Modeling Cancer Progression on the Pathway Level

by

Elena Edelman

Computational Biology & Bioinformatics Program
Duke University

Date: ____________________________
Approved:

____
Sayan Mukherjee, Supervisor

____
Lindsay Cowell

____
Phillip Febbo

____
Jonathan Mattingly

____
Joseph Nevins

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Computational Biology & Bioinformatics Program
in the Graduate School of
Duke University

2008

<u>ABSTRACT</u>

(Computational Biology & Bioinformatics)

Modeling Cancer Progression on the Pathway Level

by

Elena Edelman

Computational Biology & Bioinformatics Program
Duke University

Date: ______________________

Approved:

____
Sayan Mukherjee, Supervisor

____
Lindsay Cowell

____
Phillip Febbo

____
Jonathan Mattingly

____
Joseph Nevins

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Computational Biology & Bioinformatics Program
in the Graduate School of
Duke University

2008

Abstract

Over the past several decades, many genes have been discovered that govern important functions in the development of a variety of different cancers. However, biological insight from the list of genes is still limited and the underlying mechanisms that occur in the cell during tumorigenesis have not been well established. Studying cancer progression in terms of the oncogenic pathways that are responsible for specific actions that change normal cells into tumors is a means for bringing insight onto these issues. The work presented here will uncover mechanisms that are occurring at the pathway level that first initiate tumor formation and then continue through cancer progression and finally metastasis. This knowledge will allow for drug treatment that is better targeted towards an individual.

Microarray technology has allowed for the collection of gene expression datasets from clinical cancer and other studies. These datasets can be used to study how expression levels of individual genes or groups of related genes are altered in individuals from different phenotypic groups. Statistical methods exist which assay pathway enrichment by phenotypic class but do not describe individual variation. In order to study this individual variation, we developed a formal statistical method called ASSESS which measures the enrichment of a gene set in each sample in an expression dataset.

As cancer advances through the stages of initiation, progression, and proliferation, multiple pathways experience disruptions at various times. However, there is still much unknown on these particular pathways that evidence gene expression changes

throughout tumorigenesis. Using gene expression datasets comprised of individuals with tumors classified by location and stage, we applied ASSESS in order to study the data on the pathway level. We then utilized novel statistical methods to uncover the pathways that play a role in cancer progression and in what order the pathways become perturbed.

These analyses can give a basis for how genetic disruptions serve to alter actions in specific cell types. The results may provide insight that will lead to treatments of existing tumors and prevention of incipient cancers from forming. Treatments for existing tumors will use multiple drugs to target the pathways that show an altered state of activity.

Contents

List of Tables

List of Figures

Acknowledgements

I would first like to thank my advisor Sayan Mukherjee. I could go on and on about what a great teacher and mentor Sayan has been but since I'm thanking him it would probably be best not to annoy at the same time. Therefore, I will just say thank you for everything that you have taught me and all the opportunities you have provided me.

One of the reasons I chose to come to Duke was its emphasis on collaboration. Through my time here I have had the opportunity to work with faculty from numerous departments at the university. I would like to thank all of these people who I have worked with during my time at Duke. First to my committee members who have continually provided me with advice and guidance. Both Joe Nevins and Phil Febbo have been incredibly helpful in teaching me cancer biology. Without them I am sure I would not have the biological expertise that I have today. I would like to thank my committee members Jonathan Mattingly and Lindsay Cowell who have always had an open door whenever I needed assistance. I would also like to thank Anil Potti and Ashley Chi who have provided me with interesting biological problems to investigate and again helped me to increase my knowledge in cancer biology.

I would like to thank my fellow lab members as well as the lab members of the Ohler and Furey lab for their assistance in day-to-day problems. I would especially like to thank Justin Guinney for all of his contributions in the many projects we worked on together. The other students of the Computational Biology and Bioinformatics program have made my time at Duke more enjoyable than I thought graduate school could ever be. Thank you for your friendships and encouragement.

Thank you to my parents who have pushed me to be my best and supported me through every challenge I've taken on. Finally, thank you to my four legged family member for her wonderful ability to reduce stress.

Chapter 1

Introduction

Cancer is a complex disease defined by an extraordinary amount of heterogeneity among genes involved in tumorigenesis. The complexity is also observed by the large phenotypic variation that can occur within disease states. This genotypic and phenotypic variation makes understanding the processes driving tumorigenesis extremely challenging. One way of addressing these challenges is by studying microarray gene expression data. Gene expression datasets provide information on the expression level of genes over the entire genome which can be collected for numerous individuals of various disease states thus capturing both genotypic and phenotypic variation.

Much work has been done with microarray gene expression datasets to find genes that show significant expression differences between cancer phenotypes and disease-free individuals. However, there are many drawbacks to using individual gene analysis and it is the aim of this work to overcome these drawbacks and enhance our knowledge of cancer progression. The main contribution of this work is the development of novel methods for the understanding of the mechanisms driving tumorigenesis. This understanding is arrived at through the move from single gene analysis to pathway analysis when studying the gene expression of individual samples. In order for this to be accomplished, new methodologies must be developed which start with biological data in the form of pathway gene sets. The implementation of a methodology with the ability to assay pathway variation in individual samples will allow for the dissection of tumor samples in terms of oncogenic pathways. Mathematical models will allow for pathway delineation and will give a basis for how genetic disruptions serve to alter actions in specific cell types. Computational methods that allow for detection of the stage in cancer progression can lead to treatments of existing tumors and prevention of incipient cancers from forming. Treatments for existing tumors will use multiple drugs to target the pathways that show an altered state of activity. Additionally, these therapies can be better directed toward an individual.

1.1 Pathway Analysis

We are now at a point where gene expression profiles can easily be generated. The challenge arises in applying a method of interpretation that gleans insight into biological mechanisms at play. Pathway based methods of interpreting gene expression datasets have demonstrated advances over single gene based methods in the following ways:

1. When analyzing a gene expression dataset, there can be an overwhelming number of differentially expressed genes across sample classifications. It is hard to know which genes are relevant in class distinction, which are just caused by noise, and how to go about relating the importance of the genes to one another.

Studying gene expression on the pathway level is more robust and interpretable. Finding a set of genes which all follow some pattern of deregulation gives more confidence that these changes are not caused by noise but by some mechanism that acts on all the genes in this set [126].

2. Expression differences in the biologically significant genes may be modest and hidden by differences due to noise alone. Therefore, no genes would be found to be differentially expressed. If there is a modest decrease of the expression levels of all the genes in a certain pathway, no single gene will appear to be very significant but at the pathway level one can see that a major alteration has occurred [100].

3. There can be a lack of consensus and reproducibility in microarray experiments at the single gene level. For example, three independent research groups performed the same experiment to discover the gene signature for the functional characterization of stem cells. Only one gene was found in common among the three signatures. This demonstrates how hard it is to achieve consistency on the single gene level. Studying a group of genes is a more robust approach, as small changes to individual genes due to experimental error or noise will not have as large of an impact on the results [47].

When looking at genetic differences at the single gene level, it is often hard to understand the underlying biology that is causing the phenotypic changes. Using pathway approaches have been successful in cases where single gene methods have not. Pathway methods have identified the oxidative phosphorylation pathway as being down regulated in human diabetic muscle [100]. In this example, no single gene in the pathway was found to be differentially expressed between diabetic and normal individuals. Each gene showed only a modest decreased expression in diabetic muscle, however, taken as a group, these genes showed a significant pattern.

Additionally, pathway approaches have been used to identify MYC, RAS, and RB pathway deregulation within murine tumor models [71, 14] and KRAS deregulation in lung adenocarcinoma [133, 13]. In these studies, genetic signatures were developed which were then used to identify the activation status of the oncogenic pathways in datasets of interest.

A final example demonstrating the importance of pathway analysis comes from a study by Sjoblom et al [127]. A global analysis of 13,023 genes found many more genes to be mutated in two human tumor types than previously expected. The analysis of 11 breast and 11 colon cancer samples revealed 1,149 mutated genes, 189 of which were determined to be candidate cancer genes (CAN genes). 122 of these CAN genes were identified in breast cancer and 69 in colon cancer. This shows how little consistency there is between tumor types as breast and colon cancer samples only shared 2 out of 189 CAN genes in common. There was also much discrepancy between samples of the same cancer type. Each sample of a given tumor type had no more than six mutated CAN genes in common [127]. This recapitulates the importance of pathway analysis over single gene analysis in the study of cancer. This data suggests that there may be deregulation of specific processes that are necessary for neoplasia. Each process has many potential member genes and the alteration of any of these genes may produce the same phenotypic result. Therefore, understanding how genes implicated in cancer form functional groups will allow for insight into the mechanisms at play during neoplasia.

1.2 Methods of Pathway Analysis

Many statistical methods have been developed to analyze gene expression in terms of gene sets. In early pathway analysis, straightforward methods such as t-tests were used to determine genes which were differentially expressed between classes. These lists of genes were interpreted by looking for patterns and underlying structure. Databases such as Gene Ontology can be used to relate the individual genes in terms of their general function within the cell [128]. More specifically, gene sets and pathways defined by Gene Ontology and other databases were compared with the list of genes to find enrichment of biological themes in the list. One problem associated with such a method is determining a cutoff threshold value which is usually arbitrarily defined [108].

Several pathway analysis programs were developed to take these simple annotation methods a step further by starting with biological information and performing pathway analysis on gene expression data [132, 8, 80]. These pathway methods generally use predefined gene sets to annotate an expression dataset with samples from different classes and use cut-off free testing of the gene expression. Gene sets can be defined structurally or functionally, for instance chromosomal regions or gene regulatory pathways. Gene Set Enrichment Analysis (GSEA) is one such method which requires an expression profile with samples labeled from one of two classes (for instance normal versus tumor samples) and a database of a priori defined gene sets. The outputs are an enrichment score for each class in each gene set and an assessment of confidence on each enrichment score [132].

Numerous other gene set analysis programs have been developed which test different null hypotheses using a variety of statistical methods. Catmap [19], PAGE [80], T-profiler [17], Erminej [86], and Q1 [136] are methods aimed to test the hypothesis that a gene set has the same level of association with a phenotype as the complement of the gene set. Globaltest [56], SAFE [8], Q2 [136], PLAGE [137], the multivariate approach of Kong et al. [83], and SAM-GS [34] hypothesize that there is no gene in the gene set associated with phenotype.

A further step in pathway analysis is being taken with the development of Analysis of Sample Set Enrichment Scores (ASSESS). ASSESS tests the first hypothesis but is a novel statistical method in that it allows for the computation of enrichment scores and significance of individual samples. The major innovation of ASSESS over GSEA is the ability to calculate enrichment of each individual rather than just phenotypic groups [36].

There are also experimental approaches that will give enrichment of individual samples. One approach is to make models on cell lines with specific perturbations to discover the composition of a certain pathway [13]. These models can then be used to assay other tumor cells. In some circumstances, the gene sets made by experimental approaches can be more reliable than gene sets from other sources, for example when the gene set is defined in the same or similar context as the dataset being analyzed. Literature based gene pathways lack the genetic context that experimental pathways provide. However, the number of pathways arrived through experimental procedures is limited. The advantage of ASSESS is its ability to find enrichment of many more pathways, from a variety of platforms and sources, including the ones that are experimentally derived. Clearly, experimental approaches and ASSESS are complementary methods.

1.3 Cancer Genes and Their Pathway Interactions

Cancer is a complex genetic disease requiring alterations in multiple genes throughout its course of progression. It is important to understand some of the key genes involved in cancer progression as they are involved in many of the pathways that will be discussed later on. This section will discuss the important genes involved in cancer initiation, progression, and finally metastasis and cases in which this information has helped build a basis for a mechanistic understanding of tumorigenesis.

There are two classes of genes that are important in cancer-susceptibility [99]:

1. Oncogenes: A mutation in one allele is sufficient to confer a selective advantage on the cell.

2. Tumor-suppressor genes (TSGs): Mutations in one or both alleles causes a reduction in gene product thereby enhancing cell growth.

A quarter century of research has greatly enhanced our knowledge of oncogenes and TSGs. The focus now has moved to understanding the complex interactions of these genes in terms of the mechanisms that underly cancer initiation, progression, and eventually metastasis. Mathematical modeling must be used to first understand how the gene signaling pathways that exist within the cell make up an integrated circuit and then explain how alterations in certain genes and pathways alter the rest of the circuit. Such modeling can show how pathways effect phenotypes. A number of questions will emerge from this investigation. How many pathways must undergo mutation to change a normal cell into a cancer cell? Do perturbations always occur in the same or similar pathways in different tumors types? Can the genes that have been discovered to be important in tumorigenesis be tied together by a small number of signaling pathways [61]?

Cancer Initiation and Early Progression

One discovery that begins to give insight into these questions was the finding that almost all DNA tumor viruses inactivate both RB and P53. It may be possible that the P53 and RB pathways must be inactivated in order for a tumor to form. Studies have also shown that these pathways are disturbed in a large portion of cancer types [142]. Perhaps the cancers that do not show alterations in these pathways will eventually be shown to contain them as we learn more on the true make-up of all the genes in the pathways.

Although some information on important genes and pathways in the process of tumorigenesis is beginning to be understood, much remains elusive. For instance we know that cancer is initiated by mutations in RB1, APC, and NFI in the eye, colon, and nervous system, respectively. However, in most tumors the pathway, or potentially pathways, responsible for tumor initiation is not known [142].

Genetic Instability

After initiation, tumorigenesis can take many different courses. More aggressive tumors usually have affected stability genes. Stability genes are tumor suppressor genes or oncogenes which function in minimizing genetic mistakes due to DNA replication or exposure to mutagens. Mutations in these genes lead to genetic instability and the accumulation of genetic changes in other genes that directly control the cell cycle. There is also a large amount of uncertainty concerning genetic instability. Research continues to support the importance of genetic instability in neoplasia formation [33, 9]. Genetic instability results in an accumulation of DNA damage and loss of heterozygosity. Results of genetic instability can be studied on the pathway level. Normally the DNA damage sends signals to the DNA-damage checkpoint response pathway whose activation triggers senescence. However, if there is significant genetic instability, this will often result in ineffective DNA-damage controls. Senescence is blocked and proliferation and transformation results [33]. Further investigations into the expected hundreds of chromosomal instability genes and mechanisms underlying genetic instability will be central in understanding tumorigenesis and developing new therapeutic agents.

Metastasis

The final stage of tumorigenesis is metastasis. Although metastasis is the cause of 90% of cancer deaths, the mechanisms leading to the event remain largely unknown. Recent advances have begun to make headway into certain mechanisms responsible for metastasis; however, the regulatory pathways involved are unclear. The genes and pathways discussed above become altered early in tumorigenesis and have not been associated with the metastatic process. Metastasis is associated with a number of distinct steps, most prominently the promotion of cell growth and survival, angiogenesis, intravasation, extravasation, and colonization. The mechanisms that drive these processes are complex and as a result, metastatic cells show a great deal of heterogeneity [45]. Analyses are needed for a systematic understanding of the key metastatic pathways and whether there is a dependence on tissue type.

As in cancer initiation and progression, the order of mutations does not seem to be as important as the overall acquisition to promote metastasis. One trait that appears necessary for metastasis is an aggressive phenotype, meaning that cancer cells can survive in hostile and nutrient deprived environments. Tumor cells that have an aggressive phenotype are better suited to overcome factors in the tumor microenvironment that limit tumor progression, for example limited availability of oxygen. In low oxygen conditions, cells increase the transcription of hypoxia inducible factor-1 (HIF-1) which activates the genes involved in angiogenesis, anaerobic metabolism, cell survival, and invasion. Tumors with increased HIF-1 stabilization have a greater likelihood of metastasizing. [60].

One example of how metastasis may occur is that once cells acquire increased motility in primary tumors they naturally become attracted to blood vessels by chemoattractive gradients and extracellular tracks emanating to the vessels. After having invaded and circulated in the vasculature, metastatic cells must escape and invade the target tissue. Again, it is unclear exactly how this occurs. One thought is that VEGF, a potent vascular permeability factor, induces signals which changes the vascular permeability of blood vessels in target organs. VEGF may activate SRC family kinases in endothelial cells which disrupt cell junctions, thereby facilitating metastatic extravasation [60].

Considerable research must still be done in the area of metastatic dissemination to yield insights into mechanisms for motility, intravasation, and extravasation, as well as steps not discussed in detail here such as transit and patterns of colonization. A more complete understanding of the molecular mechanisms that control metastasis and how expression of these molecules become altered in primary and secondary sites will allow for the assembly of sets of genes whose expression levels predict specific metastatic outcomes.

1.4 Modeling Cancer Dynamics

Based on the available knowledge on cancer genes and their pathway interaction, the next step is to make sense of the information by building models to describe cancer dynamics. In 1971, Knudson used data from children with retinoblastoma to derive a two hit model for the development of cancer [81]. This initiated subsequent work which has shown that the inactivation of one TSG is rarely sufficient to cause cancers of other tissue types. Current models for cancer progression formulate equations to describe probabilities for mutation and selection. These equations have allowed us to better understand the kinetics of TSG inactivation and how they depend on compartment size, described as small, medium or large. Respectively, these require two, one, and zero rate-limiting steps [105]. While these models are useful in describing the dynamics of mutations in tumor related genes, they do not explain casual relationships nor give mechanistic insights into pathway processes.

As described in the previous section, numerous genes contributing to tumorigenesis have been discovered but work must be done to link these genes together and find shared cellular processes and pathways. One example of a mechanistic view into cancer progression comes from recent research on the RAS signaling pathway. In many human cancers, RAS is involved in tumor initiation. When RAS becomes activated, it signals to certain effector pathways within the RAS pathway. These effector pathways include PI3K, RalGEF, and MAPK. Activation of these pathways lead to events such as angiogenesis, anchorage-independent growth, and increased cell survival. In vivo experiments show that only cells in which all three of these pathways were activated turned into tumors. However, once a tumor is initiated, RAS is only needed to activate the PI3K pathway. RalGEF and MAPK are still active but no longer rely on RAS for this activation. The PI3K pathway seems to be dependent on RAS due to other mechanisms at play. The pathway is regulated by the TSG PTEN which is lost in many cancer cells. PTEN loss then leads to AKT activation which is involved in the promotion of apoptosis. RAS functions to protect the tumor cells by stopping apoptosis [90].

The RAS pathway has been found to be important in colorectal cancer progression. Colorectal cancer tends to follow a more well defined path of progression than many other cancer types and for this reason colorectal tumorigenesis has been used in early models describing causal relationships in cancer progression. Tumors at various stages of development were analyzed for gene mutations and a model was devised based on the percentage of samples showing mutations at four stages of progression. Although a main finding was that the accumulation of mutations was more important than the exact timing a mutation occurred, general trends were discovered such as a RAS mutation often occurring at the intermediate or late adenocarcinoma stage and 17p loss most often found during carcinoma [41]. This work is a first step in understanding tumorigenesis as a multistage process. Extending this to a pathway level and studying many more cancer types will add a mechanistic view of processes governing cancer development.

Research in modeling cancer dynamics has helped gain an understanding of certain genes contributing to tumorigenesis and laid a framework of the mechanisms driving cancer progression for a few cancer types such as colorectal cancer. A major aspect of the work presented here is building casual models of pathways involved in the different stages of tumorigenesis. This work will enhance our current knowledge of cancer progression by identifying mechanisms in terms of pathways which drive progression. These models will use gene expression data from individuals in varying stages in cancer progression with different cancer types. The result will be a deeper understanding of tumorigenesis along with the identification of therapeutic targets that may aid in personalized medicine.

Chapter 2 will discuss a novel methodology, ASSESS, which will be a tool necessary for the following work on modeling cancer progression on the pathway level. As the majority of the work presented here is on the pathway level, Chapter 3 discusses the idea of gene sets and their reliability. Chapter 4 then outlines the methodology used in modeling cancer progression on the pathway level. Chapters 5.1 - 7 discuss applications of the method to prostate, melanoma, and colorectal cancers. In Chapter 8, the analysis is taken a step further by studying how the pathways that emerged as significant in Chapters 5.1 - 7 have underlying dependency structures. Finally, in Chapter 9, the structure of a pathway is further analyzed using an alternative approach dealing with the mathematical characterization of the size of pathways.

Chapter 2

ASSESS

Groups of related genes, or gene sets, often describe a regulatory gene pathway. It is of interest to learn how pathway enrichment varies with individuals of certain disease states in order to understand the underlying mechanisms causing the disease. Statistical methods exist which assay pathway enrichment by phenotypic class but not do describe individual variation. To capture this individual variation, we developed ASSESS. ASSESS provides statistical evaluations and confidence assessments for the annotation of an expression data set by measuring the overlap of significantly perturbed genes in the dataset with genes in a given pathway.

2.1 Expression Based Pathway Annotation Programs

Gene expression profiling experiments have been conducted on a wide variety of cell lines and animal models with the goal of characterizing gene sets whose expression patterns characterize specific genetic or molecular perturbations. These gene sets contain candidate players in pathways, or sub-pathways, that are annotated by the experimental perturbation. The fundamental idea in pathway based analysis approaches [71, 14, 100, 133, 1, 42, 132] is that such a gene set serves as a reference base for interrogating other expression datasets. A new dataset in which a specific pathway gene set appears to be enriched, in terms of multiple genes in that set evidencing expression changes can then be annotated by that reference pathway. An analogy can be made here with sequence annotation in a BLAST search: sets of experimentally derived pathways serve as annotation reference sets for future experiments in the same way that annotated sequences serve as references in a sequence search. Statistical methods are needed and have been developed [132, 8, 80, 138] to define computational tools for such expression-based pathway annotation. Two of these methods, GSEA [132] and SAFE [8], use nonparametric statistics to provide formal statistical evaluation, and confidence assessments, for annotation of an expression data set by measuring the overlap of significantly perturbed genes with those in each pathway in a database of pathways. Gene Set Enrichment Analysis (GSEA) [132] has been successfully applied in many basic science and clinical studies [100, 133, 1, 42, 132, 13], including pathway deregulation in cancer genomics. A fundamental shortcoming of GSEA and other methods [8, 80, 138] is that they do not characterize the variation in enrichment over individual samples in the dataset.

2.2 Motivation for ASSESS

ASSESS is a statistical method that allows us to assay pathway variation in individual samples. If the enrichment of each sample in an expression dataset can be annotated, then one can assay the natural variation of pathway activity in observed gene expression datasets from clinical cancer and other studies. The ability to assay pathway variation in samples allows the implementation of a general methodology to dissect tumor samples in terms of oncogenic pathways. ASSESS extends the methodolgy of GSEA to calculate enrichment of individual samples from an expression profile in given gene sets.

A family of methods for pathway annotation was developed and used to measure pathway deregulation in breast cancer and lung cancer [71, 14, 13]. The approach involved: (a) building statistical models of pathway deregulation from cell lines where recombinant adenoviruses were used to express oncogenic activities corresponding to pathway deregulation, (b) applying the models to each sample in a dataset of tumors and estimating the probability of deregulation of the pathways. The main drawback of this methodology is that cell line perturbation data as well as tumor data are required for the analysis. For ASSESS, only the list of genes characterizing the pathway deregulation is required, the entire model and cell line data is not needed. This provides a great advantage when the gene sets are determined by literature review or a non-expression based assay, such as immunohistochemical, for which building an accurate model subsequently applicable to expression data is a very difficult challenge.

2.3 The Algorithm

Given gene sets defined by prior biological knowledge or genes co-expressed in an experiment with a specific genetic or molecular perturbation, and a dataset of expression profiles from samples belonging to two classes, ASSESS provides a measure of the enrichment of each gene set in each sample and a confidence assessment. ASSESS takes as inputs:

1. Genome-wide expression profiles consisting of p genes and n samples with each sample corresponding to one of two classes, C_1, C_2. The expression of the j-th gene in the i-th sample is x_j^i;

2. A database of m gene sets $\Gamma = \{\gamma_1, ...\gamma_m\}$ where each gene set γ_k is a list of genes (a subset of the p genes in the dataset) belonging to a pathway or other functional or structural category;

3. A ranking procedure and correlation statistic that takes the expression dataset and labels as inputs and produces correlation statistics for each sample that reflects the correlation of the p genes in that sample with respect to the the distribution of expression in the two classes. The correlation statistics for the i-th sample would be $\mathbf{c}_i = \{c_1^i, ..., c_p^i\}$ where c_j^i is calculated by any likelihood ratio statistic for measuring the correlation of a sample to one class rather than the other:

$$c_j^i = \log\left(\frac{\mathbb{P}(x_j^i \in C_1 | x_j^i, \text{data})}{\mathbb{P}(x_j^i \in C_2 | x_j^i, \text{data})}\right).$$

   An example of a parametric and nonparametric correlation statistic is described in detail in the following sections;

and produces as outputs:

1. An enrichment score for each sample in the dataset with respect to each gene set in the database. ES_i^k corresponds to the enrichment of the i-th sample

with respect to the k-th gene set;

2. A measure of confidence for each enrichment score is given by a p-value with multiplicity taken into account by Family-wise error rate (FWER) p-values and False Discovery rate (FDR) q-values.

Given the correlation statistics for the i-th sample, $\mathbf{c}_i = \{c_1^i, ..., c_p^i\}$, and a gene set γ_k, we construct the following discrete random walk over the indices of the rank-ordered correlation statistic

$$\nu(\ell) = \frac{\sum_{j=1}^{\ell} |c_{(j)}|^\tau I(g_{(j)} \in \gamma_k)}{\sum_{j=1}^{p} |c_{(j)}|^\tau I(g_{(j)} \in \gamma_k)} - \frac{\sum_{j=1}^{\ell} I(g_{(j)} \notin \gamma_k)}{p - |\gamma_k|}, \tag{2.1}$$

where $c_{(j)}$ is the rank-ordered correlation statistic, τ is a parameter (in general $\tau = 1$), γ_k is the k-th gene set, $I(g_{(j)} \in \gamma_k)$ is the indicator function on whether the j-th gene (the gene corresponding to the j-th ranked correlation statistic) is in gene set γ_k, $|\gamma_k|$ is the number of genes in the k-th gene set, and p is the number of genes in the dataset. The enrichment statistic for the i-th sample with respect to the k-th gene set is the maximum deviation of the random walk from zero

$$ES_i^k = \nu[\arg \max_{\ell=1,..,p} |\nu(\ell)|]. \tag{2.2}$$

The random walk is a tied-down Brownian bridge process and the deviation from zero is very closely related to the classical Kolmogorov-Smirnov statistic [44]. There are simpler ways to define the enrichment score such as taking the average rank of the genes in the gene set from the rank-ordered list of genes. However, using a random walk is advantageous because it allows one to see how the genes in the set are distributed in the rank-ordered list. The random walk could alternatively be solved by ranking $\left|c_{(j)}\right|$ rather than $c_{(j)}$. While in certain situations a more extreme enrichment score may be sacrificed, this ranking will not allow for the access to the additional information of which genes in the set are up or down regulated. Therefore we choose to calculate the ES by a random walk using values ranked by $c_{(j)}$, and suggest gene sets be constructed to capture genes correlated with either over expression or under expression in a class but not both.

To measure significance we assume under the null hypothesis that the labels are exchangeable and therefore we can compute the null distribution by permuting labels, ranking the genes according to the recomputed statistic $c^{\pi}_{(j)}$, and computing the "random" enrichment statistic $ES^k_i(\pi)$. This is done over many label permutations, $\pi = 1, ..., \Pi$. The p-value is computed by comparing the enrichment score to the empirical distribution generated from $\{ES^k_i(\pi)\}^{\Pi}_{\pi=1}$. Correction for multiple hypothesis testing is preformed in the same manner as in GSEA and is addressed via FWER p-values or FDR q-values (see [132] for details). Overlapping gene sets do not influence the calculation of q-values.

The key technical innovation in extending methods such as GSEA or SAFE to provide enrichment scores for individual samples is producing a correlation statistic and subsequent rankings that model how representative each gene for a given sample is with respect to the two classes. The ranking should reflect the natural variation of how each sample is correlated with class labels. We introduce two correlation statistics which reflect this variation: (1) based on a simple parametric normal model, (2) based on a nonparametric random walk model. Both methods preform well, as likely would other well-defined likelihood ratios such as a binary regression model. We choose to use the nonparametric model on the data in this paper as this method is a novel means for calculating class membership likelihoods. However, the parametric method as well as other methods not discussed here give comparable results.

2.3.1 Parametric model

The parametric model assumes that the expression of a given gene can be modeled by a mixture of two normal distributions corresponding to the two classes. The mean and standard deviations are computed from the data

$$\hat{\mu}_{j1} = \frac{1}{n_1} \sum_{i \in C_1} x_j^i, \qquad \hat{\mu}_{j2} = \frac{1}{n_2} \sum_{i \in C_2} x_j^i,$$

$$\hat{\sigma}_{j1}^2 = \frac{1}{n_1} \sum_{i \in C_1} (x_j^i - \hat{\mu}_{j1})^2, \qquad \hat{\sigma}_{j2}^2 = \frac{1}{n_2} \sum_{i \in C_2} (x_j^i - \hat{\mu}_{j2})^2,$$

where n_1 and n_2 are the number of samples in class 1 and 2. The expression of the j-th gene is modeled as $\mathbf{N}(\hat{\mu}_{j1}, \hat{\sigma}_{j1})$ or $\mathbf{N}(\hat{\mu}_{j2}, \hat{\sigma}_{j2})$ depending on whether the sample belongs to class 1 or 2. We define the class membership likelihood of expression x_j^i from the models of classes 1 and 2 as p_{j1} and p_{j2} respectively.

$$\begin{aligned}
p_{j1} &= \mathbb{P}(\xi \geq x_j^i | \xi \sim \mathbf{N}(\hat{\mu}_{j1}, \hat{\sigma}_{j1})), \quad \text{if} \;\; x_j^i \geq \hat{\mu}_{j1}, \\
p_{j1} &= \mathbb{P}(\xi < x_j^i | \xi \sim \mathbf{N}(\hat{\mu}_{j1}, \hat{\sigma}_{j1})), \quad \text{if} \;\; x_j^i < \hat{\mu}_{j1}, \\
p_{j2} &= \mathbb{P}(\xi \geq x_j^i | \xi \sim \mathbf{N}(\hat{\mu}_{j2}, \hat{\sigma}_{j2})), \quad \text{if} \;\; x_j^i \geq \hat{\mu}_{j2}, \\
p_{j2} &= \mathbb{P}(\xi < x_j^i | \xi \sim \mathbf{N}(\hat{\mu}_{j2}, \hat{\sigma}_{j2})), \quad \text{if} \;\; x_j^i < \hat{\mu}_{j2}.
\end{aligned}$$

We use the distribution function rather than the density because there is a very natural directionality assumption in this model in that if the Gaussians are well separated then the deeper inside the respective class a point x resides the higher should be the membership probability. We then use the log-likelihood ratio as the correlation statistic. Given expression, x_j^i, of the j-th gene of the i-th sample the correlation statistic is computed as:

$$
\begin{aligned}
c_j^i &= \log\left(\frac{p_{j1}}{p_{j2}}\right), \quad \text{if } \hat{\mu}_{j1} \geq \hat{\mu}_{j2} \\
c_j^i &= \log\left(\frac{p_{j2}}{p_{j1}}\right), \quad \text{otherwise.}
\end{aligned}
$$

Thus, genes are ranked based upon the differential probability of their membership in either class and because of this, genes are ranked as a continuum from those with the greatest probability of belonging to class 1 ranked at the top and genes with the greatest probability of belonging to class 2 near the bottom. As most genes will have limited differential expression between the two classes, these genes will have similar probabilities of belonging to either group and the log-likelihood ratio will be near zero.

2.3.2 Nonparametric model

The assumption of normality in the parametric model is often inappropriate for expression data. For this reason, a nonparametric model to compute class membership likelihoods is used in most applications. The class membership likelihoods are computed based upon absorption probabilities of a Brownian motion (random walk) (see Figure 2.1 for an illustration of the model).

We first estimate the densities of the j-th gene for the two classes, $\hat{p}_{j1}(x)$ and

$\hat{p}_{j2}(x)$, using a Parzen estimator [140] with a Gaussian kernel:

$$\hat{p}_{j1}(x) = \frac{1}{n_1 \sigma_{j1}\sqrt{2\pi}} \sum_{i \in C_1} e^{-|x_j^i - x|^2/2\sigma_{j1}^2},$$

$$\hat{p}_{j2}(x) = \frac{1}{n_2 \sigma_{j2}\sqrt{2\pi}} \sum_{i \in C_2} e^{-|x_j^i - x|^2/2\sigma_{j2}^2},$$

where n_1 and n_2 are the number of samples in C_1 and C_2 and bandwidths σ_{j1} and σ_{j2} are set to the average distance between points of the j-th gene in C_1 and C_2 respectively. We define two points, e_{j1}^i and e_{j2}^i, as the left or right extremes of the random walk starting at x_j^i.

$$e_{j1}^i = \min_i \{x_j^i\} \text{ if } x_j^i < \hat{\mu}_{j1}, \qquad e_{j1}^i = \max_i \{x_j^i\} \text{ if } x_j^i \geq \hat{\mu}_{j1},$$

$$e_{j2}^i = \min_i \{x_j^i\} \text{ if } x_j^i < \hat{\mu}_{j2}, \qquad e_{j2}^i = \max_i \{x_j^i\} \text{ if } x_j^i \geq \hat{\mu}_{j2}.$$

The membership likelihood of expression x_j^i for class 1 and 2 is given by the absorption probability at the points e_{j1}^i and e_{j2}^i for a Brownian motion starting at x_j^i with initial conditions distributed as $\hat{p}_{j1}(x)$ and $\hat{p}_{j2}(x)$. We again use the log-likelihood ratio as the correlation statistic. Given expression, x_j^i, of the j-th gene of the i-th sample the correlation statistic is computed as:

$$c_j^i = \log\left(\frac{\mathbb{P}(\text{absorption at } e_{j1}^i \text{ starting at } x_j^i | \hat{p}_{j1})}{\mathbb{P}(\text{absorption at } e_{j2}^i \text{ starting at } x_j^i | \hat{p}_{j2})}\right). \tag{2.3}$$

The absorption probabilities can be computed as the solution of the Dirichlet problem [35] which for the Parzen estimators results in a weighted sum of error functions and exponentials (see Appendix A for the exact form and derivation). So the correlation statistics can be computed efficiently.

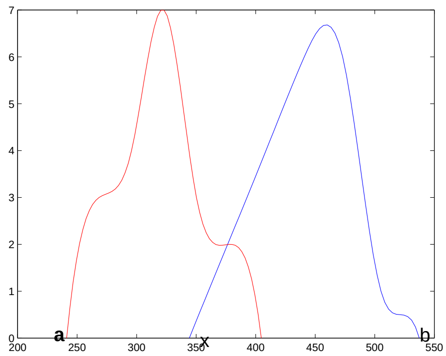


Figure 2.1: The two classes' densities are displayed by the red and blue curves, p_r, p_b. Assume we have a diffusion process (random walk) starting at x. We compute the probability of absorption at the point b if the initial conditions are distributed as p_b, $\mathbb{P}(\text{absorption at b starting at x}|p_b)$. We also compute the probability of absorption at the point a if the initial conditions are distributed as p_r, $\mathbb{P}(\text{absorption at a starting at x}|p_r)$. These two probabilities serve as a measure that an individual sample belongs to one of the two distributions.

2.4 Validation of ASSESS on Model Systems

The objective of ASSESS is to annotate each sample in an expression dataset in terms of a priori defined gene sets often constructed from model systems. In this section we validate the method by demonstrating that gene sets built from model systems or with known genetic perturbations are indeed enriched in gene expression data from the same model systems or related systems.

2.4.1 Mouse models

In [93] transgenic mice were generated that developed a highly penetrant prostatic intraepithelial neoplasia (PIN) phenotype and expressed a constitutively active AKT1 gene in the ventral prostate of the mouse. This AKT-induced PIN phenotype can be reversed with treatment of RAD001, a mTOR inhibitor. The transgenic mice were split into two groups, with one group receiving RAD001 and the other a placebo.

Tissue was taken from the prostate of both groups and DNA microarray analysis was preformed using the Affymetrix Murine U430A microarray. This resulted in two sets of expression data: samples treated with RAD001 ($n = 19$) and placebo ($n = 19$). These datasets were split into a training and test set. The training set consisted of the first 10 samples treated with RAD001 and the first 10 samples treated with the placebo. The test set was comprised of the complimentary samples. The training set was used to construct an AKT gene set using a logistic regression model (see Appendix A for details).

We applied ASSESS to the test set using the AKT gene set derived from the training data. The enrichment scores of the samples treated with RAD001 strongly indicate that genes in the AKT gene set were under expressed compared to the samples given placebo which showed enrichment in the gene set (see Figure 2.2). All samples were significantly enriched (p-value < 0.001).

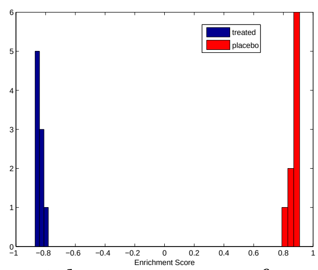


Figure 2.2: A histogram of enrichment scores for the 9 treated and 9 untreated mouse prostate samples in the test data with respect to the AKT pathway gene set computed from the training data.

2.4.2 Cell culture models

In [13] human primary mammary epithelial cell cultures (HMECs) were used to develop a series of pathway signatures which were then used to assay pathway disregulation in non-small cell lung carcinoma (NSCLC). We use this dataset to validate our method.

The data was generated by using recombinant adenoviruses to express specific oncogenes in an otherwise quiescent cell, thereby isolating the subsequent events as defined by the activation/deregulation of that single pathway. The cells were infected with adenovirus expressing either human c-Myc, activated H-Ras, human c-Src, human E2F3, or activated β-cantenin. RNA from these multiple independent infections, as well as from normal cells (with green florescent protein, GFP), was collected for DNA microarray analysis using the Affymetrix Human Genome U133 Plus 2.0 Array.

Given the independent replicates from the six conditions, the five perturbed pathways and the normal GFP cells, we split each condition into a train and test set. Thus given expression data from: 10 Myc, 10 Ras, 7 Src, 10 E2F3, 9 β-catenin, and 10 normal/GFP samples we constructed five training sets with the first half of the samples from each experimental dataset along with the first 5 normal samples. Similarly, five independent test sets were constructed using the complimentary samples (the second half of samples in the six conditions). The training sets were used to construct gene sets for each of the five pathways, Myc, Ras, Src, E2F3, and β-catenin using a logistic regression model (see Appendix A for details).

ASSESS was applied to the five test sets to calculate enrichment with respect to the five gene sets computed from the training data.

Experiment	ES for GFP cells	ES for infected cells
BCAT	-0.87($\pm$0.031)	0.88($\pm$0.042)
E2F3	-0.97($\pm$0.0069)	0.98($\pm$0.0061)
MYC	-0.89($\pm$0.018)	0.91($\pm$0.067)
RAS	-0.96($\pm$0.012)	0.91($\pm$0.021)
SRC	-0.90($\pm$0.016)	0.91($\pm$0.022)

Table 2.1: Average enrichment scores($\pm$sd) for the comparison of normal (GFP cells) to infected cells for the gene set built from the respective infected cell type.

2.4.3 Literature Based Models

The approach developed in [71, 14, 13] of building statistical models of pathway deregulation in controlled experiments and then applying these to new datasets could have been used in the previous two examples. However, this approach requires the cell line perturbation data as well as new data and that the data are on comparable platforms. This approach cannot be used for gene sets derived from literature whereas ASSESS is still applicable.

In [132] a dataset generated from mRNA expression from lymphoblastoid cell lines derived from 15 males and 17 females served as a validation set. We then sought to address the following: which gene sets were over expressed in males and which in females. Gene sets defined by cytogenetic bands and gene sets defined by pathway or functional properties were examined. As expected for males, chromosome Y as well as its two bands (Yp11 and Yq11) and a gene set corresponding to genes enriched in male reproductive tissue (testis) were overexpressed. For females two gene sets of genes that escape X-inactivation were overexpressed in addition to a gene set corresponding to genes enriched in female reproductive tissue (uterus). Genes on the X-chromosome would not be expected to be overexpressed due to dosage compensation by X-inactivation.

The enrichment of these seven gene sets with respect to the male and female samples in the lymphoblastoid cell lines is displayed in Figure 2.3. The male samples are clearly enriched with respect to: Y, Yq11, Yp11, and testis. The female samples are clearly enriched with respect to: the two escape of X-inactivation gene sets (X-inactivation and Willard X-inactivation) and the uterus gene set (labeled in the figure as Reproduction Genes). We used a Myc gene set as a control in that it is not expected to be enriched with respect to the male/female distinction and indeed this is the case.

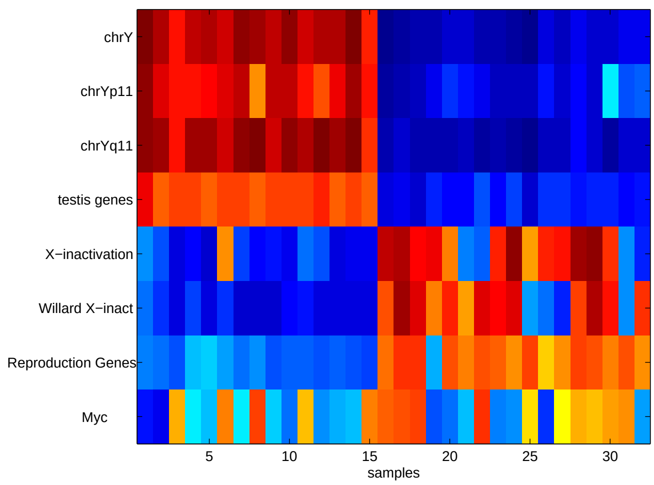


Figure 2.3: Enrichment scores for the comparison of males to females in the 8 gene sets. The male samples $(1 - 15)$ show enrichment in the Y, Yq11, Yp11, and testis gene sets. The female samples $(16 - 32)$ show enrichment in the two escape of X-inactivation gene sets and the uterus gene set. The Myc pathway shows no differential expression between males and females, as expected.

We further illustrate the procedure by plotting the random walk (Equation (2.1)) for a male and female sample with respect to one of the escape from X-inactivation gene sets and a Myc gene set (see Figure 2.4). For a female sample with respect to this gene set, the random walk increases very rapidly initially indicating that genes escaping X-inactivation appear at the top of the list of genes ordered by correlation with the female phenotype. This results in a very positive enrichment score. For the male sample the random walk is basically a mirror image of the female case indicating that genes escaping X-inactivation appear at the end of a list of genes ordered by correlation with the male phenotype. This results in a very negative enrichment score. The third case is for a female sample with respect to the Myc gene set. In this case the genes in the gene set are randomly spread over the ranked list and so the enrichment score never deviates far from zero.

2.5 Classification and Clustering in the Space of Pathways

A very natural consequence of obtaining enrichment scores for each sample in the dataset is that classification and clustering can now be performed in the space of gene sets rather than individual genes. Being able to interpret classification models using pathways offers an alternative and possibly more intuitive perspective than models using individual genes. Another aspect of building models in the space of pathways that was emphasized in [13] is knowing which pathways are disregulated with respect to outcome and how this can help suggest targeted therapeutics.

2.5.1 Clustering

In [20] a matrix factorization method (NMF) was applied to an expression dataset with acute myelogenous leukemia (AML) and acute lymphoblastic leukemia (ALL) samples [57]. The matrix factorization allowed the clustering of the samples into subsets. A parameter in this clustering method is the number of subsets k. For this dataset results with $k = 2, 3$ were computed. For the two cases the clusters comprised of $\{(25 \text{ ALL}), (11 \text{ AML}, 2 \text{ ALL})\}$ and $\{(8 \text{ ALL-T}), (17 \text{ ALL-B}), (11 \text{ AML}, 2 \text{ ALL-B})\}$, where ALL-T and ALL-B are two subtypes of ALL. We applied ASSESS to this leukemia dataset using a database of 523 gene sets [132]. We then applied NMF to this space of enrichment scores and obtained identical results. The only difference is according to the measure of confidence developed in [20], as the clustering obtained from the enrichment space had greater confidence than that from the raw expression data. The result of the clustering and the factors are displayed in Figure 2.5.

2.5.2 Classification

We examined six gene expression datasets for which single gene classification models have been built: (a) Gender – male vs. female [132], (b) cDNA Lung cancer – squamous vs. adenocarcinoma [52], (c) oligonucleotide Lung cancer – squamous vs. adenocarcinoma [112], (d) Medulloblastoma – survival vs. failure [111], (e) Prostate cancer – recurrence vs. nonrecurrence [55], and (f) Leukemia – AML vs. ALL [57].

We applied the classification using enrichment scores procedure outlined in Appendix A to compute classification accuracy on these six datasets (see Table 2.2). For all the datasets except for the Leukemia dataset the leave-one-out method was used (Algorithm 1). For the Leukemia dataset the train-test procedure was used (Algorithm 2) with the train-test split outlined in [57]. The classification accuracy was comparable or better than that for single gene classifiers for all the datasets except for the Leukemia data.

The pathways associated with recurrent prostate cancer tumors supports AS-

SESS's ability to both accurately predict outcome as well as provide biological insight. Both AKT and PTEN gene sets were found to have increased coordinate expression in samples of recurrent prostate cancer. PTEN loss is one of the most common genetic alterations seen in advanced prostate cancer resulting in activation of the PI3K-AKT pathway. Activation of this pathway is known to occur at a greater frequency in advanced prostate cancer and has prognostic significance. A "TERT-up" gene set was similarly found to be associated with recurrent prostate cancer. An essential requirement for tumor progression is avoidance of cellular senescence, telomerase restores chromosomal telomeres and is associated with the development of prostate cancer. Finally, another interesting observation is the presence of the "DNA damage signaling" and "Cell cycle checkpoint" pathways both representing common cellular processes disregulated in aggressive cancer.

Classes	Accuracy
Gender: males vs. Females	94%
Lung Cancer(cDNA): Adeno vs. Squamous	91%
Lung Cancer(oglio): Adeno vs. Squamous	84%
Medulloblastoma: survival vs. failure	72%
Prostate: recurrent vs. nonrecurrent	73%
Leukemia: ALL vs. AML	85%

Table 2.2: Classification accuracy for six datasets building classification models in the space of enrichment scores.

2.6 Cross-platform Expression Models

DNA microarray studies have been carried out on a variety of platforms for the same case-control experiment, for example both cDNA microarrays and oligonucleotide microarrays are popular in cancer genomics. The integration of data across platforms is appealing for a variety of reasons: increasing the sample size of the data, allowing for interstudy validation, mitigating platform based biases, and mitigating study based biases.

Building a model from raw expression data from one platform and applying the model to data from another platform directly will not work since the expression data from the two platforms have different distributions. One approach to normalize between the platforms is to use median rank scores and quantile discretization to map the data to a common space and then build a classification model in this space [144].

We advocate an alternative approach of applying ASSESS to expression data to map the data into the space of enrichment scores for pathways and then building models in this space. Methods such as median rank scores are no longer needed as enrichment scores between platforms are numerically comparable. There are several advantages to this approach: (1) the need to map genes using UniGene ids is avoided; (2) the problem of multiple probe mappings between platforms is avoided; (3) gene sets defined separately by probes specific to each platform can be used; (4) the enrichment statistic is much more robust than the rank of a single gene so the loss of genes between platforms is mitigated; (5) interpreting results on the level of pathways instead of single genes is appealing.

We first applied this approach to two prostate cancer studies [145, 32]. The two platforms for the studies were cDNA microarrays [32] and Affymetrix oligonucleotide microarrays [145]. The cDNA dataset contained 53 samples of which 34 were tumors and 19 were normal. The oligo dataset contained 33 samples of which 24 were tumors and 9 were normal. The catalog of human functional gene sets comprised of 433 sets annotated for both platforms [132] was used as the gene set. The error rate for using the cDNA and oligo datasets as train-test sets respectively is reported in Table 2.3, as is the error rate for a leave-one-out procedure using all the cDNA and oligo samples (see Appendix A for details on on both test-train and leave-one-out classification using gene sets). We compare these results with the leave-one-out error computed on the individual datasets (see Table 2.3).

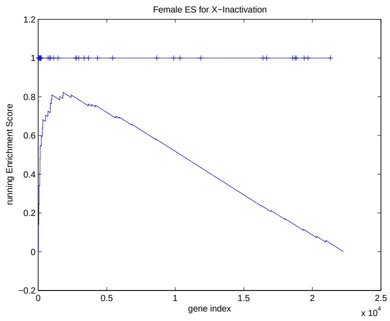
We next applied this approach to two lung cancer studies [52, 112]. The two studies involved the same platforms as the prostate example. The cDNA dataset contained 55 samples of which 38 were adenocarcinomas and 17 were squamous cell lung carcinomas [52]. The oligo dataset contained 93 samples of which 45 were adenocarcinomas and 48 were squamous cell lung carcinomas [112]. The same catalog of gene sets as used in the prostate example was used. The error rates for the cross-platform predictions as well as predictions within the individual datasets are summarized in Table 2.3.

	cDNA LOO	oligo LOO	train-test	combined LOO
prostate T/N	85.7%	76.5%	(cDNA-oligo) 73.5%	80.7%
lung A/S	88.0%	90.9%	(oligo-cDNA) 78.2%	88.5%

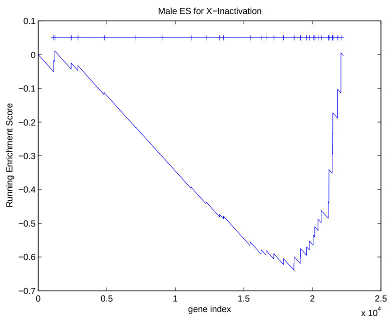
Table 2.3: Classification accuracy for cross-platform models for the prostate and lung cancer datasets.

2.7 Discussion

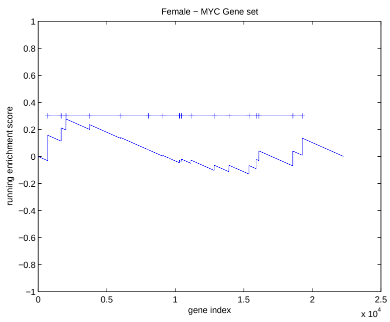
We have developed a formal statistical method to measure the enrichment of each sample in an expression dataset with respect to a priori defined gene sets. We have tested the method on a variety of model systems and demonstrated the utility by clustering and building classification models in the space of pathways or gene sets. These were in general as accurate as methods applied in the space of genes but more interpretable and robust. This robustness was illustrated by the ability to build models between different expression based technologies— cross-platform models. This is hard to do in the single gene setting. The utility will further be shown as we use ASSESS to transform gene expression data into the space of pathways for the purpose of modeling cancer progression on the pathway level.



(a)



(b)



(c)

Figure 2.4: Random walks. (a) The random walk for a female sample with respect to one of the escape from X-inactivation gene sets. The hatches of the top line indicate where the genes in the gene set fall with respect to the rank-ordering. (b) The random walk for a male sample with respect to one of the escape from X-inactivation gene sets. (c) The random walk for female sample with respect to a Myc gene set. This gene set is not significantly enriched and so the hatches appear randomly dispersed with respect to the rank-ordering.

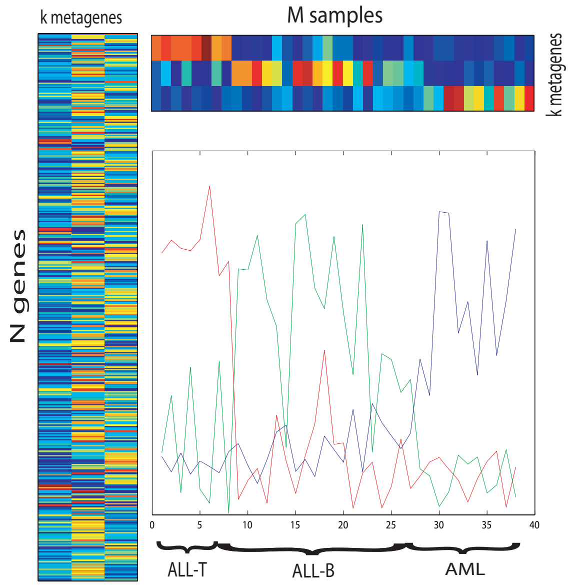


Figure 2.5: The top and left figure are the left and right matrix factors for the matrix of enrichment scores in the Leukemia data with $k = 3$. The red line is a plot of the first metapathway over the data and this metapathway selects for the ALL-T samples. The green line is the second metapathway and it selects for the ALL-B samples. The blue line is a plot of the third metapathway which selects for the AML samples.

Chapter 3

Gene Set Validation

We have developed a method, ASSESS, which interrogates a gene expression data set based on given gene sets. Gene sets can be derived from experimental perturbations, literature based studies, and a variety of other origins. A fundamental question is which of these sets is most appropriate. For example, a database of gene sets may contain 5 RAS pathways from a variety of experiments or literature surveys. For a particular analysis which is most appropriate?

The Molecular Signatures Database (MSigDB) [132] is a collection of curated gene sets from online pathway databases, publications in PubMed, and expert knowledge. A key constraint in using a priori defined gene sets and pathways is the quality of the database of gene sets and the accuracy of the annotation. Since the enrichment of gene sets or pathways is fundamental to our models, we need to validate the accuracy of these gene sets. Two points will be addressed in our validation studies: the accuracy of gene sets annotated according to known perturbations and a comparison of gene sets defined by experimental studies versus gene sets defined by expert knowledge. An affirmative answer to the first point provides confidence in the annotation and interpretations made based on the annotations. A study of the second point highlights the importance of the context of the gene set and again provides information about interpreting results. An affirmative answer to this question allows for a uniform analysis over both types of gene sets.

The validation of gene sets requires the knowledge of which pathway gene sets are deregulated in which samples according to known biology. This requirement is satisfied by studies where a model system is genetically perturbed and a gene set is defined as genes that most differentially express under the perturbation. Expression studies where the pathways driving the phenotypic distinction are known also satisfy the above requirement. Due to the difficulty in finding data satisfying the above requirements for all the gene sets in MSigDB [132], we focused on a few gene sets which we can validate: the P53 pathway, the hypoxia pathway, and the RAS pathway. The conclusion of our analysis will be that both types of gene sets seem to be accurately annotated and gene sets defined by expert knowledge may be applicable to more general conditions.

3.1 P53 Pathway

MSigDB contains 5 sets defining P53 pathways both from experimental studies and from expert sources. We first test whether these five gene sets will be significantly enriched in a dataset with a known P53 perturbation from the NCI-60 collection of cancer cell lines. This dataset includes 12 normal samples and 50 samples with a P53 mutation. Using GSEA, the normalized enrichment score (NES), p-value, and false discovery rate (FDR) for the 5 P53 pathway gene sets along with other gene sets in the MSigDB were obtained. We found that the P53 pathway gene sets were all significantly enriched in the P53 wild type cells (results summarized in Table 3.1). These results give confidence that the 5 P53 signatures, although from different sources, are accurately representing the P53 pathways.

Rank	Gene Set	NES	Pval	FDR
Enriched in Wild Type				
1	p53 Pathway	-2.38	0	0
2	Stress p53 Specific Up	-2.33	0	0
3	p53 Hypoxia	-2.19	0	0
4	p53 Genes All	-2.1	0	0.01
10	p53 Up	-1.73	0.02	0.20

Table 3.1: GSEA results for the p53 gene sets in the wild type/p53 mutant dataset. The ranks are out of 220 gene sets which are enriched in the wild type phenotype.

Further validation of the P53 pathway gene sets come from an indirect comparison based on P53 as a marker for recurrence of breast cancer [39, 124, 50]. In [91] gene expression profiles were obtained for 60 individuals with hormone receptor-positive primary breast cancer treated with adjuvant tamoxifen monotherapy. Of these individuals, 32 experienced tumor recurrence. All of the P53 pathway gene sets except 'P53 Up' differentiate the nonrecurrent from recurrent breast cancer phenotypes (see Table 3.2). The fact that 'P53 UP' did not rank highly is not surprising as it is also the gene set with the worst enrichment in the NCI-60 dataset. The ability of the other four gene sets to differentiate P53 perturbations and breast cancer phenotypes provides confidence in using these gene sets.

Rank	Gene Set	NES	Pval	FDR
Enriched in Nonrecurrent				
10	p53 Pathway	-1.27	0.15	1
15	p53 Genes All	-1.23	0.2	1
19	Stress p53 Specific Up	-1.18	0.23	1
69	p53 Hypoxia	-0.83	0.70	1
Enriched in Recurrent				
248	p53 Up	0.72	0.90	0.96

Table 3.2: GSEA results for the p53 gene sets in the recurrent/nonrecurrent breast cancer dataset. The ranks are out of 123 gene sets which are enriched in the nonrecurrent phenotype and 278 gene sets which are enriched in the recurrent phenotype.

3.2 Hypoxia Pathways

MSigDB contained 7 gene sets defining hypoxia pathways (see Table 3.3) both from experimental studies and from expert sources. Hypoxia occurs when a cell experiences an oxygen deficiency. The cell responds to the low oxygen environment by activating certain hypoxia pathways which induce numerous adaptive responses. Alternatively, cancer cells can genetically activate a hypoxia response in the setting of normal oxygen levels to activate new blood vessel formation and experience a growth advantage. Recently, much work has been done to better understand cellular responses to hypoxia conditions for therapeutic targeting of cancer cells, which have the ability to adapt to these conditions and alter certain signaling pathways to increase survival.

The enrichment of these seven gene sets were examined in an expression dataset which profiled 3 human astrocytes and 3 epithelial cells (HeLa cells) maintained under hypoxic conditions and 3 human astrocytes and 3 HeLa cells maintained under normoxia conditions [98]. GSEA was applied to this dataset to examine the enrichment of the gene sets in the database with respect to the 6 hypoxic and 6 normal cells. The enrichment scores of the seven hypoxia gene sets rank highly out of the 528 gene sets tested from the MSigDB for their ability to discriminate hypoxic and normoxic cells in both cell lineages (astrocyte and epithelial cells) (see Table 3.4). This provides further support for using gene sets defined for the same biological process from a variety of sources.

Gene Set	Source
Hypoxia Down	Manalo et al
Hypoxia Up	Manalo et al
Hypoxia Fibro Up	Kim et al
Hypoxia Reg Up	Leonard et al
Hypoxia Review	Harris
HIF Pathway	BioCarta
VEGF Pathway	BioCarta

Table 3.3: The seven hypoxia gene sets defined by experimental and expert sources.

Rank	Gene Set	NES	Pval	FDR
Enriched in Hypoxic Cells				
3	Hypoxia Up	-1.96	0.008	0.026
4	Hypoxia Review	-1.95	0	0.027
6	Hypoxia Fibro Up	-1.84	0.004	0.088
9	Hypoxia Reg Up	-1.73	0.02	0.191
10	HIF Pathway	-1.73	0.02	0.176
53	VEGF Pathway	-1.39	0.055	0.553
Enriched in Normal Cells				
17	Hypoxia Down	1.48	0.167	0.596

Table 3.4: GSEA results for the hypoxia gene sets in the hypoxia/normal dataset. The ranks are out of 323 gene sets which are enriched in the hypoxia phenotype and 205 gene sets which are enriched in the normal phenotype.

3.3 Ras Pathway

The final example illustrates the importance of context in the pathway approach. The enrichment of three RAS gene sets will be examined under two conditions or datasets. The context of the first dataset is very specific, an animal model with mutated KRAS [13]. The second data set is much more heterogeneous, non-small cell lung cancers[112]. The three gene sets correspond to two experimentally derived gene sets [13, 133] and the RAS pathway from Biocarta (http://www.biocarta.com). The RAS gene set from Biocarta is the most general with respect to context as it consists of genes thought to biochemically interact with RAS and proteins associated with RAS. The two experimentally defined gene sets are much more context specific. The first gene set corresponds to a pathway signature derived from infecting human primary mammary epithelial cell cultures with an adenovirus expressing activated HRAS [13]. The second gene set is a pathway signature derived from a mouse model with a KRAS mutation. In summary these two gene sets annotate HRAS and KRAS deregulation. Although the three sets all describe the RAS pathway, there are very few shared genes between the sets, emphasizing the importance of context when defining gene sets. The HRAS gene set and KRAS gene set share 8 genes and contain 237 and 58 unique genes respectively. The RAS gene set defined by BioCarta contains 23 genes, sharing only 1 gene with the HRAS set and having no gene in common with the KRAS gene set.

The importance of the context in which a gene set was defined is demonstrated by testing the enrichment HRAS gene set on a dataset with a KRAS mutation. The HRAS gene set was split into the genes up-regulated and down-regulated in response to RAS activity. In [13], microarrays were collected from 31 cells with tumors caused by a KRAS mutation and 19 normal cells. Using GSEA, the normalized enrichment score (NES), p-values, and false discovery rates (FDR) for the HRAS up-regulated and down-regulated gene sets were obtained. The oncogenic pathways of BCAT, E2F3, MYC, and SRC were also defined in [13] and were added to the analysis for enrichment comparisons. Results are shown in Table 3.5. The experimental HRAS gene set is narrowly defined for genes perturbed by HRAS activation and do not capture KRAS mutation specificity. On the other hand, the RAS gene set defined by BioCarta is appropriate to use regardless of the specific RAS mutation as it captures the interactions that the RAS protein can have under different conditions.

Gene Set	NES	Pval	FDR
Enriched in Tumor			
RAS Up BioCarta	1.51	0	0.05
SRC Down	1.41	0.09	0.32
MYC Up	1.25	0.15	0.35
SRC Up	1.25	0.15	0.21
HRAS Up	1.12	0.26	0.49
E2F3 Up	1.12	0.25	0.33
BCAT Up	0.81	0.74	0.80
Enriched in Normal			
RAS Down BioCarta	-1.51	0.12	0.06
E2F3 Down	-1.29	0.10	0.49
HRAS Down	-1.18	0.19	0.52
BCAT Down	-1.14	0.29	0.32
MYC Down	-0.99	0.55	0.52

Table 3.5: GSEA results for the oncogenic gene sets defined from [13] and BioCarta in the KRAS mutant/normal dataset. The BioCarta gene sets are named 'RAS Up BioCarta' and 'RAS Down BioCarta'. The remaining gene sets are experimentally derived in [13]

The context of the RAS gene set does not have as much of an impact on datasets that indirectly involve RAS perturbations, even though the different RAS gene sets share a limited number of genes. Previous studies have shown that RAS activation is linked with lung adenocarcinomas [118, 120]. The BioCarta RAS signature, the HRAS signature, and the KRAS signature were tested for their diagnostic capabilities in human lung cancer using a gene expression dataset with 45 adenocarcinoma lung cancer samples and 48 squamous lung cancer samples [112]. Using ASSESS, enrichment scores were calculated for each sample in the three RAS gene sets. The class prediction accuracy was 69.9% for the HRAS pathway, 75.3% for the KRAS pathway, and 79.6% for the BioCarta RAS pathway.

These results give confidence for the robustness of ASSESS in allowing for the use of both literature and experimental based gene sets when the dataset contains a more general perturbation. However, if a dataset is built on a perturbation of a specific proto-oncogene, enrichment will not be seen in a gene set for a different proto-oncogene, even if these two proto-oncogenes belong to the same family. Therefore, knowing the context under which both the dataset and the gene set were built is important for interpreting results of enrichment analyses.

3.4 Discussion

Programs such as GSEA and ASSESS are pathway based methods that require a priori defined gene sets. Therefore, the results of these methods rely on the accuracy of the gene sets they are given. Since our later work on modeling cancer progression requires the use of ASSESS to transform gene expression into the space of gene sets, the validation of the gene sets we will be using is extremely important. P53, hypoxia, and RAS pathways have been defined by multiple sources, both in literature and experimental settings. For each, we took data with a specific perturbation for that pathway and calculated pathway enrichment. Results suggested that gene sets are appropriate for use with careful attention to context. Context of both the dataset and gene set is important for enrichment analysis. We find this in the HRAS/KRAS example. Gene sets defined on one dataset may not be applicable for enrichment analyses when the context of the dataset changes. As such, careful attention should be paid for application of gene sets defined on one dataset to a different context.

Chapter 4

Modeling Cancer Progression Framework

Numerous DNA microarray expression datasets have been collected to profile genetic changes throughout tumor progression [139, 129, 135, 69, 91, 74]. Traditionally, gene expression profiling has been used to identify individual genes that become deregulated at distinct stages of tumorigenesis. Such analyses have shown that tumor cells have a great deal of heterogeneity as they progress through the stages of cancer development [127]. The multitude of differentially expressed genes can then be grouped together by shared biological function to uncover mechanistic alterations that may give rise to certain cancer states. This approach has resulted in the understanding of some of the genetic changes that occur during progression. However, single gene based methods do not always provide clear and accurate insight about the underlying biological processes governing tumor development since these processes involve sets of genes.

One reason for lack of insight from single gene based methods is the great deal of heterogeneity across individuals. An indication of this is that typically about 15 mutated genes drive cancer, however these genes differ greatly from individual to individual [147]. In fact, most genes are mutated in less than 5% of tumors. Wood et al. describes the genomic landscape of cancer to have a few "gene mountains" and many "gene hills". The gene mountains refer to the genes found to be mutated in almost all tumors of a given type. The gene hills are mutated less frequently and likely comprise a smaller number of pathways or functional sets since numerous gene mutations can result in the same phenotypic alteration [147, 127]. This variability in the gene hills makes it difficult to understand which genes are acting as drivers of tumorigenesis. The idea of pathway analysis is to borrow strength across genes by considering a priori defined sets of genes rather than individual genes. Modeling tumor progression at the pathway level can help bring structure to the complicated landscape of cancer. Pathway analysis has proven to be useful in providing insight about the underlying biologic processes governing tumor development and has provided a means for better functional and mechanistic insight into the cause of phenotypic differences [71, 14, 100, 133, 1, 42, 132]. Pathway analysis is particularly important for studying cancer development because of the degree of variability in mutations at the single gene level, highlighting the idea that disease development can follow various courses.

We believe that studying tumorigenesis on the pathway level will provide finer structure in modeling cancer progression by organizing the variability seen in single genes into underlying functional occurrences. It is our goal to advance the previous single gene based methods of studying cancer progression by using an integrative hierarchical analysis of tumor progression which discovers a priori defined pathways that are relevant either throughout progression or in particular steps in progression.

In addition to genetic heterogeneity acquired across individuals, there is heterogeneity across time. Building separate models for different disease stages addresses heterogeneity across time and selecting genes that are consistently mutated addresses variation across individuals. The problem with this stratification approach is loss of power due to smaller sample sizes in each separate model. A classic paradigm in addressing this problem is to borrow strength by building models jointly across all of the data. We applied this paradigm both across genes and stages of progression.

In modeling tumor progression it is vital to both analyze the transitions between individual stages and learn what is common to progression in general. The analysis of specific transitions would shed light on the molecular mechanism driving a particular stage of tumorigenesis. The analysis of progression in general would shed light on shared mechanisms.

The machine learning and statistical ideas used in the pipeline are regularized multi-task learning (RML) [40] and the ideas of learning gradients [104, 102, 103, 92] and inverse regression [89, 24]. The network inferences are based upon properties of discrete Gauss-Markov graphs [130].

In summary, our objective here is to determine which pathways are most relevant over all stages in the progression on a certain tumor type and which pathways are most relevant to only one specific stage transition. For example, individuals may be classified as having a benign (b), primary (p), or metastatic (m) tumor. The goal is then to discover which pathways are most relevant to both the $\{b \mapsto p\}$ and the $\{p \mapsto m\}$ transitions. At the same time we would like to know which pathways are only relevant to the $\{b \mapsto p\}$ transition and those relevant only to the $\{p \mapsto m\}$ transition.

4.1 Data Stratification

The expression data can be represented as a set of pairs $D = \{(x_i, y_i)\}_{i=1}^{n}$ with $x_i \in \mathbb{R}^p$ the expression over p genes and y_i is the stage of the patient. The data is stratified into T datasets corresponding to stages in progression. For example, there are three stages $y \in \{b, p, m\}$ with n_1, n_2, n_3 samples in each stage and the progression is $\{b \mapsto p \mapsto m\}$ then there are two steps in the progression, $T = 2$.

Stratifying the data with respect to these two steps results in the first dataset $D_1 = \{(x_1, y_1)\}_{i=1}^{n_1+n_2}$ consists of the n_1 samples corresponding to stage b followed by the n_2 samples corresponding to stage p with the label of the first n_1 samples labeled as 0 (less serious) and the remaining n_2 labeled as 1 (more serious). The second dataset D_2 is constructed similarly with samples corresponding to stages p and m. The first n_2 samples are labeled as 0 (less serious) and the remaining n_3 labeled as 1 (more serious) in the this dataset.

4.2 Data Mapping

Each dataset D_t is then mapped into a representation with respect to sets of genes or pathways. This is done using the pathway annotation tool ASSESS [36] that assays pathway variation in individuals. Given phenotypic or label data $Y_n = \{y_1, ..., y_n\}$, expression data $X_n = \{x_1, ..., x_n\}$ and gene sets $\Gamma = \{\gamma_1, ..., \gamma_m\}$ defined a priori ASSESS provides summary statistic $S_n = S(X_n, Y_n, \Gamma)$. The summary statistic S_n is a matrix with n columns corresponding to samples and m rows corresponding to gene sets with each element S_{ij} as the enrichment of gene expression differences in the j-th sample with respect to phenotype for genes in the i-th gene set. The application of ASSESS to the stratified datasets D_1, D_2 results in two datasets S_1, S_2.

4.3 Finding Pathways Relevant to Progression

The central statistical idea used in finding pathways relevant to progression is called multi-task learning [40] in the machine learning literature and hierarchical modeling with mixed effects in the statistics literature. We restrict ourselves in this paper to linear models and classification. The basic idea is that we have T classification problems in our case assigning a sample x_i to labels 0 (less serious) or 1 (more serious). We assume that the classification tasks are related so the conditional distributions of the phenotype given the summary statistics $\mu_t(Y|S)$ are also related. The tasks in our case are the different steps in tumor progression and the data over all tasks is $S = \{S_1, ..., S_T\}$ where $S_j = \{(y_{1j}, s_{1j}), ..., (y_{n_j j}, s_{n_j j})\}$ and n_j is the number of samples in the j-th task. We assume the generalized linear model

$$y_{it} = g[s_{it} \cdot (w_o + v_t) + b] = g[s_{it} \cdot w_t + b], \tag{4.1}$$

with $w_t = w_0 + v_t$, y_{it} is the i-th sample in task t, s_{it} are the summary statistics of the i-th sample in task t, w_0 is the baseline term over all tasks, v_t are the task specific corrections, and b is an offset. The vectors w_t correspond to the linear model for each task.

We used the RML framework developed in [40] to estimate the model parameters

w_0, v_t, b

$$\min_{w_0,v_t,b} \sum_{t=1}^{T} \sum_{i=1}^{n_T} (1 - y_{it} \cdot f(s_{it}))_+ + \lambda_1 ||w_0||^2 + \lambda_2 \sum_{t=1}^{T} ||v_t||^2, \tag{4.2}$$

where $(u)_+ = \min(u, 0)$ is the hinge loss, $f(s_{it}) = s_{it} \cdot (w_o + v_t) + b$, λ_1 and λ_2 are positive regularization parameters that trade-off between fitting the data and the smoothness or robustness of the estimates. The default parameter values are $\lambda_1 = 1$ and $\lambda_2 = 2,000$, as not to over assume dependency between tasks.

RML is then repeated with 1,000 class label permutations to obtain a null distribution of each vector: $\{\{w_0^p, w_t^p\}_{t=1}^T\}_{p=1}^{1000}$. The p-value of each gene set in w_t in $\{w_0^p\}_{p=1}^{1000}$. Gene sets relevant to the t-th step in progression, G_t are those corresponding to elements of w_t with $p \leq 0.05$.

Classification in the leave-one-out setting can be applied for an unbiased estimate of the error rate on the prediction of class labels. The dataset $\{S_t\}_{t=1}^T$ is split into s_i (the i-th data sample and S^I (the data without the i-th sample). RML is applied to the training set, S^I to build a classifier based on $\{w_0, w_t\}_{t=1}^T$ which is applied to s_i to obtain a prediction $\hat{y}_i$. Prediction accuracy is computed by applying the leave-one-out procedure to all samples in the dataset.

Chapter 5

Prostate Cancer Progression

The analysis pipeline described in Chapter 4 will now be applied to prostate cancer progression. The dataset [139] is a collection of cDNA microarray expression measurements from 22 samples of benign epithelium (b), 32 samples of primary prostate cancer (p), and 17 samples of metastatic prostate cancer (m). The progression is benign to prostate cancer (PCA) to metastasis, $\{b \mapsto p \mapsto m\}$.

At each level of the analysis we use the six hallmarks of cancer [61] to organize the analysis results. These hallmarks, thought to be necessary for the development of cancer, were defined as self-sufficiency in growth signals, insensitivity to anti-growth signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis. It was hypothesized that tumors must acquire alterations in each of these categories in order to evade the multitude of anticancer defense mechanisms in the cell. We use these hallmarks as categories to group pathways found to be relevant in our analysis. The gene sets are assigned to the biological category most fitting, although there are cases where a gene set may fall into more than one category. In addition, pathway dependence graphs inferred will illustrate the extent to which the hallmarks are interwoven. Pathway gene sets corresponding to all six hallmarks were found to be relevant.

5.1 Pathways Relevant in Prostate Cancer Progression

The first result of our analysis is a list of pathways relevant to progression in prostate cancer (see Table 11.2). In addition, the analysis indicates in which stage of progression a pathway is relevant: both stage transitions (Figure 5.1A), the first transition, (Figure 5.1B), or the second transition (Figure 5.1C). See Table 5.1 for the names and description of the gene sets in Figure 5.1.

In the three cases, a p-value has been assigned to each of the 522 gene sets in the analysis using the permutation method described in Chapter 4. This p-value denotes the probability of a gene set ranking as it did by chance alone. For each of the three stages in progression, we chose those gene sets with $p \leq 0.05$ as relevant. This resulted in 29 gene sets relevant to overall progression, 17 gene sets relevant in early progression, and 22 gene sets relevant in late progression. These relevant gene sets will be discussed in the remainder of this section.

Set ID	Gene Set Name	Source
TRANS	Transport	Brentani et al
EGF	EGF Receptor Signaling Pathway	BioCarta
GHD	MAP00361 Gamma Hexacholrocyclohexane Degradation	GenMAPP
STER	MAP00100 Sterol Biosynthesis	GenMAPP
HORM	Hormonal Functions	Brentani et al
GLY	GLYCOL	Broad Institute

Table 5.1: Gene set names for the gene set IDs in Figure 5.1

.

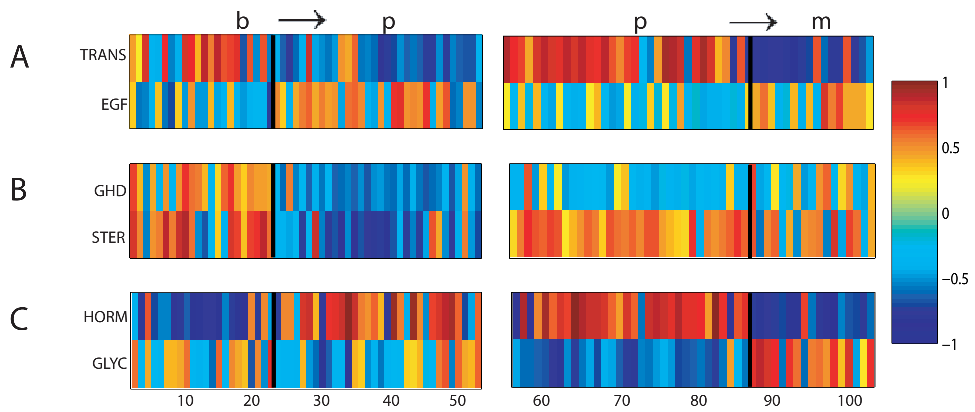


Figure 5.1: Enrichment scores of A) the 2 gene sets with the smallest p-values for relevance throughout progression $\{b \mapsto p \mapsto m\}$ based on RML, B) the 2 gene sets with the smallest p-values for relevance to the early transition $\{b \mapsto p\}$ based on RML, and C) the 2 gene sets with the smallest p-values for relevance to the late transition $\{p \mapsto m\}$ based on RML. The enrichment scores are plotted from both transitions, $\{b \mapsto p\}$ and $\{p \mapsto m\}$. Each column represents a sample. As such, the samples from the primary stage are plotted twice since they are used in the enrichment analysis of both transitions. For each transition, the samples within the primary stage are plotted in the same order. See Table 5.1 for full pathway names.

Pathway gene sets known to be involved in self-sufficiency in growth signals were found to be relevant over all stages of progression. In the list of significant gene sets relevant through both stage transitions several cell cycle gene sets are present. These gene sets correspond to general function required for the cell cycle to run normally. For instance, the cell cycle checkpoint gene set defined by Gene Ontology is found to be relevant. This gene set is comprised of genes involved in checkpoints for the spindle formation, DNA integrity, G2/M transition, G1/S transition, cell size, and meiotic recombination. The relevance of this gene set gives an idea of the general perturbations intrinsic to the tumor cell. However, a more detailed understanding of the mechanism for tumor progression and proliferation can be gained by studying relevant gene sets corresponding to particular pathways. Both the ErbB4 pathway and EGF receptor signaling pathway are found to be enriched in all stages of tumor progression. These two pathways are related. ErbB4 (or HER4) is a member of the EGF receptor family and is involved in growth signaling. Therefore, it is reasonable to hypothesize that one of the underlying mechanisms in self-sufficiency in growth signals may be the deregulation of the EGF cell surface receptor ErbB4, which results in the increased transduction of growth stimulatory signals into the cell. Further evidence of the importance of EGF signaling is seen with the ERK pathway. Like the EGF pathway, the ERK pathway is involved in RAS signaling and thereby increases proliferation and development [125]. The behavior of the Hormonal Function gene set is of great interest. The enrichment of this gene set decreased in the prostate cancer to metastatic transition. This trend has been documented in prostate cancer where during progression the tumors become androgen independent resulting in decreased androgen production in the cell [43].

Pathways corresponding to insensitivity to anti-growth signals were also found to be relevant in our analysis. Throughout progression, the PTEN pathway was shown to have decreasing enrichment and the PTDINS pathway was shown to be up-regulated. These two pathways are closely linked. PTDINS signals PTEN which is a tumor suppressor gene that is often mutated in prostate cancer. PTEN down-regulation, as well as PTDINS activation, leads to the activation of the AKT pathway and anti-apoptotic activity [43].

Pathways corresponding to the evasion of apoptosis were found to be relevant at all stages of prostate progression. The TNF and FAS Network gene set and the ROS gene set were found to be significant. FAS is a member of the TNF receptor family normally involved in cell death. The down-regulation of this pathway seen in late progression can be implicated with anti-apoptotic activity. ROS was shown to be down-regulated throughout progression. ROS generation in mitochondria activates caspase-3 and leads to a pro-apoptotic activation cycle [26]. The deregulation of ROS therefore may be an anti-apoptosis mechanism important in prostate cancer.

Although pathways corresponding to limitless growth were not as prominent as the first three hallmarks, we do see one of the classic markers of this characteristic, HTERT upregulation. The length of telomeres indicate the number of cell generations elapsed. With each generation, telomeres lose 50-100 base pairs and when the telomeres get to a critical length, the cell undergoes senescence [63]. The gene set for the telomerase enzyme HTERT is shown to be up-regulated in late progression. HTERT over-expression maintains telomere length, thereby enabling unlimited replicative potential [129].

The analysis does not reveal specific angiogenesis gene sets relevant to progression. Previous research has found hypoxia signatures associated with poor prognosis of many cancer types including breast and ovarian, yet not prostate cancer [22]. The analysis does however indicate many metabolic pathways, giving evidence for the cell's need for additional nutrients as the cell grows and proliferates.

It has been previously documented that the capacity for invasion is a multi-step process and the process of acquiring the numerous needed perturbations often begins early in the disease process [45]. The results support this finding, indicating several pathways associated with invasion to be relevant early in progression. For example, the mCalpain pathway was found to be up-regulated in the benign to PCA transition.

This pathway is associated with invasion through the regulation of integrin-mediated cell migration [12].

An interesting observation we find is that many gene sets involved in energy production rank highly in the analysis. There is a glycolysis pathway significant to the overall progression and two additional glycolysis pathways relevant to the late transition. In the early transition, the oxidative phosphorylation gene set is also up-regulated. The up-regulation of each of these gene sets describes the cell's continual need for energy as it undergoes additional replications and eventually invasion and metastasis.

5.2 Preliminary Validation of Relevant Pathways

We carried out two analyses to validate our results from the above section. The first validation metric was the leave-one-out cross-validation error. The accuracy was 96.11%, giving confidence to the robustness of the RML model.

We next performed a preliminary validation of the pathway gene sets found to be relevant in our analysis by examining their enrichment in an independent prostate cancer dataset [131]. This dataset consisted of local and metastatic prostate cancer samples. The enrichment of the same 522 gene sets was computed using GSEA with respect to the local versus metastatic class labels. The validation metric was the overlap in the top 20 gene sets from the ranked list obtained from the NES from GSEA to the top 20 gene sets from the ranked list of gene sets found to be most relevant in the prostate cancer to metastatic transition. There were 8 gene sets in the overlap corresponding to a probability of 9.2×10^{-8} chance occurrence using the hypergeometric distribution. Included in this overlap are several cell cycle control and energy production gene sets. This result gives confidence that the gene sets identified in our analysis are indeed significant for disease progression. We again see the utility of using gene sets instead of single genes as a similar analysis done with single genes resulted in no overlap.

5.3 Novel Findings

In addition to recapitulating previous findings for prostate cancer progression, new hypotheses have emerged out of our study. One interesting observation from the analysis is the relevance of ErbB4 in the early transition, suggesting the hypothesis that the ErbB4 pathway becomes deregulated in PCA individuals. The importance of EGF receptors have been documented in several cancer types, such as stomach, brain, and breast tumors [148, 121]. HER2 has been implicated to be over-expressed in prostate cancer samples and thought perhaps to be related to the development of androgen independence [66]. However, currently little is known about a connection between ErbB4/HER4 and prostate cancer. Our analysis can motivate further research on defining the role of ErbB4/HER4 and prostate cancers.

Another interesting finding is the importance of the FA gene set in the early transition. The FA protein complex is known to associate with BRAC1 and BRCA2 and be involved in cell repair machinery [106, 53]. Alterations in FA protein complex has been linked to several types of cancer including breast, brain, lung, and oral cancer [95, 106].

Chapter 6

Melanoma Progression

In this chapter, the analysis pipeline described in Chapter 4 is applied to melanoma progression. Genome-wide expression at different stages of melanoma development is available in [129]. Samples were categorized as normal (n), primary (p), or metastatic (m), with 4 individuals per group. The progression is normal to primary to metastasic, $\{n \mapsto p \mapsto m\}$.

6.1 Pathways relevant in melanoma progression

The enrichment of gene sets with the lowest p-values for relevance throughout disease progression are displayed in Figure (6.1A), those most relevant in the transition from normal to primary are displayed in Figure (6.1B), and those for the transition from primary to metastatic are displayed in Figure (6.1C). See Table 11.3 for the full listing of pathways relevant with a $p \leq 0.05$. The top two ranked gene sets for each transition are listed in Table 6.1.

The results provide an understanding at the pathway level of the deregulation occurring in a developing melanoma cell to achieve the capacity for self-sufficiency of growth. Two well documented factors influencing self-sufficiency of growth are relevant to overall progression. The oncogene E2F is continuously over-expressed, increasing cell proliferation. Additionally, the AKT pathway is up-regulated which promotes cell survival and represses apoptosis. The AKT pathway is one of the effector pathways of RAS and can be activated by PTDINS over-expression, which we also see as relevant throughout melanoma progression. Our analysis suggests that the Sprouty pathway, which normally blocks EGF signaling [38] and prevents the activation of the RAS signaling pathway [59] is down-regulated.

The second hallmark of cancer, insensitivity to anti-growth signals, is also seen in melanoma. In the early transition, the CDK5 pathway is down-regulated. This alteration has been documented to lead to the activation of MEK1, a member of the MAP Kinase RAS effector pathway, and thereby to increased cell growth [125].

Several mechanisms for the escape from apoptosis are seen to be important in progression. Repression of apoptosis occurs through D4-GDI and IGF1R pathway over-expression and P53 and FA pathway under-expression. Decreased expression of the tumor suppressor P53 pathway throughout progression brings about unregulated continuation through the cell cycle, even when DNA damage is present [114]. The FA pathway has been implicated in several cancer types [95, 106] but not yet reported in melanoma. Finally, the IGF1R pathway is found to be relevant to the late transition where the over-expression increases anti-apoptotic activity.

During melanoma progression, mechanisms involved in the defense against limitless replicative potential become deregulated. The HTERT gene set is shown to be up-regulated in early progression. HTERT over-expression maintains telomere length, thereby enabling unlimited replicative potential [129]. The fact that this alteration is seen in early disease development points to the importance of acquiring the capability for limitless growth in order to proceed into a neoplasia in addition to deregulation in other growth mechanisms.

Two important gene sets describing angiogenesis are seen during melanoma progression. The HIF pathway gene set shows over-expression throughout progression, indicating the neoplasia's continual increasing requirement for oxygen and nutrients. HIF-1 is activated during hypoxia leading to the induction of a network of genes involved in angiogenesis and glucose metabolism [123]. We again see the need for an increase in oxygen early in the disease by the Angiogenesis gene set that is found to be up-regulated at the early transition.

The final hallmark of cancer, invasion and metastasis, is acquired in melanoma progression. Cells gain the capacity for metastasis by alterations in proteins involved in cell-to-cell adhesion and cell-to-environment interactions [23]. We see proteins of this class to be altered through the deregulation of the Cytoskeleton gene set. The MTA3 pathway is down-regulated in the early transition. Decreased expression of MTA3 has been associated with ER-negative breast tumors. This down-regulation leads to the halt of E-cadherin production. E-cadherin is involved in coupling adjacent cells and transmitting anti-growth signals and is a widely observed alteration in cancers with invasion capacities [51].

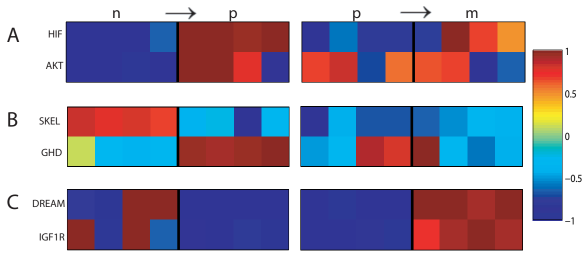


Figure 6.1: Enrichment scores of the two highest ranked gene sets from RML for A) $\{n \mapsto p \mapsto m\}$ B) $\{n \mapsto p\}$ and C) $\{p \mapsto m\}$. See Figure 1 for further details.

Set ID	Gene Set Name	Source
HIF	Hif Pathway	BioCarta
AKT	Akt Pathway	BioCarta
SKEL	Cytoskeleton	Brentani et al
GHD	MAP00361 Gamma Hexacholrocyclohexane Degradation	GenMAPP
DREAM	Dream Pathway	BioCarta
IGF1R	Igf1r Pathway	BioCarta

. **Table 6.1**: Gene set names for the gene set IDs in Figure 6.1

6.2 Prelimary Validation of Relevant Pathways

We carried out two analyses to validate our results from the above section. First, classification was applied in the leave-one-out setting. There were no errors in class prediction, giving confidence to the robustness of the RML model.

The second analysis involved an independent melanoma progression dataset [62]. This dataset consists of 9 samples of benign nevis, 6 samples of primary melanoma, and 19 samples of metastatic melanoma. The regression analysis was rerun on this dataset and several of the relevant pathways from the first analysis were again found to be significant. In both analyses we found the down-regulation of the P53 pathway gene set throughout progression and the over-expression of the D4-GDI pathway and the HTERT gene set at various stages in progression. The overlap of relevant gene sets from the two datasets [129, 62] however was not found to be significant

($p = 0.28$). This may be attributed to the small sample size in the first dataset [129]. To further explore this issue, we randomly split the second dataset [62]. The regression analysis was run on the two resulting datasets independently. For each, we obtained a list of gene sets with $p \leq 0.05$ for overall progression. Using the hypergeometric distribution, the overlap of the two lists was found to be significant, $p = 0.018$. This result gives confidence against overfitting and that the gene sets identified in the analysis are indeed significant.

Chapter 7

Colon Cancer Progression

Colorectal cancer is the second leading cause of cancer death in the United States today. In 2008, in the United States alone, it is estimated that 160,000 people will be newly diagnosed with colorectal cancer, making this an important cause of morbidity and mortality. Of newly diagnosed cases, approximately 40% are diagnosed while the cancer is in an early or localized stage [75].

Fearon and Vogelstein [41] presented a model of colorectal cancer progression, identifying times at which specific genes or chromosomal regions tend to experience perturbations during the course of the disease. This classic study laid the foundation for the concept of accumulation of genetic alterations as tumors progress to a malignant state. The study emphasized the concept of tumor heterogeneity, more fully described in recent studies that comprehensively monitored the cancer genome for genomic imbalances and associated gene expression changes [127, 147]. Amongst the multiple genetic perturbations associated with colon cancer initiations and progression, KRAS, TP53, and APC are the three most commonly described genes.

However, it is likely that the complexity of the genomic landscape of colorectal cancer far surpasses single gene abnormalities. Genome-wide expression analysis in the form of deregulated oncogenic pathways offers the opportunity to characterize and treat tumors in an individualized fashion and will likely extend the elegant genetic model presented by Fearon and Vogelstein [41, 109].

The KRAS, TP53, and APC gene mountains are difficult to target by drug treatments and using these genes as a basis for new treatments has proven unsuccessful. Analyses on the single gene level have not been able to isolate any of the less frequently mutated genes, or gene hills, as significant in colorectal cancer. Gaining a deeper understanding of the composition of gene hills into biological pathways that are relevant to colon cancer initiation and progression may lead to identification of novel targets for drug treatment. We investigated regulatory pathways that become perturbed at different stages of colorectal cancer development in order to identify new "pathway mountains", pathways that become deregulated in the majority of colorectal cancer samples. The result is the identification of biologic pathways that are not only important in colon tumorigenesis, but also represent rational targets for therapeutic strategies.

7.1 Data Collection, Standardization, and Stratification

The data consisted of 32 samples of normal colon epithelium (n), 32 samples of colon adenoma (a), 35 samples of stage 1 carcinoma (c_1), 82 samples of stage 2 carcinoma (c_2), 70 samples of stage 3 carcinoma (c_3), and 43 samples of stage 4 carcinoma (c_4), all collected from NCBI's Gene Expression Omnibus (GEO). The reference series for the data are GSE5206, GSE2138, GSE2109, GSE2461 and GSE4107. All samples were assayed on the Affymetrix Human Genome U133 Plus 2.0 Array.

These samples are a collection from studies that used a variety of normalization methods and processing. For this reason we preprocessed the data using quantile normalization so that the distribution of expression measurements over the probes on the array are comparable across samples.

Previous studies have shown that there is a genetic basis for the differences between normal, adenoma, and carcinoma colon cancer samples [147, 127, 141]. We preformed hierarchical clustering on the normal, adenoma, and stage 1 carcinoma samples. The results confirm that these three classes can be differentiated based on expression profiles (see Figure 7.1). This separation suggests that different biological processes may be relevant in different stages of progression and provides a logical basis for stratifying the samples into the categories of normal (n), adenoma (a), stage 1 carcinoma (c_1), stage 2 carcinoma (c_2), stage 3 carcinoma (c_3), and stage 4 carcinoma (c_4).

The a prior defined gene sets used in our model consist of 511 functional pathway gene sets and 286 chromosome position gene sets as defined by MSigDB [132]. Using the methods described in Chapter 4, we inferred a list of functional pathways and positional gene sets relevant across all stages of progression and those relevant to specific stages.

7.2 Stratification and Mapping to Gene Sets

The analysis pipeline closely follows that outlined in Chapter 4. However, for later validation purposes, we initially split each dataset in half where the first half was used to build the model (training data) and the second half was used to validate the model (testing data). The first two steps in the analysis are stratifying the data and mapping the data into a representation based on pathways. These steps were applied to both the training and testing datasets.

Again, the data can be represented as a set of pairs $D = \{(x_i, y_i)\}_{i=1}^{n}$ where $x_i \in \mathbb{R}^p$ is the expression over p genes and y_i is the disease stage of the patient. However, instead of three stages, there are now six stages, $y \in \{n, a, c_1, c_2, c_3, c_4\}$ with n_1, n_2, n_3, n_4, n_5, n_6 samples in each stage of the training data and n_7, n_8, n_9, n_{10}, n_{11}, n_{12} samples in each stage of the testing data. The progression is $\{n \to a \to c_1 \to c_2 \to c_3 \to c_4\}$. There are five steps in this progression, $T = 5$.

We first stratify the train and test datasets with respect to these five steps. The first dataset $D_1 = \{(x_1, y_1)\}_{i=1}^{n_1+n_2}$ consists of the n_1 training samples corresponding to stage n followed by the n_2 training samples corresponding to stage a with the label of the first n_1 samples labeled as 0 (less serious) and the remaining n_2 labeled as 1 (more serious). D_2 is constructed similarly, consisting of training samples corresponding to stage a followed by stage c_1 and again labeled as 0 and 1 respectively for less and more serious. Likewise D_3 consists of the c_1 and c_2 training samples, D_4 consists of the c_2 and c_3 training samples, and D_5 consists of the c_3 and c_4 training samples. The same procedure was applied to the testing samples to create $D_6, ..., D_{10}$.

Each dataset D_t is then mapped into a representation with respect to sets of genes or pathways. As before, this is done using the pathway annotation tool ASSESS [36] that assays pathway variation in individuals. Given phenotypic label data $Y_n = \{y_1, ..., y_n\}$, expression data $X_n = \{x_1, ..., x_n\}$ and gene sets $\Gamma = \{\gamma_1, ..., \gamma_m\}$ defined a priori, ASSESS provides the summary statistic $S_n = S(X_n, Y_n, \Gamma)$. The summary statistic S_n is a matrix with n columns corresponding to samples and m rows corresponding to gene sets with each element S_{ij} as the enrichment of gene expression differences in the j-th sample with respect to phenotype for genes in the i-th gene set. The application of ASSESS to the stratified datasets $D_1, ..., D_{10}$ results in ten datasets $S_1, ..., S_{10}$. The analysis was preformed twice, once using the gene sets in the MSigDB Functional Gene Sets and then again using the gene sets in the MSigDB Positional Gene Sets.

7.3 Validation via Mann-Whitney Test

After running the RML algorithm on the data described above, we selected the top 50 functional gene sets from each list from the training data. We tested the significance of these lists of 50 gene sets by applying the Mann-Whitney test on the lists generated from the testing datasets. The 50 top gene sets from the training data were labeled as class 1 and the remaining gene sets were labeled as class 2. The Mann-Whitney test was used to assess whether class 1 and class 2 had the same distribution in the testing data. A p-value of ≤ 0.05 validated that the top 50 gene sets from the training data were also relevant in the testing data.

As mentioned early, the same procedure was used for both the functional and positional gene sets. The only difference at this step was that the top 50 functional gene sets were selected and top 40 positional gene sets were selected. We selected these numbers as they were the smallest numbers than returned p-values ≤ 0.05 for each stage transition, see Table 7.1. See Table 11.4 and Table 11.5 for listings of the 50 significant functional and 40 significant structural gene sets respectively.

We further tested for significance by randomly selecting a list of 50 gene sets for the functional gene set analysis and 40 gene sets for the positional gene set analysis, labeling them as class 1 and the remainders as class 2, and preformed the Mann-Whitney test on the same test data as described above. We repeated this randomization 1000 times and calculated the p-value from the Mann-Whitney test for each randomization. These p-values were considered the null distribution of p-values from the Mann-Whitney test and we found where the true p-value fell within this null distribution. In all cases, the true p-value fell in the lower 5% of the null distribution, giving confidence in the significance of the results (see Table 7.2).

7.4 Model Statistics and Validation

We identified pathways and chromosomal regions deregulated at specific transitions in colorectal cancer as well as pathways and chromosomal regions deregulated over multiple stages of progression using the RML algorithm. These gene sets were identified using half of the samples in each class and their significance was then validated on the second half of the data using the Mann-Whitney test described above. The resulting p-values are displayed in Table 7.1. The gene sets selected are significant at the 0.05 level for all transitions, indicating that they are indeed relevant for modeling colorectal cancer progression.

The robustness of the inference was measured using two metrics. First, we tested the accuracy in predicting disease stage phenotype. A leave-one-out cross validation analysis resulted in 100% accuracy in the prediction of the normal, adenoma, and stage 1 carcinomas and 81.5% for prediction of carcinoma stages 2 through 4. For the second metric we tested the variability in the pathways selected as significant in each stage and across stages. We split the data in half and applied our procedure to both splits and computed the overlap between gene sets found to be significant at the 0.05 level in the two splits respectively. The number of overlapping gene sets was found to be significant for both individual stages and across all stages of progression for both positional and functional gene sets, see Table 7.3 for p-values. The p-values were computed from the hypergeometric distribution.

A confounding factor in pathway level analyses is the fact that since genes may appear in multiple pathways it is difficult to attribute relations between pathways to mechanism or function rather than how many genes overlap between pathways. This confound is typically not an issue. For example, we found that the KRAS and EGF pathways are both significant in progression of colorectal cancer without any genes overlapping. In fact of all the pathways involved in our model of progression the only ones that shared more than 2 genes in common were the Gleevec, PTEN, and PDGF pathways, which consisted of 22, 18, and 27 genes respectively. The overlaps between the EGF pathway and the Gleevec pathway was 6 genes, as was the overlap with the EGF pathway and the PDGF pathway. The Gleevec and PDGF pathways have 15 genes in common.

7.5 Comparison of Single Gene and Gene Set Models

Our gene set based model is compared to the single gene model of colorectal cancer progression proposed by Fearon and Vogelstein [41]. There is strong agreement between our gene set based model and the Fearon and Vogelstein model as shown in Figure 7.2.

The first component of the single gene model is a mutation in the KRAS gene, located on chromosome 12p12. It had been previously reported that approximately 50% of colorectal tumors show a RAS mutation by the time they reach the carcinoma stage [141] and the mutation can potentially act as an initiating event or a driver of an adenoma into a carcinoma [41]. Likewise, we found that the KRAS pathway gene set is increasingly expressed from $\{n \rightarrow a \rightarrow c_1\}$. Additionally, gene sets representing the chromosomal region of KRAS, along with surrounding regions were found to be deregulated. 12p11 and 12p13 are found to be overexpressed early in progression, $\{n \rightarrow a \rightarrow c_1\}$, followed by 12p12 over-expression as the tumor progresses during carcinoma, $\{c_2 \rightarrow c_3\}$. We also found the deregulation of pathways that are closely associated with RAS. For instance, the EGF pathway is up-regulated during the transition from $\{c_2 \rightarrow c_3\}$. Similarly associated with the RAS pathway is the PTDINS pathway which is up-regulated from $\{c_3 \rightarrow c_4\}$. PTDINS is known to act as an activator of AKT, an effector pathway of RAS. PTDINS also signals PTEN which is a tumor suppressor gene and is shown to be downregulated from $\{c_2 \rightarrow c_3\}$.

It was observed by Vogelstein et al. [141] that chromosome 17p, where TP53 is located, is deleted in 75% of colorectal cancer cases. The Fearon and Vogelstein model [41] includes a mutation to 17p in late adenoma to early carcinoma. In our analysis, we found that a P53 gene set is inactivated throughout the progression from $\{c_1 \rightarrow c_2 \rightarrow c_3 \rightarrow c_4\}$. We also found the chromosome 17p gene set to be lost in early carcinoma, $\{c_1 \rightarrow c_2\}$.

Another tumor suppressor gene included in the Fearon and Vogelstein model [41] was the APC gene, located on chromosome 5q. APC is the most commonly acquired mutation in sporadic colon cancer, occurring in 60% of colorectal carcinomas [113]. We found the gene set for chromosome 5q21 was lost at the tumor initiation stage, $\{n \rightarrow a\}$, consistent with previous studies [41]. At the pathway level we identified regulatory pathways that associated with allelic loss of 5q and APC. APC is a tumor suppressor gene that suppresses the oncogene β-catenin. APC and β-catenin are components of the WNT signaling pathway which we found up-regulated during the entire course of progression, $\{n \rightarrow a \rightarrow c1\}$ and $\{c1 \rightarrow c2 \rightarrow c3 \rightarrow c4\}$. When this pathway is activated, the APC complex which normally degrades β-catenin is inhibited, allowing for β-catenin build up and uncontrolled cell proliferation.

The final major component of the Fearon and Vogelstein model was a mutation in 18q in the late adenoma stage. Although there is not a regulatory pathway associated with this chromosomal region, we found the allelic loss of the chromosome 18q gene set in the $\{a \rightarrow c_1\}$ transition. This region is found to be lost in more than 70% of colorectal cancers by the time they reach the carcinoma stage [41].

In addition to reiterating the Fearon and Vogelstein model on the pathway level, our analysis agrees strongly with the gene mountains described by [147] where it was observed that KRAS, TP53, and APC had at least one nonsynonymous mutation in each of 131 colorectal tumors in the study, with APC showing the most mutations per sample.

7.6 Novel Pathway Findings and Therapeutic Implications

The difficulty in using single gene models is the fact that there are very few genes represented uniformly, 'gene mountains', and the vast majority of genes implicated in tumor progression, the 'gene hills', are mutated in only a fraction of individuals. It is essential to understand these mutations as they make up the greatest part of the genomic landscape of colorectal cancer. By integrating the gene hills into functional pathways or positional gene sets our objective is to identify new mountains on the pathway level, or "pathway mountains". Another strong motivation of our analysis was to find novel pathways that can be targeted with current drug therapies. The genes strongly implicated in single gene models such as KRAS, TP53, and APC have been difficult to target with current drug therapies. For these reasons we focused in greater detail on gene sets found in our analysis that have not typically been associated with colorectal cancer in order to identify new pathways as potential targets for novel therapy.

The sonic hedgehog (SHH) pathway was found to be up-regulated from $\{c_3 \rightarrow c_4\}$. The SHH pathway has been implicated in the development of several cancer types such as specific brain and skin carcinomas [10, 28], but rarely with colorectal cancer. In a recent study [115], colorectal cancer cells were treated with cyclopamine, an inhibitor of SHH signaling. Both adenomas and carcinomas experienced apoptosis. Additionally, our analysis has shown the PTC1 pathway to be deregulated in the $\{n \rightarrow a\}$ transition. The PTC1 pathway describes the genes involved as PCT1, a receptor in the SHH pathway, regulates the cell cycle. WNT, SHH, and PCT1 often act together in controlling cell growth. Our analysis identified these pathways to be deregulated, suggesting that targeting these pathways may be beneficial for colorectal cancer treatment. Our inference that the PTC1 pathway is deregulated before the SHH pathway suggests a basis for the order of events during tumorigenesis.

Finally, of interest is the upregulation of PDGF, PTEN, and Inflammatory Response Pathway during the transition into stage 1. PDGF and PTEN are components associated with PI3K signaling. Importantly, unlike the SHH, RAS, or P53 pathways, both PI3K and the inflammatory mechanisms are potential targets for therapeutic intervention. There are currently several PI3K inhibitors being tested in clinical trials and likewise anti-inflammatory drugs such as sulindac and cyclooxyenase-2 (COX2). These drugs have been shown to benefit in inhibiting colon cancer initiation and/or progression [5, 119].

7.7 Pathway Model Validation using a Recurrence Signature

The previous section has shown that a better understanding of colon cancer biology can be ascertained using a pathway approach. Such a strategy may also lead to the identification of novel targets for therapeutic intervention in early stage colon cancer such as the PI3Kinase pathway and the inflammatory respond pathway. However, it is critical to reiterate that there is still a significant unmet need to further characterize and treat tumors in an individualized fashion. Using the information from the previous section along with experimental validation can help identify those patients at the greatest risk of disease recurrence and facilitate the further deployment of therapeutic strategies. This is particularly relevant for patients diagnosed with early stage colon cancer (stages 1 and 2) who are usually considered cured after surgical resection, despite the fact that up to 25% of these pateints later develop disease recurrence [4, 68].

A separate study was performed by Garmen et al. [54] whose results help validate the pathway model of colorectal cancer. Using Bayesian regression methods, [54] identified gene expression profiles, or metagenes, constituted by 50 genes that predict the risk of recurrence in an initial training cohort of 52 patients with early-stage colon cancer (GSE5206 and GSE2138). The predictive accuracy of the 50-gene model was assessed using leave-one-out cross-validation. The 50-gene model predicted recurrence with an accuracy of 90.3% with significant difference in the predicted probability of recurrence or 'recurrence scores' between the two groups ($p < 0.001$).

The prognostic capability of the metagene model was assessed by applying it to an independent validation dataset comprising 73 colon cancer patients (GSE10402) using the predetermined Recurrence Score cut-off of 0.76. Nearly all the patients with recurrence were predicted by the model to recur resulting in a sensitivity of 90%. In a Kaplan-Meier survival analysis, statistically significant differences were seen between the cohorts predicted to recur (high recurrence score) and those predicted to be disease free (low recurrence score).

The Connectivity Map [85], a project developed at the Broad Institute to assemble a reference collection of gene-expression profiles from cells that have been treated with a variety of drugs, was used to identify drugs that might be connected to the 50-

gene colon cancer reccurence signature. The top four candidate drugs identified by this approach included Tretinoin (a retinol analog), the PI3K inhibitor LY-294002, sulindac, and celecoxib. Interestingly, COX2 inhibitors like celecoxib have been repeatedly identified in the literature as potential agents for reversing polyp growth, particularly in familial adenomatous polyposis syndromes [5, 119, 11], although their role in prevention of colon cancer recurrence has not been explored. These initial results give further confidence of our pathway model, as both methods indicate the PI3K signaling pathway and the inflammatory response pathway in early carcinoma.

In [54], the potential of these agents in the high risk population of patients with early stage colon cancer were evaluated by mapping the Recurrence Score (using the 50-gene model) to a collection of 14 colon cancer cell lines so as to classify the cell-lines as representative of the high-recurrence risk phenotype (with high Recurrence Scores) or the low-recurrence risk phenotypes (with low Recurrence Scores). Using in vitro cell proliferation assays, colon cancer cell lines were treated with celecoxib, LY294002, retinol, and sulindac. Biologically relevant differences in cell death between cell lines with high and low Recurrence Scores were observed. These results were statistically significant for celecoxib ($p=0.01$), LY294002 ($p=0.008$), and retinol ($p=0.01$).

Combining the results of the pathway model and the information from the proliferation assays can help provide more effective therapy. This study has shown that using gene expression methodologies to understand the molecular mechanisms involved in cancer progression maybe helpful beyond prognosis since this knowledge may lead to the study of drugs that target relevant, deregulated pathways in an individual patient. More importantly, we may be able to identify specific, effective agents from a repertoire of currently existing drugs.

7.8 Discussion

Colorectal tumorigenesis is an example of a complex trait as it is controlled by many genes that do not interact additively and the relationship between genetic and phenotypic variation is non-linear. Standard genetic models of tumor progression are based on a multi-step process formulated as mutations and interactions of single genes – a framework common to many models of complex traits. A fundamental motivation for the development of these models is to develop effective therapeutic strategies based on an understanding of the heterogeneity across tumors and inference of characteristics of individual tumors that are prognostic of disease outcome. Conventional methods of characterizing tumors are primarily based on visual information including tumor size, degree of metastasis, and histological tumor characterization. These methods are relatively arbitrary and do not always decompose the heterogeneity of the disease with respect to prognosis. To this extent gene expression profiles, which represent biological states in the form of patterns of gene expression, offer a more flexible characterization and stratification of predictive features or signatures which affords greater opportunity to characterize and treat tumors in an individualized fashion [57, 109].

In this chapter, the general strategy for modeling tumor progression across stages introduced in Chapter 4 was applied to colon cancer which presented opportunities for clinical intervention. Using this approach the colorectal cancer progression model presented by Fearon and Vogelstein [41] was extended from the single gene level to the pathway level. Many of the findings in the Fearon and Vogelstein model such as mutation of the "gene mountains" KRAS, P53, and APC were recapitulated at the pathway level. In addition, we inferred the stage in progression that these pathways drive tumorigenesis: KRAS pathway was up-regulated across early progression, P53 pathway was downregulated across carcinoma progression, and APC was implicated throughout all of progression through WNT signaling up-regulation.

The ability of our approach to identify a small number of pathways important to a majority of tumors resulted in the novel association of pathways to progression in colorectal cancer. An implication of our approach is that although single genes may be greatly variable across tumors, groups of genes corresponding to pathways integrate the gene hills into a pathway mountains which become deregulated in the majority of colorectal cancer samples. Some of the novel pathways we identify suggest both drug agents and at what stages of progression the agents would be most efficacious.

Further, in the clinically relevant instance of early stage disease, gene expression signatures consistent with disease recurrence are inferred. These signatures are then used to suggest therapeutic agents for individuals with the high-risk gene expression phenotype. Evidence for the efficacy of this modeling approach is provided by demonstrating the reversal of the high-risk phenotype in cell lines by application of the therapeutic agent(s) inferred by the approach. Such a strategy is fundamentally hierarchical in that the models proceed from coarse to finer levels as we proceed from modeling cancer progression to modeling risk of recurrence. Specifically, the progression model is a conjoint analysis across all stages of the tumor and the elements of the model are constitutive expression of groups of genes corresponding to pathways or chromosomal locations. Informed by the progression model a refined model focusing on particular stage specific transitions (stage 1 and 2 disease), based on the expression of individual genes is proposed for prediction of the high-risk phenotype. Finally, such an approach also provides a rational strategy for identifying effective therapeutic agents that have the potential to reverse the phenotype associated with a high recurrence risk.

Using in vitro cell proliferation assays, we demonstrate a clinically relevant relationships between sensitivity to certain therapeutic agents and a high-risk gene expression phenotype, in colon cancer. Celecoxib, LY-294002 (PI3Kinase inhibitor), retinol, and sulindac may all serve as potential therapeutic agents in the treatment of stage 2 colon cancer. The strong correlation between sensitivity to COX2 (celecoxib) and PI3Kinase inhibition (LY-294002) in tumors expressing the high risk phenotype and the ability to reverse the profile representative of the high risk phenotype in cell lines illustrates the robustness of this approach.

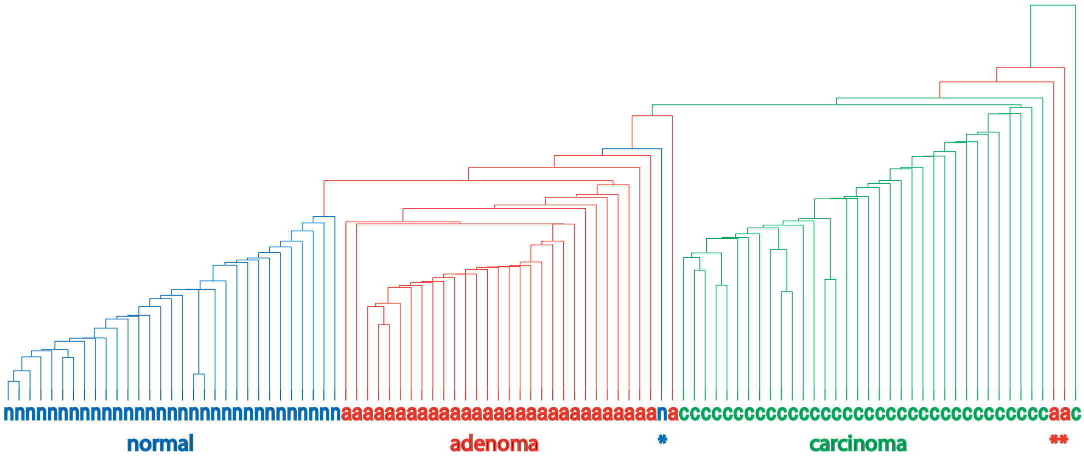


Figure 7.1: Hierarchical clustering on the gene expression of 32 normal (labeled "n", colored blue), 32 adenoma (labeled "a", colored red), and 35 stage 1 carcinoma (labeled "c", colored green) samples. With the exception of 1 normal sample and 2 adenoma samples (indicated with asterisks), all samples cluster with their respective cell type, validating that there is indeed a genetic basis for the difference between the cell types.

Stages	Positional	Functional
$n \rightarrow a \rightarrow c_1$	5.31×10^{-6}	2.11×10^{-16}
$c_1 \rightarrow c_2 \rightarrow c_3 \rightarrow c_4$	0.038	0.031
$n \rightarrow a$	5.80×10^{-9}	1.18×10^{-25}
$a \rightarrow c_1$	1.67×10^{-8}	5.05×10^{-21}
$c_1 \rightarrow c_2$	0.046	0.011
$c_2 \rightarrow c_3$	0.0043	0.049
$c_3 \rightarrow c_4$	4.01×10^{-8}	2.34×10^{-9}

Table 7.1: Significance of the gene sets identified in the analysis. The top 40 positional (chromosomal region) gene sets and the top 50 functional (pathway) gene sets were identified from the first half of the data for each stage and p-values were calculated using the Mann-Whitney rank test on the second half of the data. 286 positional gene sets and 511 functional gene sets were used in the analysis. Rows indicate the p-values of the sets identified over multiple stage transitions or individual stage transitions. Samples are classified as normal (n), adenoma (a), stage 1 carcinoma (c_1), stage 2 carcinoma (c_2), stage 3 carcinoma (c_3), and stage 4 carcinoma (c_4).

	Positional Sets	Functional Sets
N→A→C1	0	0
C1→C2→C3→C4	0.03	0.03
N→A	0	0
A→C1	0	0
C1→C2	0.01	0.01
C2→C3	0.05	0.02
C3→C4	0	0

Table 7.2: Proportion of times that the Mann-Whitney p-values from the randomized gene set lists were lower than the true p-value

Stages	Chromosomal Regions	Pathways
$n \rightarrow a \rightarrow c_1$	0.0069	2.00×10^{-4}
$c_1 \rightarrow c_2 \rightarrow c_3 \rightarrow c_4$	0.019	0.018
$n \rightarrow a$	8.23×10^{-6}	7.18×10^{-6}
$a \rightarrow c_1$	0.012	2.50×10^{-4}
$c_1 \rightarrow c_2$	0.025	0.017
$c_2 \rightarrow c_3$	0.032	0.016
$c_3 \rightarrow c_4$	0.011	0.0024

Table 7.3: P-values corresponding to significance of the overlap between the gene sets determined to be significant in two independent splits of the data. For each stage, the data is split in half and the overlap of the significant gene sets in both splits is computed. The p-value is computed using a hypergeometric distribution.

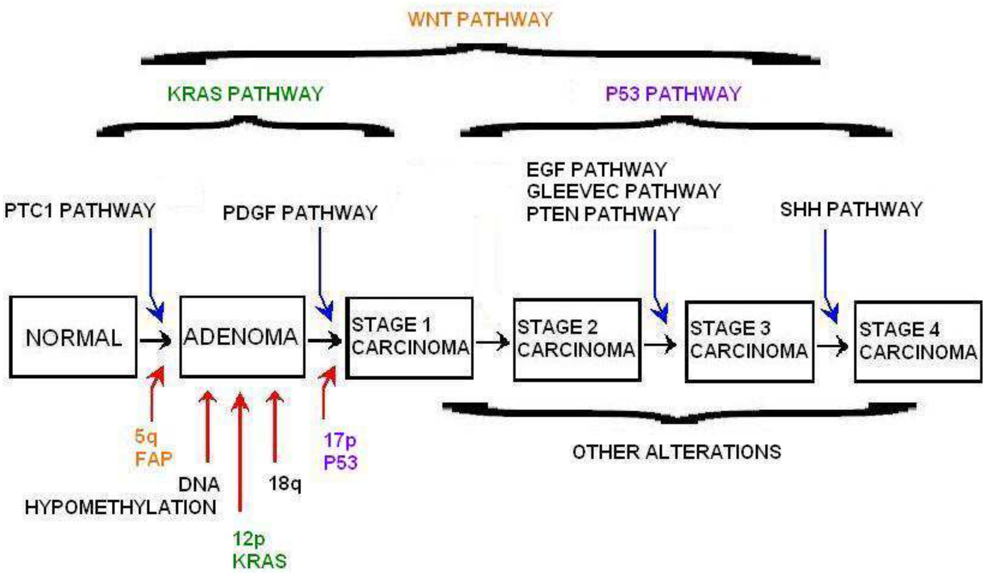


Figure 7.2: A model of the deregulation in colorectal cancer progression. Pathways are indicated at the top of the figure. Pathways identified to be relevant at specific points in progression are indicated with blue arrows. The KRAS, P53, and WNT pathways are shown to be relevant over multiple stages in progression. Single genes deregulated in the Fearon and Vogelstein model [10] are shown on the bottom of the figure with red arrows indicating the times in colorectal cancer progression when they tend to be mutated. The three gene mountains and corresponding pathways are color-coded in green, purple, and orange.

Chapter 8

Pathway Dependency Structures

In Chapters 5.1-7, pathways relevant to progression have been identified. It is of interest to know how these pathways relate to one another and how these relationships change over time. In other words, we would like to develop pathway networks for each step in tumor progression and see how the interdependence of pathways in the networks differ. We can take this analysis a step further by the refinement of relevant pathways at the gene level. Not all genes comprising a phenotypically relevant gene set may be equally important. The refinement procedure removes genes in relevant pathway gene sets that are not relevant to progression, resulting in a "refined" gene set. A gene network for each refined gene set can be inferred by estimating the interdependence between genes.

8.1 Construction of Pathway Association Graphs

The central statistical idea used in constructing association graphs as well as refining gene sets is learning gradients [104, 102, 103, 92] and inverse regression [89, 24].

These ideas apply to linear and nonlinear models but we restrict ourselves to the linear setting since we only use linear models for the cancer progression modeling.

8.1.1 Learning gradients and inverse regression

The idea of inverse regression [89, 24] is given the explanatory or input variables X and the output or response variable Y to study $X|Y$ and more specifically $\Omega_{X|Y} = \text{cov}\,(X|Y)$ where:

1. The i-th diagonal element of this covariance matrix is a measure of the relevance of the i-th variable with respect to changes in the label or output variable;

2. The ij-th off-diagonal element is a measure of the covariation between variable i and j with respect to changes in the label.

Estimating the inverse regression can be technically problematic if the covariance matrix is degenerate.

The idea of learning gradients addresses this technical problem. Given a regression or classification function f its gradient is

$$\nabla f = \left(\frac{\partial f}{\partial x^1}, ..., \frac{\partial f}{\partial x^p} \right)^T,$$

and the gradient outer product matrix (GOP) Γ is defined by its elements $\Gamma_{ij} = \left\langle \frac{\partial f}{\partial x^i}, \frac{\partial f}{\partial x^j} \right\rangle$. In [92] the following relation between the GOP matrix and the covariance of the inverse regression $\Omega_{X|Y}$ was derived for linear functions

$$\Gamma = \sigma_Y^2\, \Sigma_X^{-1} \Omega_{X|Y}\, \Sigma_X^{-1},$$

where $\Sigma_X = \operatorname{cov}(X)$ for the inputs, and $\sigma_Y^2 = \operatorname{var}(Y)$ for the outputs. This states that Γ and $\Omega_{X|Y}$ are equivalent modulo a scale parameter (the variance of the output variable) and a rotation (the precision matrix of the input variables). In [102] an efficient algorithm to estimate the gradient and GOP matrix, $\hat{\Gamma}$ given data for the classification setting was developed.

The estimate of the GOP matrix $\hat{\Gamma}$ is the covariance matrix of a multivariate Gaussian random vector by construction. The inverse of this matrix $J = \hat{\Gamma}^{-1}$ is by the theory of Gauss-Markov random fields [130] the conditional independence matrix (the pseudo-inverse is used when $\hat{\Gamma}$ is degenerate).

$$J_{ij} = \text{dependence between variables } i \text{ and } j | \text{all other input and output variables.}$$

8.1.2 Construction of pathway association graphs

Given the list G_t of pathways relevant to the t-th step in progression and the enrichment summary statistic dataset S_t, this dataset is reduced by removing the enrichment scores of all pathways not in G_t. The GOP estimate, $\hat{\Gamma}_t$, given this data is computed. This matrix is $d \times d$ where d is the number of relevant pathways in G_t. The pseudo-inverse of this matrix

$$A_t = \hat{\Gamma}_t^{-1},$$

is the pathway association graph for the t-th progression step.

8.1.3 Gene set refinement and gene association graphs

Not all genes in the relevant gene sets, G_t, are differentially expressed between the two stages of progression in step t. We reduce or refine each of the pathway gene sets to those genes most relevant in progression step t.

The following procedure is iterated for each of the relevant gene sets in G_t. Given the k-th pathway gene set, the stratified data D_t is reduced to the genes in this gene set. The GOP estimate, $\hat{\Gamma}^t$, given this data is computed. This matrix is $d \times d$ where d is the number genes in the k-th pathway in G_t. Genes corresponding to large diagonal elements of the GOP matrix, $|\hat{\Gamma}^t_{ii}| > \tau$, are those most relevant to the prediction and are the refined set R. In this paper, the threshold τ is selected such that we obtain a specific number of genes.

The GOP estimate, $\hat{\Gamma}^t$, is then reduced to only the genes in the refined gene set R. The pseudo-inverse of this matrix is the gene association matrix $\tilde{A}_k$ that provides the dependence of the refined gene set.

8.2 Pathway Dependency Structure in Prostate Cancer

Given inferences of pathways relevant to each step of tumor progression, the second objective is to understand how these pathways interact at specific stages of progression. The goal here is to infer a pathway interaction network for each stage of progression. This information will provide knowledge on the relationship between gene sets at a higher level– how certain pathways relate, interact, and influence each other with respect to the phenotype being studied, thereby creating a dependency structure of the pathways. The relationships are visualized by an undirected graphical model. We do not claim nor model causality in these graphical models, however we believe that causative relationships may be suggested that can be further experimentally tested.

We first look at three dependency relationships of interest. As previously discussed, the PTEN and PTDINS pathways are known to be closely associated. This is supported in the analysis by the dependency relationship between the two pathways. PTDINS is the 15-th ranked pathway out of $136,241$ possible relationships based on the covariance with the PTEN pathway throughout progression. Likewise, the IGF1R and ERK pathways are known to be linked through their common association with the RAS pathway. We find support for their association as the ERK pathway ranks 9-th based on the covariance with the IGF1R pathway during overall progression. Finally, PTDINS ranks 32-nd based on the covariance with IFG1R, recapitulating the known relationship between these two pathways. These findings support the ability of our analysis to identify pathways that are closely associated.

As previously discussed, p-values were calculated for each of the 522 gene sets in the analysis. We chose the ten gene sets with the lowest p-values in the benign to prostate cancer step in progression. The dependency structure of these gene sets are displayed in Figure 8.1A. Although these ten gene sets are all significant in differentiating the benign and PCA samples, the dependency structure shows that some may be more influential than others. For example, there is a group of 7 gene sets which cluster closely together, indicating that the deregulation of these genes sets play a key role in the phenotypic change from benign to PCA. We call these the core gene sets for the early step in progression. Gene sets with fewer connections are also important in phenotypic determination, however, the enrichment of these gene sets over all of the samples differ slightly from the core group. The assumption here is that one, or a few, of the gene sets in the core group are influencing the expression of the other gene sets, both in the core group and on the outside connections. The Sterol Biosynthesis and FA (Fanconi anemia) gene sets are both highly connected to the other gene sets. These are good candidates for pathways which are strongly influencing the expression pattern of other pathways.

Further information on gene set interactions can be gained by increasing the number of gene sets examined and displaying the dependence structure. For example those gene sets with $p \leq 0.05$ are shown in Figure 8.1B. Most of the gene sets discussed previously were relevant throughout progression and therefore do not appear in this network of sets describing the early transition. The few gene sets discussed that are relevant in the early transition do indeed appear in the network, such as the ERK pathway and Hormonal Function gene set. The gene sets which clustered tightly in Figure 8.1A still cluster tightly in the middle of Figure 8.1B. However, several additional gene sets are shown to have strong dependencies with these gene sets as well. The additional gene sets are important in the phenotypic transition but have slightly different enrichment patterns as the top 10 gene sets. Aside from the Sterol Biosynthesis and FA gene sets, the Hormonal Functions gene set has the largest number of connections, indicating that it may also be governing expression patterns. It is further apparent that gene sets from each of the hallmarks of cancer interact with one another. Therefore, although each hallmark represents a unique characteristic, they are not separate entities but rather influence one another.

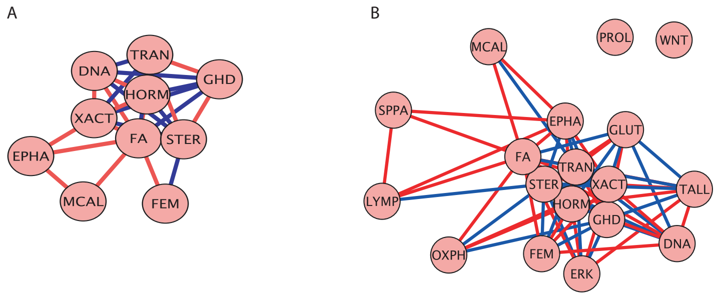


Figure 8.1: An association graph of A) the top 10 gene sets in the $\{b \mapsto p\}$ transition and B) the gene sets with $p \leq 0.05$ for the $\{b \mapsto p\}$ transition. The red lines indicate a positive dependence while the blue lines indicate a negative dependence. Distances between gene sets reflect the strength of dependencies.

8.3 Refinement of Pathways in Prostate Cancer Progression

The pathway level analysis provides us with inferences of which pathways are relevant over the two steps of prostate cancer progression as well as the interaction between pathways. The next level of analysis is the refinement at the gene level. The pathways found to be relevant in each of the steps of progression are refined using procedures described in the methods section. The HRAS/KRAS gene set validation example provides an argument for refinement. Pathways available are not always in the right context for a specific dataset of interest. The refinement procedure adapts the gene set to the context of the dataset. Adding this gene set to the database of gene sets increases the diversity of the collected gene sets. In addition, a gene network modeling the interdependence of the genes in the refined gene set is inferred.

The refinement procedure consists of culling the pathway gene set to a “refined” set corresponding to the genes in the set most relevant to the transition. We applied the refinement procedure to the top 3 pathway gene sets in Table 11.2 based on p-values in the PCA to metastatic transition as well as 2 pathway gene sets with varying ranks. The refined gene sets were added to our database of gene sets and are listed in Table 8.1.

Refined Gene Set	Genes
CR HORMONAL FUNCTIONS	THRA ESR1 THRB STAR
GLYCOL	ALDOC GRCC9 ENO2 ALDOA PFKM FBP1
TRANSPORT OF VESICLES	RAB3B RAB2L RAB27A PIK3C3 RAB7 RAB27B
ptdinsPathways	VAV2 ARF1 PRKCZ PFKP PFKM
gcrPathway	ANXA1 ADRB2 GNAS NFKB1 SYT1

Table 8.1: Refined gene sets from prostate cancer analysis

After adding the refined gene sets to our gene set database, the analysis was repeated to examine whether the new gene sets better differentiate phenotype as compared to the original sets. In all the gene sets analyzed, the refined gene sets rank higher than the original as expected since the refined gene sets represent the subset of genes most strongly correlated with phenotype (see Table 8.2).

The refined gene sets were then analyzed on an independent prostate cancer dataset [131]. Using GSEA, we again see improvement in the rankings of the NES of each of the 5 refined gene sets over the original sets.

We also applied the refinement procedure to the ERK pathway gene set in the benign to PCA step in progression. Figure 8.2A,B displays all of the genes in the ERK pathway as defined in BioCarta and visualizes the pathway network. The difference between the plots is the threshold value below which dependencies are not displayed. The threshold is higher in Figure 8.2A than Figure 8.2B so only the strongest interactions are displayed here. In Figure 8.2A,B there are 7 genes that all covary strongly with one another. These 7 genes are defined as the refined set.

MYC is shown to have strong relationships with the most genes in the gene set, being strongly dependent with respect to 8 genes. In addition to learning which genes are active in the pathway, Figure 8.2A displays numerous genes which are not relevant in this context. We conclude that the refined set of 7 genes, and in particular MYC, play an essential role in the ERK pathway deregulation seen in early prostate cancer progression.

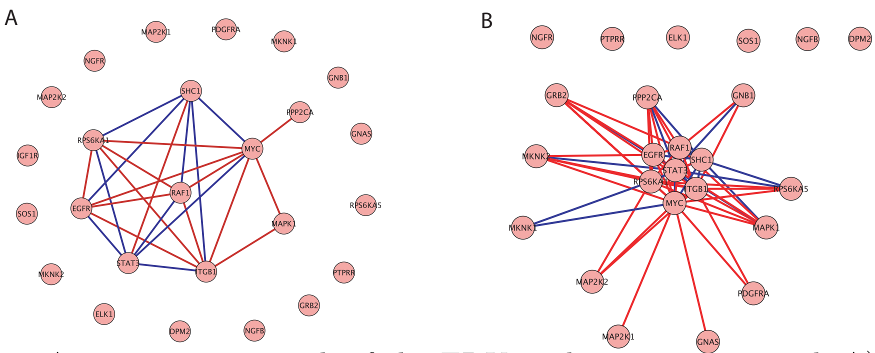


Figure 8.2: An association graph of the ERK pathway gene set with A) a strong threshold on gene dependencies and B) a weak threshold on gene dependencies. The red lines indicate a positive dependence while the blue lines indicate a negative dependence. Distances between genes reflect the strength of dependence between genes.

Prostate $\{p \mapsto m\}$ gene sets	size	rank
CR HORMONAL FUNCTIONS, original	12	6
CR HORMONAL FUNCTIONS, refined	4	3
GLYCOL, original	22	12
GLYCOL, refined	6	9
TRANSPORT OF VESICLES, original	26	18
TRANSPORT OF VESICLES, refined	6	5
ptdinsPathway, original	22	51
ptdinsPathway, refined	5	25
gcrPathway, original	19	125
gcrPathway, refined	5	39

Table 8.2: Rankings of original and refined gene sets for the PCA to metastatic state transition based on RML. Rankings are out of 527 gene sets (522 original gene sets plus 5 refined sets).

8.4 Pathway Dependency Structure in Melanoma

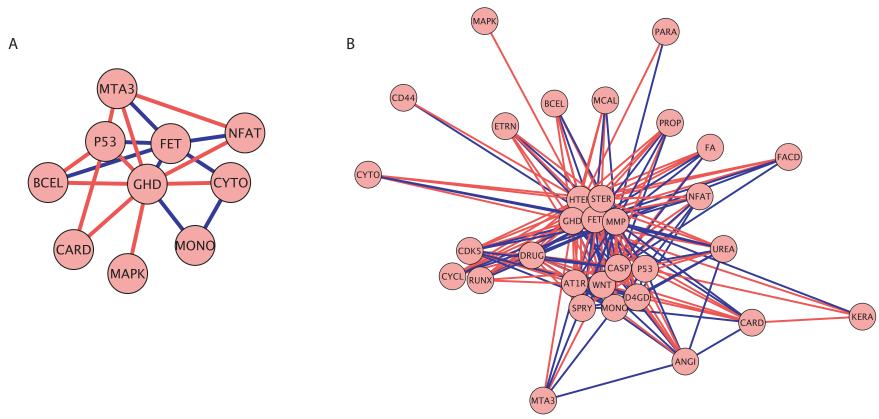


Figure 8.3: An association graph of A) the top 10 gene sets in the $\{n \mapsto p\}$ transition and B) the gene sets with $p \leq 0.05$ for the $\{n \mapsto p\}$ transition. The red lines indicate a positive dependence while the blue lines indicate a negative dependence. Distances between gene sets reflect the strength of dependencies.

The dependency structure of the 10 pathways with the lowest p-values in the normal to primary transition is shown in Figure 8.3A. Although we do not have direct validation about the level of interaction of these pathways, we have confidence they are all significant in the $\{n \mapsto b\}$ transition from the RML analysis and their strong resulting p-values. We gain further confidence in their dependency structure through documented findings of similar interactions. As in the prostate cancer progression example, the Sterol Biosynthesis gene set is highly connected. Tumor cells have long been documented to have deficiencies in sterol synthesis. One component of the Sterol Biosynthesis gene set is the mevalonate pathway which is a key intermediate in sterol production. In [58], it was found that tumor cells cannot synthesize mevalonate, but can obtain it from the host and then continue along in the sterol pathway. The dependency structure shows that the Sterol Biosynthesis gene set has a strong positive interdependence with the Fatty Acid Synthesis, Cyanoamino

Acid Metabolism, and Gamma Hexachlorocyclohexane Degradation gene sets. The interdependence of these three gene sets with the Sterol Biosynthesis gene set rank 14-th, 19-th, and 3-rd respectively out of the 523 pathways in the database. All of these gene sets may be closely tied to the inability of a tumor to synthesize certain metabolites and its increasing need for these metabolites as it grows and develops. Figure 8.3B displays the gene set dependency network for gene sets with $p \leq 0.05$. The original top 10 gene sets cluster tightly, recapitulating their close relationships.

Chapter 9

Modeling Pathway Size

In Chapter 8, it was shown that not all of the genes in a gene set are as important as others. It was also shown that while in one environmental condition a subgroup of the gene set may be active with strong interactions with each other, in a different environmental condition the active subgoup may have a different composition. As gene sets defined by sources such as BioCarta are quite general, describing all the possible gene and protein interactions in a particular pathway, it is useful to break these gene sets into a subgroup of genes that are active in a particular context. We would therefore like to look more closely at a mathematical method to characterize the size of pathways and model the dynamics of the subcomponents. This method can be used to characterize the number of subgroups in a set and also allow for insight into pathway composition, particularly if subgroups of a set are defined by core genes that control other gene interactions.

To mathematically characterize the size of pathways, we will return to the ASSESS summary statistic described in Chapter 2. Probability theory will be used to estimate a limiting distribution of the enrichment score statistic in order to: 1) better understand the properties of the summary statistic for possible improvement and 2) gain biological insight into how a pathway can be broken into groups of genes that act upon one another showing similar expression patterns. In many cases, it is not well understood if or how a pathway can be broken up into effector pathways. The calculation of the limiting distribution of the summary statistic will give an estimation on the number of subgroups that display certain expression patterns in a pathway.

9.1 The Enrichment Score Statistic

9.1.1 The Classic Kolmogorov distribution

The enrichment score used in GSEA and ASSESS is based on the Kolmogorov Distribution. The Kolmogorov test is used to find the probability that observations from a sample come from a known distribution or if observations from two samples come from the same underlying probability function. Given random samples from a distribution ordered such that $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$, the empirical distribution function (EDF) is defined as

$$F_n(x) = \frac{k}{n}, \qquad X_k \leq x \leq X_{k+1}, \quad k = 1, 2, ..., n-1.$$

Let F(x) be the true distribution of X. Then $D = \sup_x |F_n(x) - F(x)|$.

Suppose $F_n(x)$ and $G_m(x)$ are the EDF's of two independent random samples of sizes n and m respectively. Then $D = \sup_x |F_n(x) - G_m(x)|$.

Let $\Phi(\lambda) = \mathbb{P}(D < \lambda/\sqrt{n})$. Then $\Phi(\lambda)$, as $n \to \infty$ uniformly in λ, tends to

$\Phi(\lambda) = \sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2\lambda^2}$ [84].

The calculation of the running enrichment score is a tied-down Brownian bridge process and the deviation from zero is closely related to the classical Kolmogorov-Smirnov statistic. Recall from Chapter 2

$$\nu(\ell) = \frac{\sum_{j=1}^{\ell} |c_{(j)}|^{\tau} I(g_{(j)} \in \gamma_k)}{\sum_{j=1}^{p} |c_{(j)}|^{\tau} I(g_{(j)} \in \gamma_k)} - \frac{\sum_{j=1}^{\ell} I(g_{(j)} \notin \gamma_k)}{p - |\gamma_k|}, \tag{9.1}$$

$$ES_i^k = \nu[\arg\max_{\ell=1,..,p} |\nu(\ell)|]. \tag{9.2}$$

Here we have observations from two classes: genes in the gene set and genes not in the gene set. The enrichment score is a measure of the distance between the two empirical distribution functions. When τ=0, the calculation of ES_i^k is identical to the classical Kolmogorov statistic. We can use the limiting distribution of the Kolmogorov distribution as follow:

$$\Phi(ES, n_{eff}) = \Pr(D < ES) = \sum_{k=-\infty}^{\infty} (-1^k) e^{-2k^2 ES^2 n_{eff}} \tag{9.3}$$

Assuming τ=0, we can easily solve for the effective size, N_{eff}, of a gene set. The p-value of an enrichment score $p(ES_i^k)$ can be computed by comparing ES_i^k to the empirical distribution generated from $\{ES_i^k(\pi)\}_{\pi=1}^{\Pi}$. Then for gene set k

$$N_{eff} = \min_{n_{eff}} \sum_i (p(ES_i^k) - \Phi(ES_i^k, n_{eff}))^2 \tag{9.4}$$

9.1.2 The weighted Kolmogorov distribution

In practice, we generally use τ=1. This adds a weighting to the running enrichment score calculation (equation 9.1) that is proportional to the correlation statistic for each gene. The reason for such a weighting is that the genes in the gene set are by definition not independent. Gene sets are built as groups of genes that are expressed jointly, either being members of the same pathway or located in the same chromosomal region. The weighting accounts for this dependency and therefore we rarely use the classical Kolmogorov Distribution in the calculation of enrichment scores. Therefore, a weighted Kolmogorov distribution must be implemented.

In [3], a nonnegative weight function, $\psi(t)$ (0$\leq$ t $\leq$ 1), is incorporated into the Kolmogorov distribution as follows:

$$D = \sup_{-\infty<x<\infty}\sqrt{n}|F_n(x) - F(x)|\sqrt{\psi[F(x)]} \tag{9.5}$$

The limiting distribution can then be calculated by reducing Equation 9.5 to a stochastic process. First, $F(x)$ is assumed to be continuous so the transformation $u = F(x)$ can be made where $u_i = F(x_i)$ are drawn from the uniform distribution between 0 and 1. Let $G_n(u)$ be the empirical distribution of u. Then

$$K_n = \sup_{0\leq u\leq 1}\sqrt{n}|G_n(u) - u|\sqrt{\psi(u)}.$$

$Y(u) = \sqrt{(n)}|G_n(u) - u|$ is a random variable for every $0 \leq u \leq 1$

$$B_n(z) = \Pr(\sup_{0\leq u\leq 1}|Y_n(u)|\sqrt{\psi(u)} \leq z) = \Pr(K_n \leq z).$$

We want to calculate $\lim_{n\to\infty}B_n(z)$. The asymptotic process is Gaussian and the limiting process is a Gaussian process $y(u)$, $0 \leq u \leq 1$, which can be solved as a simple diffusion model.

9.1.3 Weighted correlation Kolmogorov distribution

The weighted Kolmogorov distribution is a step closer to representing the enrichment score problem, however, there is still a subtlety not accounted for. The equation for the enrichment score, Equation 9.1, has a built in weight function on the empirical distribution, $F_n^\rho(x)$. Therefore, we must solve for the limiting distribution of

$$D = \sqrt{n}|F_n^\rho(x) - F^\rho(x)|$$

where, $F_n^\rho(-\infty, x_i] = \frac{\sum_{j=1}^{i} c_{(j)}}{\sum_{j=1}^{n} c_{(j)}}$. Similarly as in the above section, a transformation must be made to reduce the equation above into a form where the limiting distribution can be found and subsequently we can solve for n. The first step is to make D into a continuous uniform distribution. This can be accomplished by again setting $u_i = F(x_i)$. Then $F^{-1}(u) = x$ and

$$D = \sqrt{n}|F_n^\rho(F^{-1}(u)) - F^\rho(F^{-1}(u))|$$

Now that D is a continuous uniform distribution, a Brownian bridge can be found to solve for the limiting distribution.

9.1.4 Simulation

We have three equations that we are interested in simulating:

$$\xi_1 = \sup_x|F_n^p(F^{-1}(u)) - F^p(F^{-1}(u))| \qquad \xi_2 = \sup_x|F_n^p(x) - F^p(x)|$$

and

$$\xi_3 = \sup_x|F_n(x) - F(x)|.$$

We would like to verify that equation 1 is the same as equation 2 but different than equation 3. We can simulate the three equations as follows.

Equation 1:

- u_i=unif[0,1] where i=1,...,n.
- x_i=$F^{-1}(u_i)$ where F(x) is the normal cumulative distribution function.
- Assign weight function w(x).
- $F_n^p(x_i) = \frac{\sum_{j=1}^{i} w(x_j)f(x_j)}{\sum_{j=1}^{n} w(x_j)f(x_j)}$ and $F^p(x_i) = \frac{\int_{-\infty}^{x_i} f(z)w(z)dz}{\int_{-\infty}^{\infty} f(z)w(z)dz}$.

Equation 2: Draw x from F(x) $\sim$ N(0,1) and follow the last three steps for Equation 1.

Equation 3:

- x_i=unif[0,1] where i=1,...,n.
- $F_n(x_i) = i/n$ and $F(x_i) = \int_{-\infty}^{x_i} f(z)dz$

Figure 9.1 shows the results from a simulation with n=100 and the weight function w(x)=x. Other weight functions can be used, such as assigning half of the genes a weighting of zero and remainder again having the weight function w(x)=x. In each case, the simulation of the three equations above indicate that Equation 1 and 2 are same but Equation 3 is different. This simulation shows that $D = \sup_x |F_n^p(x) - F^p(x)|$ can be transformed into a continuous uniform distribution.

As discussed in the previous section, the limiting distribution of a continuous uniform distribution can be found using Brownian bridges. However, we take an alternative approach and attempt to bound the weighted correlated distribution.

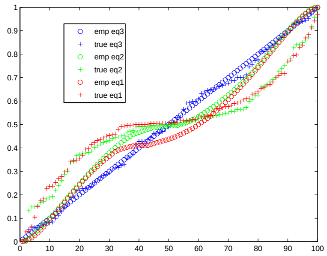


Figure 9.1: n=100, w(x)=x, ξ_1=0.1624, ξ_2=0.1387, ξ_3=0.0487

9.2 Bounding Procedure

Instead of finding the solution of the limiting distribution of Equation 9.1, we will find a close bound that can be used in further approximations. Let τ=1. We would like to find a close bound for

$$\mathbb{P}\left(\sup_{\ell\in p}\left|\frac{\sum_{j=1}^{\ell}|c_{(j)}|I(g_{(j)}\in\gamma)}{N}-\frac{\sum_{j=1}^{\ell}I(g_{(j)}\notin\gamma)}{p-s}\right|\geq\epsilon\right) \tag{9.6}$$

where γ is an a priori defined gene set, N=$\sum_{j=1}^{p}|c_{(j)}|I(g_{(j)}\in\gamma)$, p=number of genes in the data set, and s=number of genes in the gene set.

Let ϵ=$t-\left|\left(\frac{s}{p}\frac{\sum_{j=1}^{\ell}|c_{(j)}|}{N}-\frac{\ell}{p}\right)\times -1\right|$.

Then Equation 9.6 can be rewritten as

$$\mathbb{P}\left(\sup_{\ell\in p}\left|\frac{\sum_{j=1}^{\ell}|c_{(j)}|I(g_{(j)}\in\gamma)}{N}-\frac{\sum_{j=1}^{\ell}I(g_{(j)}\notin\gamma)}{p-s}\right|+\left|(-1)\left(\frac{s}{p}\frac{\sum_{j=1}^{\ell}|c_{(j)}|}{N}-\frac{\ell}{p}\right)\right|\geq t\right)$$

$$\leq \mathbb{P}\left(\sup_{\ell \in p}\left|\frac{\sum_{j=1}^{\ell}|c_{(j)}|I(g_{(j)} \in \gamma)}{N} - \frac{s}{p}\frac{\sum_{j=1}^{\ell}|c_{(j)}|}{N} - \frac{\sum_{j=1}^{\ell} I(g_{(j)} \notin \gamma)}{p-s} + \frac{\ell}{p}\right| \geq t\right)$$

$$\leq \mathbb{P}\left(\sup_{\ell \in p}\left|\frac{\sum_{j=1}^{\ell}|c_{(j)}|I(g_{(j)} \in \gamma)}{N} - \frac{s}{p}\frac{\sum_{j=1}^{\ell}|c_{(j)}|}{N}\right| + \left|\frac{\sum_{j=1}^{\ell} I(g_{(j)} \notin \gamma)}{p-s} - \frac{\ell}{p}\right| \geq t\right)$$

$$= \mathbb{P}\left(\sup_{\ell \in p}\left|\frac{\sum_{j=1}^{\ell}|c_{(j)}|I(g_{(j)} \in \gamma)}{N} - \frac{s}{p}\frac{\sum_{j=1}^{\ell}|c_{(j)}|}{N}\right| \geq \frac{t}{2}\right)$$

$$+\mathbb{P}\left(\sup_{\ell \in p}\left|\frac{\sum_{j=1}^{\ell} I(g_{(j)} \notin \gamma)}{p-s} - \frac{\ell}{p}\right| \geq \frac{t}{2}\right) \tag{9.7}$$

Let

$$|Z_j| = \left|\frac{\sum_{j=1}^{\ell}|c_{(j)}|I(g_{(j)} \in \gamma)}{N} - \frac{s}{p}\frac{\sum_{j=1}^{\ell}|c_{(j)}|}{N}\right|$$

and

$$|K_j| = \left| \frac{\sum_{j=1}^{\ell} I(g_{(j)} \notin \gamma)}{p-s} - \frac{\ell}{p} \right|.$$

Bernstein's Inequality states if $|Z_j| \leq M$,

$$\mathbb{P}\left(\sum_{j=1}^{n} > t \right) \leq \exp\left(-\frac{t^2/2}{\sum_{j=1}^{n} E[Z_j^2] + Mt/3} \right).$$

For a given ℓ, we know $|Z_j| \leq 1$ and $|K_j| \leq 1$ for all $j \in p$. We must then solve for $\sum_{j=1}^{\ell} E[Z_j^2]$ and $\sum_{j=1}^{\ell} E[K_j^2]$.

$$\sum_{j=1}^{\ell} E[Z_j^2] = \sum_{j=1}^{\ell} Var[Z_j] + \sum_{j=1}^{\ell} E[Z_j]^2.$$

By definition, $\sum_{j=1}^{\ell} E[Z_j] = 0$.

$$\sum_{j=1}^{\ell} Var[Z_j] = \sum_{j=1}^{\ell} Var\left[\frac{|c_{(j)}| I(g_{(j)} \in \gamma)}{N} \right] = \sum_{j=1}^{\ell} \frac{|c_{(j)}|^2}{N^2} Var[I(g_{(j)} \in \gamma]$$

$$= \frac{\sum_{j=1}^{\ell} c_{(j)}^2}{N^2} \frac{s}{p} \frac{p-s}{p}.$$

Also by definition, $\sum_{j=1}^{\ell} E[K_j] = 0$.

$$\sum_{j=1}^{\ell} E[K_j^2] = \sum_{j=1}^{\ell} Var[K_j] = \sum_{j=1}^{\ell} Var\left[\frac{|I(g_{(j)} \notin \gamma}{p-s} \right] = \frac{1}{(p-s)^2} \frac{s}{p} \left(\frac{p-s}{p} \right) \ell.$$

Equation 9.7 can be rewritten as $\mathbb{P}(\sup_{\ell\in p}|Z_j| \geq \frac{t}{2}) + \mathbb{P}(\sup_{\ell\in p}|K_j| \geq \frac{t}{2})$ where by Bernstein's inequality,

$$\mathbb{P}(\sup_{\ell\in p}|Z_j| \geq \frac{t}{2}) \leq \exp\left(-\frac{\frac{t^2}{8}}{\frac{\sum_{j=1}^{\ell} c_{(j)}^2}{N^2}\frac{s}{p}\frac{p-s}{p} + \frac{t}{6}}\right)$$

and

$$\mathbb{P}(\sup_{\ell\in p}|K_j| \geq \frac{t}{2}) \leq \exp\left(-\frac{\frac{t^2}{8}}{\frac{1}{(p-s)^2}\frac{s}{p}\frac{p-s}{p}\ell + \frac{t}{6}}\right).$$

Therefore, for any given ℓ

$$\mathbb{P}\left(\sup_{\ell\in p}\left|\frac{\sum_{j=1}^{\ell}|c_{(j)}|I(g_{(j)}\in\gamma)}{N} - \frac{s}{p}\frac{\sum_{j=1}^{\ell}|c_{(j)}|}{N}\right| \geq \frac{t}{2}\right)$$

$$+\mathbb{P}\left(\sup_{\ell\in p}\left|\frac{\sum_{j=1}^{\ell}I(g_{(j)}\notin\gamma)}{p-s} - \frac{\ell}{p}\right| \geq \frac{t}{2}\right)$$

$$\leq \exp\left(-\frac{\frac{t^2}{8}}{\frac{\sum_{j=1}^{\ell} c_{(j)}^2}{N^2}\frac{s}{p}\frac{p-s}{p} + \frac{t}{6}}\right) + \exp\left(-\frac{\frac{t^2}{8}}{\frac{1}{(p-s)^2}\frac{s}{p}\frac{p-s}{p}\ell + \frac{t}{6}}\right). \tag{9.8}$$

We use a union bound to make the equation applicable for all ℓ. Then Equation 9.7

$$\leq \left[\exp\left(-\frac{\frac{t^2}{8}}{\frac{\sum_{j=1}^{\ell} c_{(j)}^2}{N^2}\frac{s}{p}\frac{p-s}{p} + \frac{t}{6}}\right) + \exp\left(-\frac{\frac{t^2}{8}}{\frac{1}{(p-s)^2}\frac{s}{p}\frac{p-s}{p}\ell + \frac{t}{6}}\right)\right]p$$

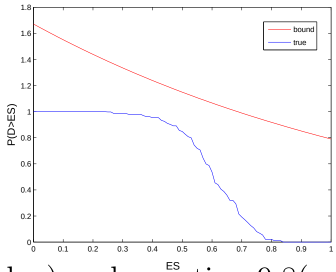


Figure 9.2: Equation 9.1(blue) and equation 9.8(red) plotted for —ES— ∈ [0,1]

Graphical representation shows that while Bernstein's Inequality does in fact bound the enrichment score (Equation 9.1), the bound is not extremely tight. N(0,1) data was generated for 100 genes in 10 samples where samples 1-5 were labeled as class 1 and samples 6-10 were labeled as class 2. Genes 1-10 comprised the gene set. Figure 9.2 displays Equation 9.6 . This represents the true probability given the simulated data using permutations of class labels. This is plotted along with the bound on the probability from Bernstein's Inequality (Equation 9.8). We can see that while we have achieved an upper bound on Equation 9.1, the bound is not a tight bound. We will simulate data to show that this is not an adequate bound to estimate net effective gene set size and then propose improvements on the bound.

9.2.1 Simulation

Data was simulated for 200 genes in 10 samples. Samples 1-5 were labeled as class 1 and samples 6-10 as class 2. Genes 1-25 in class 1 come for a N(3,1) distribution and in class 2 come from a N(1,1) distribution. These genes comprise gene set 1. Gene set 2 consists of genes 26-50. These genes have a N(5,3) distribution in samples 1-5 and a N(10,3) distribution in samples 6-10. Genes 51-75, which comprise gene set 3, have a N(0,2) distribution in samples 1-5 and a N(8,2) distribution in samples 6-10.

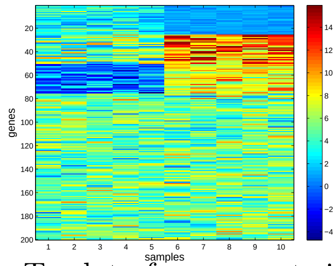


Figure 9.3: Toydata for gene set size simulation

The remaining genes in all samples come from a N(5,2) distribution. See Figure 9.3 for a heatmap of the expression values. Gene set 1, 2, and 3 are designed to have a small effective size, as all of the genes in each set similarly covary in respect to class label. Gene set 4 is comprised on genes 1-75 and was designed to have a slightly larger effective size, having three distinct expression patterns. Gene set 5 is comprised of one third of the first 75 genes. This set was designed to see if the effective size changes when the initial set is smaller than gene set 4 but still contains the same three distinct expression patterns. The hope is that gene set 4 and 5 have the same effective size. Gene set 6 should have a slightly higher effective size still, as it consists of a subgroup of genes from gene sets 1, 2, and 3 plus an addition 20 genes that do not covary with class label. Gene set 7 is a random selection of 50 genes in the data set and gene set 8 has 25 genes that do not covary with class label. Gene set 8 should have the largest net effective size, as there are no genes included that covary with class label.

We solve for the effective gene set size for each of the 8 gene sets described above. As in Equation 9.3, we find the gene set size that minimizes the difference between the cumulative distribution function (Equation 9.6) and the bound from Bernstein's Inequality (Equation 9.8). Let the cumulative distribution function for sample i in gene set k be represented as $p(ES_i^k)$ where

$$p(ES_i^k) = \mathbb{P}\left(\sup_{\ell \in p} \left| \frac{\sum_{j=1}^{\ell} |c_{(j)}| I(g_{(j)} \in \gamma_k)}{N} - \frac{\sum_{j=1}^{\ell} I(g_{(j)} \notin \gamma_k)}{p-s} \right| \geq ES \right)$$

and the equation for its bound be represented as $\Phi(ES_i^k, n_{eff})$ where

$$\Phi(ES_i^k, n_{eff}) = \exp\left(- \frac{\frac{t^2}{8}}{\frac{\sum_{j=1}^{\ell} c_{(j)}^2}{N^2} \frac{n_{eff}}{p} \frac{p-n_{eff}}{p} + \frac{t}{6}} \right)$$

$$+ \exp\left(- \frac{\frac{t^2}{8}}{\frac{1}{(p-n_{eff})^2} \frac{n_{eff}}{p} \frac{p-n_{eff}}{p} \ell + \frac{t}{6}} \right).$$

$$N_{eff} = \min_{n_{eff}} \sum_i (p(ES_i^k) - \Phi(ES_i^k, n_{eff}))^2 \tag{9.9}$$

As expected from Figure 9.2, the bounds are not tight enough to give an accurate estimation for N_{eff}. The result is that N_{eff} is estimated to be 155, 1, 1, 1, 1, 155, 155, 1 (where 155 is the maximum set size and 1 is the minimum set size) for gene sets 1-8 respectively. We therefore must find a tighter bound or directly estimate the limiting distribution of Equation 9.1. We first set out to find a tighter bound.

Instead of minimizing over just n_{eff} for each given ES_i, we now minimize over both n_{eff} and ES for all i to better fit the bound to the cumulative distribution function.

$$N_{eff} = \min_{neff} \sum_i (\min_t (p(ES_i^k)^t - \Phi(ES_i^k, n_{eff})^t)^2 \tag{9.10}$$

While, this method improves our results, in all but one of the 8 sets, the net effective gene set sizes are all still larger than the true sizes. Therefore, our bound is still not tight enough. These results provide the support that the limiting distribution must be solved for directly rather than finding a bound.

9.3 Gene Set Effective Size

The limiting distribution of the classic Kolmogorov distribution seems to be a better fit than the Bernstein's Inequality bound for Kolmogorov distribution with the built in weight function. We again solve for the gene set effective size of the 8 gene sets described in the above section on the simulated data, this time using Equation 9.4. The resulting effective set sizes are 3, 4, 4, 6, 10, 10, 7, and 19 for sets 1-8 respectively. As expected, sets 1, 2, and 3 have the smallest effective sizes and set 8 has the largest size.

The oncogenic pathways defined on model systems described in Chapter 2.4.2 all contain 50 genes which were selected by sorting the weights of a logistic regression model. It is unknown whether these genes truly make up a regulatory pathway or are involved in different mechanisms which all individually alter the cell phenotype. We can use Equation 9.4 to calculate the number of core components whose genes function together in these oncogenic gene sets. The results are shown in Table 9.1.

In Chapter 8.3, several gene sets were refined to a core set of genes that covaried with class label. Association graphs were made for strong and weak threshold. These thresholds are set depending on the information the user would like to acquire. However, with an a prior idea of the effective set size, a more accurate threshold may

Experiment	True Set Size	Effective Set Size
BCAT	50	4
E2F3	50	4
MYC	50	4
RAS	50	6
SRC	50	3

Table 9.1: Effective set size of oncogenic gene sets. In each case, many of the genes are correspondingly regulated resulting in smaller effective set sizes than true set sizes.

be selected.

The ERK pathway defined by BioCarta consists of 25 genes. The strong threshold graph clustered 9 of the genes, leaving the remaining 16 unconnected. The weak threshold graph clustered 19 of genes, leaving 6 unconnected. Equation 9.4 estimates the effective set size to be 15. Given the association graphs, this is a reasonable estimate. One representation of this gene set may be a core cluster of 11 genes with 14 unconnected. Finding the effective gene set size may therefore aid in determining the threshold when building association networks.

Chapter 10

Conclusion

10.1 Summary

The work presented here aims to solve the many problems associated with modeling complex traits. Complex traits are controlled by variation across many genes and often express variation in phenotype. This heterogeneity causes difficulty in modeling these traits. The two main sources of heterogeneity come from variation across genomic features and variation across phenotypes or disease stages. In order to account for these sources of heterogeneity we build models that use the idea of borrowing strength. We account for the variation in individual genes by building models in the gene set space where we integrate genes into a priori known gene sets defined functionally or structurally. This method is discussed in detail in Chapter 2. We then account for the phenotypic variation that results from the expression of complex traits by building conjoint models across phenotypes. In doing so, we borrow strength across phenotypes by taking advantage of the fact that there are commonalities to individuals with a specific cancer type regardless of their disease stage. Therefore, we use a shared representation as the basis for our models. These ideas are discussed fully in Chapter 4.

10.2 Future Work

The next step in this area of research is to continue to incorporate sources of cancer heterogeneity into our models. As of this point, the models of cancer progression have labeled individuals as belonging to one disease state. However, even among individuals who are classified into the same disease state, there can be a great deal of heterogeneity within each neoplasm. This diversity is a probable factor in determining which individuals progress onto more serious states and which do not. We therefore must better understand the diversity and evolutionary dynamics within neoplasms. We have already identified deregulated pathways which define specific disease states. The next objective is to define subgroups of a state by combinations of these significant pathways.

In order to determine the evolutionary dynamics within the neoplasm and discover which mechanisms define the phenotypic subgroups of a disease stage, it would be beneficial to have genetic information from the same individual as they progress through tumorigenesis. For most types of cancer, this type of data is not easily obtainable. However, for esophageal adenocarcinoma, one can get genetic data from individuals throughout there disease progression. The reason for this is that periodic biopsies are conducted on individuals with the premalignant condition Barrett's esophagus (BE). Less than 5% of people with Barrett's esophagus go on to develop esophageal adenocarcinoma. This fact, along with the high mortality and morbidity from preforming esophagectomies, has resulted in very little medical intervention for people with BE. Therefore, measures of genetic data can be collected through the periodic biopsies on the same individual throughout time. This type of data creates possibilities for modeling the nonlinearity of progression that has yet to be modeled.

Chapter 11

Appendix

11.1 Appendix A: Methods

11.1.1 Gene set construction

Given an expression data set with two class labels, we use a linear logistic regression model with regularization or shrinkage [65] to construct gene sets. We define the expression data as a matrix x_j^i with $i = 1, ..., n$ (the number of samples) and $j = 1, ..., p$ (the number of genes), the i-th sample is designated as x_i, and the class labels as $y \in \{-1, 1\}$. The logistic regression model with regularization involves solving the following optimization problem

$$\arg\min_{w,b} \left[\frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-(y_i(w \cdot x_i + b))}) + \lambda \|w\|^2 \right],$$

where λ is a model parameter that needs to be set.

Solving the above optimization problem results in a vector $\hat{w}$ and the absolute magnitude of the elements of the vector correspond to the relevance of a gene or feature. For the HMEC data sets and the AKT data set, genes corresponding to

50 elements of $\hat{w}$ most correlated with the perturbation phenotype were used to construct the gene sets. In both algorithms 1 and 2 genes corresponding to the top and bottom 50 elements of $\hat{w}$ were used.

11.1.2 Classification and Gene Set Selection

Classification using enrichment scores was applied in two settings: a train-test setting and a leave-one-out cross-validation setting. The leave-one-out setting was used for all the data sets except the leukemia data set for which we used the test-train setting. The test-train setting is a simple generalization of the leave-one-out setting.

Leave-one-out setting:

Given data set x_j^i of gene expression for $j = 1, .., p$ genes and $i = 1, ..., n$ samples where the i-th column of the matrix X correspond to the i-th sample, labels $(y_i)_{i=1}^n$, and gene sets $\Gamma = \{\gamma_1, ..., \gamma_m\}$ the leave-one-out method outlined in Algorithm 1 provides an unbiased estimate of the error rate (technically leave-one-out estimators are almost unbiased [140]).

Algorithm 1:

Leave-one-out procedure for pathway based classification

input: training data and gene sets

return: error rate

for $i = 1$ **to** n **do**

split the data into x_i (the i-th data point) and $X^{\setminus i}$ (the data with the i-th point removed); compute $\mathrm{Tr} = ES_i^k$ for $X^{\setminus i}$ (this is the enrichment of the m gene sets on the $n-1$ data in $X^{\setminus i}$); compute Test $= ES_i^k$ for x_i (this is the enrichment of the m gene sets on x_i, the label i-th point is not used in the estimation of the enrichment score by leaving this point out of the Parzen estimator); use Tr to build logistic regression with variable selection model M_i; apply M_i to Test to obtain prediction $\hat{y}_i$; **if** $y_i \neq \hat{y}_i$ **then** error rate = error rate + 1;

return error rate

Algorithm 2:

Test error estimate for pathway based classification

input: training data, test data, and gene sets

return: error rate

compute Tr $= ES_i^k$ for X (this is the enrichment of the m gene sets on the training data X); use Tr to build logistic regression with variable selection model M;

for $j = 1$ **to** n' **do**

compute Test $= ES_j^k$ for z_j (this is the enrichment of the m gene sets on the j-th test sample, use only the training data X to compute the Parzen estimator);

apply M to Test to obtain prediction $\hat{t}_j$ **if** $t_j \neq \hat{t}_j$ **then** error rate = error rate+1

return error rate

Train-test setting:

Given a training set of $X = (x_i)_{i=1}^n$ expression profiles and labels $(y_i)_{i=1}^n$, a test set of $Z = (x_j)_{j=1}^{n'}$ expression profiles with labels $(t_j)_{j=1}^{n'}$, and gene sets $\Gamma = \{\gamma_1, ..., \gamma_m\}$ the procedure outlined in Algorithm 2 provides an unbiased error estimate on the test set.

11.1.3 Computation of Absorption Probabilities

To compute the correlation coefficients in the nonparametric model we need to compute the probability that the expression of the j-th gene in the i-th sample is representative of class 1 or class 2, $\mathbb{P}(x_j^i \in C_1)$ and $\mathbb{P}(x_j^i \in C_2)$ for all samples $i = 1, ..., n$ and genes $j = 1, .., p$. We first scale the expression data for each gene to $[0, 1]$, $\hat{x}_j^i = \frac{x_j^i - \min_i(x_j^i)}{\max_i(x_j^i) - \min_i(x_j^i)}$. The class membership probabilities are the probabilities of absorption to the left or right extreme, which are $\{0, 1\}$ for the scaled data, depending on whether $\hat{x}_j^i$ is greater or less than the scaled class means (see Table 11.1). This simply reflects the directionality assumption of our model.

$\mathbb{P}(\hat{x}_j^i \in C_1)$	$\hat{x}_j^i \leq \hat{\mu}_{j1}$	$\mathbb{P}(\text{absorption at 0 starting at } \hat{x}_j^i \vert \hat{p}_{j1})$
	$\hat{x}_j^i > \hat{\mu}_{j1}$	$\mathbb{P}(\text{absorption at 1 starting at } \hat{x}_j^i \vert \hat{p}_{j1})$
$\mathbb{P}(\hat{x}_j^i \in C_2)$	$\hat{x}_j^i \leq \hat{\mu}_{j2}$	$\mathbb{P}(\text{absorption at 0 starting at } \hat{x}_j^i \vert \hat{p}_{j2})$
	$\hat{x}_j^i > \hat{\mu}_{j2}$	$\mathbb{P}(\text{absorption at 1 starting at } \hat{x}_j^i \vert \hat{p}_{j2})$

Table 11.1: Probability of class membership as a function of $\hat{x}_j^i$ and the class means.

Let

$$\begin{aligned} u(\hat{x}) &= \mathbb{P}(\text{absorption at 0 starting at } \hat{x}), \\ v(\hat{x}) &= \mathbb{P}(\text{absorption at 1 starting at } \hat{x}), \end{aligned}$$

and let $p(\hat{x})$ be supported on $[0, 1]$, then

$$\begin{aligned} \mathbb{P}(\text{absorption at 0 starting at } \hat{x}\,|p(\hat{x})) &= \int_0^{\hat{x}} u(\hat{x})\,p(\hat{x})\,d\hat{x}, \\ \mathbb{P}(\text{absorption at 1 starting at } \hat{x}\,|p(\hat{x})) &= \int_{\hat{x}}^1 v(\hat{x})\,p(\hat{x})\,d\hat{x}. \end{aligned}$$

The absorption probabilities of a Brownian motion at the end points of a line segment can be computed by solving the heat equation with appropriate boundary conditions, the Dirichlet problem [35]. In the above case we solve for

$$\frac{d^2u(\hat{x})}{d\hat{x}^2} = 0 \quad s.t. \quad u(0) = 0, \; u(1) = 1$$

$$\frac{d^2v(\hat{x})}{d\hat{x}^2} = 0 \quad s.t. \quad v(0) = 1, \; v(1) = 0.$$

This results in the solutions

$$u(\hat{x}) = \hat{x}, \qquad v(\hat{x}) = 1 - \hat{x}.$$

Given the Parzen estimates of the densities for the two classes

$$\begin{aligned} \hat{p}_{j1}(\hat{x}) &= \frac{1}{n_1 \sigma_{j1}\sqrt{2\pi}} \sum_{i \in C_1} e^{-|\hat{x}_j^i - \hat{x}|^2/2\sigma_{j1}^2}, \\ \hat{p}_{j2}(\hat{x}) &= \frac{1}{n_2 \sigma_{j2}\sqrt{2\pi}} \sum_{i \in C_2} e^{-|\hat{x}_j^i - \hat{x}|^2/2\sigma_{j2}^2}, \end{aligned}$$

we can compute the absorption probabilities as

$$\begin{aligned} \mathbb{P}(\text{absorption at 0 starting at } \hat{x}|\hat{p}_{jc}) &= \int_0^{\hat{x}} s\, \hat{p}_{jc}(s)\, ds \\ \mathbb{P}(\text{absorption at 1 starting at } \hat{x}|\hat{p}_{jc}) &= \int_{\hat{x}}^1 (1-s)\, \hat{p}_{jc}(s)\, ds \end{aligned}$$

where c denotes the classes $\{1, 2\}$. Solving the integrals results in a weighted sum of error functions and exponentials.

11.2 Appendix B: Tables

Table 11.2A

Pval	w_0	Gene Set
0	169.97	CR TRANSPORT
0	-71.15	EGF receptor signaling pathway
0	-69.07	ephA4Pathway
0	95.74	gcrPathway
0	-101.84	CR DNA MET AND MOD
0.01	-77.55	erkPathway
0.01	-65.74	MAP00010 Glycolysis Gluconeogenesis
0.01	-133.96	ranPathway
0.01	71.20	CR TRANSPORT OF VESICLES
0.01	-63.39	FA
0.02	94.30	longevityPathway
0.02	78.94	MAP00361 gamma Hexachlorocyclohexane degradation
0.02	-57.73	actinYPathway
0.02	66.43	il10Pathway
0.03	34.90	achPathway
0.03	-97.63	chrebpPathway
0.03	151.85	MAP00100 Sterol biosynthesis
0.04	-99.40	Erbb4Pathway
0.04	-54.65	SppaPathway
0.04	-104.52	ptdinsPathway
0.04	71.63	il22bpPathway
0.04	-43.06	cell proliferation
0.04	91.68	s1pPathway
0.04	-31.72	XINACT MERGED
0.04	-103.24	cell cycle checkpoint
0.04	-28.48	DnafragmentPathway
0.04	100.62	ROS
0.05	-82.60	KrebPathway
0.05	77.28	SA PTEN PATHWAY

Table 11.2B

Pval	w_1	Gene Set
0	0.09	MAP00361 gamma Hexachlorocyclohexane degradation
0.01	0.09	MAP00100 Sterol biosynthesis
0.01	-0.06	GNF FEMALE GENES
0.02	-0.07	CR DNA MET AND MOD
0.02	-0.09	FA
0.02	0.07	CR TRANSPORT
0.02	-0.07	ephA4Pathway
0.03	-0.09	XINACT MERGED
0.03	-0.06	CR HORMONAL FUNCTIONS
0.04	-0.07	mcalpainPathway
0.04	-0.05	lymphocytePathway
0.04	-0.05	cell proliferation
0.04	-0.06	tall1Pathway
0.05	-0.06	erkPathway
0.05	-0.05	MAP00190 Oxidative phosphorylation
0.05	-0.06	SppaPathway
0.05	0.07	MAP00480 Glutathione metabolism

Table 11.2C

Pval	w_2	Gene Set
0	0.08	CR HORMONAL FUNCTIONS
0	-0.06	GLYCOL
0	0.07	CR TRANSPORT OF VESICLES
0	0.07	longevityPathway
0.01	0.11	arapPathway
0.01	0.07	GNF FEMALE GENES
0.02	0.1	CR TRANSPORT
0.02	0.09	MAP00512 O Glycans biosynthesis
0.02	-0.05	MAP00010 Glycolysis Gluconeogenesis
0.02	0.06	ST Gaq Pathway
0.02	0.08	tnf and fas network
0.03	0.05	p53hypoxiaPathway
0.03	-0.07	MAP00272 Cysteine metabolism
0.03	0.06	GPCRs Class A Rhodopsin-like
0.03	-0.1	MAP00430 Taurine and hypotaurine metabolism
0.03	0.06	MAP00030 Pentose phosphate pathway
0.04	-0.07	ptdinsPathway
0.04	0.04	CR SIGNALLING
0.04	0.06	ST B Cell Antigen Receptor
0.05	0.08	aktPathway
0.05	0.04	HTERT DOWN
0.05	-0.06	GLUCO

Table 11.2: Gene sets sorted by p-value with corresponding paramters: w_0, w_1, and w_2. Negative values of these parameters indicate up-regulation and positive values indicate down-regulation.

Table 11.3A

Pval	w_0	Gene Set
0	-21.54	hifPathway
0.01	-13.41	aktPathway
0.01	-18.98	d4gdiPathway
0.01	15.7	MAP00340 Histidine metabolism
0.02	16.02	ANTI CD44 DOWN
0.03	18.85	MAP00360 Phenylalanine metabolism
0.03	-13.66	il4Pathway
0.03	-10.43	ST Interleukin 4 Pathway
0.03	14	P53 UP
0.04	17.55	fatty acid metabolism
0.04	11.36	il3Pathway
0.04	17.54	MAP00350 Tyrosine metabolism
0.04	-15.4	skp2e2fPathway
0.04	13.28	ADULT LIVER vs FETAL LIVER GNF2
0.05	16.16	CR CYTOSKELETON
0.05	-15.22	PtdinsPathway

Table 11.3B

Pval	w_1	Gene Set
0	0.01	CR CYTOSKELETON
0	0.01	MAP00361 gamma Hexachlorocyclohexane degradation
0	0.01	mapkPathway
0	-0.01	monocytePathway
0	0.01	mta3Pathway
0	0.01	nfatPathway
0	0.01	SA B CELL RECEPTOR COMPLEXES
0	-0.01	FETAL LIVER vs ADULT LIVER GNF2
0	0.01	P53 UP
0	0.01	cardiacegfPathway
0	0.01	FA
0	0.01	cdk5Pathway
0	-0.01	d4gdiPathway
0	0.01	electron transport
0.01	0.01	wntPathway
0.01	0.01	ANTI CD44 DOWN
0.01	0.01	MAP00220 Urea cycle and metabolism of amino groups
0.01	0.01	drug resistance and metabolism
0.01	0.01	pparaPathway
0.01	0.01	RUNX1 TARGETS
0.02	0.01	HTERT DOWN
0.02	0.01	MAP00640 Propanoate metabolism
0.02	0.01	MAP00100 Sterol biosynthesis
0.03	0.01	at1rPathway
0.03	0.01	SA REG CASCADE OF CYCLIN EXPR
0.03	0.01	Fatty Acid Synthesis
0.03	0.01	spryPathway
0.04	0.01	mcalpainPathway
0.05	-0.01	caspasePathway
0.05	-0.01	SA MMP CYTOKINE CONNECTION
0.05	-0.01	CR ANGIOG
0.05	0.01	KeratinocytePathway

Table 11.3C

Pval	w_2	Gene Set
0	-0.01	dreamPathway
0	-0.01	igf1rPathway
0	-0.01	ST B Cell Antigen Receptor
0	-0.02	ST Ga12 Pathway
0	-0.01	TGF Beta Signaling Pathway
0	-0.02	DOWN SOM G1ER
0	-0.01	CSR BROWN POS LIST SAM SORLIE
0	-0.01	HYPERDIPLOID DOWNING
0	-0.01	CSR BROWN SAM POS VANTVEER
0	-0.01	crebPathway
0.01	0.01	ST Interleukin 4 Pathway
0.01	-0.01	ST Fas Signaling Pathway
0.02	-0.01	XINACT
0.03	-0.01	mcalpainPathway
0.05	-0.01	carm1Pathway
0.05	-0.01	il4Pathway

Table 11.3: Gene sets sorted by p-value with corresponding paramters: w_0, w_1, and w_2. Negative values of these parameters indicate up-regulation and positive values indicate down-regulation.

Table 11.4A: N→A→C1

w_0	Gene Set	w_0	Gene Set
-0.1	MAP00230 Purine metabolism	0.08	ST adrenergic
-0.1	SA reg cascade of cyclin expr	0.07	MAP00360 Phenylalanine metabolism
-0.09	etsPathway	-0.07	brca down
-0.09	serum induced mkl indep srf	0.07	akap13Pathway
-0.09	mRNA processing	0.07	MAP00561 Glycerolipid metabolism
-0.09	Matrix Metalloproteinases	0.07	MAP00350 Tyrosine metabolism
0.09	MAP00150 Andro estro metabolism	-0.07	atmPathway
-0.08	CR cam	-0.07	MAP00790 Folate biosynthesis
-0.08	MAP00240 Pyrimidine metabolism	0.07	MAP00910 Nitrogen metabolism
-0.08	200NN RPS19 DBA BEN	-0.07	CR transcription factors
0.08	badPathway	-0.07	Wnt Signaling
-0.08	high mll amstrong	-0.07	SA g1 and s phases
-0.08	arfPathway	0.07	pgc1aPathway
0.08	shh lisa	0.07	SA fas signalling
-0.08	cell2cellPathway	0.07	anti cd44 up
-0.08	kras top100 control	-0.07	MAP03020 RNA polymerase
-0.08	fetal liver hs enriched tf jp	-0.07	melton 200 hemo
0.08	hdacPathway	0.07	agpcrPathway
0.08	crebPathway	-0.07	SA mpp cypokine connection
0.08	mapkPathway	0.07	ST p38 MAPK Pathway
-0.08	fetal liver vs. adult liver gnf2	0.07	nkcellsPathway
0.08	GPCRs Class A Rhodopsin-like 2	0.07	MAP00340 Histidine metabolism
0.08	erk5Pathway	-0.07	hoxa9 down
0.08	rabPathway	0.07	achPathway
0.08	alkPathway	0.07	calcineurinPathway

Table 11.4B: N→A

w_1	Gene Set	w_1	Gene Set
-0.01	FORTUNEL STEMNESS	-0.01	setPathway
-0.01	Cell Cycle	-0.01	g1Pathway
-0.01	LEU DOWN	-0.01	CR REPAIR
-0.01	MAP00240 Pyrimidine metabolism	-0.01	DOWN G1ER
-0.01	RAP DOWN	0.01	rarrxrPathway
-0.01	CR CELL CYCLE	0.01	TEL AML1 DOWNING ALL
-0.01	MAP00230 Purine metabolism	0.01	ZHE MOUSE WT MEG
-0.01	CSR BROWN SAM POS VANTVEER	-0.01	cell cycle checkpoint
-0.01	ptc1Pathway	-0.01	SA reg cascade of cyclin
-0.01	MAP00670 One carbon pool by folate	-0.01	CR DNA MET AND MOD
-0.01	AML NOT POLY BEN	0.01	mcalpainPathway
-0.01	CSR BROWN POS LIST SAM SORLIE	-0.01	lymphocytePathway
-0.01	DOWN SOM G1ER	-0.01	monocytePathway
0.01	muscle myosin	-0.01	atmPathway
-0.01	DNA DAMAGE SIGNALLING	-0.01	tRNA Synthetases
-0.01	SA MMP CYTOKINE CONNECTION	-0.01	mRNA splicing
-0.01	Proteasome Degradation	-0.01	ST Fas Signaling Pathway
-0.01	SA G1 AND S PHASES	-0.01	g2Pathway
-0.01	radiation sensitivity	0.01	slrpPathway
-0.01	rbPathway	-0.01	MAP00680 Methane metab
-0.01	proteasomePathway	-0.01	mRNA binding activity
-0.01	drug resistance and metabolism	0.01	myosinPathway
-0.01	cellcyclePathway	-0.01	telPathway
-0.01	MAP00970 Aminoacyl tRNA biosynthesis	-0.01	cell proliferation
-0.01	neutrophilPathway	-0.01	MELTON STEMNESS

Table 11.4C: A→C1

w_2	Gene Set	w_2	Gene Set
-0.01	DOWN MEGS	-0.01	CR CAM
-0.01	KRAS TOP100 KNOCKDOWN	0	CEBPA UPREG
-0.01	compPathway	0	cell adhesion
-0.01	circadianPathway	0	hcmvPathway
-0.01	ZHE MOUSE WT MEG	0	Matrix Metalloproteinases
-0.01	CR ANGIOG	0	cardiacegfPathway
-0.01	PML RARA DOWNING AML	0	MAP00860 Porphyrin, chlorophyll
-0.01	Inflammatory Response Pathway	0	POLY NOT AML BEN
-0.01	classicPathway	0	CEBPA DOWNREG
-0.01	Complement Activation Classical	0	DOWNREG BY HOXA9
-0.01	CR HORMONAL FUNCTIONS	0	mspPathway
0.01	tercPathway	0	achPathway
-0.01	EMT DOWN	0	HL60 ATRA BEN
-0.01	etsPathway	0	Krebs-TCA Cycle
-0.01	reckPathway	0	SA DAG1
-0.01	SERUM INDUCED MKL INDEP SRF	0	cell adhesion molecule activity
-0.01	MLL DOWNING ALL	0	parkinPathway
0.01	MAP00195 Photosynthesis	0	p38mapkPathway
-0.01	tcytotoxicPathway	0	cell2cellPathway
0.01	MAP00150 Andro estro metab	0	SA reg cascade of cyclin
-0.01	lechtinPathway	0	electron transporter activity
0.01	MAP03070 Type III secretion system	0	thelperPathway
0.01	MAP00193 ATP synthesis	0	ST Interleukin 4 Pathway
0.01	MAP00190 Oxidative phosphorylation	0	pdgfPathway
-0.01	blymphocytePathway	0	eif2Pathway

Table 11.4D: C1→C2→C3→C4

w_0	Gene Set	w_0	Gene Set
-0.09	Wnt Signaling	-0.06	tgfbPathway
-0.09	tercPathway	-0.06	HIGH MLL ARMSTRONG
0.08	il3Pathway	-0.05	mprPathway
0.07	il2rbPathway	-0.05	MAP00030 Pentose phosphate pathway
0.07	P53 DOWN	-0.05	myosinPathway
-0.07	mcalpainPathway	-0.05	MAP00530 Aminosugars metabolism
-0.07	ESC vs BM DOWN	-0.05	GLYCOL
-0.07	sppaPathway	0.05	compPathway
-0.07	telPathway	-0.05	radiation sensitivity
-0.07	nthiPathway	-0.05	p53hypoxiaPathway
-0.07	ST JNK MAPK Pathway	-0.05	rbPathway
-0.07	CR REPAIR	0.05	GATA1 WEISS
0.07	ctlPathway	0.05	alternativePathway
-0.07	HOXA9 UP	-0.05	no1Pathway
-0.06	UPREG BY HOXA9	-0.05	fbw7Pathway
-0.06	ccr3Pathway	0.05	MAP00380 Tryptophan metabolism
-0.06	FAB M7 DOWNING AML	-0.05	rarrxrPathway
-0.06	G13 Signaling Pathway	-0.05	cardiacegfPathway
-0.06	actinYPathway	0.05	il6Pathway
-0.06	ST Wnt Ca2 cyclic GMP Pathway	-0.05	cdc25Pathway
0.06	Matrix Metalloproteinases	0.05	SIG PIP3 signaling in B lymphocytes
-0.06	bcl2family and reg network	-0.05	ST Type I Interferon Pathway
0.06	lechtinPathway	-0.04	s1pPathway
0.06	pepiPathway	0.04	RUNX1 TARGETS
-0.06	parkinPathway	-0.04	chrebpPathway

Table 11.4E: C1→C2

w_1	Gene Set	w_1	Gene Set
-0.05	vitcbPathway	0.03	CSR brown pos list sam sorlie
-0.05	ace2Pathway	0.03	egfPathway
0.05	spryPathway	-0.03	gabaPathway
-0.04	ST Wnt Ca2 cyclic GMP Pathway	0.03	MAP00100 Sterol biosynthesis
-0.04	intrinsicPathway	0.03	atrbrcaPathway
0.04	ceramidePathway	0.03	tall1Pathway
0.04	igf1rPathway	0.03	mitrPathway
0.04	EGF receptor signaling pathway	-0.03	lymphocytePathway
0.04	SA DAG1	0.03	bcrPathway
0.03	rasPathway	-0.03	fxrPathway
-0.03	MAP00140 C21 Steroid hormone	0.03	GPCRis Class B Secretin-like
0.03	eea1Pathway	0.03	SA programmed cell death
0.03	raccycdPathway	-0.03	MAP00120 Bile acid biosynthesis
0.03	plk3Pathway	0.03	hbxPathway
-0.03	lairPathway	0.03	achPathway
0.03	il3Pathway	0.03	P53 DOWN
0.03	cdmacPathway	-0.03	ST T Cell Signal Transduction
-0.03	flumazenilPathway	0.03	epoPathway
-0.03	MAP00150 Andro estro metabo	-0.03	ephA4Pathway
-0.03	monocytePathway	0.02	eif2Pathway
0.03	cxcr4Pathway	0.02	aktPathway
0.03	pyk2Pathway	0.02	tpoPathway
0.03	MAP00600 Sphingoglycolipid	0.02	MAP00562 Inositol phosphate
0.03	ptdinsPathway	0.02	p38mapkPathway
-0.03	GLUCOSE UP	-0.02	cell growth and or maintenance

Table 11.4F: C2→C3

w_2	Gene Set	w_2	Gene Set
0.06	mll chimeric wo blast downing aml	0.04	HYPERDIPLOID DOWNING
0.06	lechtinPathway	-0.04	MAP00480 Glutathione metabolism
-0.05	egfr smrtePathway	0.04	PYR
0.05	CSR brown sam neg vantveer	0.04	pparaPathway
0.05	metPathway	0.04	CR TRANSPORT OF VESICLES
-0.05	CR HORMONAL FUNCTIONS	-0.04	feederPathway
0.05	GPCRs Class A Rhodopsin-like 2	-0.04	spryPathway
0.05	il2rbPathway	-0.04	SA DAG1
0.05	tpoPathway	0.04	lairPathway
0.05	compPathway	-0.04	CR REPAIR
0.05	MAP00600 Sphingoglycolipid metab	-0.04	telPathway
0.05	pdgfPathway	-0.04	rbPathway
0.05	41bbPathway	0.04	HEMO TF LIST JP
0.05	INSULIN 2F UP	0.04	KRAS TOP100 KNOCKDOWN
0.05	slrpPathway	0.04	rabPathway
0.05	ecmPathway	-0.04	ifnaPathway
0.05	classicPathway	0.04	ctlPathway
0.05	ptenPathway	0.04	HOXA9 DOWN
0.04	alternativePathway	0.04	cell adhesion receptor activity
0.04	lymphocytePathway	0.04	akap13Pathway
0.04	gleevecPathway	-0.04	eea1Pathway
-0.04	Wnt Signaling	-0.04	MAP00450 Selenoamino acid met
0.04	HTERT DOWN	0.04	ctla4Pathway
-0.04	radiation sensitivity	0.04	mll chimeric fusion downing
-0.04	Cell Cycle	0.04	tollPathway

Table 11.4G: C3→C4

w_3	Gene Set	w_3	Gene Set
-0.1	INSULIN 2F UP	-0.06	mll vs all from twomey top250
-0.09	UPREG BY HOXA9	-0.05	ptdinsPathway
-0.09	HOXA9 UP	-0.05	LEU DOWN
-0.09	HYPERDIPLOID DOWNING	-0.05	telPathway
-0.08	MELTON STEMNESS	-0.05	MAP00620 Pyruvate metabolism
-0.08	ESC vs BM DOWN	-0.05	DNA DAMAGE SIGNALLING
-0.07	Proteasome Degradation	-0.05	ST JNK MAPK Pathway
-0.07	Electron Transport Chain	-0.05	cxcr4Pathway
-0.07	gcrPathway	-0.05	rabPathway
-0.07	mll chimeric wo blast downing aml	-0.05	rarrxrPathway
-0.07	fetal liver vs adult liver gnf2	-0.05	arapPathway
-0.07	shhPathway	-0.05	gpcrPathway
-0.07	VOXPHOS	-0.05	HUMAN CD34 ENRICHED TF JP
0.06	P53 DOWN	0.05	GATA1 WEISS
-0.06	proteasomePathway	-0.05	MAP00190 Oxidative phosphorylation
-0.06	RAP DOWN	-0.05	tollPathway
-0.06	HIGH MLL ARMSTRONG	-0.05	LEM HSC
-0.06	BRCA UP	-0.05	MLL DOWNING ALL
-0.06	KRAS FINAL MIT	-0.05	Glycogen Metabolism
-0.06	fbw7Pathway	0.05	RUNX1 TARGETS
-0.06	tercPathway	-0.05	agpcrPathway
-0.06	p53hypoxiaPathway	-0.05	sppaPathway
-0.06	p35alzheimersPathway	-0.05	ccr5Pathway
-0.06	hifPathway	-0.05	plk3Pathway
-0.06	mRNA processing	-0.05	gsPathway

Table 11.4: A-C: Gene sets sorted by w_0, w_1, and w_2 parameters in early progression, N→A→C1, N→A, and A→C1 respectively. D-G: Gene sets sorted by w_0, w_1, w_2, and w_3 parameters in carcinoma progression, C1→C2→C3→C4, C1→C2, C2→C3, and C3→C4 respectively. Negative values of these parameters indicate up-regulation and positive vaules indicate down-regulation.

Table 11.5A:N→A→C1

w_0	Gene Set	w_0	Gene Set
-0.51	chr7p11	-0.14	chr12p11
-0.26	chr4q34	-0.14	chr4q12
-0.21	chr7q32	-0.14	chr9q34
-0.21	chr3p14	0.14	chr13q34
-0.2	chr5q35	-0.14	chr6p12
-0.2	chr4q13	-0.14	chr2p14
-0.2	chr3q26	0.14	chr6q21
-0.19	chr7p22	0.14	chr14q31
-0.18	chr1q24	-0.14	chr7q22
-0.18	chr5q21	-0.14	chr3q24
-0.17	chr20q11	-0.13	chr3p24
-0.17	chr12p13	-0.13	chr1p21
-0.16	chr12q13	-0.13	chr20p13
-0.16	chr6p21	0.13	chr20q13
-0.16	chr10q25	0.12	chr7q36
-0.16	chr7p13	0.12	chr1p12
-0.15	chr2q22	-0.12	chr16q24
-0.15	chr3q25	-0.12	chr4q26
-0.15	chr8q24	0.12	chr8p22
-0.15	chr18q11	0.12	chr12q24

Table 11.5B:N→A

w_1	Gene Set	w_1	Gene Set
-0.17	chr11p13	0.11	chr9p13
0.16	chr7p11	-0.11	chr6p25
0.15	chr13q12	0.11	chr7q32
0.15	chr12q13	-0.11	chr1q31
-0.14	chr4q34	0.11	chr9q34
-0.14	chr21	-0.11	chr20p13
-0.14	chr10q25	-0.11	chr8q24
-0.14	chr3p21	0.11	chr9q32
0.13	chr1q24	-0.1	chr2q33
0.13	chr7p22	-0.1	chr12q24
-0.13	chr15q15	0.1	chr2q35
-0.13	chr1q25	0.1	chr13
-0.12	chr4q22	-0.1	chr7q21
0.12	chr2p14	0.1	chr7q22
0.12	chr5q21	-0.1	chr2q22
-0.12	chr9q33	-0.1	chr4q13
0.12	chr21q21	-0.1	chr19p12
-0.12	chr2p25	0.1	chr5p14
0.12	chr6q14	0.1	chr3q26
-0.12	chr5q35	-0.1	chr21q22

Table 11.5C:A→C1

w_2	Gene Set	w_2	Gene Set
0.2	chr1p34	0.12	chr6q26
-0.2	chr7p13	0.12	chr7p15
0.18	chr2q35	0.12	chr7q31
0.18	chr12q21	-0.12	chr16p13
0.17	chr1q31	0.12	chr11p13
0.17	chr3p14	0.12	chr5q11
-0.16	chr6p23	0.12	chr20q12
0.16	chr12q22	0.12	chr1p12
-0.15	chr3p13	-0.12	chr11p12
0.15	chr18q12	-0.11	chr11q14
0.15	chr3q28	-0.11	chr5q35
-0.14	chr3q25	0.11	chr6q24
-0.14	chr7q33	-0.11	chr3q26
0.13	chr5q13	-0.11	chr6p21
0.13	chr7p12	0.11	chr4q13
-0.13	chr7q32	-0.11	chr4q25
0.13	chr4q34	0.11	chr12p11
-0.12	chr11q23	0.11	chr3p23
0.12	chr4q28	-0.11	chr14q23
0.12	chr5q12	-0.11	chrXq25

Table 11.5D:C1→C2→C3→C4

w_0	Gene Set	w_0	Gene Set
-0.51	chr14q24	-0.14	chr8q22
-0.26	chr14q11	-0.14	chr6p23
-0.21	chr6p12	-0.14	chr1p31
-0.21	chr14q12	0.14	chr12q22
-0.2	chr14q22	-0.14	chr3q27
-0.2	chr15q11	-0.14	chr7p13
-0.2	chr3q26	0.14	chr9q33
-0.19	chrXp11	0.14	chr18q21
-0.18	chr6q25	-0.14	chr12q21
-0.18	chr1p34	-0.14	chr11q23
-0.17	chrXq13	-0.13	chr14q21
-0.17	chr2p13	-0.13	chr11p15
-0.16	chr1q21	-0.13	chr14q32
-0.16	chr8q21	0.13	chr2q22
-0.16	chr10p15	0.12	chr22q12
-0.16	chr1q25	0.12	chr4q28
-0.15	chr14q23	-0.12	chr1p33
-0.15	chr8q23	-0.12	chr8q13
-0.15	chr8q11	0.12	chr9q32
-0.15	chr2q13	0.12	chr11q25

Table 11.5E:C1→C2

w_1	Gene Set	w_1	Gene Set
-0.17	chr1p31	0.11	chr15q15
0.16	chr7p11	-0.11	chr3p26
0.15	chr7p22	0.11	chr12q22
0.15	chr2p23	-0.11	chr4q32
-0.14	chr4q13	0.11	chr7q36
-0.14	chrXq23	-0.11	chr13q14
-0.14	chr13q31	-0.11	chrXp11
-0.14	chr5q12	0.11	chr2q14
0.13	chr5p12	-0.1	chr1q43
0.13	chr6p21	-0.1	chr1q24
-0.13	chr12q21	0.1	chr3p22
-0.13	chr10p12	0.1	chr11q24
-0.12	chr2p12	-0.1	chr2p16
0.12	chr11q13	0.1	chr17p11
0.12	chr16q23	-0.1	chr4q34
-0.12	chr3q24	-0.1	chr20p12
0.12	chr16p13	-0.1	chr18q22
-0.12	chr6q25	0.1	chr11q12
0.12	chr12q14	0.1	chr2p24
-0.12	chr13q21	-0.1	chr18q11

Table 11.5F:C2→C3

w_1	Gene Set	w_1	Gene Set
0.2	chr2q32	0.12	chr5q13
-0.2	chr14q11	0.12	chr16q12
-0.18	chr12p12	0.12	chr2p16
0.18	chrXq21	-0.12	chr9p21
0.17	chr1p22	0.12	chr15q22
0.17	chr2q22	0.12	chr6q22
-0.16	chr22q13	0.12	chr15q13
0.16	chr15q26	0.12	chr5q14
-0.15	chr12q13	-0.12	chr11q13
-0.15	chr6p21	-0.11	chr8q24
0.15	chr2p15	-0.11	chr9q22
-0.14	chr16q13	0.11	chrXq24
-0.14	chr11q24	-0.11	chr6q21
0.13	chr3p14	-0.11	chr11q12
0.13	chr8p22	0.11	chr4q34
-0.13	chr22q11	-0.11	chr6p12
0.13	chr15q21	0.11	chr5q15
-0.12	chr1q21	0.11	chr4p14
0.12	chr2q24	-0.11	chr3q28
0.12	chr2q31	-0.11	chr17q23

Table 11.5G:C3→C4

w_3	Gene Set	w_3	Gene Set
-0.42	chr14q24	-0.14	chr2p25
-0.2	chr5q13	-0.13	chr10p15
0.18	chr13q31	-0.13	chr8q12
-0.17	chr1p21	-0.13	chr2q14
-0.17	chr15q22	-0.13	chr7p22
-0.16	chr2p23	-0.13	chr10p11
-0.16	chr4q12	-0.13	chr15q25
-0.16	chr3p14	-0.13	chr2p14
-0.16	chr1p13	0.13	chr9q31
0.16	chr9q34	-0.12	chrXq24
-0.15	chr2p15	-0.12	chr5q14
-0.15	chr14q22	-0.12	chr1q21
-0.14	chr2p24	-0.12	chr7p15
-0.14	chr5q31	-0.12	chr8q22
-0.14	chr7q33	-0.12	chr3q13
-0.14	chr11p11	-0.12	chr2q31
-0.14	chr2p13	-0.12	chr5q15
-0.14	chr2q11	-0.12	chr8q23
-0.14	chr3p24	-0.12	chr12q24
-0.14	chr2p21	-0.12	chr2q23

Table 11.5: A-C: Postitional gene sets sorted by w_0, w_1, and w_2 parameters in early progression, N→A→C1, N→A, and A→C1 respectively. D-G: Positional gene sets sorted by w_0, w_1, w_2, and w_3 parameters in carcinoma progression, C1→C2→C3→C4, C1→C2, C2→C3, and C3→C4 respectively. Negative values of these parameters indicate up-regulation and positive vaules indicate down-regulation.

Bibliography

[1] JV Alvarez, PG Febbo, S Ramaswamy, M Loda, A Richardson, and DA Frank. Identification of a genetic signature of activated signal transducer and activator of transcription 3 in human tumors. *Cancer Res*, 65(12):5054–62, 2005.

[2] SA Amundson, KT Do, L Vinikoor, CA Paiz-Koch, ML Bittner, JM Trent, P Meltzer, and AJ Fornace. Stress-specific signatures: expression profiling of p53 wild-type and -null human cells. *Oncogene*, 24(28):4572–4579, 2005.

[3] TW Anderson and DA Darling. Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212, 1952.

[4] T Andre, E Quinaux, C Louvet, P Colin, E Gamelin, O Bouche, E Achille, P Piedbois, N Tubiana-Mathieu, A Boutan-Laroze, M Flesch, G Liedo, Y Raoul, I Debrix, M Buyse, and A Gramont. Phase iii study comparing a semimonthly with a monthly regimen of fluorouracil and leucovorin as adjuvant treatment for stage ii and iii colon cancer patients: Final results of gercor c96.1. *J Clin Oncol*, 25:3732–8, 2007.

[5] N Arber, CJ Eagle, J Spicak, I Racz, P Dite, J Hajer, M Zavoral, MJ Lechuga, P Gerletti, J Tang, RB Rosenstein, K Macdonald, P Bhadra, R Fowler, J Wittes, AG Zauber, SD Solomon, and B Levin. Celecoxib for the prevention of colorectal adenomatous polyps. *N Enlg J Med*, 355:885–95, 2006.

[6] E Bandres, R Malumbres, E Cubedo, B Honorato, R Zarate, A Labarga, U Gabisu, JJ Sola, and J Garcia-Foncillas. A gene signature of 8 genes could identify the risk of recurrence and progression in dukes'b colon cancer patients. *Oncol Rep*, pages 1089–94, 2007.

[7] A Barrier, P Boelle, F Roser, J Gregg, C Tse, D Brault, F Lacaine, S Houry, M Huguier, B Franc, A Flahault, A Lemoine, and S Dudoit. Stage ii colon cancer prognosis prediction by tumor gene expression profiling. *J Clin Oncol*, 24:4685–91, 2006.

[8] WT Barry, AB Nobel, and FA Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–9, 2005.

[9] J Bartkova, N Rezaei, M Liontos, P Karakaidos, D Kletsas, N Issaeva, L Vassil-

iou, E Kolettas, K Niforou, V Zoumpourlis, M Takaoka, H Nakagawa, F Tort, K Fugger, F Johansson, M Sehested, C Anderson, L Dryskjot, T Orntoft, C Kittas, T Helleday, T Halazonetis, J Bartek, and V Gorgoulis. Oncogene-induced senescence is part of the tumorigenesis barrier imposed by dna damage checkpoints. *Nature*, 444:633–637, 2006.

[10] DM Berman, N Desai, X Wang, SS Karhadkar, M Reynon, C Abate-Shen, PA Beachy, and MM Shen. Roles for hedgehog signaling in androgen production and prostate ductal morphogenesis. *Developmental Biology*, 267(2):387–398, 2004.

[11] MM Bertagnolli, CJ Eagle, AG Zauber, M Redston, SD Solomon, K Kim, J Tang, RB Rosenstein, J Wittes, D Corle, TM Hess, GM Woloj, F Boisserie, WF Anderson, JL Viner, D Bagheri, J Burn, DC Chung, T Dewar, R Foley, N Hoffman, F Macrae, RE Pruitt, JR Saltzman, B Salzberg, T Sylwestrowicz, GB Gordan, and ET Hawk. Celecoxib for prevention of sporadic colorectal adenomas. *N Engl J Med*, 355:873–84, 2006.

[12] K Bialkowska, S Kulkarni, X Du, DE Goll, TC Saido, and JE Fox. Evidence that beta3 integrin-induced rac activation involves the caplain-dependent formation of integrin clusters that are distint from the focal complexes and focal adhesions that form as rac and rhoa become active. *J Cell Biol*, 151(3):685–696, 2000.

[13] AH Bild, G Yao, JT Chang, Q Wang, A Potti, D Chasse, MB Joshi, D Harpole, JM Lancaster, A Berchuck, JA Olson, JR Marks, HK Dressman, M West, and J Nevis. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439(19):353–357, 2006.

[14] EP Black, E Huang, H Dressman, R Rempel, N Laakso, SL Asa, S Ishida, M West, and JR Nevins. Distinct gene expression phenotypes of cells lacking Rb and Rb family members. *Cancer Res*, 63(13):3716–3723, 2003.

[15] CD Blanke, BL Eisenberg, and MC Heinrich. Gastrointestinal stromal tumors. *Current Treatment Options in Oncology*, 2(6):485–491, 2001.

[16] W Bodmer. Familial adenomatous polyposis (fap) and its gene, apc. *Cytogenetics and Cell Genetics*, 86:99–104, 1999.

[17] A Boorsma, BC Foat, D Vis, F Klis, and HJ Bussemaker. T-profiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Res*, 33:W592–5, 2005.

[18] JL Bos, ER Fearon, SR Hamilton, MV Vries, JH van Boom, AJ van der Eb, and B Vogelstein. Prevalence of ras gene mutations in human colorectal cancers. *Nature*, 327(6120):293–297, 1987.

[19] T Breslin, P Eden, and M Krogh. Comparing functional annotation analyses with catmap. *BMC Bioinformatics*, 5:193, 2004.

[20] JP Brunet, P Tamayo, TR Golub, and JP Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, 101(12):4164–9, 2004.

[21] JK Chen, J Taipale, KE Young, T Maiti, and PA Beachy. Small molecule modulation of smoothened activity. *Proceedings of the National Academy of Sciences*, 99(22):14071, 2002.

[22] JT Chi, Z Wang, D Nuyten, EH Rodriguez, ME Schaner, A Salim, Y Wang, GB Kristensen, A Helland, A Borresen-Dale, A Giaccia, MT Longaker, T Hastie, GP Yang, M Vijver, and PO Brown. Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS Medicine*, 3(3):e47–e47, 2006.

[23] G Christofori and H Semb. The role of cell-adhesion molecule E-cadherin as a tumour-suppressor gene. *Trends Biochem Sci*, 24(2):73–76, 1999.

[24] RD Cook and S Weisberg. discussion of "sliced inverse regression for dimension reduction". *Journal of the American Statistical Association*, 86:328–332, 1991.

[25] MK Cooper, JA Porter, KE Young, and PA Beachy. Teratogen-mediated inhibition of target tissue response to shh signaling. *Proc. Natl. Acad. Sci. USA*, 92:5510, 1995.

[26] S Corda, C Laplace, E Vicaut, and J Duranteau. Rapid reactive oxygen species production by mitochondria in endothelial cells exposed to tumor necrosis factor-alpha is mediated by ceramide. *Am. J. Respir. Cell Mol. Biol.*, 24(6):762–768, 2001.

[27] R Dagher, M Cohen, G Williams, M Rothmann, J Gobburu, G Robbie, A Rahman, G Chen, A Staten, and D Griebel. Approval summary: Imatinib mesylate in the treatment of metastatic and/or unresectable malignant gastrointestinal stromal tumors. *Clinical Cancer Research*, 8(10):3034, 2002.

[28] N Dahmane, J Lee, P Robins, P Heller, and A Ruiz i Altaba. Activation of the transcription factor gli1 and the sonic hedgehog signalling pathway in skin tumours. *Nature*, 389(6653):876–81, 1997.

[29] M Del Rio, F Molina, C Bascoul-Mollevi, V Copois, F Bibeau, P Chalbos, C Bareil, A Kramer, N Salvetat, C Fraslon, E Conseiller, V Granci, B Leblanc, B Pau, P Martineau, and M Ychou. Gene expression signatures for the prediction of response to leucovorin, fluorouracil, and irinotecan in advanced colorectal cancer patients. *J Clin Oncol*, 25:773–80, 2007.

[30] GD Demetri. Targeting c-kit mutations in solid tumors: scientific rationale and novel therapeutic options. *Semin Oncol*, 28(5 Suppl 17):19–26, 2001.

[31] GD Demetri, M von Mehren, CD Blanke, AD Van den Abbeele, B Eisenberg, PJ Roberts, MC Heinrich, DA Tuveson, S Singer, and M Janicek. Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. *New England Journal of Medicine*, 347(7):472, 2002.

[32] SM Dhanasekaran, TR Barrett, D Ghosh, R Shah, S Varambally, K Kurachi, KJ Pienta, MA Rubin, and AM Chinnaiyan. Delineation of prognostic biomakers in prostate cancer. *Nature*, 412(6849):822–826, 2001.

[33] R Di Micco, M Fumagalli, A Cicalese, S Piccinin, P Gasparini, C Luise, C Schurra, M Garre, PG Nuciforo, A Benismon, R Maestro, PG Pelicci, and F Fagagna. Oncogene-induced senescence is a dna damage response triggered by dna hyper-replication. *Nature*, 444:638–642, 2006.

[34] I Dinu, JD Potter, T Mueller, Q Lui, AJ Adewale, GS Jhangri, G Einecke, KS Famulski, P Halloran, and Y Yasui. Improving gene set analysis of microarray data by sam-gs. *BMC Bioinformatics*, 8:242, 2007.

[35] R Durrett. *Stochastic Calculus: A Practical Introduction*. CRC Press, Boca Raton, FL, 1996.

[36] E Edelman, A Porrello, J Guinney, B Balakumaran, A Bild, P Febbo, and S Mukherjee. Analysis of sample set enrichment scores: assaying the enrichment of sets of gene for individual samples in genome-wide expression profiles. *Bioinformatics*, 22(14):e108–e116, 2006.

[37] EJ Edelman, J Guinney, JT Chi, PG Febbo, and S Mukherjee. Modeling cancer progression via pathway dependencies. *PLoS Comput Biol*, 4(2):e28, 2008.

[38] JE Egan, AB Hall, BA Yatsula, and D Bar-Sagi. The bimodal regulation of epidermal growth factor signaling by human sprouty proteins. *Proc Natl Acad Sci U S A*, 99(9):6041–6046, 2002.

[39] RM Elledge and CD Allred. Prognostic and predictive value of p53 and p21 in breast cancer. *Breast Cancer Research and Treatment*, 52:79–98, 1998.

[40] T Evgeniou, CA Micchelli, and M Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

[41] ER Fearon and B Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61:759–767, 1990.

[42] PG Febbo, JP Richie, DJ George, M Loda, J Manola, S Shankar, AS Barnes, C Tempany, W Catalona, PW Kantoff, and WK Oh. Neoadjuvant docetaxel before radical prostatectomy in patients with high-risk localized prostate cancer. *Clin Cancer Res*, 11(14):5233–40, 2005.

[43] BJ Feldman and D Feldman. The development of androgen-independent prostate cancer. *Nature*, 1:34–45, 2001.

[44] W Feller. *An Introduction to Probability Theory and Its Applications, Vol 1.* John Wiley & Sons, New York, 1971.

[45] IJ Fidler. The pathogenesis of cancer metastasis: the seed and soil hypothesis revisited. *Nature Reviews Cancer*, 3:453–458, 2003.

[46] K Forrester, C Almoguera, K Han, WE Grizzle, and M Perucho. Detection of high incidence of k-ras oncogenes during human colon tumorigenesis. *Nature*, 327(6120):298–303, 1987.

[47] NO Fortunel, HH Otu, HH Ng, J Chen, X Mu, T Chevassut, X Li, M Joseph, C Bailey, and JA Hatzfeld. Stemness: Transcriptional profiling of embryonic and adult stem cells and a stem cell molecular signature. *Science*, 302(393):393b, 2003.

[48] J Fox. *An R and S-PLUS Companion to Applied Regression.* Sage Publications, Inc., Thousand Oaks, California, 2002.

[49] M Frank-Kamenetsky, XM Zhang, S Bottega, O Guicherit, H Wichterle, and H Dudek. Bimcrot, d. fy wang, s. jones, j. shulok, ll rubin and ja porter, small-

molecule modulators of hedgehog signaling: identification and characterization of smoothened agonists and antagonists. *J. Biol*, 1(1):10, 2002.

[50] K Friedrichs, S Gluba, H Eidtmann, and W Jonat. Overexpression of p53 and prognosis in breast cancer. *Cancer*, 72(12):3641–3647, 1993.

[51] N Fujita, DL Jaye, M Kajita, C Geigerman, CS Moreno, and PA Wade. MTA3, a Mi-2/NuRD complex subunit, regulates an invasive growth pathway in breast cancer. *Cell*, 113(2):207–219, 2003.

[52] ME Garber, OG Troyanskaya, K Schluens, S Petersen, Z Thaesler, M Pacyna-Gengelbach, M van de Rijn, GD Rosen, CM Perou, RI Whyte, RB Altman, PO Brown, D Botstein, and I Petersen. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A*, 98(24):13784–9, 2001.

[53] I Garcia-Higuera, T Taniguchi, S Ganesan, M Meyn, C Timmers, J Hejan, M Grompe, and A Andrea. Interaction of the fanconi anemia proteins and brca1 in a common pathway. *Mol. Cell*, 7(2):249–262, 2001.

[54] KS Garman, CR Acharya, M Grade, J Gaedcke, S Sud, Shivani, K Walters, G Ginsburg, W Barry, AM Diehl, D Provenzale, BM Ghadimi, T Ried, D Hsu, JR Nevins, and A Potti. A genomic approach to dissecting colon cancer progression yields biological insights into therapeutic opportunities. *Pub Natl Acad Sci*, In publication.

[55] GV Glinsky, AB Glinskii, AJ Stephenson, RM Hoffman, and WL Gerald. Gene expression profiling predicts clinical outcome of prostate cancer. *J Clin Invest*, 113(6):913–23, 2004.

[56] JJ Goeman, SA van de Geer, F de Kort, and HC van Houwelingen. A global test for groups of genes: testing assocation with a clinical outcome. *Bioinformatics*, 20:93–9, 2005.

[57] TR Golub, DK Slonim, P Tamayo, C Huard, M Gaasenbeek, JP Mesirov, H Coller, ML Loh, JR Downing, MA Caligiuri, CD Bloomfield, and ES Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7, 1999.

[58] IY Gore and G Popjak. Sterol biosynthesis in neoplastic cells: Utilization of [14C] acetate and [2-14C] mevalonate. *Biochem. J.*, 84:93–99, 1962.

[59] I Gross, B Bassit, M Benezra, and JD Licht. Mammalian sprouty proteins inhibit cell growth and differentiation by preventing ras activation. *J Biol Chem*, 276(49):46460–46468, 2001.

[60] GP Gupta and J Massague. Cancer metastasis: Building a framework. *Cell*, 127:679–695, 2006.

[61] D Hanahan and RA Weinberg. The hallmarks of cancer. *Cell*, 100:57–70, 2000.

[62] C Haqq, M Nosrati, D Sudilovsky, J Crothers, D Khodabakhsh, BL Pulliam, S Federman, JR Miller, RE Allen, MI Singer, SPL Leong, B Ljung, RW Sagebiel, and M Kashani-Sabet. The gene expression signatures of melanoma progression. *Proc Natl Acad Sci U S A*, 102(17):6092–6097, 2005.

[63] C Harley, AB Futcher, and CW Greider. Telomeres shorten during ageing of human fibroblasts. *Nature*, 346:866–868, 1990.

[64] AL Harris. Hypoxia – a key regulatory factor in tumour growth. *Nature Reviews Cancer*, 2(1):38–47, 2002.

[65] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning, Data Mining, Inference and Prediction.* Springer Verlag, New York, 2000.

[66] CA Heinlein and C Chang. Androgen receptor in prostate cancer. *Endocrine Reviews*, 25(2):276–308, 2004.

[67] MC Heinrich, CL Corless, M Von Mehren, H Joensuu, GD Demetri, CD Blanke, S Dimitrijevic, B Kiese, CDM Fletcher, and JA Fletcher. Pdgfra and kit mutations correlate with the clinical responses to imatinib mesylate in patients with advanced gastrointestinal stromal tumors (gist). *Proc Am Soc Clin Oncol*, 22:A3274, 2003.

[68] SB Ho, A Hyslop, R Albrecht, A Jacobson, M Spencer, DA Rothenberger, GA Niehans, J D'Cunha, and RA Kratzke. Quantification of colorectal cancer mircometastases in lymph nodes by nested and real-time reverse transcriptase-pcr analysis for carcinoembryonic antigen. *Clin Cancer Res*, 10:5777–84, 2004.

[69] KS Hoek, NC Schlegel, P Brafford, A Sucker, S Ugurel, R Kumar, B Weber, KL Nathanson, DJ Philips, M Herlyn, D Schadendorf, and R Dummer. Metastatic potential of melanomas defined by specific gene expression profiles with no BRAF signature. *Pigment Cell Res*, 19:290–302, 2006.

[70] Y Hong, KS Ho, KW Eu, and PY Cheah. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clinical Cancer Research*, 13(4):1107–1114, 2007.

[71] E Huang, S Ishida, J Pittman, H Dressman, A Bild, M Kloos, M D'Amico, RG Pestell, M West, and JR Nevins. Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat Genet*, 34(2):226–230, 2003.

[72] JP Incardona. The teratogenic veratrum alkaloid cyclopamine inhibits sonic hedgehog signal transduction. *Development*, 125(18):3553–3562, 1998.

[73] A Inga, F Storici, TA Darden, and MA Resnick. Differential transactivation by the p53 transcription factor is highly dependent on p53 level and promoter target sequence. *Molecular and Cellular Biology*, 22(24):8612–8625, 2005.

[74] AV Ivshina, J George, O Senko, B Mow, TC Putti, J Smeds, T Lindahl, Y Pawitan, P Hall, H Nordgren, JFL Wong, ET Liu, J Bergh, VA Kuznetsov, and LD Miller. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Research*, 66:10292–10301, 2006.

[75] A Jemal, R Siegel, E Ward, Y Hao, J Xu, T Murray, and M Thun. Cancer statistics, 2008. *CA Cancer J Clin*, 58:71–96, 2008.

[76] A Jemal, R Siegel, E Ward, T Murray, J Xu, and MJ Thun. Cancer statistics, 2007. *CA: A Cancer Journal for Clinicians*, 57(1):43–66, 2007.

[77] D Kalderon. Similarities between the hedgehog and wnt signaling pathways. *Trends in Cell Biology*, 12(11):523–531, 2002.

[78] K Kannan, N Amariglio, G Rechavi, J Jakob-Hirsch, I Kela, N Kaminski, G Getz, E Domany, and D Givol. Dna microarrays identification of primary and secondary target genes regulated by p53. *Oncogene*, 20(18):2225–2234, 2001.

[79] H Kim, DK Lee, JW Choi, JS Kim, SC Park, and HD Youn. Analysis of the effect of aging on the response to hypoxia by cdna microarray. *Mechanisms of Ageing and Development*, 124:941–949, 2003.

[80] SY Kim and DJ Volsky. PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics*, 6(144), 2005.

[81] AG Knudson. Mutation and cancer: Statistical study of retinoblastoma. *Proc. Nat. Acad. Sci. USA*, 68(4):820–823, 1971.

[82] K Koinuma, Y Yamashita, W Liu, H Hatanaka, K Kurashina, T Wada, S Takada, R Kaneda, YL Choi, SI Fujiwara, Y Miyakura, H Nagai, and H Mano. Epigenetic silencing of axin2 in colorectal carcinoma with microsatellite instability. *Oncogene*, 25(1):139–146, 2006.

[83] SW Kong, WT Pu, and PJ Park. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, 22:2373–80, 2005.

[84] S Kotz and NL Johnson. *Breakthroughs in Statistics Volume II.* Springer-Veriag, New York, NY, 1992.

[85] J Lamb, ED Crawford, D Peck, JW Modell, IC Blat, MJ Wrobel, J Lerner, J Brunet, A Subramanian, KN Ross, M Reich, H Hieronymus, G Wei, SA Armstrong, SJ Haggarty, PA Clemons, R Wei, SA Carr, ES Lander, and TR Golub. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313:1929–35, 2006.

[86] HK Lee, W Braynen, K Keshav, and P Pavlidis. Erminej: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, 6:93–9, 2005.

[87] Y Lee and JA Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society*, 58(4):619–678, 1996.

[88] MO Leonard, DC Cottell, C Godson, HR Brady, and CT Taylor. The role of hif-1 alpha in transcriptional regulation. *Journal of Biological Chemistry*, 278(41):40296–40394, 2003.

[89] KC Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86:316–342, 1991.

[90] KH Lim and CM Counter. Reduction in the requirement of oncogenic ras signaling to activation of PI3K/AKT pathway during tumor maintenance. *Cancer Cell*, 8:381–392, 2005.

[91] XJ Ma, R Salunga, T Tuggle, J Gaudet, E Enright, T McQuary, T Payette, M Pistone, K Stecker, BM Zhang, YX Zhou, H Varnholt, B Smith, M Gadd,

E Chatfield, J Kessler, TM Baer, MG Erlander, and DC Sgroi. Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci U S A*, 100(10):5974–5979, 2003.

[92] M Maggioni, S Mukherjee, Q Wu, and J Guinney. Learning gradients: predictive models that reflect geometry and dependencies, 2007. Submitted to Journal of Machine Learning Research.

[93] PK Majumder, PG Febbo, R Bikoff, R Berger, Q Xue, LM McMahon, J Manola, J Brugarolas, TJ McDonnell, TR Golub, M Loda, HA Lane, and WR Sellers. mTOR inhibition reverses Akt-dependent prostate intraepithelial neoplasia through regulation of apoptotic and HIF-1-dependent pathways. *Nat Med*, 10(6):594–601, 2004.

[94] DJ Manalo, A Rowan, T Lavoie, L Natarajan, BD Kelly, SQ Ye, JG Garcia, and GL Semenza. Transcriptional regulation of vascular endothelial cell responses to hypoxia by hif-1. *Blood*, 105(2):659–69, 2005.

[95] JC Marsit, M Liu, HH Nelson, M Posner, M Suzuki, and KT Kelsey. Inactivation of the fanconi anemia/brca pathway in lung and oral cancers: implications for treatment and survival. *Oncogene*, 23(4):1000–1004, 2004.

[96] D Matei, DD Chang, and MH Jeng. Imatinib mesylate (gleevec) inhibits ovarian cancer cell growth through a mechanism dependent on platelet-derived growth factor receptor alpha and akt inactivation. *Clin Cancer Res*, 10(2):681, 2004.

[97] MJ Mauro and BJ Druker. Sti571: Targeting bcr-abl as therapy for cml. *The Oncologist*, 6(3):233, 2001.

[98] SM Mense, A Sengupta, M Zhou, C Lan, G Bentsman, DJ Volsky, and L Zhang. Gene expression profiling reveals the profound upregulation of hypoxia-responsive genes in primary human astrocytes. *Physiological Genomics*, 25:435–449, 2006.

[99] F Michor, Y Iwasa, and MA Nowak. Dynamics of cancer progression. *Nature Reviews*, 4:197–206, 2004.

[100] VK Mootha, CM Lindgren, KF Eriksson, A Subramanian, S Sihag, J Lehar, P Puigserver, E Carlsson, M Ridderstrale, E Laurila, N Houstis, MJ Daly, N Patterson, JP Mesirov, TR Golub, P Tamayo, B Spiegelman, ES Lander,

JN Hirschhorn, D Altshuler, and LC Groop. Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267–273, 2003.

[101] A Mukherjee, P Tamayo, S Rogers, R Rifkin, A Engle, C Campbell, TR Golub, and JP Mesirov. Estimating dataset size requirements for classifying dna microarray data. *Journal Computational Biology*, 10(2):119–142, 2003.

[102] S Mukherjee and Q Wu. Estimation of gradients and coordinate covariation in classification. *Journal of Machine Learning Research*, 7:2481–2514, 2006.

[103] S Mukherjee, Q Wu, and D-X Zhou. Gradient learning and feature selection on manifolds. *Under revision in Annals of Statistics*, 2006.

[104] S Mukherjee and D-X Zhou. Learning coordinate covariances via gradients. *Journal of Machine Learning Research*, 7:519–549, 2006.

[105] MA Nowak, F Michor, NL Komarova, and Y Iwasa. Evolutionary dynamics of tumor suppressor gene inactivation. *PNAS*, 101(29):10635–10638, 2004.

[106] K Offit, O Levran, B Mullaney, K Mah, K Nafa, S Batish, R Diotti, H Schneider, A Deffenbaugh, T Scholl, V Proud, M Robson, L Norton, N Ellis, H Hanenberg, and A Auerbach. Shared genetic susceptibility to breast cancer, brain tumors, and fanconi anemia. *Journal of the National Cancer Institute*, 95(20):1548–1551, 2003.

[107] M Olivier, R Eeles, M Hollstein, M Khan, CC Harris, and P Hainaut. The iarc tp53 database: New online mutation analysis and recommendations to users. *Human Mutation*, 19(6):607–614, 2002.

[108] KH Pan, CJ Lih, and SN Cohen. Effects of threshold choice on biological conclusions reached during analysis of gene expression by dna microarrays. *Proc Natl Acad Sci*, 102:8961–8965, 2005.

[109] CM Perou, T Sorlie, MB Eisen, M van de Rijn, SS Jeffrey, CA Rees, JR Pollack, DT Ross, H Johnsen, LA Akslen, O Fluge, A Pergamenschikov, C Williams, SX Zhu, PE Lonning, AL Borresen-Dale, PO Brown, and D Botsein. Molecular portraits of human breast tumours. *Nature*, 406:747–52, 2000.

[110] K Pestell. Anti-cancer drug success emerges from molecular biology origins. *Trends in Pharmacological Sciences*, 22(7):342–342, 2001.

[111] SL Pomeroy, P Tamayo, M Gaasenbeek, LM Sturla, M Angelo, ME McLaughlin, JY Kim, LC Goumnerova, PM Black, C Lau, JC Allen, D Zagzag, JM Olson, T Curran, C Wetmore, JA Biegel, T Poggio, S Mukherjee, R Rifkin, A Califano, G Stolovitzky, DN Louis, JP Mesirov, ES Lander, and TR Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–42, 2002.

[112] A Potti, S Mukherjee, R Petersen, HK Dressman, A Bild, J Koontz, R Kratzke, MA Watson, M Kelley, GS Ginsburg, M West, DH Jr. Harople, and JR Nevins. A genomic strategy to refine prognosis in early stage non-small cell lung carcinoma. *N Engl J Med*, 355(6):570–580, 2006.

[113] SM Powell, N Zilz, Y Beazer-Barclay, TM Bryan, ST Hamilton, SN Thibodeau, B Vogelstein, and KW Kinzler. Apc mutations occur early during colorectal tumorigenesis. *Nature*, 395:235–237, 1992.

[114] C Prives and PA Hall. The p53 pathway. *J Pathol.*, 187(1):112–126, 1999.

[115] D Qualtrough, A Buda, W Gaffield, AC Williams, and C Paraskeva. Hedgehog signalling in colorectal tumour cells: Induction of apoptosis with cyclopamine treatment. *International Journal of Cancer*, 110(6):831–837, 2004.

[116] MD Radmacher, R Simon, R Desper, R Taetle, AA Schaffer, and MA Nelson. Graph models of oncogenesis with an application to melanoma. *J. theor. Biol.*, 212:535–548, 2001.

[117] J Rahnenfuhrer, N Beerenwinkel, WA Schulz, C Hartmann, A Deimling, B Wullich, and T Lengauer. Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics*, 21(10):2438–2446, 2005.

[118] S Rodenhuis, RJC Slebos, AJM Boot, SG Evers, WJ Mooi, SS Wagenaar, PC Bodegom, and JL Bos. Incidence and possible clinical significance of k-ras oncogene activation in adenocarcinoma of the human lung. *Cancer Research*, 48:5738–5741, 1998.

[119] A Rostom, C Dube, G Lewin, A Tsertsvadze, N Barrowman, C Code, M Sampson, and D Moher. Nonsteroidal anti-inflammatory drugs and cyclooxygenase-2 inhibitors for primary prevention of colorectal cancer: A systematic review prepared for the u.s. preventive services task force. *Ann Intern Med*, 146:376–89, 2007.

[120] R Salgia and AT Skarin. Molecular abnormalities in lung cancer. *Journal of Clinical Oncology*, 16:1207–1217, 1998.

[121] DS Salomon, R Brandt, F Ciardiello, and N Normanno. Epidermal growth factor-related peptides and their receptors in human malignancies. *Crit Rev Oncol Hematol*, 19:183–232, 1995.

[122] M Sattler and JD Griffin. Mechanisms of transformation by the bcr/abl oncogene. *International Journal of Hematology*, 73(3):278–291, 2001.

[123] GL Semenza. Hypoxia-inducible factor 1: oxygen homeostasis and disease pathophysiology. *Trends Mol Med*, 7(8):345–350, 2001.

[124] ZM Shao, J Wu, ZZ Shen, and M Nguyen. p53 mutation in plasma dna and its prognostic value in breast cancer patients. *Clinical Cancer Research*, 7:2222–2227, 2001.

[125] P Sharma, Veeranna, M Sharma, N Amin, RK Sihag, P Grant, N Ahn, AB Kulkarni, and HC Pant. Phosphorylation of MEK1 by cdk5/p35 down-regulates the mitogen-activated protein kinase pathway. *J Biol Chem*, 277(1):528–534, 2002.

[126] D Singh, PG Febbo, K Ross, DG Jackson, J Manola, C Ladd, P Tamayo, AA Renshaw, AV D'Amico, JP Richie, ES Lander, M Loda, PW Kantoff, TR Golub, and WR Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

[127] T Sjoblom, S Jones, LD Wood, DW Parsons, J Lin, TD Barber, D Mandelker, RJ Leary, P Janine, N Silliman, S Szabo, P Buckhaults, C Farrell, P Meeh, SD Markowitz, J Willis, D Dawson, J Wilson, AF Gazder, J Hartigan, L Wu, C Lui, G Parmigiani, BH Park, KE Bachman, N Papadoloulos, B Vogelstein, KW Kinzler, and VE Velculescu. The consensus coding sequences of human breast and colorectal cancers. *Science*, 314:268–274, 2006.

[128] AP Smith, K Hoek, and D Becker. Whole-genome expression profiling of the melanoma progression pathway reveals marked molecular differences between nevi/melanoma in situ and advanced-stage melanomas. *Cancer Biology and Therapy*, 4(9):1018–1029, 2005.

[129] LL Smith, HA Coller, and JM Roberts. Telomerase modulates expression of growth-controlling genes and enhances cell proliferation. *Nat. Cell Biology*, 5(5):474–479, 2003.

[130] T Speed and H Kiiveri. Gaussian Markov distributions over finite graphs. *Annals of Statistics*, 14:138–150, 1986.

[131] M Stanbrough, GJ Bubley, K Ross, TR Golub, MA Rubin, TM Penning, PG Febbo, and SP Balk. Increased expression of genes converting adrenal androgens to testosterone in androgen-independent prostate cancer. *Cancer Research*, 66(5):2815–2825, 2006.

[132] A Subramanian, P Tamayo, VK Mootha, S Mukherjee, BL Ebert, MA Gillette, A Paulovich, SL Pomeroy, TR Golub, ES Lander, and JP Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, 2005.

[133] A Sweet-Cordero, S Mukherjee, A Subramanian, H You, JJ Roix, C Ladd-Acosta, J Mesirov, TR Golub, and T Jacks. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat Genet*, 37(1):48–55, 2005.

[134] J Taipale, JK Chen, MK Cooper, B Wang, RK Mann, L Milenkovic, MP Scott, and PA Beachy. Effects of oncogenic mutations in smoothened and patched can be reversed by cyclopamine. *Nature*, 406(6799):1005–9, 2000.

[135] D Talantov, A Mazumder, JX Yu, T Briggs, Y Jiang, J Backus, D Atkins, and Y Wang. Novel genes associated with malignant melanoma but not benign melanocytic lesions. *Clinical Cancer Research*, 11(20):7234–7242, 2005.

[136] L Tian, SA Greenberg, SW Kong, J Altschuler, IS Kohane, and PJ Park. Discovering statistically significant pathways in expression data in terms of gene sets: methodological issues. *Proc Natl Acad Sci U S A*, 102:13544–9, 2005.

[137] J Tomfohr, J Lu, and TB Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6:225, 2005.

[138] J Tomfohr, J Lu, and TB Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6(225), 2005.

[139] SA Tomlins, R Mehra, DR Rhodes, X Cao, L Wang, SM Dhanasekaran, S Kalyana-Sundaram, JT Wei, MA Rubin, KJ Pienta, RB Shah, and AM Chinnaiyan. Integrative molecular concept modeling of prostate cancer progression. *Nature Genetics*, 39(1):41–51, 2007.

[140] VN Vapnik. *Statistical learning theory.* J. Wiley and Sons, 1998.

[141] B Vogelstein, ER Fearon, SR Hamilton, SE Kern, AC Preisinger, M Leppert, Y Nakamura, R White, AM Smits, and JL Bos. Genetic alterations during colorectal-tumor development. *New England Journal of Medicine*, 319(9):525532, 1988.

[142] B Vogelstein and KW Kinzler. Cancer genes and the pathways they control. *Nature Medicine*, 10(8):789–799, 2004.

[143] Y Wang, T Jatkoe, Y Zhang, MG Mutch, D Talantov, J Jiang, HL McLeod, and D Atkins. Gene expression profiles and molecular markers to predict recurrence of dukes'b colon cancer. *J Clin Oncol*, 22:1564–1571, 2004.

[144] P Warnat, R Eils, , and B. Brors. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6(265), 2005.

[145] M Welsh, LM Sapinoso, AI Su, SG Kern, J Wang-Rodriguez, CA Moskaluk, HF Jr Frierson, and GM Hampton. Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res.*, 61(16):5974–5978, 2001.

[146] JA Williams. Hedgehog signaling pathway as a target for therapeutic intervention in basal cell carcinoma. *Drug News Perspect*, 16(10):657–662, 2003.

[147] LD Wood, DW Parsons, S Jones, J Lin, T Sjoblom, RJ Leary, D Shen, SM Boca, T Barber, and J Ptak. The genomic landscapes of human breast and colorectal cancers. *Science*, 318:1108–1113, 2007.

[148] Y Yarden and A Ullrich. EGF and erbB2 receptor overexpression in human tumors. Growth factor receptor tyrosine kinases. *Annu. Rev. Biochem*, 57:443–478, 1988.

[149] Y Zhang, Y Xiong, and WG Yarbrough. Arf promotes mdm2 degradation and stabilizes p53: Arf-ink4a locus deletion impairs both the rb and p53 tumor suppression pathways. *Current Biology*, 10(9):539–542, 2000.

[150] L Zhou, N An, RC Haydon, Q Zhou, H Cheng, Y Peng, W Jiang, HH Luu, P Vanichakarn, and JP Szatkowski. Tyrosine kinase inhibitor sti-571/gleevec down-regulates the -catenin signaling activity. *Cancer Letters*, 193(2):161–170, 2003.

Biography

Elena Edelman was born July 8, 1982 and grew up in Bethesda, Maryland. Elena completed her undergraduate education at Duke University on May 2004. She received a Bachelor of Science in Mathematics and a Bachelor of Arts in Biology, graduating Cum Laude. In the fall of 2004 she matriculated into the Computational Biology and Bioinformatics program at Duke University. She joined Sayan Mukherjee's lab in the spring of 2005 and began her work in the lab by developing a statistical methodology to measure pathway enrichment in an individual. In August 2006 she was invited to give an oral presentation about this work at the Intelligent Systems for Molecular Biology (ISMB) conference in Brazil. She extended this work to modeling cancer progression on the pathway level. In April 2008 she was awarded an NSF grant to present her work at the Isaac Netwon Institute in Cambridge, UK. Her research has resulted in the scientific publications listed below:

1. Edelman E, Porrello A, Guinney J, Balakumaran B, Bild A, Febbo P, and Mukherjee S. (2006) Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics* 22:e108-e116.

2. Edelman E, Guinney J, Chi J, Febbo P, and Mukherjee S. (2008) Modeling cancer progression via pathway dependencies. *PLoS Computational Biology* 4(2):e28.

3. Garman KS, Acharya CR, Edelman E, Grade M, Gaedcke J, Sud S, Walters

K, Ginsburg G, Barry W, Dieh AM, Povenzale D, Ghadimi BM, Ried T, Hsu D, Nevins JR, Mukherjee S, Potti A. (2008) A genomic approach to dissecting colon cancer progression yields biological insights into therapeutic opportunities. *Proc Natl Acad Sci* 105(49):19432-19437.