 2.6: **Effect of the choice of the overlap significance cutoff on the number of genes, samples and links in the graphs.** Left to right: Plots showing the number of genes added to the graphs, the number of samples in the graphs, and the total number of links in the graphs for incremental decreases in the significance level by an order of magnitude, respectively. Panels A, B, and C correspond to the TCGA, METABRIC, and GEO data sets (described in Chapter 3), respectively. Figure reproduced from [1].

a modest gain in information in terms of the addition of new genes or samples to the biclusters themselves. Thus, the choice of the p-value cutoff is dictated by the consideration of this trade-off between the gain of new information (genes and samples) in the biclusters, and the number of edges that get added to our graph. We propose the following heuristic for choosing the overlap significance cutoff value: the cutoff for the significance level of overlap should be such that a decrease in the significance level by an order of magnitude leads to an $40 - 60\%$ increase in the number of edges that get added to the graph (in panel B of Fig. 2.5, observe the divergence in the total number of edges in the graph as the significance is lowered below 10^{-20}).

2.8 Implementation and availability of TuBA

TuBA was implemented using R, which is an open source programming language and environment for statistical computing [55]. In addition to using the base packages in R, we relied on the following packages to perform all the required computations to analyze the gene expression data sets:

1. **data.table** [56]
2. **plyr** [57]
3. **igraph** [58]
4. **ggplot2** [59]
5. **survival** [60, 61] - for survival analysis

The graphs that show the co-expressed genes in the biclusters were made with the help of the *Cytoscape* software [62].

The source code for TuBA (currently in the form of R functions) is licensed under the GNU GPL v3 protocol and can be downloaded from https://github.com/KhiabanianLab/TuBA

Chapter 3
Application to Breast Invasive Carcinoma

"To confront cancer is to encounter a parallel species, one perhaps more adapted to survival than even we are."

– Siddhartha Mukherjee, *The Emperor of All Maladies*

"Ever tried. Ever failed. No matter. Try again. Fail again. Fail better."

– Samuel Beckett, *Worstward Ho*

3.1 Breast Cancers - An overview

Breast cancer is one of the most common malignancies that affects millions of women across the globe. In 2018, more than 260,000 cases were diagnosed in the United States (US) alone. Moreover, the number of deaths from breast cancer of women in the US in 2018 is estimated to be more than 40,000 women [63]. It is the most common cancer amongst women in the US, and it alone accounts for 30% of all new cancer diagnoses.

Breast cancer is genetically, clinicopathologically, as well as clinically, heterogeneous. Over the last several decades, a large number of clinically relevant factors have been identified that have associations with response to therapy and/or patient survival. These include factors such as the size of the tumor, tumor histology, cellular proliferation rate, lymph node status, age at diagnosis. Quite significantly, the expression of specific molecular markers such as oestrogen (ER), progesterone (PR), and the human epidermal growth factor receptor 2 (HER2) are crucial aids in determining the choice of therapy for the patient. Encouraged by the improvements in patient management decisions based on the molecular markers together with the other clinical factors, several studies have been undertaken in the past two decades aimed at classifying breast cancer into robust subtypes. These studies have revolved around comprehensive gene profiling of hundreds of breast tumor samples. In an extremely influential work published in 2000, Perou *et al* [64] identified groups of co-expressed genes that exhibited substantial variations in their expression levels between subsets of tumors in

Subtype	Suggested Therapy	Notes
Lum A	Endocrine	Some need cytotoxic
Lum B (HER2−)	Endocrine + Cytotoxic	No cytotoxic for some
Lum B (HER2+)	Endocrine + Cytotoxic + Anti-HER2	
HER2+ (non-luminal)	Anti-HER2+Cytotoxic	
Basal-like (ductal)	Cytotoxic	

Table 3.1: **Treatment recommendations for BRCA based on PAM50 subtypes**. This table has been adapted from Table 3 in [69]. *Cytotoxic* refers to cytotoxic drugs that inhibit cell division and are used to destroy cancer cells, *Endocrine* refers to hormonal agents (such as tamoxifen) that block natural hormones to inhibit tumor growth, and *Anti-HER2* refers to monoclonal clonal antibodies that specifically bind to the HER2 and induce immune-mediated response.

their data set. Based on these gene sets (which they called the *intrinsic* gene set), they classified their breast invasive carcinoma (BRCA) samples into 4 subtypes -

- **Luminal-like** - tumors within the ER+ subtype that also expressed breast luminal cell markers
- **Her2-enriched** - tumors with ER−, but HER2+
- **Basal-like** - tumors within the ER−/HER2− subtype that exhibited high expression of keratins 5,6 and 17
- **Normal-like** - tumors within the ER−/HER2− subtype that exhibit gene expression signatures similar to the ones expressed by normal breast tissue

With the addition of more samples to their analysis, they later identified two sub-classes within the luminal-like subtype [65] - (i) luminal A (Lum A), and (ii) luminal B (Lum B). Lum B breast tumors are known to exhibit higher risks of recurrence compared to Lum A tumors [66].
There have been a few criticisms of this classification scheme based on the observation that the subtypes are relatively unstable, and exhibit dependence on the original genes and samples used by the authors [67, 68]. Despite this criticism, this subtype classification has been widely adopted in the breast cancer research community, and even serves a crucial role in clinical decision making. Table 3.1 gives an overview of the treatment recommendations based on these subtypes (adapted from Table 3 in Goldhirsch [69]).

3.2 The datasets

3.2.1 TCGA

The Cancer Genome Atlas (TCGA) is an ambitious cancer genomics program that was initiated in 2006 as a joint effort by the National Cancer Institute and the National Human Genome Research

Institute. The program brought together a large number of institutions and researchers from diverse backgrounds to enable the molecular characterization of more than 20,000 primary tumors and matched normal samples spanning 33 different cancer types.

BRCA was one of the 33 types that was investigated by the TCGA study. In fact, BRCA is the one with the most number of studied samples (from more than a thousand patients). The TCGA study relied on the Illumina HiSeq 2000 RNA sequencing platform to measure gene expression which provides better dynamic range than microarray data. The $log_2(x + 1)$ transformed RSEM normalized counts of Level 3 data (2016-08-16 version), the clinical data (including relapse status and PAM50 subtype annotation from the 2012 Nature study [70]) (2016-04-27 version), and gene-level copy number variation (CNV) data were all downloaded from the UCSC Xena Portal (http://xena.ucsc.edu). Genes with zero expression in all samples, as well as the samples with NA values for any gene were removed from the analysis.

3.2.2 METABRIC

Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) is a collaborative project jointly supported by Canada and UK to specifically investigate breast tumors for identification of novel subcategories. The expectation is that based on the molecular signatures of the subcategories it will be possible to refine and plan the optimal course of treatment for the particular forms of breast tumors.

In a paper published in 2012 [71], the authors of the study classified breast cancer into 10 subtypes based on grouping by common genetic features, which they further demonstrated to correlate with survival. The gene expression data used in this study was generated using the the Illumina HT-12 v3 microarray platform. We downloaded the normalized gene expression data, copy number data, and the file containing the clinical information from the cBioPortal (http://www.cbioportal.org) on 2017-05-14 [72, 73]. Gene expression dataset of 1,970 samples that had both relapse status and PAM50 subtype annotation were used in this study

3.2.3 GEO

The Gene Expression Omnibus (GEO) data set actually consists of data from 6 independent cohorts that relied on the same microarray platform (Affymetrix HGU133A) to measure gene expression in the breast tumor samples in these studies. The GEO accession numbers of the data corresponding to these studies are: GSE1456, GSE2034, GSE3494, GSE4922, GSE6532, and GSE7390. Normalized gene expression data and the clinical data with relapse status were downloaded from the supplementary data of Gyorrfy *et al.* [74] on 2017-05-10.

Dataset	No. of Samples	No. of Genes	Percentile Set	Overlap Significance	No. of Biclusters
TCGA	908	20241	5%	1.00E-16	353
METABRIC	1970	24368	5%	1.00E-26	340
GEO	1062	13031	5%	1.00E-10	369
TCGA (PAM50)	522	20207	5%	1.00E-07	480
TCGA - Low	908	20241	Bottom 5%	1.00E-20	202
TCGA - Low (PAM50)	522	20207	Bottom 5%	1.00E-07	445

Table 3.2: **Summary of BRCA data sets analyzed by TuBA.** The table also contains the details of the two parameters (percentile set size and overlap significance cutoff) used to apply TuBA to each data set. Separate data sets for the TCGA data (TCGA & TCGA (PAM50) were created due to the fact that in this data PAM50 subtype annotation was available for only a subset (522 tumors) of primary tumors.

Table. 3.2 summarizes the basic information of all 3 data sets, along with the respective parameters chosen for TuBA.

3.3 Permutation test confirms gene pair associations in TuBA's graphs are significant

How likely is it for gene pair associations to be identified in our graphs due to chance alone? We decided to investigate this question by performing a permutation test on the METABRIC dataset (1970 samples) with the top percentile set size cutoff set to 5%. For each gene, we permuted the labels of the samples prior to preparing the list of samples that corresponded to the top 5%, respectively. The significance values for overlaps between every pair of genes were computed using the one-sided Fisher's exact test with contingency tables similar to the one shown in panel B of Fig. 2.1. The histogram for the distribution of the $p-$values is shown in Fig. 3.1. After adjusting for multiple hypothesis testing, none of the gene pair $p-$values were found to be significant (panel B in Fig. 3.1). We performed 100 iterations of these permutations for the entire data set, and did not identify a single significant gene pair association in any of those iterations.

3.4 TuBA's proximity measure benchmarked against standard pairwise correlation measures

Even before we perform biclustering on gene expression data sets using TuBA. It is important to assess the performance of its proximity measure against other well-known pairwise proximity measures. Pearson's correlation coefficient and Spearman's correlation coefficient are two of the most commonly used pairwise proximity measures for clustering.

We expect that genes that are co-expressed across all samples would also have significant overlaps between the subsets of samples that correspond to their top (or bottom) percentile sets. Also,

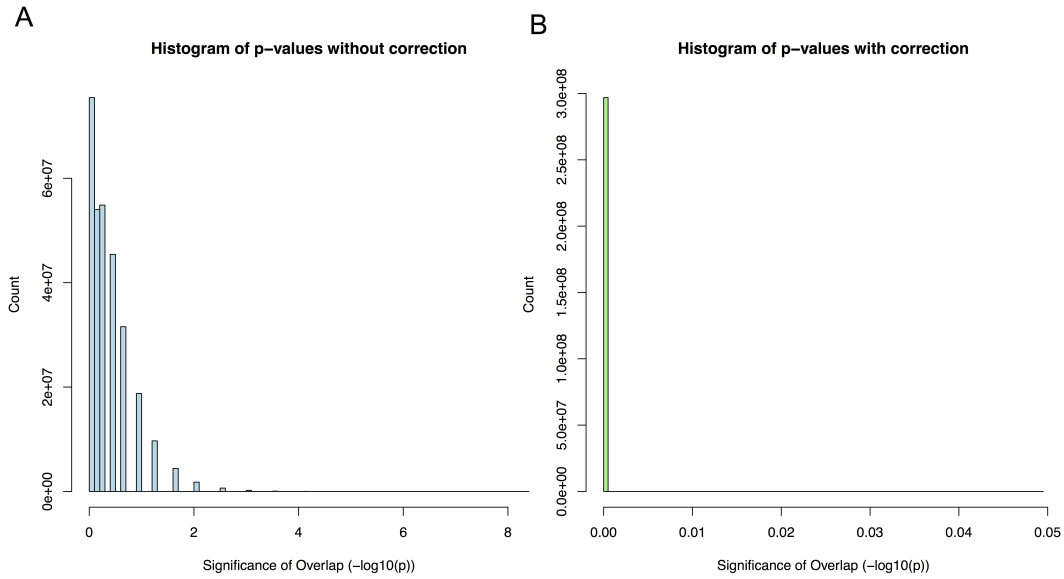

Figure 3.1: **Permutation test shows that it is extremely unlikely to observe gene pair associations at the overlap significance cutoffs chosen for all 3 datasets.** Panel A shows the histogram for the number of gene pairs with significance of overlaps (more accurately $-log_{10}(p)$) shown along the horizontal axis; panel B shows the histogram after correcting the $p-$values for multiple hypothesis testing. Figure reproduced from [1].

Pearson's correlation coefficient, is susceptible to the influence of outliers, which would mean that co-expression signatures associated with these outliers that are captured by our proximity measure should also show higher correlation coefficient values.

Given these observations, we tested the hypothesis that gene sets identified by global proximity measures have significant overlap with those identified by TuBA within its biclusters. We computed the Pearson's correlation coefficients between all possible pairs of genes in the TCGA and METABRIC data sets, respectively. The gene pairs with the correlation coefficient greater than or equal to 0.6 were identified. Graphs using these gene pairs were made, where in this case the edges just indicate that the gene pair has correlation coefficient greater than or equal to 0.6. We then employed our graph-based algorithm to identify gene co-expression modules within the graphs.

For TCGA and METABRIC, we obtained 569 and 298 gene co-expression modules, respectively. We investigated the association between the gene sets in biclusters discovered by TuBA's proximity

	Genes in module j	Genes not in module j
Genes in bicluster i	**a**	**b**
Genes not in bicluster i	**c**	**d**

Table 3.3: **Contingency table for testing enrichment between biclusters and co-expression modules based on their gene sets.**

measure and the gene co-expression modules identified by these two global correlation metrics by relying on the one-sided Fisher's exact test. The form of the 2×2 contingency table for the tests is shown in Table 3.3. Since we calculated significance values between all pairs of biclusters and co-expression modules, we corrected them for multiple hypothesis testing to get FDRs.

The quantities **a**, **b**, **c** and **d** in the contingency table respectively represent:

- **a**: the subset of genes shared between bicluster i and co-expression module j
- **b**: the subset of genes in bicluster i not present in co-expression module j
- **c**: the subset of genes in co-expression module j not present in bicluster i
- **d**: the set of all the genes in the data sets that are not members of bicluster i and co-expression module j

We observed that more than 89% (316 out of 353 biclusters) of the biclusters discovered by TuBA in the TCGA dataset were made up of gene sets that were enriched in at least one gene co-expression module ($FDR < 0.001$), while 86% (293 out of 340 biclusters) of the biclusters discovered by TuBA in the METABRIC dataset were enriched in at least one co-expression module.

We performed a similar analysis using the Spearman's rank correlation, with the same cutoff of 0.6 for the correlation coefficient as well. We obtained 524 and 232 gene co-expression modules for TCGA and METABRIC, respectively. More than 80% (285 out of 353 biclusters) of TuBA's biclusters in the TCGA data set were made up of gene sets that were enriched in at least one gene co-expression module ($FDR < 0.001$), while 73% (249 out of 340 biclusters) of the biclusters discovered by TuBA in the METABRIC data set were enriched in at least one co-expression module. Thus, we observed a strong agreement between gene sets in TuBA's biclusters and the co-expression modules based on the two global proximity measures.

This agrees with our expectation. As noted earlier, due to samples that exhibit aberrant/outlier expression of some genes, the linear correlation coefficients can often get skewed to reflect greater pairwise correlations between such sets of genes. These are captured by our graph-based algorithm, however the global nature of these proximity measures means that the graphs made with these correlation measures lack any information about the samples that might be associated with the

aberrant expression of these sets of genes; unlike the graphs based on our proximity measure, the edges in these graphs do not represent any subset of samples. Thus, the simple yet novel design of our proximity measure not only makes it possible to identify co-expressed sets of genes, but also enables us to discern the subsets of samples that exhibit higher (or lower) expression levels of those genes relative to the rest of the samples.

3.5 TuBA's biclusters are enriched in extremal sample sets of the bicluster genes

We discussed the nature of TuBA's biclusters in section **2.6**, where we explained that TuBA's biclusters are associated with the sets of samples that belong to the extremal (top or bottom) sets for the genes in the bicluster. In subsection **2.6.2**, we explained with the help of an example how we can perform the enrichment tests for each gene and sample in biclusters to determine whether they are associated with these extremal sets.

We applied these test for each of TuBA's biclusters in the TCGA, METABRIC, and GEO data sets. For high expression, we observed that all the genes in all (353) the biclusters from TCGA showed significant enrichment ($FDR < 0.001$). In case of METABRIC, we observed 2 biclusters out of 340 biclusters with only 1% of their constituent genes not exhibiting enrichment, while in case of GEO we observed only 1 bicluster out of 369 with 1% of its constituent genes not exhibiting enrichment. A key observation we made for all 3 data sets was that even in the few biclusters that included a few genes with enrichment $FDR > 0.001$, none of those genes were constituents of the seeds of those biclusters. Based on this observation we decided to rely on the subsets of genes in our biclusters that make up the seeds for future gene set enrichment tests. These enrichments would reveal the core functional signatures of the biclusters.

A similar analysis for the samples revealed that 95% of biclusters (336 out of 353) from the TCGA, 97% of biclusters (329 out of 340) from the METABRIC, and 89% of biclusters (328 out of 369) from the GEO data set, respectively, had more than 95% of samples enriched in the top sample sets of the bicluster genes ($FDR < 0.001$).

3.6 TuBA consistently discovers biclusters made up of similar gene sets within a data set

We investigated whether TuBA could consistently discover biclusters within the same data set. For this, we picked the TCGA data set that consisted of 908 samples. These 908 samples were split

	No. of matching biclusters in data set 1	No. of matching biclusters in data set 2
Split 1	225 out of 306 (73%)	230 out of 303 (76%)
Split 2	220 out of 313 (70%)	219 out of 299 (73%)
Split 3	207 out of 278 (74%)	224 out of 310 (72%)
Split 4	235 out of 323 (73%)	214 out of 305 (70%)
Split 5	221 out of 298 (74%)	225 out of 305 (74%)

Table 3.4: **Consistency of TuBA's biclusters obtained from subsets of the TCGA data set.**

randomly into two groups of 454 samples each. This was done 5 times to generate 5 pairs of data sets. TuBA was applied to all 5 pairs of data sets with the same choice of parameters to minimize bias - the percentile cutoff was set at 5%, and the overlap significance cut-off was fixed at $FDR \leq 10^{-08}$. For each pair of split data sets, we compared the biclusters obtained from the application of TuBA using the one-sided Fisher's exact test. To illustrate the setup of the contingency tables for the tests, we take the aid of the Venn diagram in panel B of Fig. 2.1. For this particular case, the grey rectangular box in the Venn diagram represents the set of all the genes present in the data sets, the red circle represents the set of genes present in bicluster i from data set 1, and the blue circle represents the set of genes present in bicluster j from data set 2. The regions labeled by **a**, **b**, **c** and **d** in the Venn diagram then respectively represent:

- **a**: the subset of genes shared between biclusters i and j
- **b**: the subset of genes in bicluster i not present in bicluster j
- **c**: the subset of genes in bicluster j not present in bicluster i
- **d**: the set of all the genes in the data sets that are not members of biclusters i and j

Table 3.3 shows the 2×2 contingency table for these tests. Once again, the $p-$values obtained from each test were corrected for multiple hypothesis testing to get $FDRs$.

Table 3.4 presents summary of the results from the pairwise comparisons (between sets of genes) of all the biclusters between the 5 pairs of split data sets. On average, 73% biclusters from one data set in each pair were enriched ($FDR < 0.001$) in at least one bicluster from the other data set. We also investigated whether there was a significant difference in the sizes (in terms of number of genes) of the biclusters that matched, compared to the biclusters that did not match between the data set pairs. We used the Mann-Whitney U test [75], which is a non-parametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample. We found a significant difference ($p < 10^{-05}$) in the number of genes contained in biclusters that matched among the pairs, compared

Dataset (Compared With)	No. of Genes	Percentile Set Size	Overlap Significance Cutoff
TCGA (METABRIC)	17209	5%	1.00E-15
METABRIC (TCGA)	17209	5%	1.00E-25
TCGA (GEO)	11979	5%	1.00E-12
GEO (TCGA)	11979	5%	1.00E-09
METABRIC (GEO)	11186	5%	1.00E-22
GEO (METABRIC)	11186	5%	1.00E-09

Table 3.5: **Details of the data sets used for comparing TuBA's biclusters.** The table also contains the details of the two parameters (percentile set size and overlap significance cutoff) used to apply TuBA to each data set.

to the number of genes in biclusters that did not - the biclusters that did not match contained fewer genes than the biclusters that matched; the median size of biclusters that matched was 20 (range: $3-840$), while the median size of biclusters that did not match was 3 (range: $3-18$) (Note, 3 is the minimum size of the biclusters discovered by TuBA). Thus, TuBA robustly identified co-expression signatures that involved larger numbers of genes. Overall, TuBA was able to consistently identify matching sets of co-expressed genes from randomly sampled subsets of data within a data set.

3.7 TuBA's biclusters are consistent across independent datasets

It is quite reasonable to expect consistency of biclusters discovered within the same data. However, what is of greater interest is to see whether the algorithm can discover biclusters in data sets of independent cohorts are similar to each other. While breast cancers in particular are quite heterogeneous, we still expect to observe gene co-expression signatures that are consistent due to shared alterations, or common underlying mechanisms in subsets of tumors within these independent cohorts. We compared biclusters discovered by TuBA between the following data sets:

1. TCGA and METABRIC
2. TCGA and GEO
3. METABRIC and GEO

For each case, we prepared each data set to ensure that each member of the pair had the same set of genes. The details of the data sets and the parameters used for TuBA for each data set are summarized in Table 3.5. We performed pairwise comparisons of biclusters between each pair of data sets to identify the ones that share a significant proportion of their genes. For these comparisons, we used the one-sided Fisher's exact test with contingency tables similar to the one shown in Table 3.3. We made the following observations for each comparison:

1. **TCGA vs. METABRIC** - We found 64% of the biclusters obtained for one dataset enriched ($FDR < 0.001$) in at least one bicluster from the other

Percentile Cutoff	Overlap Significance Cutoff	Number of Matching Biclusters
5 (Reference)	1.00E-16	353
7.5	1.00E-19	334
10	1.00E-23	327
12.5	1.00E-26	322
15	1.00E-30	304

Table 3.6: **Robustness of TuBA's biclusters to different choices of its two parameters for the TCGA data set.**

2. **TCGA vs GEO** - We found 69% of the biclusters obtained for one dataset enriched ($FDR < 0.001$) in at least one bicluster from the other

3. **METABRIC vs GEO** - We found 76% of the biclusters obtained for one dataset are enriched ($FDR < 0.001$) in at least one bicluster from the other.

Thus, TuBA's biclusters were consistent to a remarkable extent in data sets from three independent sources which also involved distinct technologies for measuring gene expression (RNA-seq and Microarray).

3.8 TuBA's biclusters are robust over a range of choices of its tunable parameters

The principal goal of TuBA is to identify subsets of genes that are co-expressed at high (or low) levels within subsets of samples. The exact number of biclusters is not biologically meaningful; it is possible to have some differences in the total number of biclusters as the knobs of TuBA (the two parameters) are varied. We investigated the robustness of TuBA's results by making multiple different choices of its parameters and inquiring whether there was agreement between the respective sets of biclusters obtained for a given data set. We decided to keep the biclusters obtained for TCGA (high expression) with the following choices for the parameters: (i) percentile set size: 5%, and (ii) overlap significance cutoff: 10^{-16}, as our reference. We then applied TuBA to the TCGA dataset with the following choices for the the sizes of the top percentile sets - 7.5%, 10%, 12.5%, 15% (note, all the percentile set sizes are larger than the reference percentile set size). For each of these choices of the percentile set size, we also had to make different choices for the overlap significance cutoffs that were consistent with our proposed heuristic. All the choices of TuBA's parameters are summarized in Table 3.6. Pairwise comparisons (using one-sided Fisher's exact test) between the reference biclusters and the biclusters obtained by application of TuBA with the other sets of parameters, identified those reference biclusters that shared a significant proportion of their genes and samples

Percentile Cutoff	Overlap Significance Cutoff	Number of Matching Biclusters
5 (Reference)	1.00E-16	353
10	1.00E-23	327
10	1.00E-24	327
10	1.00E-25	317
10	1.00E-26	303
10	1.00E-27	290
10	1.00E-28	288

Table 3.7: **Robustness of TuBA's biclusters to different choices of the overlap significance cutoff for the TCGA data set.**

with biclusters obtained from other parameter choices. We observed that 95% of our reference biclusters (334 out of 353 biclusters) were enriched ($FDR < 0.001$) in at least one bicluster in the set of biclusters corresponding to the percentile cutoff of 7.5%, and the overlap significance cutoff of 10^{-19}. However, the agreement steadily decreased as we increased the size of the percentile sets - only about 86% of the reference biclusters (304 out of 353 biclusters) were enriched ($FDR < 0.001$) in at least one bicluster in the set of biclusters corresponding to the percentile cutoff of 15%, and the overlap significance cutoff of 10^{-30}. Once again, we observed that there was a significant difference (Mann-Whitney U test $p < 10^{-05}$) in the number of genes contained in the biclusters that matched, compared to the ones that did not match; the median size of biclusters that matched was 20 (range: $3 - 1012$), while the median size of biclusters that did not match was 4 (range: $3 - 60$). Thus, we can conclude that some of the smaller biclusters that correspond to alterations/deregulation in small subsets of tumors were not identified as we increased the size of the percentile sets. Nonetheless, we were able to validate most of our reference biclusters in the sets of biclusters obtained with other choices of the parameters, thereby demonstrating their robustness over a range of choices for the percentile set size cutoff.

Another aspect that we investigated was the level of agreement between the reference biclusters and the biclusters obtained for different choices of just the overlap significance cutoff. Our reference biclusters once again corresponded to the ones obtained for TCGA (high expression) with the following choices of parameters for TuBA: (i) Percentile Cutoff: 5%, and (ii) Overlap Significance Cutoff: 10^{-16}. We compared the reference biclusters to the ones obtained for the choice of the percentile cutoff fixed at 10%, and the overlap significance cutoffs ranging between 10^{-23} to 10^{-28} (in decrements of orders of 10) (see Table 3.7). For the overlap significance cutoff of 10^{-23}, we observed that 93% (327 out of 353 biclusters) of the reference biclusters were enriched ($FDR < 0.001$) in at least one bicluster in the set of biclusters corresponding to the percentile cutoff of 10%. However, as we decrease the overlap significance cutoff (higher significance) the agreement decreases - at the overlap significance of 10^{-28}, only 81% (288 out of 353 biclusters) of the reference biclusters had

corresponding matches. This is reasonable, since an increase in overlap significance would lead to a decrease in the total number of gene-pairs in the entire graph(s). This leads to fewer genes and samples in the graphs overall, and would therefore lead to an omission of gene-pair associations that correspond to some of our reference biclusters. However, inspite of a five-fold difference in the significance level of overlap, the agreement between the reference biclusters and the biclusters obtained with other choices of the overlap significance cutoff is quite respectable.

3.9 Utility of TuBA's tunable knobs

The results of the robustness analyses may appear to indicate that TuBA's biclusters are impervious to changes in the choices of its parameters. That is certainly not the case. In fact, the tunable aspect of our algorithm is precisely due to the influence of our parameters (knobs) on the final biclusters. The influence of the choices of the parameters on the determination of the final biclusters can best be illustrated with help of two real examples from the application of TuBA to the TCGA data set. For this data set, with the choice of top 5% percentile cutoff and overlap significance cutoff of $p \leq 10^{-16}$, one of the biclusters we discovered was made up exclusively of genes from the Cancer-Testis Antigen family: *MAGEA2*, *MAGEA3*, *MAGEA6*, *MAGEA10*, *CSAG1*, *CSAG2*. However, this bicluster was not identified for percentile set cutoff choices of top 10% and 20% with overlap significance cutoffs of $p \leq 10^{-25}$ and $p \leq 10^{-36}$, respectively. This clearly suggests that these genes are only expressed in approximately 5% of all BRCAs. Thus, an increase in the percentile cutoff results in the omission of aberrant co-expression signatures found in comparatively small subsets of tumors within the population, such as this one. In sharp contrast, we observed another small bicluster made up of genes exclusively from the Cancer-Testis Antigen family (*CTAGE4*, *CTAGE6*, *CTAGE9*) for all three upper percentile cutoffs (5%, 10% and 20%) and the respective overlap cutoffs, suggesting that these genes exhibit aberrant co-expression in a larger proportion of tumors within the population (at least 10%).

The second example is that of the bicluster corresponding to the HER2 (*ERBB2* gene) amplicon ($17q12$). Panel A in Fig.3.2 shows the number of samples present in the bicluster for the top percentile cutoff of 5%, as the overlap cutoff is lowered from 10^{-20} to 10^{-16}. We can observe that no new samples get added to the bicluster, even though there is a reduction in the significance level of overlap by four orders of magnitude. We can infer that the choice of the upper percentile cutoff at 5% put a cap on the maximum number of samples that can belong to the bicluster; lowering of the overlap cutoff does not lead to an increase in the number of samples in the bicluster precisely because almost all of them were identified at the higher significance level of 10^{-20}. However, when

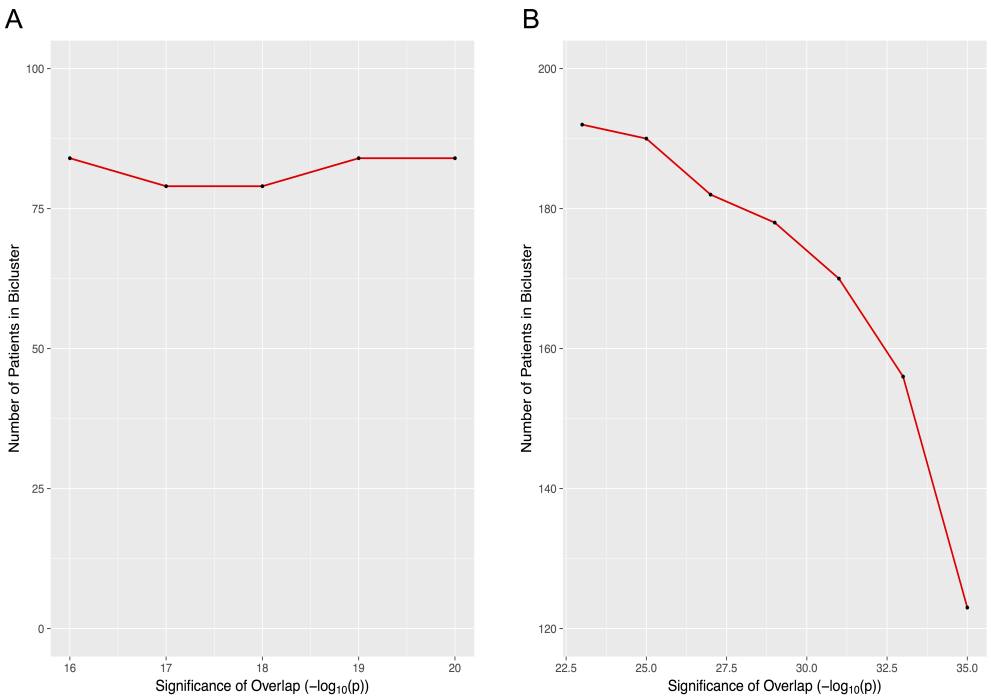

Figure 3.2: **Effect of the choice of percentile cutoff on the bicluster associated with the HER2 amplicon.** Panel A shows the number of samples in the bicluster as the overlap significance is lowered from 10^{-20} to 10^{-16} for percentile set size of 5%, and panel B shows the number of samples in the bicluster as the overlap significance is lowered from 10^{-35} to 10^{-23} for percentile set size of 10%. Figure reproduced from [1].

the top percentile cutoff is chosen at 10%, we see a steady increase in the number of samples in the bicluster as we lower the overlap cutoff from 10^{-35} to 10^{-23}, with a gradual reduction in the number of samples that get added as we approach $p = 10^{-23}$ (panel B in Fig. 3.2). Further reduction in the overlap cutoff results in the addition of only a handful of samples to the bicluster, however, it leads to a significant increase in the number of edges in the overall graph. This is the trade-off that we need to keep in mind while making our choice for the second parameter.

3.10 TuBA can be used for RNA-seq data to find biclusters associated with low expression levels of the associated genes

One of the biggest advantages of RNA sequencing over microarray assays is the reliable measurement of transcripts at low levels of expression. Theoretically, only the depth of sequencing limits the dynamic range of RNA-seq data [76, 77]. Out of the three data sets, only TCGA used an RNA-seq platform for measuring transcript abundances. Given that the TCGA data has adequate sequencing depth, we can expect a reliable quantification of transcripts that are expressed at low levels in subsets of tumors. We therefore applied TuBA to the TCGA data set to explore transcriptional profiles associated with low expression.

The protocol followed to identify relevant gene pairs using our proximity measure is slightly different to the one used for identifying gene pairs corresponding to high expression. For low expression, (i) the grey rectangular box in the Venn diagram in panel B of Fig. 2.1 represents the set of all the samples in the data set, (ii) the red circle represents the set of bottom (5%,10%, etc.) percentile samples for gene 1 *or* the set of all samples with zero expression for gene 1, whichever is larger, and (iii) the blue circle represents the set of bottom (5%, 10% etc.) percentile samples for gene 2 *or* the set of all samples with zero expression for gene 2, whichever is larger. Consequently, the regions labeled by **a**, **b**, **c** and **d** in the Venn diagram respectively represent:

- **a**: the set of samples that are found in the bottom percentile sets (or the subset of samples with expression value zero) of both gene 1 and gene 2
- **b**: the set of samples that are found only in the bottom percentile (or the subset of samples with expression value zero) set of gene 1
- **c**: the set of samples that are found only in the bottom percentile (or the subset of samples with expression value zero) set of gene 2
- **d**: Set of samples that are neither found in the bottom percentile set of gene 1 (or the subset of samples with expression value zero), nor in the bottom percentile set (or the subset of samples

with expression value zero) of gene 2

The contingency table corresponding to this would be exactly the same as the one shown in panel B of Fig. 2.1. Once again, since we are calculating significance values for all possible gene pairs, we need to correct them for multiple hypothesis testing to get FDRs.

3.11 TuBA identifies biclusters enriched in known subtypes of BRCA

3.11.1 Enrichment in ER/HER2 based subtypes

Based on the expression levels of the $ESR1$ (ER) and the $ERBB2$ (human epidermal growth factor receptor 2 [HER2]) genes, breast tumors can be classified into four known subtypes: (i) ER−/HER2−, (ii) ER+/HER2−, (iii) ER−/HER2+, and (iv) ER+/HER2+ (where + corresponds to over expressed and − corresponds to under expressed). The classification of breast tumors based on the expression levels of these genes allows for systemic treatment protocols to be adopted based on the subtype.

To find whether some of our biclusters were associated with one or more of these subtypes, we relied on the following information: (i) For METABRIC, we used the ER and HER2 status available in the clinical file, and (ii) for TCGA and GEO, we used the expression levels of the transcripts to classify the samples into one of the four subtypes described above. Once again we relied on the one-sided Fisher's exact test to determine whether there is enrichment of the biclusters within sets of samples that belong to certain subtypes. In the Venn diagram in panel B of Fig. 2.1, the grey rectangular box (containing both the circles) in this case represents the set of all samples that have an unambiguous ER/HER2 subtype status. The red circle represents the set of all the samples belonging to the bicluster, the blue circle represents the set of samples that belong to a given subtype. Therefore, the regions labeled by **a**, **b**, **c**, and **d** respectively represent:

- **a**: the set of samples in the bicluster that also belong to the given ER/HER2 subtype
- **b**: the set of samples in the bicluster that do not belong to the given ER/HER2 subtype
- **c**: the set of samples for the given subtype that are not present in the bicluster
- **d**: the set of samples that are neither present in the bicluster, nor are they of the given subtype

For each of the four ER/HER2 subtypes, we prepared such contingency tables for every bicluster and tested for enrichment within the subtype. The generic form of the contingency tables for these tests is shown in Table. 3.8.

	Samples belonging to subtype X	Samples not belonging to subtype X
Samples in bicluster	**a**	**b**
Samples not in bicluster	**c**	**d**

Table 3.8: **Contingency table for calculating enrichment of biclusters in subtypes.**

High expression

Quite remarkably, we observed that a majority of biclusters for all three data sets were enriched in the ER-/HER2- subtype (Fig. 3.3 shows the subtype enrichments of biclusters for TCGA and METABRIC; the association with copy number is explained in a following section). The overall result of the enrichments for each data set is summarized below -

- 53% of the biclusters (180 out of 340 biclusters) for METABRIC were enriched in ER−/HER2−
- 54% of the biclusters (191 out of 353 biclusters) for TCGA were enriched in ER−/HER2−
- 40% of the biclusters (148 out of 369 biclusters) for GEO were enriched in ER−/HER2−

Low expression (TCGA)

Similar to what we observed for biclusters corresponding to high expression for TCGA, a majority of the biclusters corresponding to low expression were enriched in the ER-/HER2- subtype - to be precise, 46% of the biclusters (94 out of 203 biclusters) from TCGA were enriched in the ER-/HER2- subtype. Fig. 3.4 shows the subtype enrichments of biclusters for TCGA (low); the association with copy number is explained in a separate section below.

3.11.2 Enrichment in the PAM50 subtypes

According to the subtype classification based on the Prosigna Breast Cancer Prognostic Gene Signature Assay (PAM50), there are five subtypes of BRCA: (i) Basal-like, (ii) Her2-enriched, (iii) Luminal A, (iv) Luminal B, and (v) Normal-like [78]. To determine whether our biclusters were associated with one or more of these subtypes, we used the following information - for METABRIC, we used the PAM50 subtype labels for the samples provided in the clinical file, while for TCGA we used the PAM50 calls available for a subset of samples in the clinical file from the paper published by the TCGA group in 2012 [70]. For GEO, no such information was available, thus the GEO data set was not included in the subtype enrichment analysis for PAM50. For both TCGA and METABRIC, we identified the samples that had been assigned to at least one of the five PAM50 subtypes and prepared new data sets that only contained samples with the PAM50 subtype information.
The setup for the one-sided Fisher's exact test is similar to the one for the ER/HER2 subtypes.

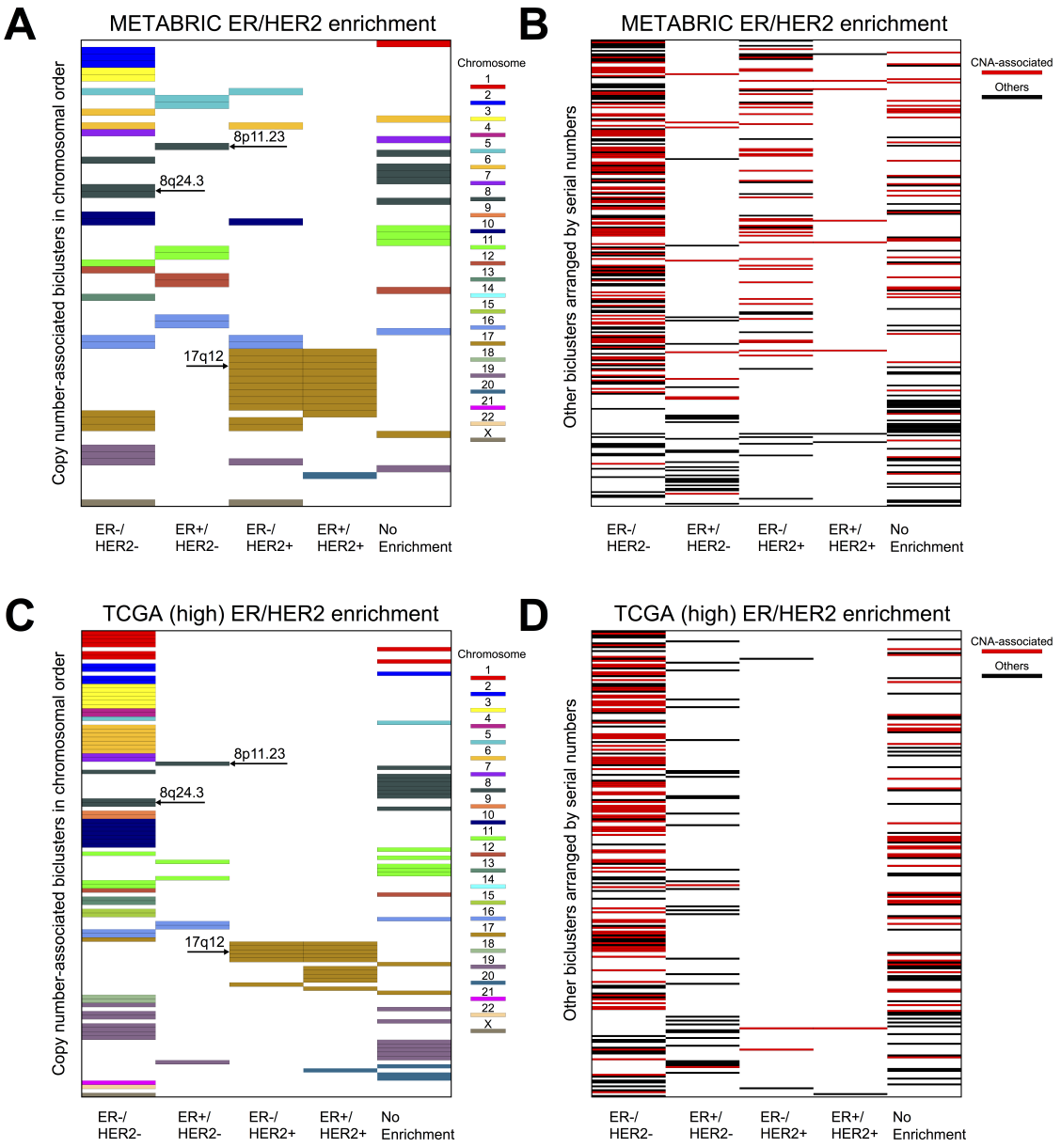

Figure 3.3: **TuBA's biclusters are enriched in ER/HER2 subtypes and are also associated with copy number gains.** The biclusters are represented by horizontal bars in each panel. For CNA associated biclusters with proximally located genes (panels A and C) the bars are color-coded according to the chromosome number of their constituent genes for METABRIC and TCGA data sets, respectively. Panels B and D show the remaining biclusters arranged according to their serial numbers. The ones associated with copy number gains of genes (not located proximally) are shown in red, while the rest are shown in black. The thickness of the bar in each panel depends on the total number of biclusters displayed in the given panel and does not represent its chromosomal extent. Figure reproduced from [1].

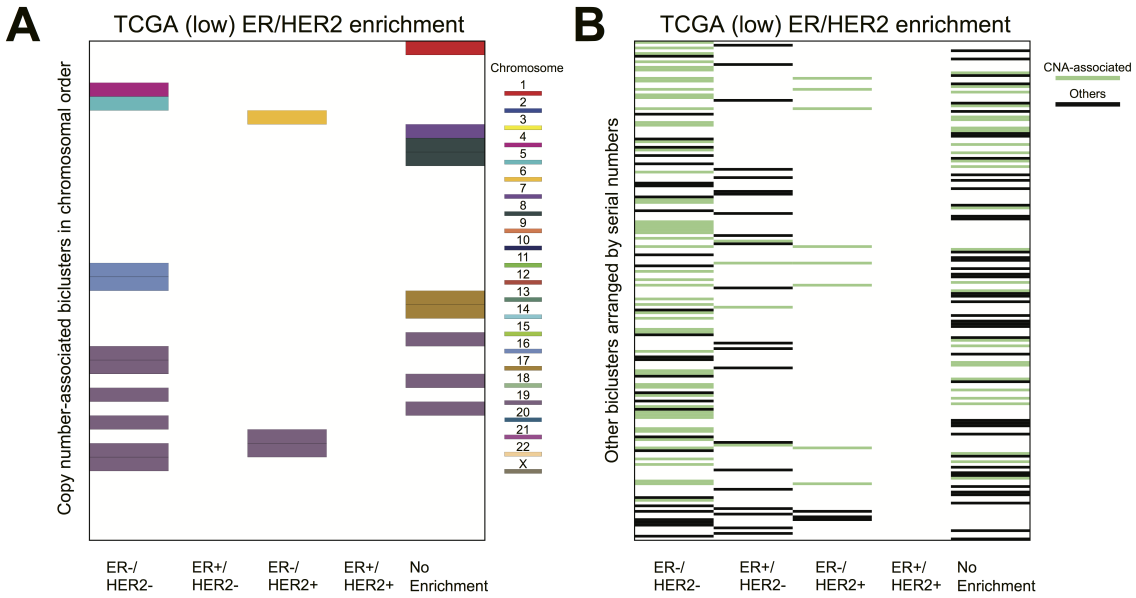

Figure 3.4: **TuBA's biclusters are enriched in ER/HER2 subtypes and are also associated with copy number losses.** The biclusters are represented by horizontal bars in each panel. For copy number loss associated biclusters with proximally located genes (panel A) the bars are color-coded according to the chromosome number of their constituent genes. Panels B and D show the remaining biclusters arranged according to their serial numbers. The ones associated with copy number losses of genes (not located proximally) are shown in green, while the rest are shown in black. The thickness of the bar in each panel depends on the total number of biclusters displayed in the given panel and does not represent its chromosomal extent. Figure reproduced from [1].

For the case of PAM50, in the Venn diagram in panel B of Fig. 2.1: (i) the grey rectangular box represents the set of all samples that have an unambiguous PAM50 subtype assignment available, (ii) the red circle represents the set of all the samples belonging to the bicluster, and (iii) the blue circle represents the set of samples that belong to the given PAM50 subtype. Therefore, the regions labeled by **a**, **b**, **c** and **d** respectively represent:

- **a**: the set of samples in the bicluster that also belong to the given PAM50 subtype
- **b**: the set of samples in the bicluster that do not belong to the given PAM50 subtype
- **c**: the set of samples for the given subtype that are not present in the bicluster
- **d**: the set of samples that are neither present in the bicluster, nor are they of the given subtype

The forms of the contingency tables are exactly the same as the one shown in Table. 3.8. After the significance values were calculated for every bicluster, they were corrected for multiple hypothesis testing.

High expression

In the case of PAM50 subtypes, we observed that the majority of biclusters, for both TCGA and METABRIC, were enriched in the Basal-like subtype -

- 52% of the biclusters (177 out of 340 biclusters) for METABRIC were enriched in the Basal-like subtype
- 55% of the biclusters (264 out of 480 biclusters) for TCGA were enriched in the Basal-like subtype

Low expression (TCGA)

For the low expression analysis of the TCGA (PAM50) data set, we observed that 48% of the biclusters (231 out of 480 biclusters) from TCGA were enriched in the Basal-like subtype (panels E and F in Fig. 3.5). Therefore, a substantial proportion of biclusters were enriched in the Basal-like subtype for both high and low expression.

This is a truly remarkable observation, especially given that ER−/HER2− and/or Basal-like subtypes only account for $15\% - 20\%$ of all BRCAs. Thus, TuBA is able to uncover several biclusters that are associated with altered transcriptional programs within tumors of this subtype, further highlighting their tremendous heterogeneity.

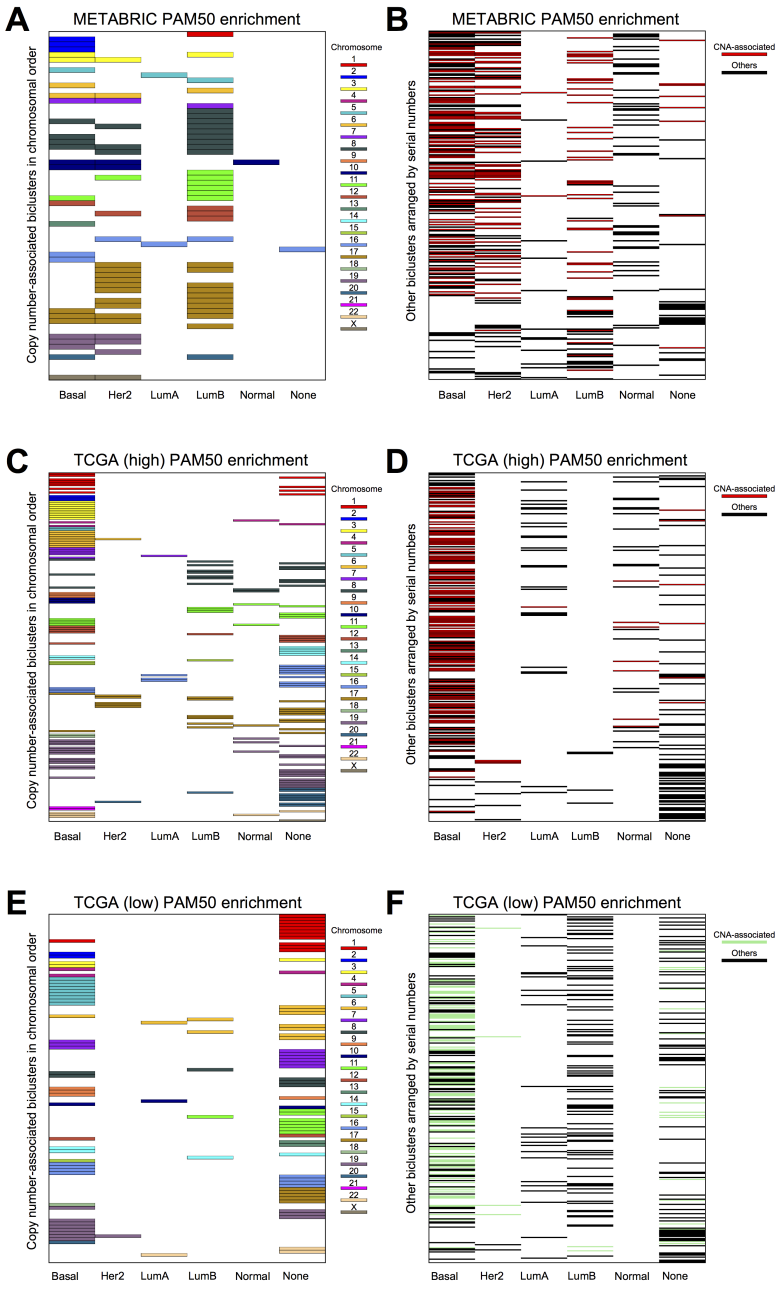

Figure 3.5: **TuBA's biclusters are enriched in PAM50 subtypes and are also associated with copy number gains and losses.** The biclusters are represented by horizontal bars in each panel. For copy number alteration associated biclusters with proximally located genes (panels A, C and F) the bars are color-coded according to the chromosome number of their constituent genes for METABRIC and TCGA data sets, respectively. Panels B and D show the remaining biclusters arranged according to their serial numbers. The ones associated with CNA of genes (not located proximally) are shown in red, the ones associated with while the rest are shown in black. Panel F shows the biclusters arranged according to their serial numbers, with the ones associated with copy number loss of genes (not located proximally) shown in green. Figure reproduced from [1].

3.12 TuBA discovers biclusters with proximally located genes

In all three data sets, TuBA discovered several biclusters almost exclusively made up of genes that are known to be located near each other on the chromosomes. One of the underlying mechanisms responsible for such co-expression signatures could be copy number alterations. There are two kinds of copy number of alterations that influence the expression levels in completely opposite ways:

1. *Copy Number Amplification (CNA)* - CNA refers to significant gains in the number of copies of the affected genes beyond the 2 copies that are normally present in the genome. If the affected site(s) is transcriptionally active, the expression levels of the affected genes would be up-regulated.

2. *Copy Number Deletion or Loss* - This refers to the loss of one or both the copies of the affected genes in the genome. If the affected site(s) is transcriptionally active, the expression levels of the affected genes would be down-regulated.

In order to determine the associations of copy number alterations with our biclusters, we used thresholded copy number data for both TCGA and METABRIC. In these copy number data sets, for any given gene the values for a sample could be: (i) +2: high level amplification, (ii) +1: copy number gain, (iii) 0: neutral, or no change, (iv) -1: hemizygous deletion (deletion of one copy of given gene), or (v) -2: homozygous deletion (deletion of both copies of the given gene in the genome). In the following subsections we describe the results of the associations between copy number altered sites and our observed biclusters for the three data sets.

3.12.1 Biclusters associated with CNA

We adopted the following stepwise process in order to identify whether gain in copy number might be the underlying mechanism for some of our biclusters -

1. List the genes that constitute the bicluster.

2. Pick a gene from the list prepared in step 1.

3. List all the samples that have a gain in copy number for the chosen gene (sample values > 0) *and* are also present in the gene expression dataset.

4. List all the samples that are in the top percentile set for the given gene and have copy number data available.

	Samples with CNA	Samples without CNA
Top samples in bicluster	**a**	**b**
Samples not in bicluster	**c**	**d**

Table 3.9: **Contingency table for testing bicluster enrichment in copy number gain associated samples.**

We used the one-sided Fisher's exact test to look for enrichment of samples with copy number gains for the given gene in the bicluster. To understand the setup of the contingency tables for the test, we again refer to the Venn diagram in panel B of Fig. 2.1. The grey rectangular box (containing both the circles) now represents the set of all the samples that have both copy number and gene expression data available, the red circle represents the set of top percentile samples corresponding to the given gene present in the bicluster, and the blue circle represents the set of samples that have a gain in copy number for the given gene. Thus, the regions labeled by **a**, **b**, **c** and **d** in the Venn diagram respectively represent:

- **a**: the set of top percentile samples in the bicluster that also have a gain in copy number for the given gene
- **b**: the set of top percentile samples in the bicluster that do not have a copy number gain for the given gene
- **c**: the set of samples that have a gain in copy number for the given gene but are not present in the bicluster
- **d**: the set of samples that are neither present in the bicluster nor do they have a gain in copy number for the given gene

The corresponding 2×2 contingency table is of the form shown in Table. 3.9. Note, there is a contingency table set up for each individual gene in all the biclusters discovered by TuBA. Thus, after calculating the $p-$values for each gene, we corrected them for multiple hypothesis testing to obtain FDRs.

In order to discern the biclusters associated with CNA, we decided to calculate the overall enrichment of the biclusters in copy number gain samples. We did this by combining the FDRs of all the genes within the biclusters using Fisher's method [79] to yield a single significance value for every bicluster. Based on this, we observed that 56% and 64% of biclusters from the METABRIC and TCGA datasets were associated with CNA ($p < 0.001$), respectively. These biclusters included those that contained genes from multiple locations across different chromosomes. In order to shortlist those biclusters that were composed exclusively of genes that were located near each other, we

imposed the following constraint: only those CNA-associated biclusters were kept, whose seeds were made up exclusively of genes located on the same chromosome. Based on this criteria, only 60 (18%) biclusters from METABRIC were associated exclusively with CNA of proximally located genes (see panel A in Fig. 3.3), the remaining biclusters associated with CNA were enriched in genes from distant chromosomal locations (panel B in Fig. 3.3). Similarly, 112 (32%) biclusters from TCGA were associated with CNA of proximally located genes (panel C in Fig. 3.3C). Several of these were associated with chromosomal loci shown to exhibit copy number gains in BRCAs in previous studies [80, 81].

3.12.2 Biclusters associated with copy number loss

In order to explore the association between the biclusters obtained from the low expression analysis and loss of copy number, we repeated the copy number analysis described above, except that this time, the samples in row 1 of the contingency table in Table 3.9 corresponded to the bottom percentile set, and the samples in the column 1 corresponded to samples with copy number deletion or loss.
We observed that 52% biclusters from the TCGA dataset were enriched in copy number losses (Fig. 3.4). However, only 21 biclusters contained genes located near each other on the same chromosome (panel A in Fig. 3.4), the remaining biclusters associated with copy number loss contained genes located in different chromosomes.

3.12.3 Biclusters not associated with copy number gains or losses

We also observed some biclusters with proximally located genes that were not associated with gain or loss in copy number. We focused on the ones discovered for high expression, since these may be of greater prognostic value.
For TCGA, 14 biclusters out of 353 consisted of genes located next to each other within the chromosome, while 18 biclusters out of 340 for METABRIC consisted of genes located near each other. Details of the genes and subtype-specific enrichments for some of these biclusters are summarized in Table 3.10. Examples of biclusters from this category include the biclusters consisting of genes from the Cancer-Testis antigens family - *MAGEA2*, *MAGEA3*, *MAGEA6*, *MAGEA10*, *CSAG1*, *CSAG2*, *CSAG3* (Xq28)/*CT45A3*, *CT45A5*, *CT45A6* (Xq26.3). Previous studies have shown these genes to exhibit aberrant expression in ER-/HER2- breast tumors [82], as well as in a few other tumor types [83].

Cytoband Locus	Genes	Subtype Enrichment
1p13.3	**NBPF4, NBPF6, NBPF22P (5q14.3)**	ER+/HER2- for TCGA
1p31.1	**PRKACB, SAMD13, DNASE2B**	ER+/HER2+ for TCGA
		ER+/HER2-, LumA & LumB for METABRIC
1q21.2	**PPIA (7p13), PPIAL4C, PPIAL4G**	ER-/HER2- for TCGA
1q21.3	**S100A7, S100A8, S100A9**	ER-/HER2+ for METABRIC
		Her2-enriched (PAM50) for METABRIC
1q23.1	**CD1B, CD1C, CD1E**	ER-/HER2- for TCGA
2q21.1	**TUBA3C (13q12.11), TUBA3D, TUBA3E**	ER+/HER2- for TCGA
		ER+/HER2- for METABRIC
4q13.2	**UGT2B7, UGT2B10, UGT2B11, UGT2B28**	ER-/HER2+ for TCGA
		Her2-enriched (PAM50) for METABRIC
4q32.2	**NAF1, NPY1R, NPY5R, TKTL2**	ER+/HER2- for TCGA
7p15.2	**HOXA2, HOXA3, HOXA5, HOXA6, HOXA7, HOXA9, HOXA10**	None in TCGA
		Basal-like for METABRIC
12p12.3	**WBP11, C12orf60**	ER+/HER2- for TCGA
		ER+/HER2- & LumB for METABRIC
12p13.1	**HEBP1, HTR7P1**	ER+/HER2- for TCGA
12q13-q13.13	**KRT81, KRT83, KRT86, KRT6A, KRT6B, KRT6C**	ER-/HER2- for TCGA
		ER-/HER2- & Basal-like for METABRIC
15q21.1	**DUOX1, DUOX2, DUOXA1, DUOXA2**	ER-/HER2- for TCGA
16p13.3	**HBA1, HBA2, HBB (11p15.4)**	None
19q13-q13.41	**KLK5, KLK6, KLK7, KLK8**	ER-/HER2- for TCGA
Xq22.1-q22.2	**TCEAL1, TCEAL3, TCEAL4, TCEAL6**	ER+/HER2- for TCGA
		ER+/HER2-, LumA & LumB for METABRIC
Xq28	**CSAG1, CSAG2, CSAG3, MAGEA2, MAGEA3, MAGEA6, MAGEA10, MAGEA12**	ER-/HER2- for TCGA
		ER-/HER2- & ER-/HER2+ for METABRIC
		Basal-like & Her2-enriched (PAM50) for METABRIC

Table 3.10: **Biclusters with genes that are located near each other but are not associated with copy number changes.**

3.13 TuBA identifies biclusters associated with the immune and stromal cells present in the tumor samples

One of the most commonly used methods to validate the results of clustering or biclustering analysis is Gene Ontology (GO). It is based on a hierarchical graph structure in which the nodes represent terms dealing with biological processes, molecular functions, cell components etc., and the edges connecting the nodes indicate dependency/association between them. The usual approach is to perform a statistical analysis (hypergeometric test) to look for over-representation in GO. By that we mean, that given a subset of genes (genes in the bicluster) from a larger population (the set of all the genes in the data set), we are interested in knowing if the frequency of an annotation to a GO term is more than what would be expected from chance alone, given the overall population. We obtain a $p-$value for all those terms in GO in which the genes of the bicluster are enriched. We used GeneSCF [84], a functional enrichment tool, to perform GO term enrichment for our biclusters. We relied on the Benjamini-Hochberg FDRs ($FDR < 0.001$) to identify GO terms enriched in our biclusters.

Based on the GO term enrichment analysis, we discovered that multiple biclusters identified by TuBA appear to be associated with non-tumor cells. For instance, some of the largest biclusters independently identified in all three data sets were associated with immune response. The top five Gene Ontology - Biological Processes (GO-BP) terms for the biclusters associated with immune response were: T cell co-stimulation, T cell receptor signaling pathway, T cell activation, regulation of immune response, and positive regulation of T cell proliferation. This indicated immune cell infiltration in a significant number of tumor samples. In order to corroborate these results, we

	Samples in group (i)	Samples not in group (i)
Samples in bicluster	**a**	**b**
Samples not in bicluster	**c**	**d**

Table 3.11: **Contingency table for determining enrichment of bicluster samples in samples with high immune infiltration/high stromal scores.**

took the help of a tool called ESTIMATE (Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data) developed by Yoshihara *et al* [85] to predict tumor purity, and the presence of infiltrating stromal/immune cells in tumor tissues using gene expression data. The algorithm of ESTIMATE is based on Gene Set Enrichment Analysis (GSEA) and generates 3 scores for a given sample:

1. Stromal score - captures the presence of stroma in the tumor tissue.
2. Immune score - captures the extent of infiltration of immune cells in the tumor tissue.
3. Estimate score - uses the two scores above to infer tumor purity.

We calculated the immune scores for the samples in the TCGA data set, and stratified the samples based on these scores into three groups: (i) top 25 percentile (samples with highest immune infiltration), (ii) intermediate 50 percentile, and (iii) bottom 25 percentile (samples with lowest immune infiltration). We then prepared contingency tables of the form shown in Table. 3.11 for the one-sided Fisher's exact test. The quantities **a**, **b**, **c**, and **d** in the table correspond to:

- **a**: the set of samples in the bicluster that also belong to group (i)
- **b**: the set of samples in the bicluster that do not belong to group (i)
- **c**: the set of samples in group (i) not present in the bicluster
- **d**: the set of samples neither present in group (i), nor in the bicluster

Based on these tests, we verified that samples in the biclusters enriched in GO-BP terms associated with immune response were indeed enriched in samples with the highest levels of immune infiltration as calculated by ESTIMATE ($FDR < 0.001$).

Additionally for all three datasets, we also observed a bicluster associated with the stromal adipose tissue. The top 5 GO-BP terms for this bicluster were: response to glucose, triglyceride biosynthetic process, triglyceride catabolic process, retinoid metabolic process, and retinol metabolic process. Once again based on a test similar to the one described above for immune infiltration (but with stromal scores this time), we confirmed that these biclusters were enriched within the top 25 percentile samples based on stromal scores determined by ESTIMATE.

	Gene pairs enriched in GTEx	Gene pairs not enriched in GTEx
Gene pairs in bicluster	**a**	**b**
Gene pairs not in bicluster	**c**	**d**

Table 3.12: **Contingency table for determining enrichment of biclusters in gene co-expression signatures associated with normal mammary tissues.**

Thus, to summarize it appears that TuBA can identify the subsets of samples that exhibit greater presence of non-tumor cells within the biopsied tumor tissues. In this process, it also identifies the same gene sets that ESTIMATE relies on to arrive at its scores for immune infiltration, stromal presence, and overall tumor purity.

3.14 TuBA identifies gene co-expression signatures associated with normal tissue

We were curious about how many of TuBA's biclusters represented co-expression signatures associated with normal breast tissues. In order to investigate this we decided to use gene expression data obtained from normal tissues in the Genotype-Tissue Expression (GTEx) database available at www.gtexportal.org/home/. The aim of the GTEx project is to collect and analyze multiple human tissues from donors in order to assess genetic variation within their genomes, and to seek expression quantitative trait loci (eQTLs) that explain variations in gene expression based on the genetic variants.

Instead of biclustering the GTEx gene expression data set consisting of 214 normal breast tissue samples, we simply used our proximity measure to compare the top 10 percentile samples between all those pairs of genes that are also present in the graph analyzed by TuBA for the TCGA dataset. We thus calculated the significance of overlaps to obtain a matrix of $p-$values for all theses pairs of genes. For each bicluster, we listed all its constituent gene pairs and identified the number of gene pairs that were found to have a level of overlap significance above a preset cutoff ($p \leq 10^{-05}$). We then set up contingency tables for each bicluster of the form shown in Table 3.12 to test for the significance of the proportion of gene pairs that are also enriched in GTEx. The quantities **a**, **b**, **c**, and **d** in Table 3.12 respectively represent:

- **a**: the set of gene pairs present in bicluster as well as enriched in GTEx
- **b**: the set of gene pairs present present in bicluster from TCGA but not enriched in GTEx
- **c**: the set of gene pairs enriched in GTEx but not present in bicluster

- **d**: the set of gene pairs not enriched in GTEx and not present in bicluster

We observed that only 6.75% of biclusters obtained for the TCGA versus GTEx comparison were enriched in gene-pair associations identified in the GTEx dataset. The bicluster associated with the adipose tissue signature was one of the biclusters found enriched in GTEx. Another group of biclusters enriched in the three cancer datasets as well as in GTEx, were those associated with translation and ribosomal assembly. The top 5 GO-BP terms for these biclusters were: translation, rRNA processing, ribosomal small subunit biogenesis, ribosomal large subunit assembly, and ribosomal large subunit biogenesis. These biclusters were enriched in the ER−/HER2− subtype ($FDR < 0.001$).

3.15 TuBA identifies biclusters of clinical relevance

Given that the core objective of identification of biomarkers of prognostic, or predictive value is to improve the clinical outcome for the patients. It is only natural that emphasis be placed on those molecular signatures that show the strongest associations with the outcomes of the disease.
To ascertain if some of the discovered biclusters showed differential clinical outcomes for the patients whose tumors exhibited gene co-expression signatures captured by the bicluster, we performed a Kaplan-Meier (KM) survival analysis using recurrence free survival (RFS) times for the patients in the METABRIC and GEO data sets (the TCGA cohort had insufficient number of patients with incidence of recurrence for this kind of simplistic survival analysis to be statistically robust). The idea of the survival analysis was quite simple. For each bicluster, we generated survival curves for two groups -

1. Patients whose tumor samples belong to the bicluster
2. Patients whose tumors samples do not belong to the bicluster

We then used the logrank test [86] to test the null hypothesis that there is no difference between the two sets of patients in the probability of a recurrence at any time point.

One of the biclusters reaffirmed what is already known to the community. For METABRIC, patients in the bicluster associated with the HER2 amplicon (17q12) had significantly shorter RFS times compared to the rest (Fig. 3.6). This is because patients in the METABRIC study were enrolled before the general availability of trastuzumab [71].
In addition to the HER2 amplicon, We also observed biclusters associated with CNA at the 8q24.3 locus in all three datasets. The patients belonging to these biclusters also exhibited significantly shorter RFS times compared to those patients whose tumors did not have amplification at this locus (panels A, B and C in Fig. 3.7).

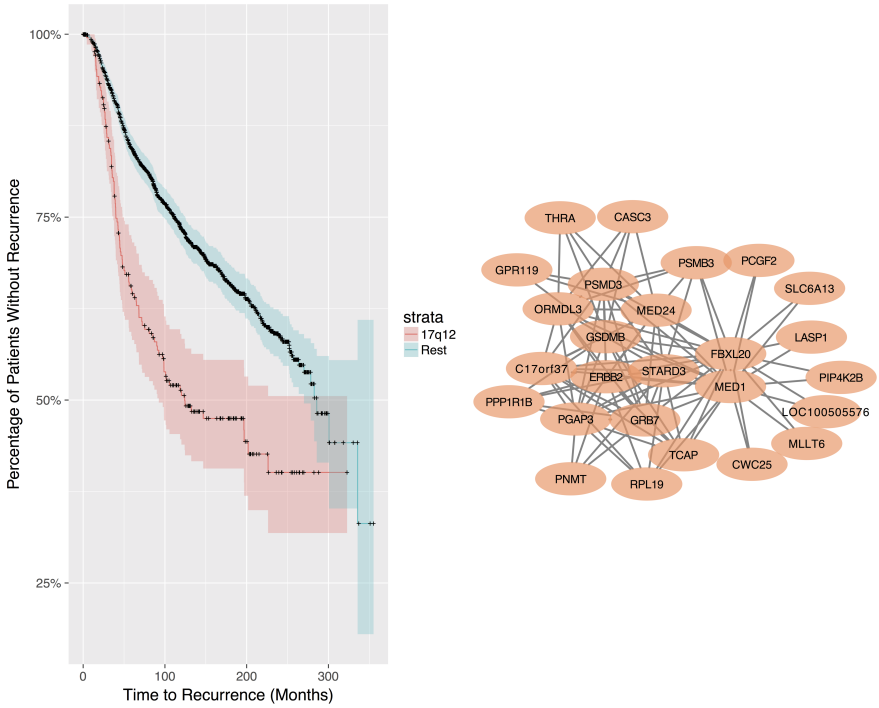

Figure 3.6: **HER2 amplicon is associated with poor prognosis of patients in the METABRIC data set.** On the left: The KM survival curve (red) for the patients belonging to the bicluster associated with the HER2 amplicon for the METABRIC data set compared to the survival curve for the rest of the patients (blue). On the right: the graph of the bicluster associated with the HER2 amplicon for the METABRIC data set. Figure reproduced from [1].

We observed a similar result when we restricted the set of samples to include just the ER+/HER2- tumors, which validated an observation made earlier that copy number gain of the 8q24.3 locus may confer resistance to ER targeted therapy [87]. However, it must be pointed out that in all three data sets, the biclusters associated with amplification of the 8q24.3 locus were enriched in the ER−/HER2− subtype ($p < 0.001$).

Based on the degrees of the genes in the biclusters associates with copy number gain at 8q24.3, a few promising candidates would include *PUF60*, *EXOSC4*, *COMMD5*, and *HSF1*. Specifically, PUF60 is an RNA-binding protein known to contribute to tumor progression by enabling increased MYC expression and greater resistance to apoptosis [88].

In all 3 data sets, we also observed a robust bicluster exhibiting co-expression of genes located at the 8p11.22-p11.23 locus. For both METABRIC and GEO, patients in biclusters associated with copy number gains of the 8p11.21-p11.23 locus had significantly shorter RFS times compared to patients that did not have copy number gains at this locus (panels D, E and F in Fig. 3.7). We found that patients in this bicluster were enriched in the luminal B subtype. Patients with tumors of luminal

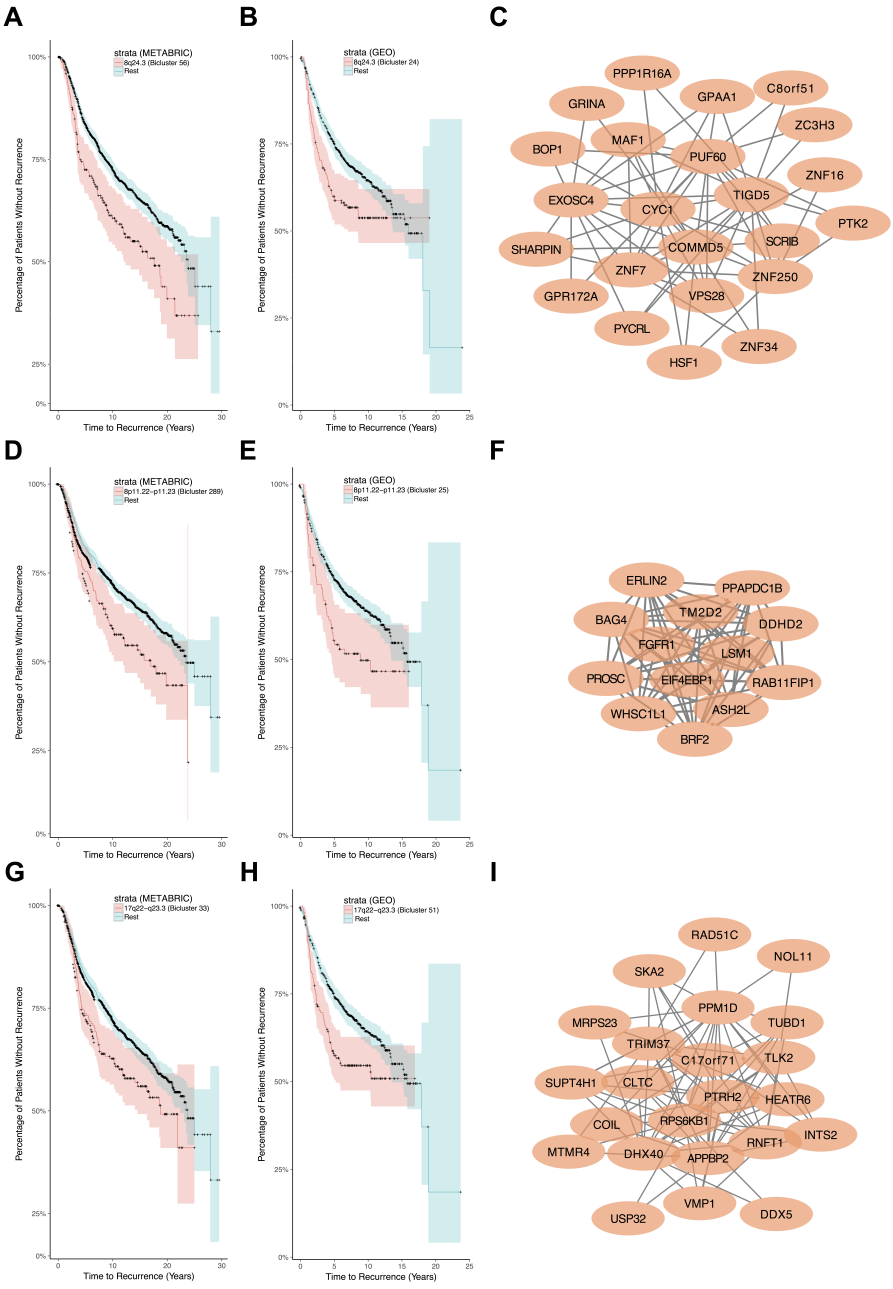

Figure 3.7: **Transcriptionally active copy number amplified sites associated with higher risk of recurrence identified by TuBA.** Panel on the left show the KM curves for the groups of patients (red curve) exhibiting copy number amplifications at (top to bottom) (i) 8q24.3, (ii) 8p11.22-p11.23, and (iii) 17q22-q23.3, respectively, compared to the rest of the patients (blue curves) in the METABRIC data set. corresponding patients belonging to the bicluster associated with the HER2 amplicon for the METABRIC data set compared to the survival curve for the rest of the patients (blue). On the right: the graph of the bicluster associated with the HER2 amplicon for the METABRIC data set. Figure reproduced from [1].

B subtype are known to have poorer prognosis than patients with the luminal A subtype among ER+/HER2− tumors [66]. This suggests to us that amplification of the 8p11.21-p11.23 loci may be another marker of higher risk of recurrence post ER targeted therapy.

The third bicluser we wish to highlight is the one associated with copy number gains at the 17q22-q23.3 locus. In both METABRIC and GEO data sets, patients belonging to the associated biclusters had significantly shorter RFS times compared to patients whose tumors did not exhibit such copy number gains at this locus (panels G, H, and I in Fig. 3.7). For METABRIC, the samples in these biclusters were enriched in the luminal B (PAM50), ER+/HER2+, and ER−/HER2+ subtypes ($FDR < 0.001$), while for GEO, the samples in these biclusters were enriched in the ER+/HER2+ and ER−/HER2+ subtypes ($FDR < 0.05$). Therefore, copy number gains at this locus may confer added risk of recurrence in HER2+ breast cancers.

There were several other biclusters that exhibited differential relapse outcomes (not all associated with CNA). For METABRIC, 61 biclusters out of 340 were found to exhibit differential relapse outcomes (poor prognosis) for the patients present in the biclusters. Out of these 61 biclusters, 69% were enriched in the ER−/HER2− subtype (64% for basal-like). Out of these, (67%) were associated with copy number gains. For GEO, there were 48 such biclusters (13%) that exhibited differential relapse outcomes (poor prognosis), 25% of these were enriched in the ER−/HER2− subtype.

We also looked at other clinically relevant variables such as tumor grade and the lymph node status to determine if some of our biclusters were particularly enriched in tumors of advanced stages and/or positive lymph node status. Enrichment tests for bicluster samples in tumors of higher grades revealed that 8 biclusters from TCGA were enriched in tumors of grade 3C. GO-BP enrichment of the gene sets in these biclusters revealed that these were associated with terms related to angiogenesis, vasculogenesis, blood vessel maturation etc. For METABRIC, 4 biclusters were enriched in tumors of grade 3. Out of these, 2 were associated with the HER2 amplicon (17q12). For GEO, 68 biclusters were enriched in tumors of grade 3, including one associated with CNA at the HER2 amplicon. Enrichment tests for lymph node status of patients in our biclusters revealed that 4 biclusters in TCGA that were enriched in positive lymph node status - one associated with the HER2 amplicon, others associated with CNA at the 8q22.1-q22.3 locus, the 17q23.1-q23.3 locus, and the 19q13.43 locus, respectively. For METABRIC, we also observed 4 biclusters enriched in samples with positive lymph node status in the corresponding patients - 2 of them were associated with copy number gains at the HER2 amplicon, the other 2 were associated with copy number gains at the 19q13.11-q13.12 locus and the 1q21.3-q25.1 locus, respectively. What is interesting is that, biclusters associated with CNA at the 8q24.3 locus, the 8p11.21-p11.23 locus, and 17q22-q23.3 locus were not enriched in tumors of higher grades or in patients with positive lymph node status in any of the 3 datasets.

Using the METABRIC data set, we confirmed that none of these biclusters (associated with 8q24.3, or 8p11.21-p11.23, or 17q23.1-q23.3) were among the 36 biclusters found to be enriched in samples with the poorest expected 5-year survival outcome (Nottingham Prognostic Index ($NPI > 5.4$) [89, 90]. This underscores the importance of including these transcriptionally active CNA sites into gene prognostic signature assays so as to reclassify patients with these alterations into categories with higher risk of recurrence.

3.16 Putting bicluster signatures together - clustering of biclusters reveals shared mechanisms within subsets of tumors

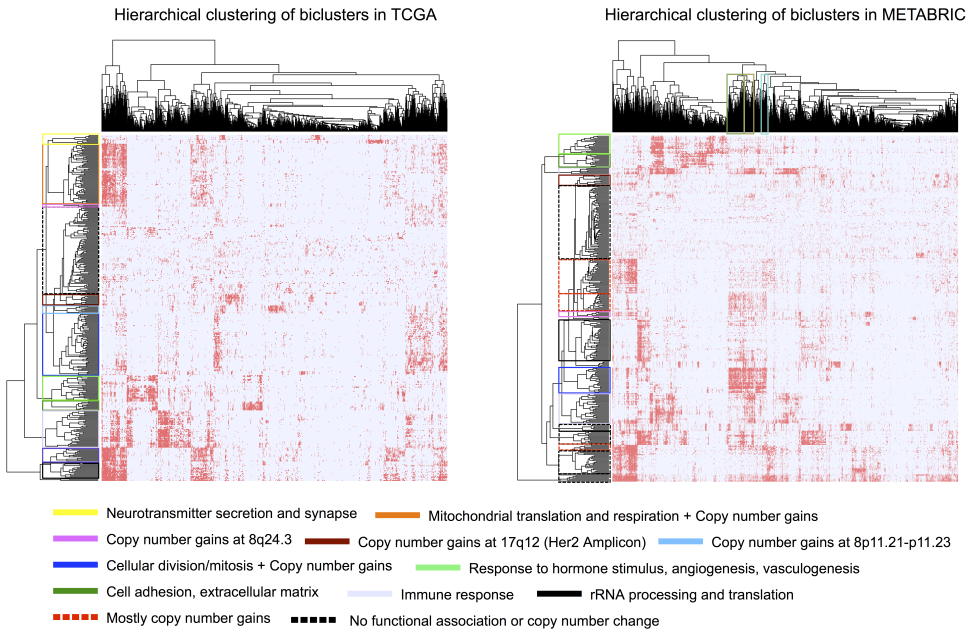

Figure 3.8: **Hierarchical clustering of biclusters and samples reveals shared mechanisms within subsets of tumors.** The bicluster (rows)-samples (columns) binary matrix was clustered using Hamming distance. Panel A on the left shows the clustered matrix for TCGA; panel B shows the clustered matrix for METABRIC. Figure reproduced from [1].

Sample membership based hierarchical clustering of biclusters revealed distinct groups of biclusters that presumably share common functional mechanisms (Fig. 3.8). These included clusters associated with cell cycle and proliferation, immune response, cell adhesion (extracellular matrix), translation, mitochondrial translation, and ribosomal RNA processing pathways. Since a significant fraction of our biclusters were associated with copy number alterations, we also found distinct groups of biclusters associated with significant copy number changes such as the ones associated with the

HER2 amplicon, the 8p11.21-p11.23 loci, or the 8q24.3 locus. Similarly, we used hierarchical clustering to group samples that were enriched in similar sets of biclusters, highlighting differential clinical outcomes. In particular, we observed 2 sets of samples enriched in biclusters associated with CNA at the 8q24.3 locus. In one group, the samples were enriched in biclusters related to immune response; this group showed significantly lower incidence of recurrence compared to those without enrichment in immune response-related biclusters. Both of these sets of samples were enriched in biclusters associated with cell division and proliferation. In contrast, we observed a cluster of samples enriched in biclusters associated with 8q24.3 copy number gain and a number of other loci, however these were not enriched in biclusters associated with cell division and proliferation. This group exhibited low incidence of recurrence. We also observed a cluster of samples with significantly poor RFS that were enriched in biclusters associated with CNA at 17q25.1-q25.3, and in biclusters associated with cell division and proliferation.

Clustering analysis of biclusters and samples based on the membership of samples within biclusters allowed us to identify the sites that were altered concomitantly within the same subsets of samples. It also improved our perspective on the tumor microenvironment in the subsets of samples that exhibit non-tumor associated signatures (such as immune, extracellular matrix, etc.). For instance, we noticed a difference in RFS outcomes between two groups of patients that exhibit copy number gains at 8q24.3; the group that was additionally associated with an immune response signature was observed to have better RFS outcomes compared to the group that did not exhibit a strong association with the immune response. Such differences in disease progression due to distinct microenvironments in tumors with similar transcriptional alterations can help us better understand the potential role of the microenvironment within the context of tumors harboring these specific alterations.

3.17 Summary

3.17.1 TuBA's relevance to cancer data sets

TuBA is based on a proximity measure specifically designed to extract gene co-expression signatures that correspond to the extremes of expression (both high and low for RNA-seq data. This enables it to preferentially identify aberrant gene co-expression signatures associated with the heterogeneous disease states of tumors. Identification of such altered transcriptional profiles can be particularly relevant for those tumors that have so far eluded targeted drug development for therapy. The best example to illustrate TuBA's ability to identify such alterations is that involving tumors belonging to the Basal-like, and/or triple negative subtypes for BRCA. Although these tumors account for

only 15% of all BRCAs in the population, TuBA identified a large number of biclusters that are associated with alterations within tumors of these subtypes.

A locus of particular interest is the one that exhibits copy number gain at 8q24.3. We observed that 3040% of all biclusters were enriched in samples with copy number gains at the 8q24.3 locus ($FDR < 0.001$). Additionally, 51% of all biclusters obtained from the low expression analysis of TCGA were enriched in the samples corresponding to the 8q24.3 bicluster. In sharp contrast, the samples in the biclusters corresponding to CNA at 8p11.21-p11.23 or 17q12 (HER2 amplicon) were independently enriched ($FDR < 0.001$) in only about 5% of all biclusters, for both TCGA and METABRIC respectively. An earlier study has also identified 8q24.3 by Representational Difference Analysis as a location of oncogenic alterations in breast cancer that can occur independent of neighboring MYC amplifications [91]. Although the 8q24.3 bicluster itself is enriched within the ER$-$/HER2$-$ samples, these observations highlight this locus as a promising prognostic molecular marker for BRCAs, irrespective of subtype.

3.17.2 A surprising absence - ER

Somewhat to our surprise, we observed that none of the biclusters for all 3 data sets contained the $ESR1$ gene, which codes for the oestrogen receptor (ER). The vast majority ($70\% - 75\%$) of BRCAs overexpress ER, therefore it was surprising that we did not observe the $ESR1$ gene as a member of any of our biclusters. Upon closer investigation, we found that $ESR1$ had statistically significant associations with several genes, however its level of significance of overlap with these genes was much lower ($FDR > 10^{-7}$) than the chosen cutoffs for all three datasets. Thus, it appears that over-expression of ER may not be a sufficient condition to strongly drive the co-expression of genes involved in other pathways discovered by TuBA.

3.17.3 Identification of potential biomarkers

Change in copy number is often not a sufficient condition for elevated (or suppressed) expression levels of transcripts, as there are multiple layers of regulation of transcription in cells [92, 93]. TuBA specifically identifies sets of genes with copy number changes that are transcriptionally active (or inactive), filtering out the ones that are unlikely to influence disease progression. Moreover, the graph-based approach allows us to infer the relative importance of each gene within a bicluster, based on its degree. In the case of high expression analysis, the degree of each gene is an indicator of how frequently it is expressed aberrantly at high levels by the subset of samples that comprise any given bicluster. As an example, consider the CNA-associated bicluster from TCGA associated with gains at the 8q22.1-q22.3 loci. The bicluster exhibited enrichment in lymph node positive

patients (the corresponding bicluster in METABRIC has a significance level of FDR = 0.052 for patients with positive lymph node status). The gene with the highest degree in the bicluster was $MTDH$ (metadherin), which has been shown to be associated with increased chemo resistance and metastasis in BRCA [94, 95, 96].

3.17.4 Limitations

Unlike most biclustering methods, TuBA does not allow arbitrary overlaps between its biclusters. This is because it is designed to discover biclusters enriched in samples that correspond to the extremals for the corresponding gene set. It does not consider biclusters with other conditions for the same gene set. That being said, TuBA's biclusters are not exclusive, some overlap between their genes and samples is permitted. For example, in case of an ER−/HER2− BRCA sample that exhibits CNA at 8q24.3, because of high immune-cell infiltration in the tumor, the same sample may also be present in the biclusters enriched in genes associated with the immune cells.

Another limitation of TuBA is that it can only be applied reliably for large datasets (containing more than several tens of samples). This is because depending on cohort heterogeneity, some of the overlaps between percentile sets may not be significant in smaller datasets. Thus, this limitation is imposed by the particular nature of our proximity measure that leverages the size of the datasets. However, in data sets with sufficient number of samples this proximity measure offers a significant benefit - it not only enables the identification of the plethora of gene aberrant co-expression signatures associated with the tumors, but also enables the estimation of the extent or prevalence of the identified alterations in the population. This is where the tunable aspect of TuBA becomes relevant - the two knobs should be viewed as valuable aids that help estimate the extents of the prevalence of various alterations in the tumor population and their clinical relevance.

Chapter 4
Comparison with other biclustering methods

"Comparison is the thief of joy."

– Theodore Roosevelt

"Oh, the cleverness of me!"

– J. M. Barrie, *Peter Pan*

4.1 Necessary context

TuBA is designed to identify biclusters with samples that correspond to the extremals for the corresponding sets of genes, and does not consider other subset of conditions for the same sets of genes for biclustering. In contrast, most biclustering methods seek sub-matrices with constant, or coherent gene expression patterns. Given this key difference, only those biclusters that exhibit such expression patterns in the extremal (top or bottom) subsets of samples for some subsets of genes, are expected to have agreement with the biclusters identified by TuBA.

In earlier studies that compared biclustering algorithms [46, 97], synthetic datasets that contained constant, shifting, and/or scaling patterns of expression values for subsets of conditions and genes were relied on to evaluate how well the algorithms were able to identify biclusters with these known patterns. TuBA is not based on a mathematical model of expression values, hence a comparison based on synthetic datasets based on explicitly defined patterns is not feasible. For real data sets, the most common approach to seek validation for the identified biclusters is to perform GO term enrichment. Biclustering algorithms have been assessed based on how many of their identified biclusters were enriched in GO terms. However, this approach is not entirely satisfactory. Ideally, just as in the case for synthetic data sets, if some gene co-expression signatures are already known to exist within some subsets of conditions in real data sets, biclustering algorithms could be assessed based on whether they are able to identify such signatures.

In case of tumor related gene expression data sets, we have the benefit of complementary genomic data that do provide us with truth-known scenarios for validation. For example, we know that a

significant proportion of tumors across multiple tumor types frequently exhibit genomic alterations such as gains or losses in the copy numbers of genes. Quite often, these alterations are not limited to a single gene but include multiple genes located at neighboring chromosomal locations. If such alterations are located at transcriptionally active sites in the chromosomes, then we should expect to observe co-expression of the genes located in these CNA regions. In BRCA for instance, approximately 20% of tumors possess extensive gains in copy numbers of genes at the 17q12 cytoband locus (includes *ERBB2* (Her2), *STARD3*, *GRB7*, *PNMT*, *PGAP3*, *MED1* etc.). Identification of co-expression of genes at this locus in the subset of samples that are histologically HER2-positive (HER2+) represents a simple truth-known scenario that can be used to verify whether a given biclustering algorithm identifies the co-expression of these genes in the subset of samples that exhibit this alteration.

4.2 Other biclustering methods

In a paper by Serin and Vingron [98], a novel biclustering method (DeBi) that identifies differentially expressed biclusters was described and applied to both synthetic and real gene expression data sets. One of the real gene expression data sets they applied DeBi to, was that for diffuse large B-cell lymphoma (DLBCL), which consisted of 661 genes and 180 samples [99]. Apart from DeBi, they applied ISA [100], OPSM [101], QUBIC [47], and SAMBA [102] biclustering algorithms to this data set. We applied TuBA to the DLBCL data set and used the biclustering results obtained by the authors to compare TuBA against these methods. To ensure a uniform and unbiased comparison between the enrichment results for biclusters from different algorithms, we used GeneSCF [84] to perform GO-BP enrichment on the biclusters found by all the methods.

We provide brief descriptions for each of the biclustering methods TuBA was compared to, in the following subsections below (the descriptions have been adapted from Eren *et al.* [97] and Pontes *et al.* [43]).

4.2.1 BIMAX

Binary Inclusion-Maximal (BIMAX) biclustering algorithm was proposed by Prelic *et al.* as a simple biclustering method to serve as a reference for comparative purposes [46]. The first step in the algorithm involves binarizing the gene expression data, which is done by thresholding such that expression values higher than the chosen threshold are set to 1, the other values are set to 0. The algorithm then proceeds to identify all the submatrices made up completely of 1s. Coupled with the thresholding criteria described above, BIMAX can find only up-regulated biclusters.

4.2.2 DeBi

The Differentially Expressed Biclusters (DeBi) algorithm developed by Serin and Vingron [98], is based on a well known data mining approach called frequent itemset. The gene expression data is first binarized for both up-regulation and down-regulation based on a predefined threshold. DeBi's biclusters have the following two properties: (i) a bicluster is a maximum homogenous gene set where each gene in the bicluster should be highly or lowly expressed over all the bicluster samples, and (ii) each gene in the bicluster shows statistical difference in expression between the samples in the bicluster and the samples not in the bicluster.

4.2.3 ISA

Iterative Signature Algorithm (ISA) was developed by Bergmann *et al.* [100]. It defines its biclusters as transcription modules that consist of sets of genes that are co-regulated most stringently within specific sets of experimental conditions. It relies on two symmetric requirements: each column in the bicluster must have an average value above some threshold T_C, and each row must have an average value above some threshold T_R. The algorithm starts with a set of randomly selected genes or conditions (called seeds), iteratively refining the genes and conditions until they match the definition of a transcription module depending on the chosen thresholds. Initial seeds are randomly chosen without any overlap restriction, therefore, different biclusters may contain overlapped genes and/or conditions. ISA can find up-regulated or down-regulated biclusters.

4.2.4 OPSM

The Order Preserving Submatrix Method (OPSM) was developed by Ben-Dor *et al.* [101]. According to this method, biclusters are defined as order-preserving submatrices in which the columns are linearly ordered such that the expression values in the rows of the submatrices increase linearly. Thus, constant columns, shifting, scaling and shift-scale expression patterns can be identified by this method. OPSM biclusters are constructed by iteratively growing partial biclusters, each time assigning scores based on the probability that it will grow to some fixed target size. The best partial biclusters (the ones with the highest probability) are kept for the next iteration till the scores converge.

4.2.5 QUBIC

The Qualititative Biclustering (QUBIC) is a graph-based biclustering method developed by Li *et al.* [47]. The main idea of their method revolves around finding heavy subgraphs in the bipartite graph

representation of the data. In these graphs, genes are represented as vertices, and edges connect every pair of genes that are similar each other for given subset of conditions (the edges are weighted based on the level of similarity). Biclusters are identified in these graphs starting with the heaviest unused edge as a seed. In the first iteration, it basically seeks biclusters with nonzero constant columns in the discretized data. Subsequent expansion steps relax this constraint and allow the addition of rows that are not totally consistent. The qualitative (or semi-qualitative) representation allows the algorithm to detect different kind of patterns including shifting and scaling. It can also find both positively and negatively correlated expression patterns.

4.2.6 SAMBA

The Statistical-Algorithmic Method for Bicluster Analysis (SAMBA) developed by Tanay *et al.* [102] relies on probabilistic modeling of data and graph-based methods to identify subsets of genes that *respond* jointly within a subset of conditions, where a gene is said to respond to a condition if there is a significant change in its expression level (with respect to normal) in that condition. They first model the gene expression data as a bipartite graph whose two parts correspond to conditions and genes, respectively, and the edges refer to significant expression changes. Vertex pairs in the graph are assigned weights according to a probabilistic model, so that heavy subgraphs correspond to biclusters with high likelihood (the weight of a subgraph is the sum of the weights of the gene-condition pairs in it). SAMBA can find up-regulated or down-regulated biclusters.

4.3 Results of the comparison based on GO term enrichment

4.3.1 DLBCL

We applied TuBA to the DLBCL dataset, and performed a GO-BP enrichment analysis on the seeds of the resulting biclusters using GeneSCF. TuBA discovered 94 biclusters in total (45 biclusters for high, and 49 biclusters for low expression), DeBi discovered 127 biclusters in total (68 biclusters with up-regulated genes, and 59 biclusters with down-regulated genes), ISA discovered 49 biclusters, OPSM discovered just 12 biclusters, QUBIC discovered 100 biclusters (the default number), and SAMBA discovered 128 biclusters. Panel A in Fig. 4.1 shows the proportions of GO-BP enriched biclusters for 5 different significance levels (FDRs) - 0.001%, 0.1%, 0.5%, 1%, and 5%. For the FDR cutoff of 5%, almost all the biclusters for every biclustering algorithm were enriched in at least one GO-BP term.

At higher levels of significance, TuBA had fewer enriched biclusters compared to other algorithms. This can be partly attributed to the fact that several of TuBA's biclusters include proximally located

genes with aberrant expression due to copy-number changes, which may not show enrichment in GO-BP terms. Another reason is that the other algorithms discover biclusters that can have arbitrary overlaps between their genes. TuBA, on the other hand, does not permit any overlap between the genes of the seeds of its biclusters. We identified the biclusters discovered by other algorithms did not share genes between them. For this, we filtered the biclusters to exclude the ones that have significant overlaps between their genes (i.e. biclusters with hypergeometric test $FDR < 0.001$). DeBi had only 9 biclusters out of a total 127 biclusters that did not have significant overlaps of their genes with any other bicluster (7 biclusters with up-regulated genes, and 2 biclusters with down-regulated genes), ISA had 15 biclusters out of a total 49 biclusters that did not share genes, OPSM had 3 biclusters out of 12, QUBIC had 29 out of 100, and SAMBA had just 10 biclusters out of a total of 128 that did not share genes.

Since most of the biclusters discovered by these algorithms share genes with other biclusters, we expected redundancy in the enriched GO-BP terms as well. To take this redundancy into account, we identified the top 5 GO-BP terms for every bicluster obtained by each algorithm, and prepared a list comprising all the unique GO-BP terms for the entire set of biclusters. For the 5 levels of significance of enrichment: (i) TuBA identified sets comprising 337, 218, 98, 51, and 25 distinct GO-BP terms, (ii) DeBi identified sets comprising 259, 146, 120, 68, and 34 distinct GO-BP terms, (iii) ISA identified sets with 172, 100, 69, 39, and 25 distinct GO-BP terms, (iv) OPSM identified sets with 41, 26, 24, 13, and 8 distinct GO-BP terms, (v) QUBIC identified sets with 237, 172, 106, 54, and 22 distinct GO-BP terms, and (vi) SAMBA identified sets with 250, 138, 120, 72, and 37 distinct GO-BP terms, respectively. Panel C in Fig. 4.1 shows the ratios of the number of elements in these sets to the total number of biclusters for each algorithm, for the 5 different significance levels.

4.3.2 TCGA - BRCA

We also investigated the TCGA BRCA dataset with the following biclustering algorithms: (i) BIMAX, (ii) ISA, (iii) QUBIC, and (iv) SAMBA. We used the biclust package in R for BIMAX [103], the isa2 package in R for ISA [104], the QUBIC package in R for QUBIC [105], and the Expander software for running SAMBA [106]. We used the respective default parameters for all four biclustering algorithms. TuBA discovered 556 biclusters in total (353 biclusters for high, and 203 biclusters for low expression), BIMAX discovered 100 biclusters (default), ISA discovered 244 biclusters, QUBIC discovered 100 biclusters (default), and SAMBA discovered 405 biclusters. Panel B in Fig. 4.1 shows the proportion of GO-BP terms enriched biclusters of each algorithm for 5 different significance levels. This time, when we filtered the biclusters obtained from the other algorithms to

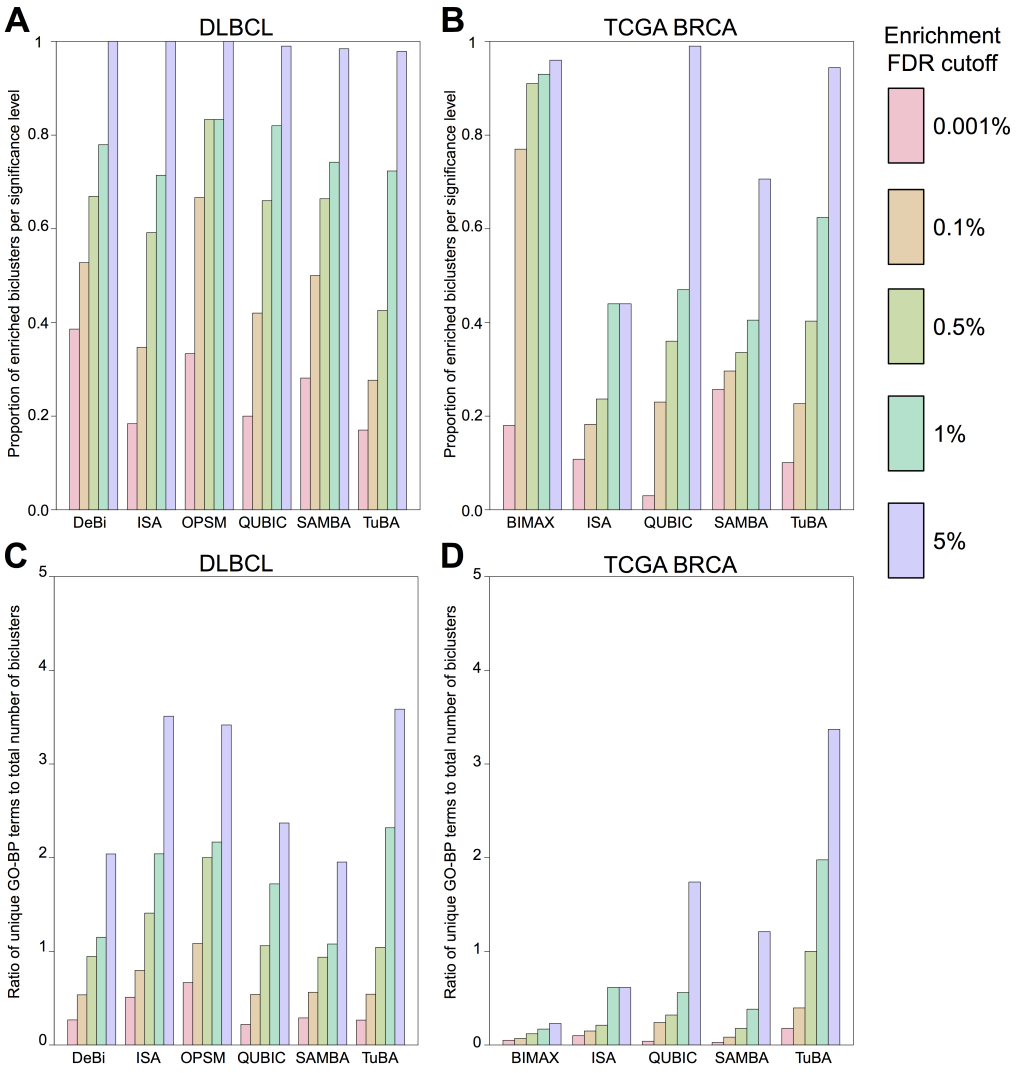

Figure 4.1: **TuBA compared to other biclustering methods based on GO term enrichment of biclusters.** Panels A and B show the proportions of GO-BP terms enriched biclusters for each biclustering method at 5 different significance levels for the DLBCL dataset and the TCGA BRCA dataset, respectively. Panels C and D show the ratios of no. of unique GO-BP terms and total no. of biclusters at 5 different significance levels for the DLBCL dataset and the TCGA BRCA dataset, respectively. Figure reproduced from [1].

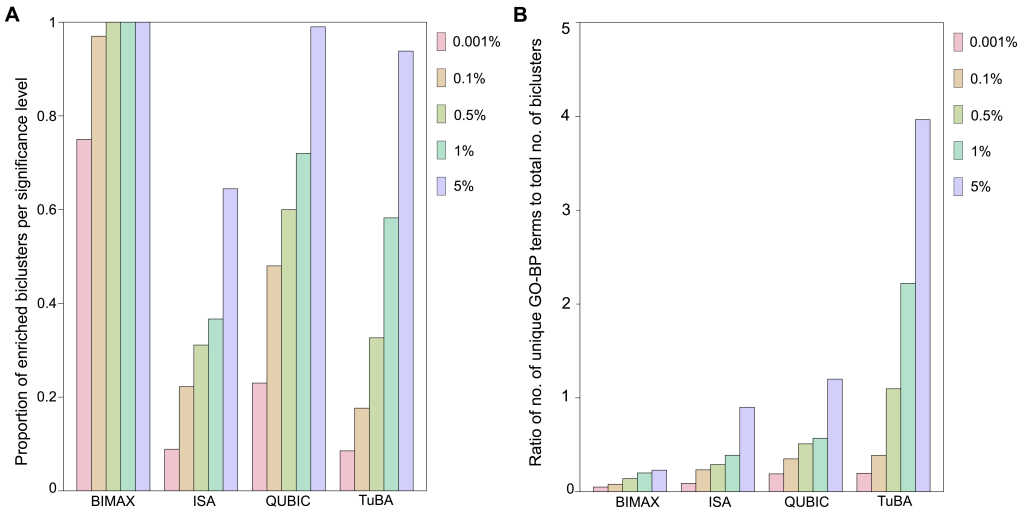

Figure 4.2: **TuBA compared to other biclustering methods based on GO term enrichment of biclusters for METABRIC.** Panels A shows the proportions of GO-BP terms enriched biclusters for each biclustering method at 5 different significance levels for the METABRIC data set. Panel B shows the ratios of no. of unique GO-BP terms and total no. of biclusters at 5 different significance levels for the METABRIC data set. Figure reproduced from [1].

exclude the ones that have significant overlaps between genes (i.e. biclusters with hypergeometric test $FDR < 0.001$), we discovered that none of the four algorithms identified a single bicluster that did not have significant overlap of its genes with at least one other bicluster.

Once again, we identified unique sets of GO-BP terms for the results of each biclustering algorithm. For the 5 levels of significance of enrichment, TuBA identified unique sets with 1874, 1099, 556, 220, and 99 distinct GO-BP terms, respectively. In sharp contrast, BIMAX identified sets with just 23, 17, 12, 7, and 5 GO-BP terms, ISA identified sets with 148, 148, 51, 36, and 24 distinct GO-BP terms, QUBIC identified sets with 174, 56, 32, 24, and 4 distinct GO-BP terms, while SAMBA identified sets with 490, 155, 72, 34, and 11 distinct GO-BP terms, respectively. Panel D in Fig. 4.1 shows the ratios of the number of elements in these sets to the total number of biclusters for each algorithm, for the 5 different significance levels.

4.3.3 METABRIC

Finally, we investigated the METABRIC dataset with the following biclustering algorithms (apart from TuBA): (i) BIMAX (ii) ISA, and (iii) QUBIC. We used the respective default parameters for all three biclustering algorithms. TuBA discovered 340 biclusters (high expression), BIMAX discovered 100 biclusters (default), ISA discovered 90 biclusters, and QUBIC discovered 100 biclusters (default). Panel A in Fig. 4.2 shows the proportion of GO-BP terms enriched biclusters of each algorithm for 5 different significance levels. When we filtered the biclusters obtained from the other

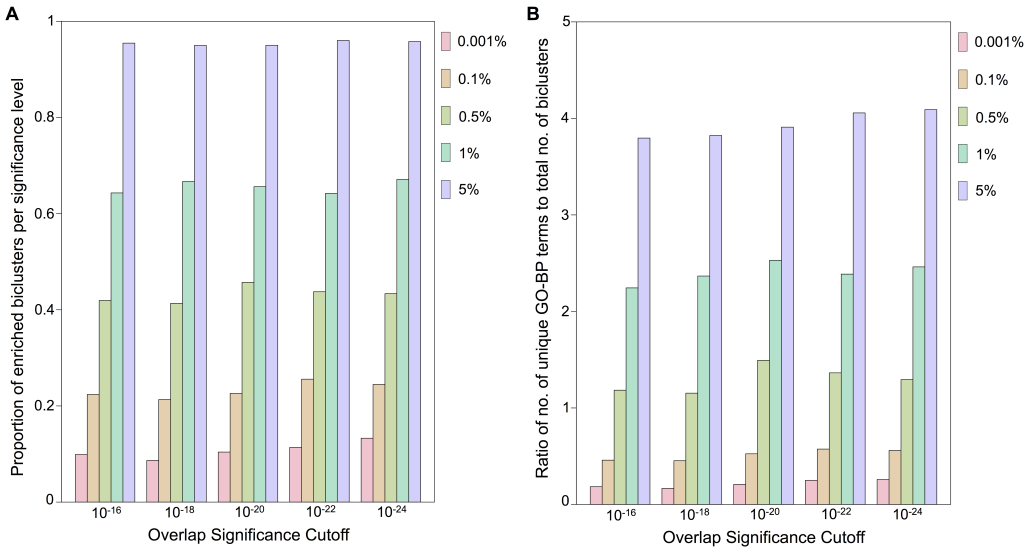

Figure 4.3: **Proportions of GO-BP term enriched biclusters found by TuBA remain consistent across different choices of the overlap significance cutoff for the TCGA data set.** Panel A shows the proportions of GO-BP termenriched biclusters obtained by TuBA at various significance levels for 5 different choices of the overlap significance cutoff for the TCGA dataset. Panel B shows the ratios of number of unique GO-BP terms to the total number of biclusters at different significance levels for the 5 choices of the overlap significance cutoff for the TCGA data set. Figure reproduced from [1].

algorithms to exclude the ones that have significant overlaps between genes (i.e. biclusters with hypergeometric test $FDR < 0.001$), we discovered that none of the three algorithms identified a single bicluster that did not have significant overlap of its genes with at least one other bicluster. We identified unique sets of GO-BP terms for the results of each biclustering algorithm. For the 5 levels of significance of enrichment, TuBA identified unique sets with 1348, 755, 373, 132, and 67 distinct GO-BP terms, respectively. BIMAX identified sets with just 23, 20, 14, 8, and 5 GO-BP terms, ISA identified sets with 81, 35, 26, 21, and 8 distinct GO-BP terms, while QUBIC identified sets with 120, 57, 51, 35, and 19 distinct GO-BP terms, respectively. Panel B in Fig. 4.2 shows the ratios of the number of elements in these sets to the total number of biclusters for each algorithm, for the 5 different significance levels.

For both TCGA and METABRIC, TuBA compares quite favorably with respect to the other algorithms, especially when we account for the redundancy of the GO term enrichment of the biclusters obtained from the other methods.

4.4 GO term enrichment of TuBA's biclusters is not impacted by the choice of its parameters

For most biclustering algorithms, the choice of their parameters play a crucial role in determining their performance. It is possible that different (possibly better) results could be obtained by optimizing the choices of parameters for the other algorithms in the comparisons above. In case of TuBA, there is no concept of optimal (or default) choice of its two parameters; the biclusters obtained for any given choice of the parameters simply satisfy the basic requirements laid down by those choices. Depending on those choices some biclusters may vary, however the most robust co-expression signatures would be observed for large ranges of choice of its tunable parameters.

We looked at GO-BP-term enrichments for TuBAs biclusters for TCGA for 5 different choices of the overlap significance cutoffs - 10^{-16}, 10^{-18}, 10^{-20}, 10^{-22}, 10^{-24}. For these 5 choices of the overlap cutoff the number of biclusters discovered by TuBA were - 353, 300, 221, 176, and 143, respectively. Although, the total number of biclusters obtained differed for each choice, the proportion of enriched biclusters at different significance levels remained similar, irrespective of the parameter choice (panel A in Fig.4.3). Similarly, the ratios of the number of unique GO-BP terms to the total number of biclusters were consistent across all 5 choices of the overlap cutoffs (panel B in Fig. 4.3).

4.5 Results for a truth-known scenario

We discussed earlier that in tumors of some cancer types (such as BRCA) genomic alterations such as gains or losses in the copy numbers of genes is quite common. Our expectation is that it should be possible to find co-expression of the genes located in CNA regions that are transcriptionally active sites. We chose the HER2 amplicon as our reference truth-known CNA region.

We identified HER2+ samples in the TCGA dataset, and for each biclustering algorithm selected those biclusters that were enriched in these samples (hypergeometric test $FDR < 0.001$). BIMAX and SAMBA did not discover any, but ISA identified two biclusters enriched in HER2+ samples. Although the genes from the 17q12 amplicon - *ERBB2*, *STARD3*, *GRB7*, *PNMT*, *PGAP3*, *MED1* etc. were present in them, they made up a tiny subset within the total set of genes in these biclusters. QUBIC also found 4 biclusters that were enriched in HER2+ samples, however they did not contain any genes from the HER2 amplicon itself (not even *ERBB2*). In contrast, not only did TuBA identify a bicluster exclusively associated with genes located at the HER2 amplicon, it identified many other biclusters associated exclusively with CNA of genes located near each other.

Thus, apart from TuBA, only ISA identified co-expression of the genes located at the HER2 amplicon. However, ISA's co-expression modules corresponding to the HER2 amplicon were embedded within

large sets of genes. In the absence of information about copy number gain of the ERBB2 gene, it would be extremely difficult to identify the co-expression module exclusively associated with the amplicon. In conclusion, TuBA successfully uncovered co-expression signatures of genes that are associated with CNA of neighboring sites on the chromosome, and is particularly efficient at identifying transcriptionally active copy number gains.

4.6 TuBA identifies differential co-expression signatures in an unsupervised manner

The nature of our proximity measure allows us to determine differential co-expression signatures without the need to specify conditions in advance. Gao *et al.* [107] proposed a biclustering method - *Bicmix* - based on a Bayesian statistical model to infer subsets of co-regulated genes that covary in all samples, or in only a subset of samples. They also developed a method to recover context-specific gene co-expression networks from the sparse biclustering matrices obtained by Bicmix. They applied Bicmix to a breast cancer data set obtained from the studies by van't Veer *et al.* [108] and van de Vijver *et al.* [109]. We downloaded the data set in the form of an eSet using the breastCancerNKI package in R [110]. We cleaned the data by removing probes with $> 10\%$ missing values, and imputing the missing values for the included probes.

Using Bicmix, Gao *et al.* identified 432 genes that were differentially co-expressed in ER+, and ER− samples. Out of these 432 genes, 430 were up-regulated in ER− samples and down-regulated in ER+ samples, while 2 genes are down-regulated in ER− samples and up-regulated in ER+ samples. We applied TuBA (for high expression) to the same dataset with the following choice of parameters: (i) percentile set size: 10%, and (ii) overlap significance cutoff: $FDR \leq 10^{-08}$. We obtained 549 biclusters, several of which comprised solely of probes associated with a single gene. This is reasonable, since probes that correspond to the same gene are expected to demonstrate higher expression levels in the same set of samples.

We corroborated the differential co-expression signature between ER+ and ER− samples identified by Bicmix using one-sided Fisher's exact tests. We found that the set of 430 genes up-regulated in ER− samples and down-regulated in ER+ samples were enriched in 30 different biclusters discovered by TuBA. Interestingly, the genes that had the highest degrees in the co-expression network discovered by Bicmix - *CD247*, *CD53*, *IL10RA*, and *CXCR3* - were among the ones with the highest degrees in the bicluster with the maximum enrichment found by TuBA. The two genes (*SFRP2* and *COL12A1*) that were up-regulated in ER+ samples and down regulated in ER− samples were also found to be co-expressed in a bicluster found by TuBA.

TuBA also identified biclusters corresponding to amplicons at 17q12 (HER2), enriched in ER− samples ($FDR = 0.02$); 8q24.3, enriched in ER− ($FDR = 0.003$) samples; 17q25-q25.3, enriched in ER− samples ($FDR = 7.09 \times 10^{-05}$). Thus, in addition to the 2 differential co-expression networks identified by Bicmix, TuBA recovered biclusters associated with genomic alterations such as CNA, several of which were differentially expressed between ER+ and ER− samples.

4.7 Summary

TuBA is explicitly designed to identify biclusters with samples that correspond to the extremals for the corresponding sets of genes. Thus, in principle it can only find a subset of the entire set of biclusters that the other algorithms may find in a data set. However, we were able to demonstrate that TuBA not only finds biclusters associated with relevant biological processes (based on GO term enrichment), but it clearly outperforms the other algorithms when it comes to finding co-expression signatures of genes located in transcriptionally active copy number altered regions.

Additionally, TuBA offers an advantage over other differential co-expression analyses methods, since no prior specification of subsets of samples (context) is necessary. Once again this is ensured by our proximity measure which preferentially identifies such differential co-expression signatures. Given these considerations, TuBA offers great promise as a biclustering method that can identify biologically relevant gene co-expression signatures that are not successfully captured by other unsupervised approaches. These co-expression signatures revealed by TuBA can complement the biclusters obtained using other biclustering methods, which can further improve our understanding of the underlying alterations and shared mechanisms in subsets of tumors.

Chapter 5
Ongoing projects and future work

"You must go on. I can't go on. I'll go on."

– Samuel Beckett, *The Unnamable*

5.1 TCGA - Other cancer types

TuBA has been applied to the TCGA data sets of 23 cancer types apart from BRCA. The basic information about these data sets is summarized in Table 3.2. Only those cancer types were chosen which had at least 100 primary tumor samples in the RNA-seq gene expression data sets.
For each cancer type, we identified biclusters made up exclusively of proximally located CNA genes. The method used to identify such biclusters was identical to the one we used in subsection **3.11.1** in Chapter 3. We determined what proportion of the total number of biclusters did these biclusters constitute. The results are summarized in Fig. 5.1. Ovarian and breast cancers are the top 2 cancer types when it comes to transcriptionally active CNA sites identified by TuBA. They were followed by the two kinds of lung cancers which had slightly lesser proportions of biclusters associated with transcriptionally active CNA sites. Note, that in all these cancer types they may have many more copy number alterations. However, most of them would be transcriptionally inactive (since they were not picked up by TuBA), and therefore unlikely to have any impact on disease progression.
We also looked at which transcriptionally active CNA sites are frequently altered in more than 1 cancer type. A few of the most common ones are listed in Table 5.2. CNA at 1q21.3-q23 seems to be the most common alteration, with genes located within this region found in TuBA's biclusters across 10 cancer types.

Currently, work is underway to find clinically relevant associations of the patients in the biclusters for each of the 24 cancer types. In the following subsection, we briefly discuss some interesting observations made for the TCGA bladder cancer data set.

Cancer Type	No. of genes	No. of samples	No. of biclusters
Acute Myeloid Leukemia (LAML)	19939	173	588
Bladder (BLCA)	20240	407	345
Breast (BRCA)	20247	1097	421
Cervical (CESC)	20126	303	286
Colon (COAD)	20035	286	209
Esophagal (ESCA)	20277	184	492
Brain - Glioblastoma (GBM)	19989	154	352
Brain - Lower Grade Glioma (LGG)	20223	516	158
Head and Neck (HNSC)	20261	520	260
Kidney - Clear Cell (KIRC)	20244	533	317
Kidney - Papillary (KIRP)	20205	290	431
Liver (LIHC)	20153	371	429
Lung Adenocarcinoma (LUAD)	20192	515	388
Lung Squamous Cell (LUSC)	20242	502	554
Ovarian (OV)	20184	304	685
Pancreatic (PAAD)	20049	178	421
Pheochromocytoma & Paraganglioma (PCPG)	20034	179	437
Prostate Adenocarcinoma (PRAD)	20223	497	192
Sarcoma (SARC)	20220	259	417
Stomach Adenocarcinoma (STAD)	20290	415	163
Testicular Germ Cell (TCGT)	20168	150	406
Thymoma (THYM)	19998	120	306
Uterine Corpus Endometrial (UCEC)	20145	176	310

Table 5.1: **Summary of TCGA data sets TuBA was applied to.** We also show the number of biclusters TuBA found for each data set for high expression.

Cytoband Locus	Cancer Types
1q21.3 - q23	BLCA, BRCA, CESC, COAD, HNSC, LUAD, LUSC, OV, SARC, STAD
8p11.22 - p11.23	BLCA, BRCA, ESCA, HNSC, LUAD, LUSC, OV, SARC,
12q15 - q21.1	BLCA, BRCA, ESCA, HNSC, LUAD, LUSC, OV, SARC, STAD
13q32 - q34	BLCA, BRCA, COAD, ESCA, LIHC, LUAD, LUSC, OV, SARC, STAD

Table 5.2: **A few transcriptionally active CNA sites common across multiple cancer types.**

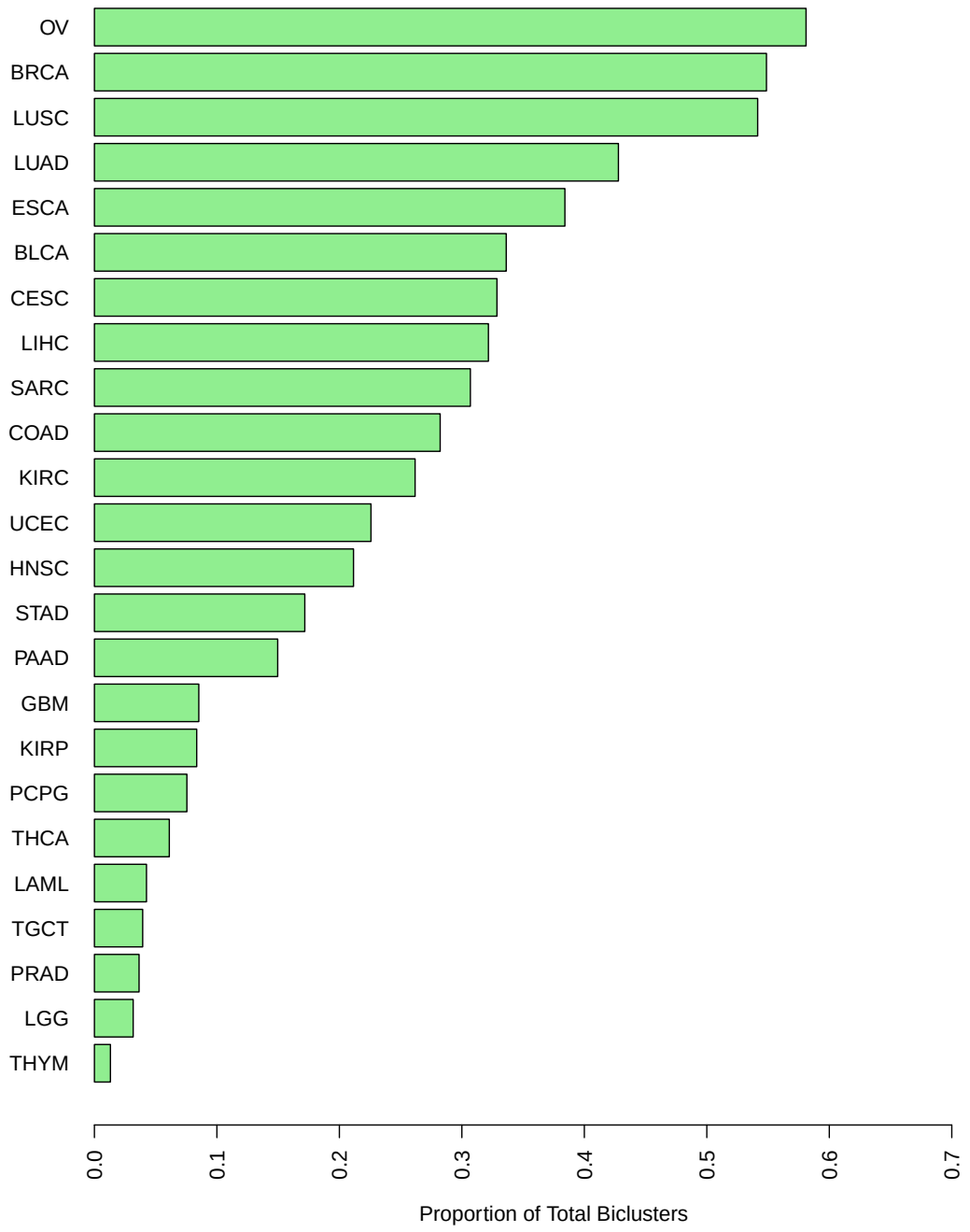

Figure 5.1: **Proportions of total number of biclusters associated exclusively with proximally located CNA genes for 24 cancer types.**

5.1.1 Bladder Urothelial Carcinoma (BLCA)

We applied TuBA to the RNA-seq gene expression data set of TCGA BLCA that consisted of 407 samples and 20240 genes with the following choices of parameters: (i) percentile set size: 5%, and (ii) overlap significance cutoff: 10^{-06}. We obtained 345 biclusters. We performed a simple KM survival analysis similar to the one described in Chapter 3, i.e., for each bicluster, we stratified the set of patients in the data set into two groups - (i) set of patients that belonged to the bicluster, and (ii) set of patients that did not belong to the bicluster. We compared the survival curves of the two groups using the logrank test and corrected the resultant $p-$values for multiple hypothesis testing. We found 3 biclusters that were associated with higher risk of recurrence for the group of patients that belonged to these biclusters. We looked at the GO term enrichments of the gene sets in these biclusters and found that 2 out of these 3 biclusters were associated with terms related to adhesion, extracellular matrix organization etc. In addition, we observed that the gene sets in these biclusters were enriched in the hallmark gene set associated with epithelial-mesenchymal transition (EMT) obtained from the Molecular Signature Database (MSigDB). EMT describes the process wherein epithelial cells take on the mesenchymal phenotype. In case of cancers, it is considered to be one of the ways in which tumor cells gain the capability to invade surrounding tissues and metastasize [111].

This observation is consistent with the one made by Wang *et al.* [112] for the TCGA data set (albeit using a supervised approach). They noted that although there is a positive correlation between infiltrating T-cell abundance (ITA) and EMT-related gene expression, their impact on prognosis is disparate. Higher expression levels of EMT-related genes in tumors with ITA was associated with poor overall survival. In their own study cohort, they observed that in patients with metastatic BLCA treated with nivolumab (an immune checkpoint therapy drug), the ones with higher expression of EMT-related genes showed lower response rates and greater risk of recurrence *despite* the presence of T-cells in their tumors. Quite interestingly, they showed for tumor samples from their own study that non-hematopoietic stromal cells are the major source of the EMT-related gene expression in the bulk transcriptomes of these samples. This raises questions about the EMT-related gene expression being the agent for aggressiveness in these tumors.

We decided to investigate whether early stage bladder cancers also exhibited EMT-related gene signatures. For this, we applied TuBA to a data set that was generated as part of a large-scale study (called UROMOL) to identify molecular markers that could predict the likelihood of progression in patients with Ta or T1 bladder tumors [113]. The gene expression data set consisted of 35154 transcripts and 476 tumor samples. The parameters chosen for TuBA were: (i) percentile set

size: 5%, and (ii) overlap significance cutoff: 10^{-15}. We obtained 259 biclusters, out of which 15 were enriched in the EMT-related gene signature. A closer comparison of these biclusters with biclusters obtained from the TCGA data set revealed that the ones obtained from TCGA contained a few additional genes which are known to be associated with increasing tumor invasiveness. For example, Tks4 is an adaptor protein necessary for formation of podosome/invadopodia by tumor cells [114]. It may be possible that the aggressiveness of tumors with EMT-related gene signature is not a consequence of the EMT signature (which may be associated purely with non-hematopoietic stromal cells), but is in fact due to these genes that enable the tumor cells to invade surrounding tissues. We plan to investigate additional data sets of advanced and early-stage BLCA in order to further validate these findings.

5.2 GTEx - Ribosomal gene modules

Ribosomes are widely considered to be highly conserved and monolithic in composition across all tissues. However, a few recent studies have suggested that based on certain stimuli (both extra- and intra-cellular) the ribosomal composition can vary in order to fulfill cell or tissue specific objectives [115, 116, 117]. In this section, we describe some results that are part of a collaborative study (for which the manuscript is currently under preparation) which investigates the question of tissue specificity of ribosomal compositions. One of the questions asked was whether there are distinct co-expression modules of subsets of RP genes in different tissues across the body.

For this analysis, we used RNA-seq gene expression data from the GTEx data set for 78 RP genes across 53 normal tissue types. In all, the data set had 11,688 samples from 714 non-diseased individuals. The RP transcript read counts were normalized by gene length, and then rescaled so that the sum of the transcript levels were the same for each sample. This eliminates the total variation in RP transcript levels among tissues which is not relevant to our study since we are interested in analyzing inter-tissue variations based on relative RP transcript levels among tissues (the data was prepared by Anshuman Panda who is the lead author of the study).

We applied TuBA to this data set with the following choices for its parameters: (i) percentile set size: 2%, and (ii) overlap significance cutoff: 0.01. We observed two biclusters made up of completely distinct RP genes, enriched in whole blood and brain tissues (brain - cerebellum/cerebellar hemisphere), respectively (see Fig. 5.2). These biclusters correspond to RP gene co-expression modules that have the highest relative frequencies in the two tissues respectively. In order to find out if there are such RP gene co-expression modules associated specifically with other tissues, we iteratively removed the samples in the biclusters after each iteration to create new data sets (the parameter

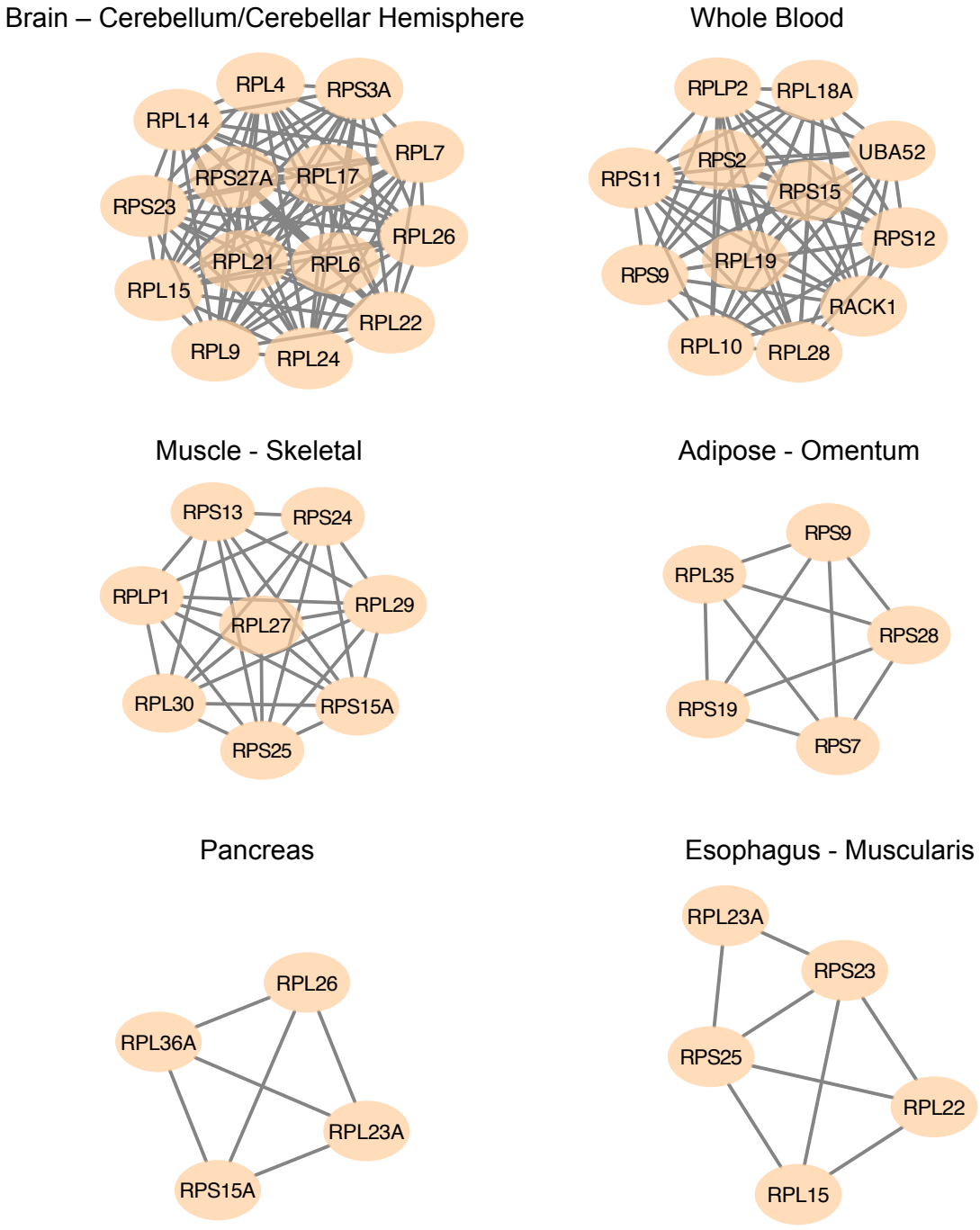

Figure 5.2: **Tissue specific ribosomal genes co-expression modules with high relative frequencies within given tissue types.**

choices were kept the same). We observed several more biclusters with RP gene modules enriched in specific tissue types, i.e., based on one-sided Fisher's exact tests we found these biclusters to be enriched in samples corresponding to certain tissue types. We show the seeds of some of these biclusters associated with specific tissue types in Fig. 5.2 .

5.3 Future application - DNA methylation data

As discussed in Chapter 1, DNA methylation plays a key role in regulating gene expression. The methylation profiles are frequently altered in tumors. While hypermethylation of CpG sites in the promoter regions of tumor suppressor genes leads to lower expression levels of the affected genes, hypomethylation of CpG sites in promoters of genes that are not usually expressed in the specific tissue type leads to aberrant expression of genes downstream. A good example of the latter case is that of genes of the Cancer-Testis family (which we briefly discussed in Chapter 3). These genes are usually expressed in the testis during early development. However, aberrant expression of these genes has been observed in multiple cancer types with demethylation (or hypomethylation) of CpGs in the promoter regions understood to be the underlying mechanism [118, 119].

TCGA has genome-wide methylation data available for a large number of tumor samples. For instance for breast cancer, there are 890 samples (including matched normals) whose methylation profiles were measures using the Illumina Human Methylation 450 platform which targets more than 450,000 methylation sites. The methylation data is made up of *beta-values* which are defined in terms of the intensity of signals from flourescent probes associated with methylated sites, with respect to the signal intensity of unmenthylated sites,

$$\beta = \frac{Intensity\ of\ methylated\ probe}{Total\ intensity}$$

where total intensity is the sum of the methylated and unmethylated probe intensities. Thus, ideally for a given probe the beta-value should be either 0 (unmethylated) or 1 (methylated). However, in reality it takes values between 0 and 1; closer to 0 when the given site is unmethylated, and closer to 1 when the site is methylated.

We can apply TuBA to this data in the same way that we apply it to gene expression data sets, except that the interpretation of the biclusters and their relation to gene expression would be quite different. The biggest hurdle to application of TuBA to methylation data is the sheer size of the data sets. The number of rows in these data sets is in excess of 480,000. Graphs based on gene pair associations for these many probes would invariably lead to unreasonably long computation times. We propose to first analyze these data sets in a chromosome-by-chromosome manner to identify altered CpG islands that may be associated with aberrant expression of some sets of genes.

Subsequent to such an analysis, the samples in the biclusters across chromosomes could be used to identify which CpG islands across different chromosomes show coordinated alterations. In summary, integration of methylation profile alterations with gene co-expression signatures found by TuBA can significantly enhance our understanding of the alterations in tumor cells that may be susceptible to therapeutic interventions.

Bibliography

[1] Amartya Singh, Gyan Bhanot, and Hossein Khiabanian. TuBA: Tunable biclustering algorithm reveals clinically relevant tumor transcriptional profiles in breast cancer. *GigaScience*, 8(6), 06 2019. ISSN 2047-217X. doi: 10.1093/gigascience/giz064. URL `https://doi.org/10.1093/gigascience/giz064`.

[2] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J.D. Watson. *Molecular Biology of the Cell*. Garland, 4th edition, 2002.

[3] Matt Ridley. *Genome*. Harper Perennial, 2006.

[4] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, Aug 1970.

[5] M. Esteller. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.*, 8(4):286–298, Apr 2007.

[6] M. Esteller. Epigenetics in cancer. *N. Engl. J. Med.*, 358(11):1148–1159, Mar 2008.

[7] Joseph F Costello and Christoph Plass. Methylation matters. *Journal of Medical Genetics*, 38(5):285–303, 2001. ISSN 0022-2593. doi: 10.1136/jmg.38.5.285. URL `https://jmg.bmj.com/content/38/5/285`.

[8] H. Cedar and Y. Bergman. Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.*, 10(5):295–304, May 2009.

[9] S. D. Briggs, T. Xiao, Z. W. Sun, J. A. Caldwell, J. Shabanowitz, D. F. Hunt, C. D. Allis, and B. D. Strahl. Gene silencing: trans-histone regulatory pathway in chromatin. *Nature*, 418 (6897):498, Aug 2002.

[10] B. Modrek and C. Lee. A genomic view of alternative splicing. *Nat. Genet.*, 30(1):13–19, Jan 2002.

[11] Nuno L. Barbosa-Morais, Manuel Irimia, Qun Pan, Hui Y. Xiong, Serge Gueroussov, Leo J. Lee, Valentina Slobodeniuc, Claudia Kutter, Stephen Watt, Recep Çolak, TaeHyung Kim, Christine M. Misquitta-Ali, Michael D. Wilson, Philip M. Kim, Duncan T. Odom, Brendan J. Frey, and Benjamin J. Blencowe. The evolutionary landscape of alternative splicing in

vertebrate species. *Science*, 338(6114):1587–1593, 2012. ISSN 0036-8075. doi: 10.1126/science.1230612. URL https://science.sciencemag.org/content/338/6114/1587.

[12] J. S. Mattick and I. V. Makunin. Non-coding RNA. *Hum. Mol. Genet.*, 15 Spec No 1:17–29, Apr 2006.

[13] C. C. Mello and D. Conte. Revealing the world of RNA interference. *Nature*, 431(7006): 338–342, Sep 2004.

[14] J. Krol, I. Loedige, and W. Filipowicz. The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.*, 11(9):597–610, Sep 2010.

[15] E. Barillot, L. Calzone, P. Hupé, J-P. Vert, and A. Zinovyev. *Computational Systems Biology of Cancer*. Chapman & Hall/CRC, 1st edition, 2012.

[16] B. Vogelstein and K. W. Kinzler. Cancer genes and the pathways they control. *Nat. Med.*, 10 (8):789–799, Aug 2004.

[17] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nat. Rev. Cancer*, 4(3):177–183, Mar 2004.

[18] James G. Herman and Stephen B. Baylin. Gene silencing in cancer in association with promoter hypermethylation. *New England Journal of Medicine*, 349(21):2042–2054, 2003. doi: 10.1056/NEJMra023075. URL https://doi.org/10.1056/NEJMra023075. PMID: 14627790.

[19] R. Mayor, L. Casadome, D. Azuara, V. Moreno, S. J. Clark, G. Capella, and M. A. Peinado. Long-range epigenetic silencing at 2q14.2 affects most human colorectal cancers and may have application as a non-invasive biomarker of disease. *Br. J. Cancer*, 100(10):1534–1539, May 2009.

[20] J. Jovanovic, J. A. R?nneberg, J. Tost, and V. Kristensen. The epigenetics of breast cancer. *Mol Oncol*, 4(3):242–254, Jun 2010.

[21] Roscoe Klinck, Anne Bramard, Lyna Inkel, Geneviève Dufresne-Martin, Julien Gervais-Bird, Richard Madden, Éric R. Paquet, ChuShin Koh, Julian P. Venables, Panagiotis Prinos, Manuela Jilaveanu-Pelmus, Raymund Wellinger, Claudine Rancourt, Benoit Chabot, and Sherif Abou Elela. Multiple alternative splicing markers for ovarian cancer. *Cancer Research*, 68(3):657–663, 2008. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-07-2580. URL http://cancerres.aacrjournals.org/content/68/3/657.

[22] Julian P. Venables. Aberrant and alternative splicing in cancer. *Cancer Research*, 64 (21):7647–7654, 2004. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-04-1910. URL http://cancerres.aacrjournals.org/content/64/21/7647.

[23] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan 2000.

[24] Douglas Hanahan and RobertA. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646 – 674, 2011. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2011.02.013. URL http://www.sciencedirect.com/science/article/pii/S0092867411001279.

[25] D. Pellman. Cell biology: aneuploidy and cancer. *Nature*, 446(7131):38–39, Mar 2007.

[26] H. Rajagopalan and C. Lengauer. Aneuploidy and cancer. *Nature*, 432(7015):338–341, Nov 2004.

[27] T. Boveri. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *J. Cell. Sci.*, 121 Suppl 1:1–84, Jan 2008.

[28] H. Satzinger, T. Boveri, and M. O. Boveri. Theodor and Marcella Boveri: chromosomes and cytoplasm in heredity and development. *Nat. Rev. Genet.*, 9(3):231–238, 03 2008.

[29] Sarah J Pfau and Angelika Amon. Chromosomal instability and aneuploidy in cancer: from yeast to man. *EMBO reports*, 13(6):515–527, 2012. doi: 10.1038/embor.2012.65. URL https://www.embopress.org/doi/abs/10.1038/embor.2012.65.

[30] Z. Storchova and D. Pellman. From polyploidy to aneuploidy, genome instability and cancer. *Nat. Rev. Mol. Cell Biol.*, 5(1):45–54, Jan 2004.

[31] K. Chandramouli and P. Y. Qian. Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Hum Genomics Proteomics*, 2009, Dec 2009.

[32] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.*, 19(3):1720–1730, Mar 1999.

[33] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, 36(16):e105, Sep 2008.

[34] Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, 12 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp692. URL https://doi.org/10.1093/bioinformatics/btp692.

[35] James H. Bullard, Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11(1):94, Feb 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-94. URL `https://doi.org/10.1186/1471-2105-11-94`.

[36] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010. doi: 10.1186/gb-2010-11-10-r106. URL `http://genomebiology.com/2010/11/10/R106/`.

[37] Mark D. Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, 11(3):R25, Mar 2010. ISSN 1474-760X. doi: 10.1186/gb-2010-11-3-r25. URL `https://doi.org/10.1186/gb-2010-11-3-r25`.

[38] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 95(25):14863–14868, Dec 1998.

[39] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, Feb 2000.

[40] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct 1999.

[41] C. M. Perou, S. S. Jeffrey, M. van de Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. Lee, D. Lashkari, D. Shalon, P. O. Brown, and D. Botstein. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. U.S.A.*, 96(16):9212–9217, Aug 1999.

[42] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform*, 1(1):24–45, 2004.

[43] B. Pontes, R. Giraldez, and J. S. Aguilar-Ruiz. Biclustering on expression data: A review. *J Biomed Inform*, 57:163–180, Oct 2015.

[44] J. A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972. doi: 10.1080/01621459.1972.10481214. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1972.10481214.

[45] G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. U.S.A.*, 97(22):12079–12084, Oct 2000.

[46] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, May 2006.

[47] G. Li, Q. Ma, H. Tang, A. H. Paterson, and Y. Xu. QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res.*, 37(15):e101, Aug 2009.

[48] R. A. Fisher. On the interpretation of ¡sup¿2¡/sup¿ from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922. ISSN 09528385. URL http://www.jstor.org/stable/2340521.

[49] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101.

[50] S. van Dam, U. Vosa, A. van der Graaf, L. Franke, and J. P. de Magalhaes. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinformatics*, 19(4): 575–592, 07 2018.

[51] Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger, editors, *Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations, held March 20–22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, and sponsored by the Office of Naval Research, Mathematics Program, IBM World Trade Corporation, and the IBM Research Mathematical Sciences Department*, pages 85–103. Springer US, Boston, MA, 1972. ISBN 978-1-4684-2001-2. doi: 10.1007/978-1-4684-2001-2_9. URL https://doi.org/10.1007/978-1-4684-2001-2_9.

[52] Coen Bron and Joep Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, September 1973. ISSN 0001-0782. doi: 10.1145/362342.362367. URL http://doi.acm.org/10.1145/362342.362367.

[53] F. Cazals and C. Karande. A note on the problem of reporting maximal cliques. *Theoretical Computer Science*, 407(1):564 – 568, 2008. ISSN 0304-3975. doi: https://doi.org/10.1016/j.tcs.2008.05.010. URL http://www.sciencedirect.com/science/article/pii/S0304397508003903.

[54] David Eppstein, Maarten Löffler, and Darren Strash. Listing all maximal cliques in sparse graphs in near-optimal time. *CoRR*, abs/1006.5440, 2010. URL http://arxiv.org/abs/1006.5440.

[55] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

[56] Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*, 2019. URL https://CRAN.R-project.org/package=data.table. R package version 1.12.2.

[57] Hadley Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011. URL http://www.jstatsoft.org/v40/i01/.

[58] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL http://igraph.org.

[59] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

[60] Terry M Therneau. *A Package for Survival Analysis in S*, 2015. URL https://CRAN.R-project.org/package=survival. version 2.38.

[61] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. ISBN 0-387-98784-3.

[62] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, Nov 2003.

[63] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, 68(1):7–30, 2018. doi: 10.3322/caac.21442. URL https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21442.

[64] C. M. Perou, T. S?rlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu,

P. E. L?nning, A. L. B?rresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, Aug 2000.

[65] T. S?rlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. L?nning, and A. L. B?rresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.*, 98(19): 10869–10874, Sep 2001.

[66] M. C. Cheang, S. K. Chia, D. Voduc, D. Gao, S. Leung, J. Snider, M. Watson, S. Davies, P. S. Bernard, J. S. Parker, C. M. Perou, M. J. Ellis, and T. O. Nielsen. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J. Natl. Cancer Inst.*, 101(10): 736–750, May 2009.

[67] S. Guiu, S. Michiels, F. Andre, J. Cortes, C. Denkert, A. Di Leo, B. T. Hennessy, T. Sorlie, C. Sotiriou, N. Turner, M. Van de Vijver, G. Viale, S. Loi, and J. S. Reis-Filho. Molecular subclasses of breast cancer: how do we define them? The IMPAKT 2012 Working Group Statement. *Ann. Oncol.*, 23(12):2997–3006, Dec 2012.

[68] B. Weigelt, A. Mackay, R. A'hern, R. Natrajan, D. S. Tan, M. Dowsett, A. Ashworth, and J. S. Reis-Filho. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol.*, 11(4):339–349, Apr 2010.

[69] A. Goldhirsch, W. C. Wood, A. S. Coates, R. D. Gelber, B. Thrlimann, H.-J. Senn, and Panel members. Strategies for subtypesdealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Annals of Oncology*, 22(8):1736–1747, 06 2011. ISSN 0923-7534. doi: 10.1093/annonc/mdr304. URL `https://doi.org/10.1093/annonc/mdr304`.

[70] D. C. Koboldt, R. S. Fulton, M. D. McLellan, H. Schmidt, J. Kalicki-Veizer, J. F. McMichael, L. L. Fulton, D. J. Dooling, L. Ding, E. R. Mardis, R. K. Wilson, A. Ally, M. Balasundaram, Y. S. Butterfield, R. Carlsen, C. Carter, A. Chu, E. Chuah, H. J. Chun, R. J. Coope, N. Dhalla, R. Guin, C. Hirst, M. Hirst, R. A. Holt, D. Lee, H. I. Li, M. Mayo, R. A. Moore, A. J. Mungall, E. Pleasance, A. Robertson, J. E. Schein, A. Shafiei, P. Sipahimalani, J. R. Slobodan, D. Stoll, A. Tam, N. Thiessen, R. J. Varhol, N. Wye, T. Zeng, Y. Zhao, I. Birol, S. J. Jones, M. A. Marra, A. D. Cherniack, G. Saksena, R. C. Onofrio, N. H. Pho, S. L. Carter, S. E. Schumacher, B. Tabak, B. Hernandez, J. Gentry, H. Nguyen, A. Crenshaw, K. Ardlie, R. Beroukhim, W. Winckler, G. Getz, S. B. Gabriel, M. Meyerson, L. Chin, P. J. Park, R. Kucherlapati,

K. A. Hoadley, J. Auman, C. Fan, Y. J. Turman, Y. Shi, L. Li, M. D. Topal, X. He, H. H. Chao, A. Prat, G. O. Silva, M. D. Iglesia, W. Zhao, J. Usary, J. S. Berg, M. Adams, J. Booker, J. Wu, A. Gulabani, T. Bodenheimer, A. P. Hoyle, J. V. Simons, M. G. Soloway, L. E. Mose, S. R. Jefferys, S. Balu, J. S. Parker, D. Hayes, C. M. Perou, S. Malik, S. Mahurkar, H. Shen, D. J. Weisenberger, T. Triche, P. H. Lai, M. S. Bootwalla, D. T. Maglinte, B. P. Berman, D. J. Van Den Berg, S. B. Baylin, P. W. Laird, C. J. Creighton, L. A. Donehower, G. Getz, M. Noble, D. Voet, G. Saksena, N. Gehlenborg, D. DiCara, J. Zhang, H. Zhang, C. J. Wu, S. Y. Liu, M. S. Lawrence, L. Zou, A. Sivachenko, P. Lin, P. Stojanov, R. Jing, J. Cho, R. Sinha, R. W. Park, M. D. Nazaire, J. Robinson, H. Thorvaldsdottir, J. Mesirov, P. J. Park, L. Chin, S. Reynolds, R. B. Kreisberg, B. Bernard, R. Bressler, T. Erkkila, J. Lin, V. Thorsson, W. Zhang, I. Shmulevich, G. Ciriello, N. Weinhold, N. Schultz, J. Gao, E. Cerami, B. Gross, A. Jacobsen, R. Sinha, B. Aksoy, Y. Antipin, B. Reva, R. Shen, B. S. Taylor, M. Ladanyi, C. Sander, P. Anur, P. T. Spellman, Y. Lu, W. Liu, R. R. Verhaak, G. B. Mills, R. Akbani, N. Zhang, B. M. Broom, T. D. Casasent, C. Wakefield, A. K. Unruh, K. Baggerly, K. Coombes, J. N. Weinstein, D. Haussler, C. C. Benz, J. M. Stuart, S. C. Benz, J. Zhu, C. C. Szeto, G. K. Scott, C. Yau, E. O. Paull, D. Carlin, C. Wong, A. Sokolov, J. Thusberg, S. Mooney, S. Ng, T. C. Goldstein, K. Ellrott, M. Grifford, C. Wilks, S. Ma, B. Craft, C. Yan, Y. Hu, D. Meerzaman, J. M. Gastier-Foster, J. Bowen, N. C. Ramirez, A. D. Black, R. E. Pyatt, P. White, E. J. Zmuda, J. Frick, T. M. Lichtenberg, R. Brookens, M. M. George, M. A. Gerken, H. A. Harper, K. M. Leraas, L. J. Wise, T. R. Tabler, C. McAllister, T. Barr, M. Hart-Kothari, K. Tarvin, C. Saller, G. Sandusky, C. Mitchell, M. V. Iacocca, J. Brown, B. Rabeno, C. Czerwinski, N. Petrelli, O. Dolzhansky, M. Abramov, O. Voronina, O. Potapova, J. R. Marks, W. M. Suchorska, D. Murawa, W. Kycler, M. Ibbs, K. Korski, A. Spycha?a, P. Murawa, J. J. Brzezi?ski, H. Perz, R. ?a?niak, M. Teresiak, H. Tatka, E. Leporowska, M. Bogusz-Czerniewicz, J. Malicki, A. Mackiewicz, M. Wiznerowicz, X. V. Le, B. Kohl, V. T. Nguyen, R. Thorp, V. B. Nguyen, H. Sussman, D. P. Bui, R. Hajek, P. H. Nguyen, V. T. Tran, Q. T. Huynh, K. Z. Khan, R. Penny, D. Mallery, E. Curley, C. Shelton, P. Yena, J. N. Ingle, F. J. Couch, W. L. Lingle, T. A. King, A. M. Gonzalez-Angulo, G. B. Mills, M. D. Dyer, S. Liu, X. Meng, M. Patangan, F. Waldman, H. Stoppler, W. Rathmell, L. Thorne, M. Huang, L. Boice, A. Hill, C. Morrison, C. Gaudioso, W. Bshara, K. Daily, S. C. Egea, M. Pegram, C. Gomez-Fernandez, R. Dhir, R. Bhargava, A. Brufsky, C. D. Shriver, J. A. Hooke, J. L. Campbell, R. J. Mural, H. Hu, S. Somiari, C. Larson, B. Deyarmin, L. Kvecher, A. J. Kovatich, M. J. Ellis, T. A. King, H. Hu, F. J. Couch, R. J. Mural, T. Stricker, K. White, O. Olopade, J. N. Ingle, C. Luo, Y. Chen, J. R. Marks, F. Waldman, M. Wiznerowicz, R. Bose, L. W. Chang, A. H. Beck,

A. M. Gonzalez-Angulo, T. Pihl, M. Jensen, R. Sfeir, A. Kahn, A. Chu, P. Kothiyal, Z. Wang, E. Snyder, J. Pontius, B. Ayala, M. Backus, J. Walton, J. Baboud, D. Berton, M. Nicholls, D. Srinivasan, R. Raman, S. Girshik, P. Kigonya, S. Alonso, R. Sanbhadti, S. Barletta, D. Pot, M. Sheth, J. A. Demchok, K. R. Shaw, L. Yang, G. Eley, M. L. Ferguson, R. W. Tarnuzzer, J. Zhang, L. A. Dillon, K. Buetow, P. Fielding, B. A. Ozenberger, M. S. Guyer, H. J. Sofia, and J. D. Palchik. Comprehensive molecular portraits of human breast tumours. *Nature*, 490 (7418):61–70, Oct 2012.

[71] C. Curtis, S. P. Shah, S. F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Graf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, A. Langer?d, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A. L. B?rresen-Dale, J. D. Brenton, S. Tavare, C. Caldas, S. Aparicio, C. Caldas, S. Aparicio, C. Curtis, S. P. Shah, C. Caldas, S. Aparicio, J. D. Brenton, I. Ellis, D. Huntsman, S. Pinder, A. Purushotham, L. Murphy, C. Caldas, S. Aparicio, C. Caldas, H. Bardwell, S. F. Chin, C. Curtis, Z. Ding, S. Graf, L. Jones, B. Liu, A. G. Lynch, I. Papatheodorou, S. J. Sammut, G. Wishart, S. Aparicio, S. Chia, K. Gelmon, D. Huntsman, S. McKinney, C. Speers, G. Turashvili, P. Watson, I. Ellis, R. Blamey, A. Green, D. Macmillan, E. Rakha, A. Purushotham, C. Gillett, A. Grigoriadis, S. Pinder, E. de Rinaldis, A. Tutt, L. Murphy, M. Parisien, S. Troup, C. Caldas, S. F. Chin, D. Chan, C. Fielding, A. T. Maia, S. McGuire, M. Osborne, S. M. Sayalero, I. Spiteri, J. Hadfield, S. Aparicio, G. Turashvili, L. Bell, K. Chow, N. Gale, D. Huntsman, M. Kovalik, Y. Ng, L. Prentice, C. Caldas, S. Tavare, C. Curtis, M. J. Dunning, S. Graf, A. G. Lynch, O. M. Rueda, R. Russell, S. Samarajiwa, D. Speed, F. Markowetz, Y. Yuan, J. D. Brenton, S. Aparicio, S. P. Shah, A. Bashashati, G. Ha, G. Haffari, and S. McKinney. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, Apr 2012.

[72] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, 2(5):401–404, May 2012.

[73] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*, 6(269):pl1, Apr 2013.

[74] B. Gyorffy and R. Schafer. Meta-analysis of gene expression profiles related to relapse-free survival in 1,079 breast cancer patients. *Breast Cancer Res. Treat.*, 118(3):433–441, Dec 2009.

[75] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 18(1):50–60, 03 1947. doi: 10.1214/aoms/1177730491. URL `https://doi.org/10.1214/aoms/1177730491`.

[76] S. Marguerat and J. Bahler. RNA-seq: from technology to biology. *Cell. Mol. Life Sci.*, 67 (4):569–579, Feb 2010.

[77] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7):621–628, Jul 2008.

[78] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, 27(8):1160–1167, Mar 2009.

[79] R.A. Fisher. *Statistical methods for research workers.* Edinburgh Oliver & Boyd, 1925.

[80] A. Kallioniemi, O. P. Kallioniemi, J. Piper, M. Tanner, T. Stokke, L. Chen, H. S. Smith, D. Pinkel, J. W. Gray, and F. M. Waldman. Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proc. Natl. Acad. Sci. U.S.A.*, 91(6):2156–2160, Mar 1994.

[81] J. Kao, K. Salari, M. Bocanegra, Y. L. Choi, L. Girard, J. Gandhi, K. A. Kwei, T. Hernandez-Boussard, P. Wang, A. F. Gazdar, J. D. Minna, and J. R. Pollack. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLoS ONE*, 4(7):e6146, Jul 2009.

[82] G. Curigliano, G. Viale, M. Ghioni, A. A. Jungbluth, V. Bagnardi, G. C. Spagnoli, A. M. Neville, F. Nole, N. Rotmensz, and A. Goldhirsch. Cancer-testis antigen expression in triple-negative breast cancer. *Ann. Oncol.*, 22(1):98–103, Jan 2011.

[83] A. J. Simpson, O. L. Caballero, A. Jungbluth, Y. T. Chen, and L. J. Old. Cancer/testis antigens, gametogenesis and cancer. *Nat. Rev. Cancer*, 5(8):615–625, Aug 2005.

[84] S. Subhash and C. Kanduri. GeneSCF: a real-time based functional enrichment tool with support for multiple organisms. *BMC Bioinformatics*, 17(1):365, Sep 2016.

[85] K. Yoshihara, M. Shahmoradgoli, E. Martinez, R. Vegesna, H. Kim, W. Torres-Garcia, V. Trevino, H. Shen, P. W. Laird, D. A. Levine, S. L. Carter, G. Getz, K. Stemke-Hale, G. B.

Mills, and R. G. Verhaak. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun*, 4:2612, 2013.

[86] J. M. Bland and D. G. Altman. The logrank test. *BMJ*, 328(7447):1073, May 2004.

[87] E. Bilal, K. Vassallo, D. Toppmeyer, N. Barnard, I. H. Rye, V. Almendro, H. Russnes, A. L. B?rresen-Dale, A. J. Levine, G. Bhanot, and S. Ganesan. Amplified loci on chromosomes 8 and 17 predict early relapse in ER-positive breast cancers. *PLoS ONE*, 7(6):e38575, 2012.

[88] J. Wang, Q. Liu, and Y. Shyr. Dysregulated transcription across diverse cancer types reveals the importance of RNA-binding protein in carcinogenesis. *BMC Genomics*, 16 Suppl 7:S5, 2015.

[89] J. L. Haybittle, R. W. Blamey, C. W. Elston, J. Johnson, P. J. Doyle, F. C. Campbell, R. I. Nicholson, and K. Griffiths. A prognostic index in primary breast cancer. *Br. J. Cancer*, 45 (3):361–366, Mar 1982.

[90] M. H. Galea, R. W. Blamey, C. E. Elston, and I. O. Ellis. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res. Treat.*, 22(3):207–219, 1992.

[91] D. Mu, L. Chen, X. Zhang, L. H. See, C. M. Koch, C. Yen, J. J. Tong, L. Spiegel, K. C. Nguyen, A. Servoss, Y. Peng, L. Pei, J. R. Marks, S. Lowe, T. Hoey, L. Y. Jan, W. R. McCombie, M. H. Wigler, and S. Powers. Genomic amplification and oncogenic properties of the KCNK9 potassium channel gene. *Cancer Cell*, 3(3):297–302, Mar 2003.

[92] T. I. Lee and R. A. Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251, Mar 2013.

[93] K. M. Lelli, M. Slattery, and R. S. Mann. Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.*, 46:43–68, 2012.

[94] L. Wan and Y. Kang. Pleiotropic roles of AEG-1/MTDH/LYRIC in breast cancer. *Adv. Cancer Res.*, 120:113–134, 2013.

[95] Z. Song, Y. Wang, C. Li, D. Zhang, and X. Wang. Molecular Modification of Metadherin/MTDH Impacts the Sensitivity of Breast Cancer to Doxorubicin. *PLoS ONE*, 10(5): e0127599, 2015.

[96] X. Shi and X. Wang. The role of MTDH/AEG-1 in the progression of cancer. *Int J Clin Exp Med*, 8(4):4795–4807, 2015.

[97] K. Eren, M. Deveci, O. Kucuktunc, and U. V. Catalyurek. A comparative analysis of biclustering algorithms for gene expression data. *Brief. Bioinformatics*, 14(3):279–292, May 2013.

[98] A. Serin and M. Vingron. DeBi: Discovering Differentially Expressed Biclusters using a Frequent Itemset Approach. *Algorithms Mol Biol*, 6(1):18, Jun 2011.

[99] A. Rosenwald, G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland, J. M. Giltnane, E. M. Hurt, H. Zhao, L. Averett, L. Yang, W. H. Wilson, E. S. Jaffe, R. Simon, R. D. Klausner, J. Powell, P. L. Duffey, D. L. Longo, T. C. Greiner, D. D. Weisenburger, W. G. Sanger, B. J. Dave, J. C. Lynch, J. Vose, J. O. Armitage, E. Montserrat, A. Lopez-Guillermo, T. M. Grogan, T. P. Miller, M. LeBlanc, G. Ott, S. Kvaloy, J. Delabie, H. Holte, P. Krajci, T. Stokke, and L. M. Staudt. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.*, 346(25):1937–1947, Jun 2002.

[100] S. Bergmann, J. Ihmels, and N. Barkai. Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys Rev E Stat Nonlin Soft Matter Phys*, 67(3 Pt 1):031902, Mar 2003.

[101] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini. Discovering local structure in gene expression data: the order-preserving submatrix problem. *J. Comput. Biol.*, 10(3-4):373–384, 2003.

[102] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. U.S.A.*, 101(9):2981–2986, Mar 2004.

[103] Sebastian Kaiser, Rodrigo Santamaria, Tatsiana Khamiakova, Martin Sill, Roberto Theron, Luis Quintales, Friedrich Leisch, and Ewoud De Troyer. *biclust: BiCluster Algorithms*, 2018. URL `https://CRAN.R-project.org/package=biclust`. R package version 2.0.1.

[104] Sven Bergmann, Jan Ihmels, and Naama Barkai. Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys Rev E Nonlin Soft Matter Phys*, page 031902, 2003.

[105] Yu Zhang, Juan Xie, Jinyu Yang, Anne Fennell, Chi Zhang, and Qin Ma. QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data. *Bioinformatics*, 33(3):450–452, 2017. doi: 10.1093/bioinformatics/btw635.

[106] R. Shamir, A. Maron-Katz, A. Tanay, C. Linhart, I. Steinfeld, R. Sharan, Y. Shiloh, and R. Elkon. EXPANDER–an integrative program suite for microarray data analysis. *BMC Bioinformatics*, 6:232, Sep 2005.

[107] C. Gao, I. C. McDowell, S. Zhao, C. D. Brown, and B. E. Engelhardt. Context Specific and Differential Gene Co-expression Networks via Bayesian Biclustering. *PLoS Comput. Biol.*, 12 (7):e1004791, 07 2016.

[108] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, Jan 2002.

[109] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, 347(25):1999–2009, Dec 2002.

[110] Markus Schroeder, Benjamin Haibe-Kains, Aedin Culhane, Christos Sotiriou, Gianluca Bontempi, and John Quackenbush. *breastCancerNKI: Genexpression dataset published by van't Veer et al. [2002] and van de Vijver et al. [2002] (NKI).*, 2018. URL `http://compbio.dfci.harvard.edu/`. R package version 1.20.0.

[111] Raghu Kalluri and Robert A. Weinberg. The basics of epithelial-mesenchymal transition. *The Journal of Clinical Investigation*, 119(6):1420–1428, 6 2009. doi: 10.1172/JCI39104. URL `https://doi.org/10.1172/JCI39104`.

[112] L. Wang, A. Saci, P. M. Szabo, S. D. Chasalow, M. Castillo-Martin, J. Domingo-Domenech, A. Siefker-Radtke, P. Sharma, J. P. Sfakianos, Y. Gong, A. Dominguez-Andres, W. K. Oh, D. Mulholland, A. Azrilevich, L. Hu, C. Cordon-Cardo, H. Salmon, N. Bhardwaj, J. Zhu, and M. D. Galsky. EMT- and stroma-related gene expression and resistance to PD-1 blockade in urothelial cancer. *Nat Commun*, 9(1):3503, 08 2018.

[113] J. Hedegaard, P. Lamy, I. Nordentoft, F. Algaba, S. H?yer, B. P. Ulh?i, S. Vang, T. Reinert, G. G. Hermann, K. Mogensen, M. B. H. Thomsen, M. M. Nielsen, M. Marquez, U. Segersten, M. Aine, M. Hoglund, K. Birkenkamp-Demtroder, N. Fristrup, M. Borre, A. Hartmann, R. Stohr, S. Wach, B. Keck, A. K. Seitz, R. Nawroth, T. Maurer, C. Tulic, T. Simic, K. Junker,

M. Horstmann, N. Harving, A. C. Petersen, M. L. Calle, E. W. Steyerberg, W. Beukers, K. E. M. van Kessel, J. B. Jensen, J. S. Pedersen, P. U. Malmstrom, N. Malats, F. X. Real, E. C. Zwarthoff, T. F. ?rntoft, and L. Dyrskj?t. Comprehensive Transcriptional Analysis of Early-Stage Urothelial Carcinoma. *Cancer Cell*, 30(1):27–42, 07 2016.

[114] M. D. Buschman, P. A. Bromann, P. Cejudo-Martin, F. Wen, I. Pass, and S. A. Courtneidge. The novel adaptor protein Tks4 (SH3PXD2B) is required for functional podosome formation. *Mol. Biol. Cell*, 20(5):1302–1311, Mar 2009.

[115] Z. Shi, K. Fujii, K. M. Kovary, N. R. Genuth, H. L. Rost, M. N. Teruel, and M. Barna. Heterogeneous Ribosomes Preferentially Translate Distinct Subpools of mRNAs Genome-wide. *Mol. Cell*, 67(1):71–83, Jul 2017.

[116] N. R. Genuth and M. Barna. Heterogeneity and specialized functions of translation machinery: from genes to organisms. *Nat. Rev. Genet.*, 19(7):431–452, Jul 2018.

[117] N. R. Genuth and M. Barna. The Discovery of Ribosome Heterogeneity and Its Implications for Gene Regulation and Organismal Life. *Mol. Cell*, 71(3):364–374, 08 2018.

[118] R. Kim, P. Kulkarni, and S. Hannenhalli. Derepression of Cancer/testis antigens in cancer is associated with distinct patterns of DNA hypomethylation. *BMC Cancer*, 13:144, Mar 2013.

[119] W. Zhang, C. J. Barger, P. A. Link, P. Mhawech-Fauceglia, A. Miller, S. N. Akers, K. Odunsi, and A. R. Karpf. DNA hypomethylation-mediated activation of Cancer/Testis Antigen 45 (CT45) genes is associated with disease progression and reduced survival in epithelial ovarian cancer. *Epigenetics*, 10(8):736–748, 2015.