IDENTIFYING GENETIC DRIVERS OF CANCER MORPHOLOGY

Pang Wei Koh
Stanford University

An Undergraduate Honors Thesis Submitted to
the Department of Computer Science
Stanford University

Principle Advisor
Daphne Koller
Professor, Department of Computer Science
Stanford University

Co-Advisor
Andrew H. Beck, MD
Assistant Professor, Department of Pathology
Harvard Medical School

Abstract

Cancer is a leading cause of mortality worldwide, claiming the lives of nearly 8 million people in 2008 alone. To effectively treat cancer, we need a holistic understanding of how aberrations in key cellular pathways can drive tumor formation. Current research, however, remains predominantly focused on molecular data, despite the fact that clinical diagnosis and prognostication rely primarily on the morphologic analysis of histologic data.

In this thesis, we develop 1) an image processing pipeline capable of extracting clinically-relevant morphological features from whole-slide tissue samples, and 2) a system of multi-task regressions to robustly and efficiently associate gene expression levels with transformations in specific morphological traits. These allow us to distill massive amounts of histological and molecular data into a set of unbiased and testable hypotheses regarding the effect of specific genes on particular clinically-relevant aspects of tumor morphology.

We demonstrate our system on matching histological and molecular data from a total of 574 breast cancer patients from two independent cohorts: 248 from the Netherlands Cancer Institute, and 326 from the Cancer Genome Atlas. Our results corroborate many associations between known onco- and tumor-suppressor- genes and tumor morphology, including the recently discovered role of *CDC6* in epithelial-mesenchymal transition. We also identify several putative and previously unknown key genes in breast carcinoma, together with their purported role in tumor morphology, e.g., the role of *VIPR2* in promotion of the stromal environment. These promising results pave the way for future investigative work into these genes, and show the viability of our integrative analysis of morphological and molecular data.

Acknowledgements

This project would not have been possible without the guidance and mentorship of my advisors: Daphne, for giving me the courage to try all sorts of random ideas because I knew that there'd be somebody to tell me if I were being completely crazy; and Andy, for putting up with questions like "Is that purple thing a cell?". Thank you both - you are real inspirations.

Zhenghao, my fellow office-mate, was the bouncing board and inspiration for many of the ideas in this thesis; likewise, Sara Mostafavi introduced me to the messy world of gene expression and patiently explained for the umpteenth time how module networks actually worked.

This thesis is the culmination of a research trajectory that really started with us sitting starry-eyed in Andrew Ng's office, during freshman orientation. Along the way, I have been very fortunate to work with and learn from Andrew and his grad students Quoc, Ngiam, and Andrew Saxe; they have all been inspirations, and in many ways, what (little) I now know of research is thanks to them. Big thanks also goes to my fellow students-of-Quoc and machine learning buddies Zhenghao, Marc, Daniel, and Jean, without whom I would still be stuck trying to figure out variational Bayes. At least now we can be stuck together!

Last but certainly not least, thanks goes to Deanna and my family for being wonderfully supportive, and looking interested whenever I ramble on about machine learning. My mother taught me biology as a kid, and my father computers; none of this would have come together if not for that. Thank you.

Contents

List of Algorithms

List of Figures

Chapter 1

Introduction

Cancer is a disease of the genome, and to treat it we must first understand and characterize the underlying genetic pathways that drive cancer growth. But every tumor is unique, and every cancer cell contains an overwhelming number of molecular aberrations and genetic lesions, the majority of which are non-essential "passenger" mutations. How can we sift through the chaff to find the important "driver" mutations - and the crucial molecular pathways - that contain the key to cancer?

Our answer to that question is predicated on one central hypothesis: that changes in the expression levels of key driver genes will manifest as corresponding changes in the underlying morphology of the tumor. In other words, genes whose expression levels are strongly associated with clinically-relevant aspects of tumor morphology are likely to be the driver genes that we seek. Critical genetic lesions in tumors must manifest themselves through changes in cell morphology; if they do not have any discernible effect on the number, size, or shape of the cancer cells, then they cannot, by definition, contribute to the growth of the tumor. It is for this reason that the review of histological slides remains the gold standard in cancer diagnosis and treatment today. While morphological grading of tumor cell morphology is known to be a good prognostic for many different types of cancer, in contrast, molecular methods such as genotyping are still in their infancy and are not yet accurate or convenient enough for widespread deployment.

This thesis explores in detail the use of both morphological and molecular data to identify putative driver genes and pathways underlying tumor formation, and represents what

is to the best of our knowledge the first such integrative morpho-molecular analysis in cancer biology. Central to the thesis is our development of 1) an image processing pipeline capable of extracting high-level, clinically-relevant morphological features from digitized 10,000-megapixel whole-slide tissue samples, and 2) a system of multi-task regressions to robustly and efficiently associate gene expression levels with transformations in specific morphological traits. Taken together, these allow us to distill massive amounts of histological and molecular data into a set of unbiased and testable hypotheses regarding the effect of specific genes on particular clinically-relevant aspects of tumor morphology. In subsequent sections, we will flesh out the details of our system, and demonstrate it on half a terabyte of data from a total of 574 breast cancer patients across a spectrum of races, ages, and other demographics.

The advent of modern targeted therapeutics underscores the importance of elucidating the key functional mechanisms driving carcinogenesis, and it is widely believed that the key to accomplishing this lies in so-called "Big Data", with particular reference to the massive explosion in data from the genomics revolution. Yet genomic data alone cannot be a silver bullet. To paint a comprehensive picture, we need intelligent integration and analysis of the diversity of data modalities available to us: raw DNA sequences, levels of gene expression, gene methylation status, clinical data, patient demographic information, and so on. This thesis, by integrating histological and molecular data, represents our attempt at a step in that direction.

Chapter 2

Background and Related Work

2.1 Cancer: Biology and Treatment

> "All substances are poisonous, there is none that is not a poison; the right dose differentiates a poison from a remedy."
>
> Paracelsus, alchemist and physician, 1538 (Weinberg, 2007)

Cancer is a disease of the genome – an uncontrollable growth of cells driven by mutations and other genetic abnormalities accumulated over an organism's lifetime. Yet to call it a single disease would be to miss the reason why cancer remains a leading cause of death in the world, despite the billions of dollars poured into cancer research each year. Because any case of unchecked cellular replication falls under the umbrella of cancer, no two cancers are the same, in the way that all cases of Huntington's disease or other simple genetic disorders are similar. Instead, behind each tumor is a bewildering and unique array of genetic lesions - mutations, insertions, deletions, and even large scale chromosomal aberrations, all of which serve to throw the carefully-orchestrated cell cycle off-kilter.

Faced with the chaotic heterogeneity of cancers, it is no surprise that modern chemotherapy largely consists of drugs that are unable to effectively discriminate between cancerous cells and normal cells. (Mukherjee, 2010) These drugs are basically cellular poisons that blithely destroy all cells that come into contact with them. For example, methotraxate and fluorouracil, two of the most common chemotherapy drugs in use today, work

by inhibiting DNA synthesis; the efficacy of these drugs stem from the fact that cancerous cells divide far more rapidly than normal cells, and are thus more strongly affected by the drugs. (Ahmad et al., 1998) As expected, these cytotoxic drugs have adverse systemic effects, which give rise to observable and debilitating side effects: hair loss, nausea, fatigue, and so on. These side effects turn chemotherapy into a delicate balancing act between trying to kill off cancerous cells while keeping the patient's normal cells alive. Unfortunately, in a large majority of cases, such chemotherapy does not result in a complete cure.

In the last couple of decades of cancer research, however, researchers have started to identify a small set of cellular pathways that appear persistently dysregulated in a large proportion of cancers, giving rise to the hope that there might be some method in the madness that is the cancer genome. Indeed, it is now strongly suspected that out of the thousands of genetic lesions in any given cancer cell, only a handful are so-called "driver mutations" that are powering the growth of the tumor, with the other lesions being "passenger mutations" that arise from damaged DNA repair mechanisms but do not directly contribute to tumorigenesis. (Greenman et al., 2007; Beroukhim et al., 2010) Moreover, these driver mutations tend to reoccur in the same genes and cellular pathways; for example, a recent study published by the Cancer Genome Atlas Research Network showed that 23% of the 206 glioblastoma tumors they analyzed contained inactivating mutations in the NF1 gene. (Network, 2008)

These discoveries meant that cancers were in some sense more homogeneous than had been thought, and spurred researchers to develop interventions that could specifically target these driver mutations without damaging healthy cells. The first success story of this new wave of "targeted therapeutics" came out in the mid 90s and featured imatinib (Gleevec), which selectively targets the *bcr-abl* fusion protein characteristic of chronic myelogenous leukemia (CML). (O'Brien et al., 2003) In a series of landmark clinical trials, imatinib was shown to produce stunning remissions in CML patients, raising the 5-year survival rate from a previously dismal 30% to a far more optimistic 89%. (Pray, 2008)

With the success of other targeted therapeutics like Avastin or Herceptin has come a shift in cancer research: from trying to find a single silver bullet that would cure all of cancer to finding a large set of silver bullets, each of which is exceptionally effective against a small subset of cancers. In order to do this, we first have to understand and carefully

characterize the molecular networks and pathways underlying each subtype of cancer. This in itself is a herculean task, as each cell consists of approximately 25,000 genes interacting in extraordinarily intricate ways, and dysregulation of any of these genes could potentially lead to cancer. Indeed, despite the successes of drugs like Gleevec and Avastin, there have been many failed attempts at creating targeted therapeutics: we are only scratching at the surface of the biology underlying the targeted pathways, and oftentimes the tumors that we try to treat by blocking particular pathways develop alternative and unforeseen ways of compensating. Indeed, follow-up studies showed that a significant proportion of patients developed resistance to Gleevec after many years. (Gambacorti-Passerini et al., 2003) While the progress made so far has been promising, it is clear that we still have a long way to go.

2.2 Searching for Driver Mutations

Cancer biologists have traditionally employed methods like genetic screens to tease apart the mechanisms underlying tumorigenesis, but researchers have increasingly turned to high-throughput methodologies made possible by the completion of the Human Genome Project and by the general increase in computational power and decrease in the cost of large-scale assays and sequencing. In contrast to traditional genetic screens, these high-throughput methodologies involve automatically sifting through large volumes of data - genome sequences, gene expression microarrays, patient demographics, etc. - to pick out promising patterns for further study.

What are some common methods of finding mutations and shared pathways that drive cancer formation? Perhaps the most straightforward approach is to simply sequence many tumors and identify recurring somatic mutations and chromosomal abnormalities; this has been repeatedly shown to be successful to some extent. (Campbell et al., 2010; Jones et al., 2008; Sjöblom et al., 2006) While this approach is useful for picking up the more obvious driver mutations, we are rapidly exhausting the "low-hanging fruit" of driver mutations that are common to a large proportion of cancers, such as those in genes like *TP53* or *PI3K*. Instead, in recognition of the heterogeneity of cancer, we need to start looking for driver mutations which affect a smaller proportion of cancers. But it is not easy to distinguish

between driver and passenger mutations, and the noise from the fact that passenger mutations are often significantly more numerous than the driver mutations makes it difficult for somatic mutation analysis to pick these "smaller" driver mutations up. Moreover, somatic mutation analysis typically focuses only on gene-coding areas of the genome, for statistical and computational reasons. This means that such analyses will not pick up on situations such as mutations in noncoding regulatory RNAs which could in turn affect the expression levels of downstream genes.

Sequencing is still expensive to perform in scale, though costs are dropping rapidly. Today, by far the most common approach to finding driver mutations and pathways is to work with gene expression microarray data alone (Bild et al., 2006) or more typically, in combination with data on methylation status, copy number variation obtained via comparative genomic hybridization (Adler et al., 2006; Akavia et al., 2010; Beroukhim et al., 2007; Berger et al.; Taylor et al., 2010), protein-protein interaction networks (Mani et al., 2008; Cerami et al., 2010), etc. The issue with such approaches is a lack of a supervised signal (i.e., an external phenotypic measurement that can be used to determine which genes/pathways are important), which forces them to largely utilize unsupervised learning techniques, for example, clustering gene expression data to find common cancer subtypes (Verhaak et al., 2010) or expression signatures for metastasis (Ramaswamy et al., 2003). While such techniques have clearly had their own share of successes, the lack of supervision (as in the case of somatic mutation analysis) makes it hard to pick up on more subtle signals in the data.

Where supervision has been used, it has mostly come from two sources: survival data (Vandin et al., 2012), or prior knowledge obtained from studying animal models or other cancer types in humans (Boehm et al., 2007). Survival data is arguably the "gold standard", because we are ultimately concerned with finding mutations and aberrations in pathway regulation that result in increased mortality. However, it is notoriously noisy, because survival times can be influenced by a wide variety of other factors, and as a result has not seen much success in terms of being able to identify driver mutations. The use of prior knowledge is also problematic; while it is undeniably useful to investigate whether a driver mutation in a particular type of cancer can drive tumorigenesis in a different type of cancer,

relying solely on prior knowledge precludes the possibility of identifying novel pathways and mechanisms.

Where can we find a better - and easily obtainable - source of clinically-relevant phenotypic markers that we can use as a supervised signal for the discovery of genetic drivers? The answer lies in tumor histology and morphology, which as of today still remains a largely untapped resource in this line of research.

2.3 Quantitative Tumor Morphology

Today, the diagnostic standard of care for virtually all types of (solid) cancer treatment remains the review of histology images by a pathologist following surgical resection, together with radiological data in the form of X-ray or MRI scans; this is true in particular of breast cancer, which we study in this thesis. (Le Doussal et al., 1989; Pereira et al., 1995; Bloom and Richardson, 1957) This practice of histological examination has been in place for nearly a hundred years for the simple reason that morphological characteristics of tumors can convey important information about prognosis and possible responses to treatment. (Patey et al., 1928)

Previous approaches to measuring and using morphological data from tumors have largely consisted of using a small set of image features hand-picked by pathologists. These features tend to be "low-level" features such as nuclear size and area, because it is difficult to obtain features that are more contextual (e.g., distance between nuclei, or between epithelial nuclei and the stroma). (Teverovskiy et al., 2004; Donovan et al., August 20, 2008; Aaltomaa et al., 1991; Cooper et al., 2012) Moreover, large quantities of digitized tumor slides and the computational resources needed to process these have been difficult to obtain. As a result, there has been a relative dearth of high-impact studies using large-scale morphometric data.

However, recent advances - both in the release of high-throughput data sets, such as the Cancer Genome Atlas, as well as in computational techniques, have allowed tumor morphology to be quantitated on a large-scale and in an unbiased manner. In a landmark study, (Beck et al., 2011) showed that one can produce clinically relevant histological features from digitized tissue microarray cores by first building a classifier to recognize different

types of tissue (e.g., epithelium vs. stroma), and then extracting both standard morphological features and higher-level relational features that take these tissue types etc. into account.

In Chapter 3, we will extend the system in (Beck et al., 2011) to handle whole-slide tissue samples that are orders of magnitude larger than the tissue microarray cores used in that study; tissue microarray cores are typically 0.6mm in diameter, whereas our whole-slide tissue samples can be as large as 30x26mm in size, a difference in area of a factor of 2750x. First, however, we will study some of the current methods that can be used to associate gene expression levels with whatever morphological features we might extract.

2.4 Gene-Morphology Regression

Given morphological data for a set of patients, coupled with the corresponding gene expression data for each patient, how can we find the genes that are mostly strongly associated with cell morphology? This can be cast as an instance of a *regression* problem, and in this section we will cover some mathematical preliminaries and standard approaches to the problem.

Assume that we have G genes, P morphological traits (where a trait could be, e.g., the average size of a cell), and N patients. We can represent our gene expression data with a matrix $X \in \mathbf{R}^{N \times G}$, and our morphological data with a matrix $Y \in \mathbf{R}^{N \times P}$. Our goal is then to find a matrix of regression coefficients $W \in \mathbf{R}^{G \times P}$, such that if we only had access to the gene expression data, the "predicted" morphological data XW would be as close to Y as possible. Note that each row $w_i \in \mathbf{R}^{1 \times P}$ is the vector of regression coefficients corresponding to gene i across all P morphological traits.

We can attempt to do this by solving the least-squares regression problem

$$\text{minimize}_W \quad \|Y - XW\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius matrix norm. This is a common distance metric to use, as performing ordinary least-squares regression has the property of being the uniform

minimum variance estimator of W under a certain set of assumptions, such as the residuals $Y - XW$ being independently and identically distributed according to a normal distribution.

The issue, however, is that in our case $G \gg N$, and therefore W is underspecified: there will be an infinite number of possible W's that result in zero residual. To get around this issue, we will encode our prior assumption that only a small subset of genes are going to have an effect on cancer cell morphology in the form of a *sparsity penalty*:

$$\text{minimize}_W \quad \|Y - XW\|_F^2 + \lambda \sum_i \text{card}(\|w_i\|_1),$$

where $\lambda \in \mathbf{R}$ is a hyperparameter, and the cardinality function $\text{card}(x)$ is 0 when $x = 0$, and 1 otherwise.

Unfortunately, the cardinality function is non-convex, which makes actually solving the above optimization problem extremely difficult. In standard practice, one simply replaces $\text{card}(x)$ with $\|x\|_1$, which is its convex envolope over $[-1, 1]$. This results in an ℓ_1-minimization problem that is equivalent to a standard *Lasso* model (Tibshirani, 1996):

$$\text{minimize}_W \quad \|Y - XW\|_F^2 + \lambda \sum_i \|w_i\|_1.$$

This standard ℓ_1 model can be extended in two different ways, which we will cover separately in the subsequent subsections.

2.4.1 Iteratively Reweighted ℓ_1 Minimization (IRLM)

One issue with the optimization problem above is that $\|x\|_1$ is at best a very crude approximation to $\text{card}(x)$, chosen only because of its convexity. Instead, we can follow the approach of (Candès et al., 2007), who proved in a landmark paper that one can often obtain better solutions by using the function $\log(\varepsilon + |x|)$ to approximate $\text{card}(x)$, and then running concave-convex programming to obtain a series of reweighted ℓ_1-minimization problems, each of which is easy to solve. This approach is described in detail in Algorithm 2.1.

Algorithm 2.1 Original IRLM (based on Section 2.2 in (Candès et al., 2007))

1. Set the iteration count $t = 0$ and the sparsity penalty weights $r_i^{(0)} = 1, i = 1, \ldots, n$.
2. Solve the weighted ℓ_1 minimization problem
$$\text{minimize}_{W^{(t)}} \|Y - XW^{(t)}\|_F^2 + \lambda r_i \sum_i \|w_i^{(t)}\|_1.$$
3. Update the penalty weights: for each $i = 1, \ldots, n$,
$$r_i^{(t+1)} = \frac{1}{|w_i^{(t)}| + \varepsilon}.$$
4. Terminate on convergence or when t hits a max number of iterations. Otherwise, increment t and go to step 2.

2.4.2 Elastic Net Regression

Another issue with the standard ℓ_1 minimization problem above is that if groups of genes have highly correlated expression levels (which would occur, for example, among genes along the same pathway), the optimization would tend to randomly select only one gene from each group to use as a predictor. To get around this, (Zou and Hastie, 2005) introduce *elastic net regression*, which introduces a small amount of ridge regression into the standard ℓ_1 problem.

To simplify notation, we first note that our original optimization problem can be decomposed into P independent subproblems, each one of which involves regressing against a single morphological trait. Let $w_{:,j}$ be the vector corresponding to any particular trait j; in the original problem, $w_{:,j}$ would be the j-th column of W. We can then write:

$$\text{minimize}_{w_{:,j}} \quad \|Y_j - Xw_{:,j}\|_F^2 + \lambda_j \|w_{:,j}\|_1,$$

where we have split up the original single hyperparameter λ into separate hyperparameters for each phenotype, for greater accuracy (because different phenotypes would have

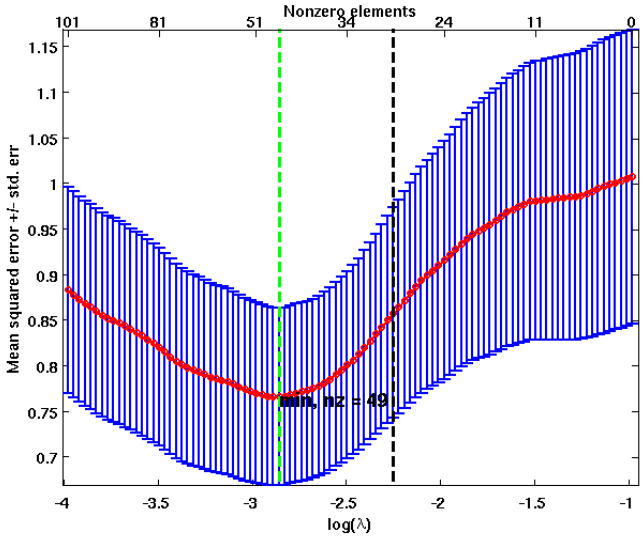

Figure 2.1: Example of an elastic net regularization path.

a different number of genes involved in predicting them). Note that the ℓ_1 norm on W is invariant to whether we apply it to its columns instead of its rows.

Elastic net regression would then give us:

$$\text{minimize}_{w_{:,j}} \quad \|Y_j - Xw_{:,j}\|_F^2 + \lambda_j \left(\tfrac{1-\alpha}{2} \|w_{:,j}\|_2^2 + \alpha \|w_{:,j}\|_1 \right),$$

where the elastic net hyperparameter $\alpha \in [0,1]$ can be set to 0 to obtain the standard ridge regression penalty, or 1 to obtain the standard lasso penalty.

Elastic net regression also has the computationally desirable property of being able to set each λ_j efficiently, because the magnitude of λ_j is in monotonic correspondence with the number of zero entries in $w_{:,j}$. This allows us to calculate the full regularization path for each λ_j (i.e., finding cross-validation prediction error for all possible values of λ_j), as detailed in (Zou and Hastie, 2005). An example of this calculation, generated by the software package `glmnet` (Friedman et al., 2010a), is shown in Fig. 2.1.

Chapter 3

Methods

We are now well-positioned to describe the methodology we adopted to 1) obtain accurate morphological measurements from breast cancers, and 2) integrate these with molecular data to find putative clinically-relevant genes that drive carcinogenesis.

3.1 Extracting Morphological Features

We developed an image processing pipeline to process digitized whole-slide breast tumor samples, within the Definiens Developer XD image analysis environment. We faced two challenges here: 1) typical images contain around 10 billion pixels, far too many to work with directly, and 2) in order to produce a clinically-relevant set of higher-level features, we have to reliably distinguish between epithelium and stroma, and nucleus and cytoplasm, across all patients. This is complicated by the large degree of variation between patients due not only to the intrinsic differences between patients in terms of age, race, etc., but also to the fact that different patients were processed in different centers.

To address these issues, we performed the analysis at two different resolutions, which allowed us to use an initial low-resolution pass to identify the relevant areas of each image, before zooming in for more detail. Additionally, we use a large set of hand-labeled slides to train an epithelium-stroma classifier, allowing us to build a robust classifier that can automatically take into account variation between patients without the need for us to explicitly hard-code a set of rules.

Our pipeline comprises three main stages, which we describe in the following subsections.

3.1.1 Tissue Segmentation

The aim of the first stage is to segment the slide into small, contiguous, and visually coherent sections called superpixels. This involves four steps:

1. Create a low-resolution map at a 0.5% scale and a medium-resolution map at a 3% scale from the original image.

2. On the low-resolution map, create large superpixels and threshold these based on their total green pixel intensity to find superpixels that contain a sufficient quantity of tissue. Discard the other superpixels, i.e., the ones that are almost entirely white.

3. Synchronize the low-resolution and medium-resolution maps, i.e., discard portions of the medium-resolution map that were discarded in the low-resolution map.

4. Segment the remainder of the medium-resolution map into small superpixels.

Figure 3.1 shows a graphical depiction of these steps.

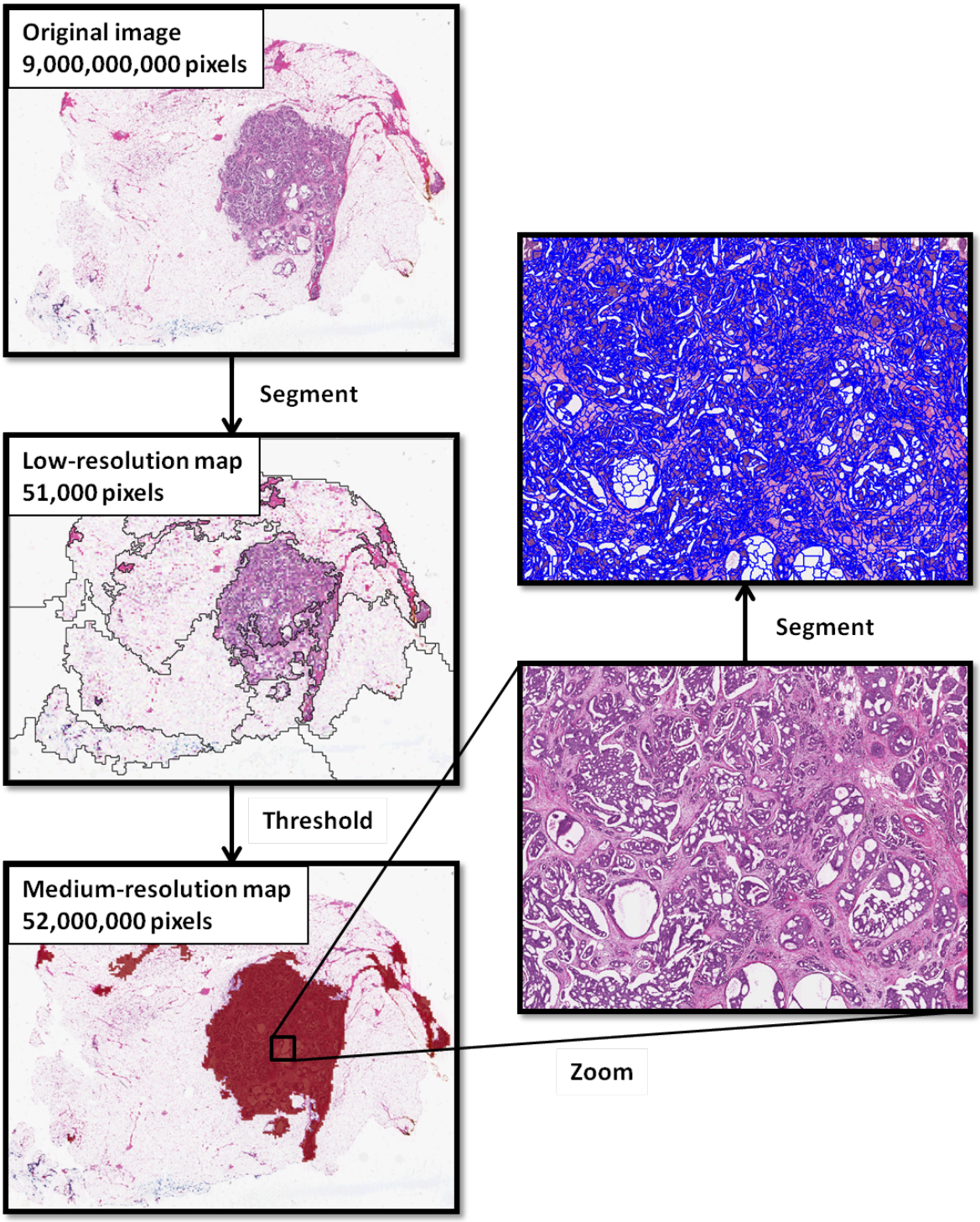

Figure 3.1: Pipeline for tissue segmentation.

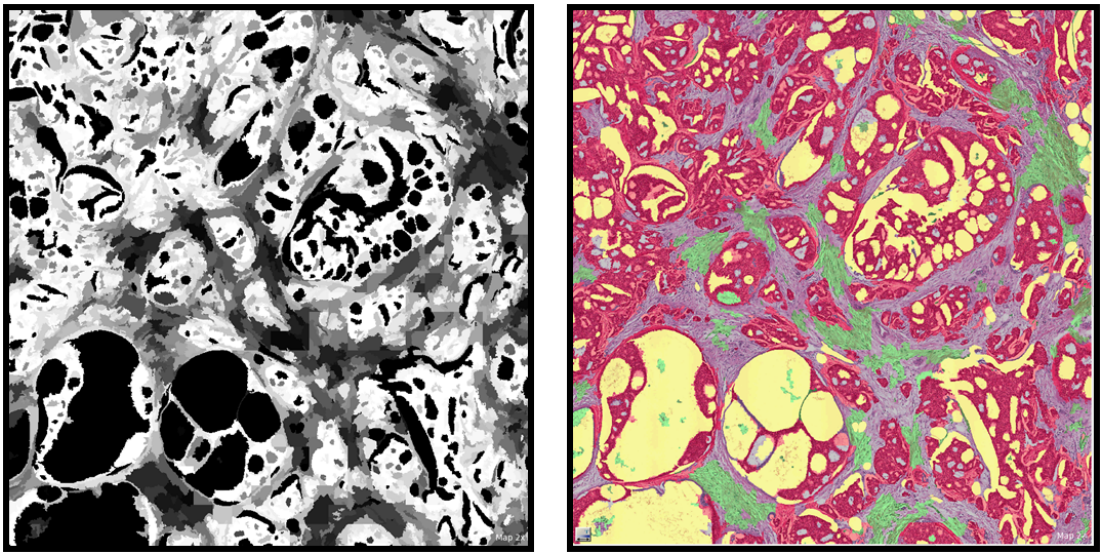

Figure 3.2: Left: Map of $P(\text{Epithelium})$, with white = probability 0. Right: Classified tissue segments. Red = Epithelium, Purple = EMT, Green = Stroma, Yellow = Adipose.

3.1.2 Tissue Classification

In the second stage, we classify the superpixels found in the first stage into four clinically-relevant categories: epithelium, stroma, epithelial-mesenchymal transition (EMT), and adipose. This is accomplished through the following steps:

1. Within each superpixel, identify nuclei and cytoplasm based on color; we divide each superpixel up into a fine grid, and apply a simple threshold on each grid square.

2. Label empty superpixels as adipose.

3. For each remaining superpixel, create a 150-dimensional feature vector, incorporating size, shape, texture, intensity, and other characteristics of intra-superpixel nuclei and cytoplasm.

4. Use these feature vectors (together with the hand-labeled set of tumor slides) to train a binary epithelium-stroma classifier. We use ℓ_1-regularized logistic regression, with cross-validation to calculate regularization paths. (Friedman et al., 2009) The result is a classifier that assigns to every superpixel a probability that it is epithelial. (Figure 3.2-Left)

5. Classify superpixels with $P(\text{epithelium}) \geq 0.75$ as epithelium, superpixels with $P(\text{epithelium}) \leq 0.25$ as stroma, and the rest as EMT. (Figure 3.2-Right)

3.1.3 Feature Construction

The third and final stage involves calculating morphological features based on the tissue categories computed in the preceding stage. For each patient, we generate 19 high-level summary features from the corresponding tumor tissue slide, including:

1. Total area covered by epithelial tissues, and likewise for stroma, EMT, and adipose.
2. Total number of epithelial objects, and likewise for stroma, EMT, and adipose.
3. Total area covered by nuclei, and total area covered by cytoplasm.
4. Total number of nuclei, and total number of cytoplasmic objects.
5. Nuclei staining intensity and heterogeneity.
6. Degree of epithelial-ness, as measured by $\sum_{s \in \text{Superpixels}} P(s = \text{epithelium}) * \text{Size}(s)$, normalized by the total tissue area.

These features are then concatenated into a single 19-dimensional vector, for use in the subsequent regression analysis.

3.2 Gene-Morphology Regression

Recall from Section 2.4 that we can denote our gene expression data with a matrix $X \in \mathbf{R}^{N \times G}$, and our morphological data with a matrix $Y \in \mathbf{R}^{N \times P}$. Our goal is to find a matrix of regression coefficients $W \in \mathbf{R}^{G \times P}$, such that if we only had access to the gene expression data, the "predicted" morphological data XW would be as close to Y as possible. By examining the entries of W, we could then find out which genes are significantly predictive of which morphological traits, and use that to inform further research and biological validation.

The dominant issue in any attempt to work with high-throughput molecular and molecular data, however, is the unfortunate fact that $N \ll G$. It is difficult and costly to obtain matching tumor tissue samples and gene expression data for patients suffering from cancer, which in our case means that N is in the low hundreds. On the other hand, gene expression microarrays means that we have access to expression levels of approximately $20,000$ genes. Compounding this issue is the problem of measurement noise; both morphological and molecular measurements are extremely noisy and not very precise. For example, the former depends on how representative the extracted tissue sample is of the entire tumor, while the latter is an aggregate over many cells and typically exhibits a large degree of variance in measurements, even among technical replicates. Together, these factors make it difficult to robustly estimate W using straightforward techniques.

To overcome these obstacles, we develop a series of extensions to the regression techniques outlined in Section 2.4.

3.2.1 Transfer Learning via IRLM

Our first observation is that morphological traits are highly interlinked - we would expect that, for example, that related mechanisms underlie the size, shape, and texture of a cell. In particular, this means we would expect genes that are strongly predictive of one morphological trait to also be predictive of other morphological traits. Conversely, if a gene is only predictive of a single morphological trait and is not correlated with any other, then it is more likely that that single association is the result of statistical noise rather than true signal. This can be viewed as a form of *multi-task* or *transfer learning*, where the coefficients we learn for one morphological trait informs the coefficients we learn for another morphological trait, and vice versa.

We can encode this with a *group sparsity prior* that acts on the elements of individual rows W_i, encouraging them to go to zero together. One recent and popular method for doing that is to use *sparse group lasso*, which amounts to adding a ℓ_2 penalty on the rows W_i. (Friedman et al., 2010b) Unfortunately, this loses us the decomposability property described in Section 3.2.2, because it couples together the individual subproblems (of finding

genes that can effectively predict a single phenotype). This makes it computationally infeasible to run over large datasets comprising thousands of genes and morphological traits. Moreover, as discussed in Section 2.4.1, the ℓ_1 and ℓ_2 penalties are not ideal approximations of the cardinality function.

Instead of using the sparse group lasso, we can derive a multi-task regression algorithm in the spirit of (Candès et al., 2007). The transfer learning problem can be stated as trying to solve:

$$\text{minimize}_{\text{W}} \quad \|XW - Y\|_F^2 + \lambda \sum_i \text{card}\left(\|W_i\|_1\right),$$

where we now care about the cardinality of entire rows W_i, which corresponds to the vector of coefficients for a single gene across all the morphological traits. As a reminder, in Section 2.4, we were concerned about minimizing the cardinality of each individual element of W. In the following discussion, we omit the hyperparameter λ for simplicity.

Now, as before, we can approximate this cardinality term with $\sum_{i=1}^{G} \log(\varepsilon + \|W_i\|_1) = \sum_{i=1}^{G} \log(\varepsilon + \sum_{j=1}^{P} |W_{ij}|)$, which gives us:

$$\text{minimize}_{\text{W}} \quad \|XW - Y\|_F^2 + \sum_{i=1}^{G} \log(\varepsilon + \sum_{j=1}^{P} |W_{ij}|).$$

Because $\log(\cdot)$ is monotonic, we can introduce dummy variables U_{ij} to obtain an equivalent optimization problem:

$$\begin{aligned} \text{minimize}_{\text{W,U}} \quad & \|XW - Y\|_F^2 + \sum_{i=1}^{G} \log(\varepsilon + \sum_{j=1}^{P} U_{ij}) \\ & |W_{ij}| \leq U_{ij}, \ \forall i, j \end{aligned} \quad .$$

The log term is unfortunately concave, but we can apply standard sequential convex programming techniques to solve this efficiently. The idea is to solve a sequential set of convex optimization problems, with each iteration guaranteed to produce solutions that are better (or not worse) than the last, with respect to the original objective function. (Yuille and Rangarajan, 2003; Sriperumbudur and Lanckriet, 2009) We first linearize this objective

at a point $\tilde{U}$ to obtain:

$$\begin{aligned} \text{minimize}_{\text{W,U}} \quad & \|XW - Y\|_F^2 + \sum_{i=1}^{G} \log(\varepsilon + \sum_{j=1}^{P} \tilde{U_{ij}}) \quad + \sum_{i=1}^{G} \sum_{j=1}^{P} \frac{1}{\varepsilon + \sum_{k=1}^{P} \tilde{U_{ik}}} (U_{ij} - \tilde{U_{ij}}) \\ & |W_{ij}| \leq U_{ij}, \, \forall i, j \qquad . \end{aligned}$$

and then remove constant terms, substitute W back for U, and simplify to get:

$$\text{minimize}_{\text{W}} \quad \|XW - Y\|_F^2 \quad + \sum_{i=1}^{G} \frac{1}{\varepsilon + \|\tilde{W}_i\|_1} \|W_i\|_1 .$$

In this formulation, $\tilde{W}$ represents the weight matrix at the previous iteration (i.e., the previous problem in the sequence of convex problems). This leads to the following algorithm for approximating our original transfer learning task:

Algorithm 3.1 Multi-task IRLM

1. Set the iteration count $t = 0$ and the sparsity penalty weights $r_i^{(0)} = 1, i = 1, \ldots, n$.
2. Solve the weighted ℓ_1 minimization problem

$$\text{minimize}_{\text{W}^{(\text{t})}} \quad \|XW^{(t)} - Y\|_F^2 \quad + \sum_{i=1}^{G} r_i^{(t)} \|W_i^{(t)}\|_1 .$$

3. Update the weights: for each $i = 1, \ldots, n$,

$$r_i^{(t+1)} = \frac{1}{\|W_i^{(t)}\|_1 + \varepsilon}.$$

   This uses the sum of the absolute coefficients of each gene across all of the related phenotypes to determine the new regularization weightage.

4. Terminate on convergence or when t hits a max number of iterations. Otherwise, increment t and go to step 2.

This has the advantage that the subproblems are all decomposable (because the ℓ_1 norm is decomposable), making the problem easily parallelizable in a Map-Reduce fashion: a

typical run of the algorithm would involve solving all the subproblems (individual phenotypes) separately, gathering the weights from all of them and recomputing the ℓ_1 weights, and then solving all the subproblems separately again. In practice, this not only results in significant speed-up over group lasso methods, but also allows us to tackle very large datasets via distributed computing.

3.2.2 Elastic Net Regression

The decomposable-by-phenotypes property holds even when we introduce the elastic net penalty from Section 3.2.2. Recall that elastic net regression involves solving problems of the form:

$$\text{minimize}_W \quad \|Y - XW\|_F^2 + \sum_j \lambda_j \left(\tfrac{1-\alpha}{2} \|w_{:,j}\|_2^2 + \alpha \|w_{:,j}\|_1 \right).$$

We can simply introduce this into step 2 of algorithm 3.1, using the parameters $r_i^{(t)}$ to weight the vectors $w_{:,j}$ accordingly. Our regression problem thus involves two distinct forms of regularization: elastic net penalties over the columns of W to encourage correlated genes to co-predict phenotypes, and reweighted ℓ_1 penalties on the rows of W to encourage group sparsity. As before, we use fast regularization path algorithms in combination with 5-fold cross-validation on the vanilla elastic net problem described in Section 3.2.2 to determine the hyperparameters λ_j.

3.2.3 Pathway Regularization and Filtering

Furthermore, we can input some additional prior knowledge into our regression pipeline by looking at sets of curated genetic pathways, such as the Biocarta or Reactome collections. (Subramanian et al., 2005) For both statistical and computational reasons, we first reduce the number of genes used as regressors by selecting only the top 20% of the available genes, based on the variance in their expression levels across all the patients; the idea is that genes with relatively constant expression levels will not be particularly helpful in predicting morphological traits. Next, out of these genes, we only consider genes that belong to at least one of the pathways in our curated set, the assumption being that genes that are

biologically important in one way or another are more likely than not to have already been at least cursorily studied (in a non-cancer setting), and placed into one of these curated sets. We note that this is still a largely unbiased process, for the curated gene sets are large and span many different areas of biological research: for example, the Biocarta collection comprises 1267 genes over 217 pathways, while the Reactome collection comprises 4159 genes over 430 pathways. (BioCarta, 2012; Joshi-Tope et al., 2005)

With this reduced set of genes, we can extend the transfer learning system developed above by incorporating regularization over the pathways in our collection of choice. The intuition behind this is as follows: genes in the same pathway act together to bring about phenotypical change, and therefore if a particular gene is strongly predictive of some aspect of tumor morphology, it is likely that other genes in the same pathway also affect tumor morphology. Concretely, we can encode this by simply replacing $\|W_i^{(t)}\|_1$ in step 3 of algorithm 3.1 with $\frac{1}{|\text{Pathway}(i)|}\sum_{j\in\text{Pathway}(i)}\|W_j^{(t)}\|$, where $\text{Pathway}(i)$ is the set of all genes that are in the same pathway(s) as gene i; dividing by the total number of such genes ensures that we do not over-penalize genes that belong to large pathways.

3.2.4 Regression with Active Genes

The estimates obtained by running the multi-task regression described in the preceding sections are biased towards 0, because of the nature of the ℓ_1 penalty. (Friedman et al., 2001) We can therefore improve on these estimates by using multi-task regression to find the sparse set of genetic predictors for each morphological trait, and then running unpenalized least-squares regression using only these active genes. This is typically possible because the number of active genes is almost always smaller than the number of patients (due to the sparsity prior); if not, we add a small amount of ridge regression.

3.2.5 Summary

We are finally in a position to assemble all the pieces:

Algorithm 3.2 Overall gene-morphology regression pipeline.

Input: Gene expression matrix $X \in \mathbf{R}^{N \times G}$, morphological trait matrix $Y \in \mathbf{R}^{N \times P}$.
Output: Matrix of regression coefficients $W \in \mathbf{R}^{G \times P}$.
Procedure:

1. For each phenotype j, determine the regularization hyperparameter λ_j by calculating the regularization path for the appropriate single-task elastic net regression (using 5-fold cross-validation). (Section 3.2.2)

2. Run multi-task IRLM with pathway regularization to find the (biased) coefficient matrix $\tilde{W}$. (Sections 3.2.1, 3.2.3)

3. For each phenotype j, find the unbiased coefficient vector $w_{:,j}$ by running unpenalized least-squares regression using only the genes with non-zero coefficients in $\tilde{w}_{:,j}$. (Section 3.2.4)

We can then examine W to realize our original goal of finding genes that are strongly predictive of morphology.

Chapter 4

Results

4.1 Synthetic Data

We first tested out our multi-task IRLM algorithm on two synthetic datasets, each with 6 morphological phenotypes, 1000 genes and 100 patients. The gene expression values were all drawn IID from a standard Gaussian distribution.

4.1.1 Dataset 1 - Single group of phenotypes

In the first dataset, we aimed to simulate a case of having a group of largely related phenotypes, each pair of phenotypes differing by only one gene. As such, for $i = 1, 2, \ldots 10$, we formed phenotype i by taking a linear combination of genes $1, \ldots, i-1, i+1, \ldots, 10$, with coefficients drawn uniform from $[-1, -0.3] \bigcup [0.3, 1]$. For example, phenotype 1 is a linear combination of genes 2 to 10. Noise drawn from $N(0, 4)$ was added to each phenotype value. The idea is that our multi-task strategies should not force any given phenotype to include genes that are strong predictors of other phenotypes, but which are not good predictors of itself.

We measure the relative error on each phenotype by taking the ratio of the predicted mean-squared test set error to the original mean-squared error (i.e., without using any gene expression values as regressors), the latter being equal to the variance of the phenotype in the test set. These are the results:

Method \ Phenotype	1	2	3	4	5	6
Single-task Elastic Net	0.67	0.40	0.47	0.37	0.50	0.61
Single-task IRLM	0.71	0.48	0.36	0.38	0.53	0.59
Multi-task IRLM	0.32	0.30	0.23	0.26	0.33	0.38
Multi-task Group Lasso	2.87	3.11	6.19	5.14	2.99	4.39

Table 4.1: Relative test errors on Dataset 1

Group lasso did surprisingly badly; this is probably due to inadequate cross-validation of the regularization coefficient. IRLM, on the other hand, is very robust to different values of the regularization coefficient (a default setting of 1 works well). We see that single-task elastic net / IRLM perform comparably, while multi-task IRLM does far better. As desired, in the multi-task IRLM setting, gene 1 was not picked up as a significant predictor of phenotype 1, even though it is involved in the other phenotypes, and likewise for the other genes.

4.1.2 Dataset 2 - Two groups of phenotypes

In this dataset, phenotypes 1-3 were linear combinations of genes 1-10, while phenotypes 4-6 were linear combinations of genes 11-20, with coefficients and noise drawn from the same distributions as before. The aim of this dataset is to see if the multi-task methods would be robust to very heterogenous groups of phenotypes, where one would expect transfer learning to be largely inapplicable across groups (though still helpful within each group). These are the results:

Method \ Phenotype	1	2	3	4	5	6
Single-task Elastic Net	0.29	0.70	0.53	0.30	0.84	0.32
Multi-task IRLM	0.22	0.27	0.38	0.29	0.39	0.27

Table 4.2: Relative test errors on Dataset 2

As before, single-task IRLM performed comparably to elastic net, and multi-task group lasso did poorly; we omit these results from the table for brevity. In contrast, multi-task IRLM was able to correctly partition the genes and phenotypes into two distinct sets, and accurately infer their relationships.

4.2 Tissue Microarrays: Netherlands Cancer Institute

We next ran our gene-morphology association pipeline on data from a cohort of 248 patients from the Netherlands Cancer Institute (NKI), using 6,642 tissue-microarray-derived morphological features from (Beck et al., 2011) and clinical data and expression levels of 11,040 genes from (van de Vijver et al., 2002). To increase the saliency of our results, we first use univariate Cox regression to pick out the 132 morphological features most strongly associated with patient survival, and filter the genes by taking only the top 20% by variance.

We then ran 3 regression experiments: 1) using the entire variance-filtered gene list, without any pathway-based filtering; 2) using the Reactome collection of pathways for filtering and regularization; and 3) using the Biocarta collection of pathways. Each experiment consists of the same 10 random splits of the 248 patients into a training cohort of 198 patients and a held-out test cohort of 50 patients. In each random split, we use our multi-task IRLM method to obtain regression coefficients on the training set, and then measure the relative error for each phenotype on the test set, as we did with synthetic data in the previous section. For each phenotype, we then take the median relative error over the 10 random train/test splits.

As can be seen from the aggregated results in Figure 4.1, all 3 methods are able to explain away a surprisingly large fraction of the variance in morphological phenotype. We also performed a control experiment in which we permuted the order of the phenotypes in the test dataset; as expected, we were unable to consistently explain any single phenotype in this permuted dataset (median relative test set errors were all greater than 1).

4.2.1 Interpretation

Analysis of the estimated regression coefficients yielded striking results. Our experiments with the Reactome and Biocarta collections of pathways yielded 37 putative genes which were significantly associated with tumor morphology, including the following genes with experimentally-validated roles in tumorigenesis:

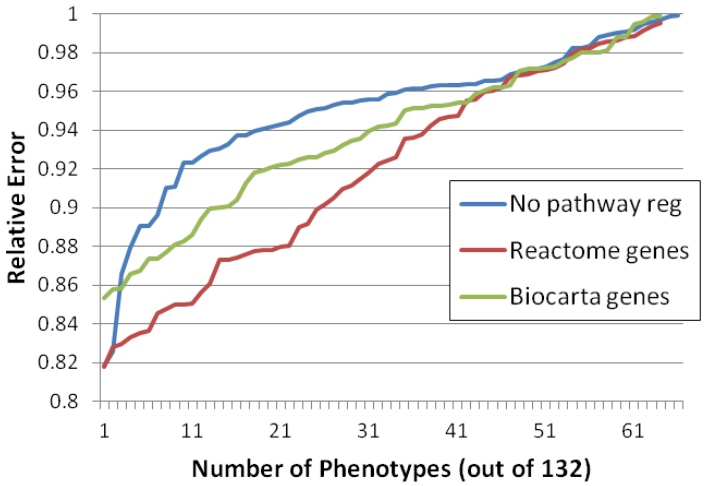

Figure 4.1: Median relative test set errors on the NKI dataset. To visualize these, we sort the relative errors on each phenotype and plot them as a straight line. For example, the graph above tells us that about 20 phenotypes are predicted with < 0.88 error by Reactome genes. We omit phenotypes that have median errors > 1.

- **Transforming Growth Factor Beta-3.** TGF-β acts as a gateway gene in normal cells, regulating the cell cycle by stopping at the G1 phase when the cell is not ready to proliferate. In cancer cells, however, TGF-β is often mutated and loses its regulatory function, allowing the cell cycle to run unchecked. (Djonov et al., 1997; Elliott and Blobe, 2005) In our experiments, TGF-β3 was associated with a whole host of cytoplasmic and nuclear morphological features. In particular, lower levels of TGF-β3 were consistently associated with larger cell sizes and higher levels of nuclear staining intensity, which reflects an increased density of nuclear chromatin.

- **Kirsten Rat Sarcoma Viral Oncogene Homolog.** KRAS is a well-known oncogene that serves to activate signal transduction pathways involving growth factors and other established oncogene products such as c-Raf. (Bos, 1989; Kranenburg, 2005) In our experiments, KRAS is consistently associated with the morphological properties of the stromal matrix, with elevated KRAS levels contributing to increased stromal morphological heterogeneity.

- **Fibroblast Growth Factor 18.** FGFs are pleiotropic growth factors that stimulate proliferation across a variety of tissues. FGF-18 in particular has been shown to be active in a significant fraction of colorectal cancers**(Shimokawa et al., 2003; Hu et al., 1998)**. In our experiments, elevated FGF-18 levels are associated with a large variety of changes in cytoplasmic morphology, including an increase in cellular heterogeneity.

A full list of the top genes can be found in Appendix A.1, including many other known or suspected oncogenes/tumor-suppressor genes, such as the mitotic checkpoint genes CENPA and BUB1 (Kops et al., 2005), the GEF protein VAV3 (Lee et al., 2008), the anchor protein Akap9 (Ciampi et al., 2005), the progesterone receptor PGR (Horwitz and McGuire, 1978), and DSG1 from the cadherin superfamily(Chidgey and Dawson, 2007). It is interesting to note that many of these genes were discovered in a context other than breast cancer (e.g., colon cancer for FGF18), and are currently not associated with breast cancer in particular. The fact that they are implicated in other forms of cancer strongly suggest that they are indeed molecular drivers of cancer morphology in the breast as well, and open these genes up for further study in the context of breast carcinoma and its treatment.

4.2.2 Coefficient Stability

Because a large thrust of our effort is to find scientifically-plausible hypotheses between gene function and particular aspects of cancer morphology, it is important that we not only find associations between genes and morphology, but also ascribe a consistent directionality to these interactions. We find that the directionality of our coefficients are stable to a very large extent. For example, Figure 4.2 shows representative results on three different genes against a morphometric quantization of cell shape.

4.3 Whole-slide Tissue: the Cancer Genome Atlas

Finally, we ran our entire image processing and analytic pipeline on image and molecular data from 326 patients with breast carcinomas from the Cancer Genome Atlas. For each

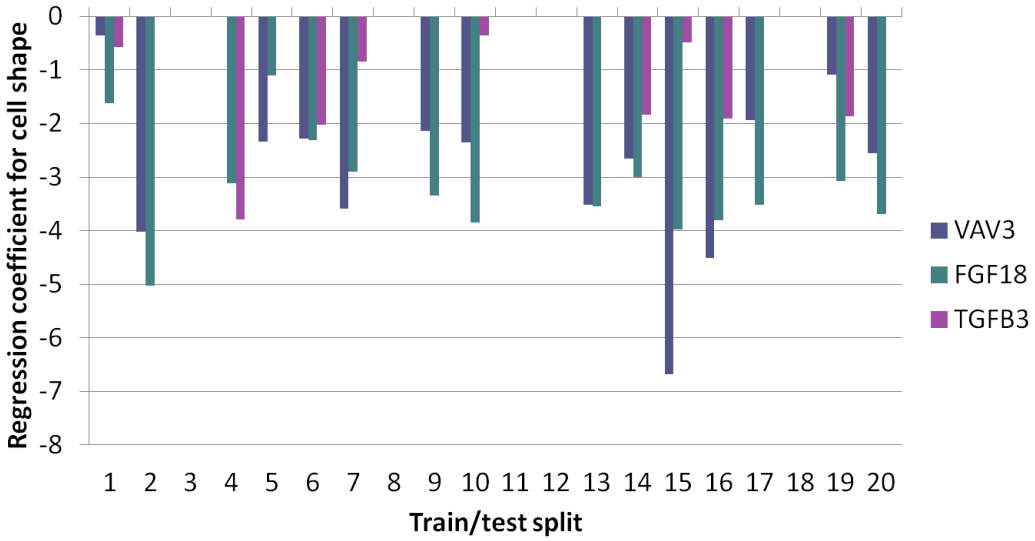

Figure 4.2: Regression coefficients for VAV3, FGF18, and TGFB3 against cell shape, across 20 different train/test splits.

patient, we obtained a digitized whole-slide tumor tissue sample (originally used for primary diagnosis in the hospital), together with gene expression data on 17,580 genes. We used the morphological feature extraction pipeline described in Section 3.1, together with a epithelium-stroma training set of 365,000 hand-labeled superpixels from a set of 90 slides, to derive a set of 19 morphological features for each patient.

As with the NKI dataset, we ran 3 regression experiments: 1) using the entire variance-filtered gene list; 2) using Reactome pathways for filtering and regularization; and 3) doing likewise with Biocarta pathways. Figure 4.3 shows the median relative test errors on 10 random splits of the 326 patients into a training cohort of 261 patients and a held-out test cohort of 65 patients. We performed a similar control experiment in which we permuted the order of the phenotypes in the test dataset. As before, we were unable to accurately predict any single phenotype in this permuted dataset (median relative test set errors were all greater than 1).

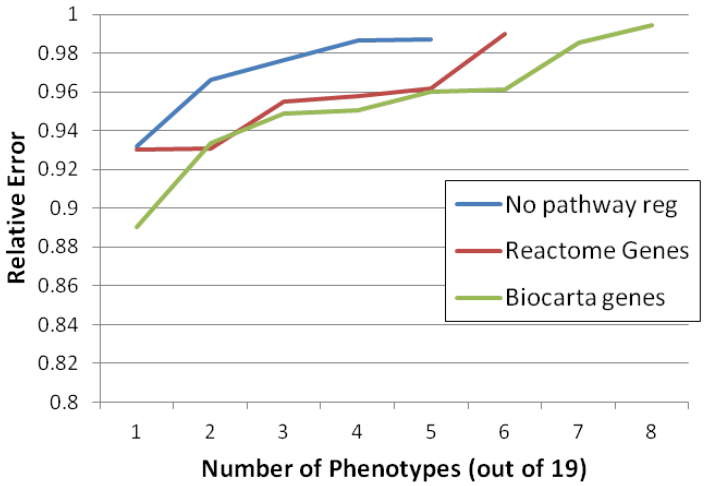

Figure 4.3: Median relative test set errors on the TCGA dataset. To visualize these, we sort the relative errors on each phenotype and plot them as a straight line. For example, the graph above tells us that about 5 phenotypes are predicted with < 0.96 error by Reactome genes. We omit phenotypes that have median errors > 1.

4.3.1 Interpretation

Together, the experiments with the Reactome and Biocarta pathways turned up a set of 21 putative morphology-driving genes (Appendix A.2). As with the NKI dataset, we find many genes that have already been validated as, or are suspected to be, heavily involved in cancer: KIT and FGF18 as in the NKI dataset, and more notably the well-known genes BRCA2 (breast cancer 2 susceptibility protein) and EGFR (epidermal growth factor receptor). Mutations in BRCA2 have stunning correlations with early-onset breast cancer (Wooster et al., 1995), while EGFR is a cell-surface receptor that mediates signal transduction pathways for growth and has been, unsurprisingly, linked to cancer. (Nicholson et al., 2001)

Our experiments also show that elevated levels of early growth response factor 1 (EGR1) and leptin (LEP) are associated with an increase in the number of adipose objects in the tumor. This is highly consistent with the role of insulin-regulated EGR1 in adipose differentiation and of leptin in energy metabolism; indeed, leptin and insulin are the major adiposity signals in the human body. (Havel, 2000)

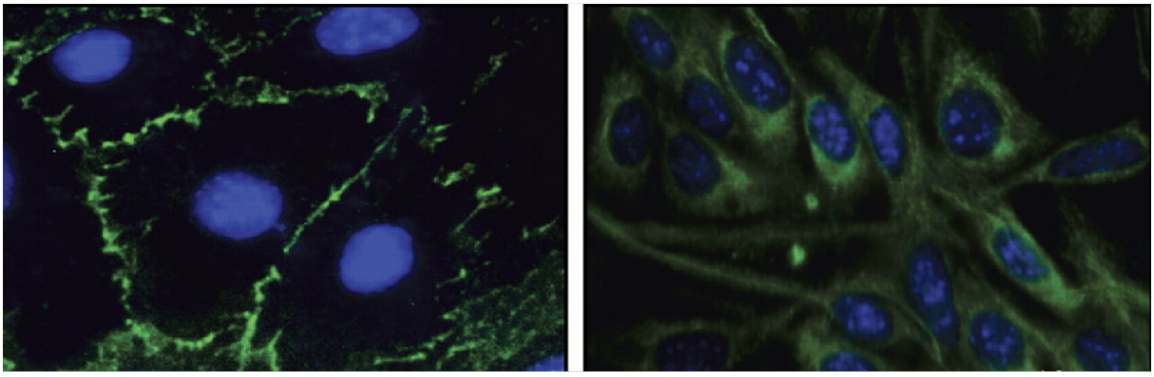

Figure 4.4: Left: Normal cell line. Right: After expression of CDC6, showing a loss of contact inhibition and a transition towards invasion and metastasis. Reproduced from (Sideridou et al., 2011).

For brevity, instead of enumerating the list, we next focus on two of the more striking associations found.

4.3.1.1 CDC6 affects epithelial-mesenchymal transition

Our experiments indicated that elevated expression levels of cell division control protein 6 homolog (CDC6) were consistently associated with an increase in the area of tissue classified as undergoing epithelial-mesenchymal transition (EMT). This is clinically important, because EMT is a critical step of the transition from local invasion to metastasis; if we can prevent EMT from occurring, we can therefore prevent metastasis.

EMT often results from a loss in contact inhibition between cells and a concurrent transformation into a more spindle-like shape, which allows the normally-adhesive cancer cells to slip out from the epithelial layer and into the circulation. Less than a year ago, (Sideridou et al., 2011) produced a striking experimental demonstration of our finding, by showing that expression of CDC6 repressed E-cadherin expression in epithelial cells, allowing them to morph into more invasive variants. (Figure 4.4) This was an unexpected finding, as CDC6 was previously thought to be predominantly involved in the regulation of DNA replication.

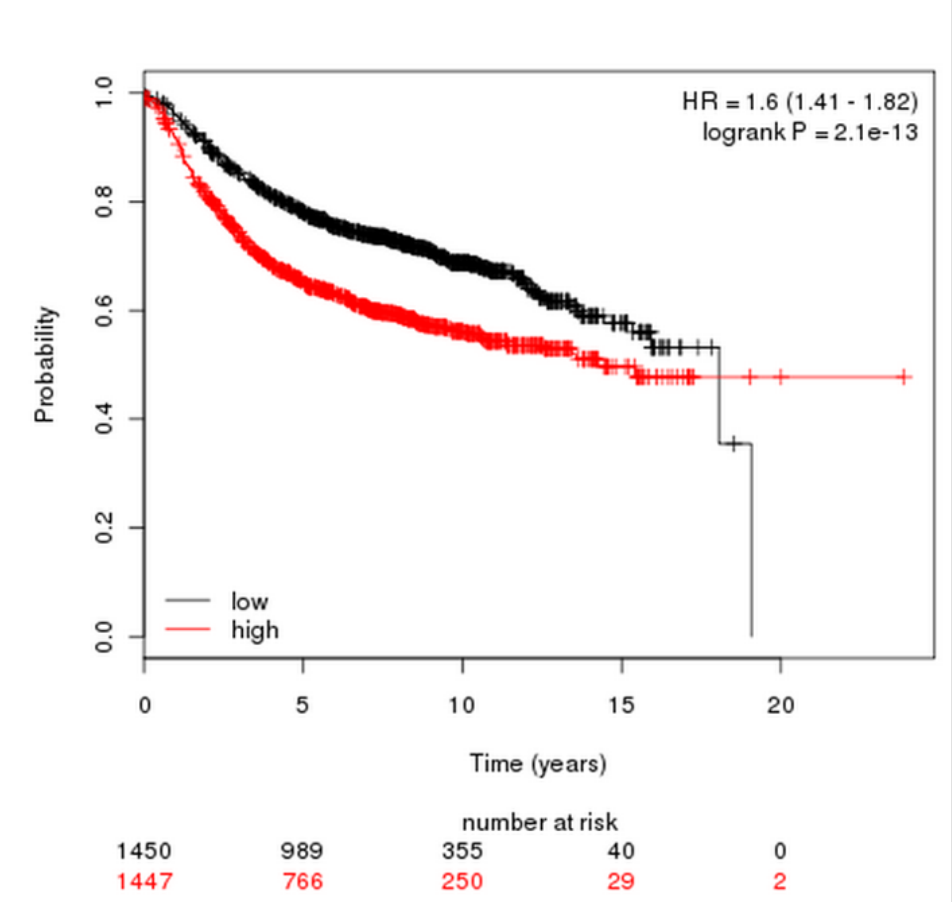

Figure 4.5: Kaplan-Meier plot comparing the survival of 2977 breast cancer patients with low (black) and high (red) levels of CDC6 expression.

We further investigated this association by stratifying a cohort of 2,977 breast cancer patients into two groups, based on whether they had CDC6 expression levels above or below the median, and studying the corresponding Kaplan-Meier plots. (Győrffy et al., 2010) As shown in Figure 4.5, the observed survival stratification between these two groups is extremely significant (logrank p-value of 2.1×10^{-13}); patients with higher levels of CDC6 have a markedly worse prognosis, consistent with the hypothesis that CDC6 expression gives rise to EMT and promotes metastasis, and is consequently bad for prognosis. (Beck et al., 2011)

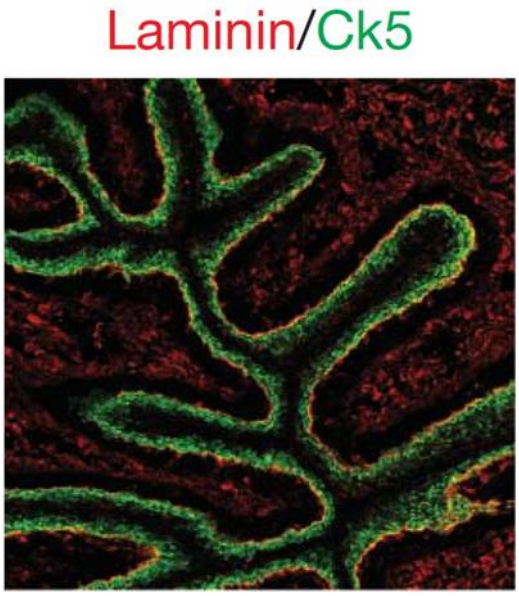

Figure 4.6: Laminin (red) immunostaining in the stroma. Reproduced from (Shin et al., 2011).

4.3.1.2 VIPR2 and LAMA2 are associated with epithelium-stroma proportions

The next set of experimental results involves two genes, the vasoactive intestinal polypeptide receptor 2 (VIPR2) and laminin, alpha 2 (LAMA2); our experiments show that elevated levels of both of these genes translate into an increase in the proportion of stroma vs. epithelium.

Laminin is an protein that is secreted in the extracellular space and extracellular matrix, and forms a integral component of the basement membrane; see Figure 4.6 for a graphical depiction of the levels of laminin expression in the stroma. Our finding that laminin is related to the proportion of stroma vs. epithelium is therefore a validation of our experimental methodology; indeed, laminin was recently found to play a key role in breast cancer. (Spencer et al., 2011)

Our experimental findings regarding VIPR2, however, are unexpected. Indeed, VIPR2 is not even known to be expressed in the breast (Mosca et al., 2010), and it has only been tenously linked to other cancers as a cell-surface marker on gastrointestinal tumor cells, without much functional elucidation. (Virgolini et al., 1996) To corroborate our findings,

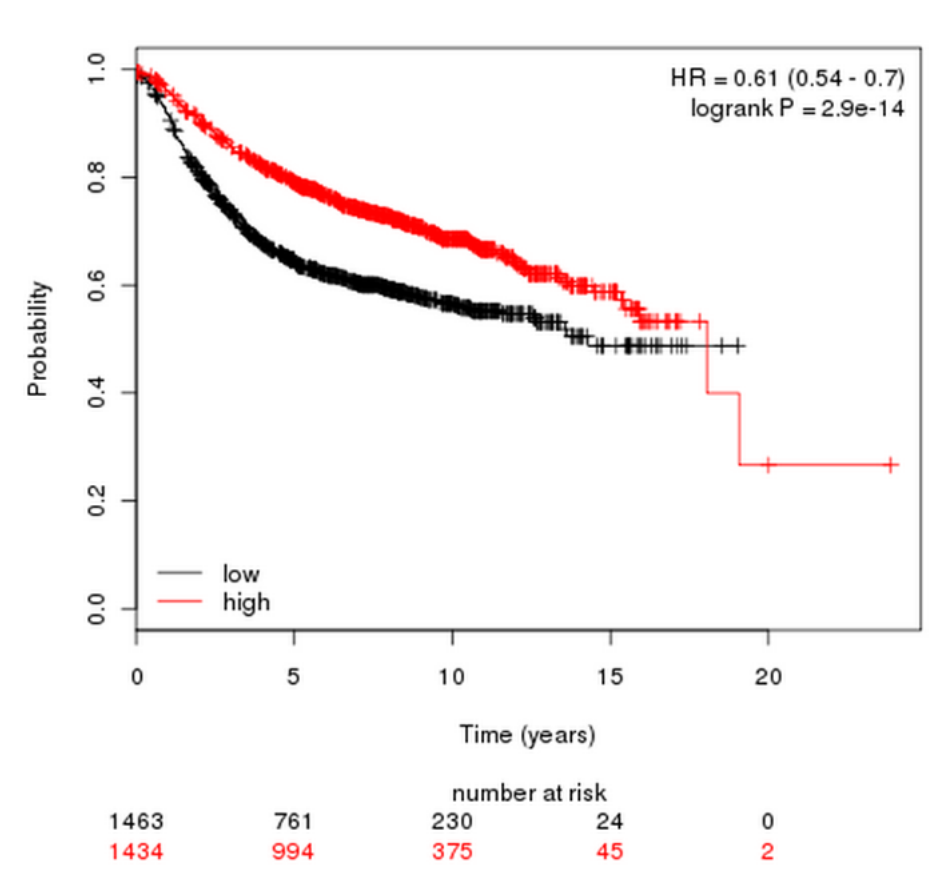

Figure 4.7: Kaplan-Meier plot comparing the survival of 2977 breast cancer patients with low (black) and high (red) levels of VIPR2 expression.

we repeat our survival analysis (as in the previous section) on the same cohort of 2,977 breast cancer patients, using VIPR2 expression as the basis for stratification. As shown in Figure 4.7, patients with higher levels of VIPR2 expression have a significantly better prognosis (logrank p-value of 2.9×10^{-14}) than those with lower levels of VIPR2 expression. This is consistent with previous findings that higher stromal proportions signify better prognosis in breast cancer (Beck et al., 2008), and strongly suggests that VIPR2 plays a more important role in breast carcinogenesis than previously thought.

Chapter 5

Discussion and Future Directions

Our results show the viability of running a integrative analysis of morpho-molecular data. In particular, we successfully re-discovered several known and experimentally-validated relations between gene activity and tumor morphology, and uncovered other putative genetic drivers that could lead to fruitful investigative follow-ups. These putative drivers include not only genes that are known to be important in other (non-breast) cancers, e.g., FGF-18, but also genes with suspected novel functions, such as VIPR2.

The recent discovery that CDC6 had a malignant and previously-unexpected function in relation to tumor morphology and metastasis underscores the need for high-throughput, unbiased studies that can leverage data modalities *other* than gene expression or mutation data. In that vein, we believe that the work in this thesis is a step in the right direction.

What are some promising avenues for future work? The logical next step is to perform in-vitro or in-vivo validations of the most promising gene-morphology associations that we have generated in the context of breast cancer. Besides these, there are two large and inviting areas of future work:

- **Transfer learning across different cancer types.** The Cancer Genome Atlas, from which we obtained half of the data used in this work, currently contains data on 6,000+ patients across 20+ different cancer types, and is growing rapidly. While each type - or sub-type - of cancer is bound to have some of its own unique mechanisms, the underlying machinery behind carcinogenesis is likely to be largely similar. We can use this to our advantage by jointly analyzing data from disparate cancer types,

in the same way that we used multi-task learning to combine data from disparate morphological traits, in order to paint a more complete picture of the biology of cancer.

- **Transfer learning across different data modalities.** Besides gene expression data, there is a plethora of complementary data sources available in the Cancer Genome Atlas, from somatic mutation calls to microRNA transcription levels. Generalizing the system we have developed in this thesis to handle these additional data modalities is a topic of both algorithmic and clinical interest.

At the same time, there are several more technical ideas to pursue, for example:

- **Data-driven phenotype grouping.** The current multi-task setup presumes some sort of inductive transfer across *all* of the separate tasks (morphological traits). While this is a reasonable assumption to make in the context of breast cancer alone, it might be overly restrictive, especially if we enrich our morphological feature set with, e.g., many tissue-specific features. A different approach is to assume that there is some sort of latent structure within the tasks that we need to discover: for example, the morphological features could be in reality divided into two different groups, one related to epithelium and one related to stroma, with little transfer learning occuring across group boundaries. We have been exploring various techinques of automatically learning this grouping from data, e.g., by performing unsupervised clustering on the coefficient weights at each iteration in the multi-task IRLM.

- **Performance-based phenotype regularization.** We have been careful to restrict our attention to a trusted set of morphological traits, e.g., by performing Cox regression to pick the top survival-related morphological features in the NKI dataset, or by laboriously obtaining a hand-labeled set of epithelial and stromal tissues to train a robust classifier for the TCGA dataset. As with the previous point, this is because our regressions are setup for transfer learning across *all* of the tasks; this means that the addition of noisy morphological features can significantly impact the estimates for other features. One possibility for circumventing this is to explore a

performance-based regularization system, where the amount that a particular morphological feature contributes towards shared learning is based on how well we can predict that feature on a held-out validation set.

Regardless of the future directions that the work presented here takes, it is clear that we, and cancer researchers in general, have only just begun to scratch the surface of integrative analyses that span and synthesize disparate data modalities. We eagerly look forward to the next wave of advances in cancer biology and therapy that we hope will fall into place, as scientists puzzle out the explosion of data that has characterized the last decade of cancer research.

Appendix A

List of Putative Genes Driving Cancer Morphology

A.1 NKI Experiments

1. agt
2. Akap9
3. BIRC5
4. BUB1
5. C1QB
6. Cdc25b
7. CENPA
8. cfd
9. CPB2
10. DSG1
11. FGF18
12. FGG

13. KIT
14. Kras
15. ldhb
16. pfkp
17. PGR
18. PLA2G7
19. SLC26A3
20. SLC40A1
21. Stat1
22. TGFB3
23. Vav3

A.2 TCGA Experiments

1. ADH1C
2. BRCA2
3. CDC6
4. CFD
5. CHRNA1
6. CKMT2
7. COL17A1
8. CPB2
9. EGFR
10. EGR1

11. FGF18
12. GPLD1
13. KIT
14. LAMA2
15. LEP
16. NGFR
17. PPP1R1B
18. SYN2
19. TAC1
20. TLR7
21. VIPR2

Bibliography

S. Aaltomaa, P. Lipponen, M. Eskelinen, E. Alhava, and K. Syrjänen. Nuclear morphometry and mitotic indexes as prognostic factors in breast cancer. *The European journal of surgery = Acta chirurgica*, 157(5):319–324, May 1991. ISSN 1102-4151. URL http://view.ncbi.nlm.nih.gov/pubmed/1678644.

A. S. Adler, M. Lin, H. Horlings, D. S. A. Nuyten, M. J. van de Vijver, and H. Y. Chang. Genetic regulators of large-scale transcriptional signatures in cancer. *Nature genetics*, 38(4):421–30, Apr. 2006. ISSN 1061-4036. doi: 10.1038/ng1752. URL http://dx.doi.org/10.1038/ng1752.

S. I. Ahmad, S. H. Kirk, and A. Eisenstark. Thymine metabolism and thymineless death in prokaryotes and eukaryotes. *Annual Review of Microbiology*, 52(1):591–625, 1998. doi: 10.1146/annurev.micro.52.1.591. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.micro.52.1.591.

U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. a. Garraway, and D. Pe'er. An Integrated Approach to Uncover Drivers of Cancer. *Cell*, 143(6):1005–1017, Dec. 2010. ISSN 1097-4172. doi: 10.1016/j.cell.2010.11.013. URL http://www.ncbi.nlm.nih.gov/pubmed/21129771.

A. Beck, I. Espinosa, C. Gilks, M. van de Rijn, and R. West. The fibromatosis signature defines a robust stromal response in breast carcinoma. *Laboratory investigation*, 88(6):591–601, 2008.

A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn, and D. Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, 3(108):108ra113, 2011. doi: 10.1126/scitranslmed.3002564. URL `http://stm.sciencemag.org/content/3/108/108ra113.abstract`.

J. A. Berger, S. Hautaniemi, S. K. Mitra, and J. Astola. Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 3(1):2–16. ISSN 1545-5963. doi: 10.1109/TCBB.2006.10. URL `http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?reload=true&arnumber=1588842`.

R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. C. Lee, J. H. Huang, S. Alexander, J. Du, T. Kau, R. K. Thomas, K. Shah, H. Soto, S. Perner, J. Prensner, R. M. Debiasi, F. Demichelis, C. Hatton, M. a. Rubin, L. a. Garraway, S. F. Nelson, L. Liau, P. S. Mischel, T. F. Cloughesy, M. Meyerson, T. a. Golub, E. S. Lander, I. K. Mellinghoff, and W. R. Sellers. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50):20007–12, Dec. 2007. ISSN 1091-6490. doi: 10.1073/pnas.0710052104. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2148413&tool=pmcentrez&rendertype=abstract`.

R. Beroukhim, C. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. Boehm, J. Dobson, M. Urashima, and Others. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905, 2010. ISSN 0028-0836. doi: 10.1038/nature08822.The. URL `http://www.nature.com/nature/journal/vaop/ncurrent/abs/nature08822.html`.

A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M.-B. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, J. a. Olson, J. R. Marks, H. K. Dressman, M. West, and J. R. Nevins. Oncogenic pathway signatures in human cancers as a guide to

targeted therapies. *Nature*, 439(7074):353–7, Jan. 2006. ISSN 1476-4687. doi: 10.1038/nature04296. URL http://www.ncbi.nlm.nih.gov/pubmed/16273092.

BioCarta. BioCarta - Charting Pathways of Life. 2012. URL http://www.biocarta.com/genes/index.asp.

H. J. Bloom and W. W. Richardson. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. *British journal of cancer*, 11(3):359–377, Sept. 1957. ISSN 0007-0920. URL http://view.ncbi.nlm.nih.gov/pubmed/13499785.

J. S. Boehm, J. J. Zhao, J. Yao, S. Y. Kim, R. Firestein, I. F. Dunn, S. K. Sjostrom, L. a. Garraway, S. Weremowicz, A. L. Richardson, H. Greulich, C. J. Stewart, L. a. Mulvey, R. R. Shen, L. Ambrogio, T. Hirozane-Kishikawa, D. E. Hill, M. Vidal, M. Meyerson, J. K. Grenier, G. Hinkle, D. E. Root, T. M. Roberts, E. S. Lander, K. Polyak, and W. C. Hahn. Integrative genomic approaches identify IKBKE as a breast cancer oncogene. *Cell*, 129(6):1065–79, June 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.03.052. URL http://www.ncbi.nlm.nih.gov/pubmed/17574021.

J. Bos. Ras oncogenes in human cancer: a review. *Cancer research*, 49(17):4682, 1989.

P. J. Campbell, S. Yachida, L. J. Mudie, P. J. Stephens, E. D. Pleasance, L. a. Stebbings, L. a. Morsberger, C. Latimer, S. McLaren, M.-L. Lin, D. J. McBride, I. Varela, S. a. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, C. a. Griffin, J. Burton, H. Swerdlow, M. a. Quail, M. R. Stratton, C. Iacobuzio-Donahue, and P. A. Futreal. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319):1109–1113, Oct. 2010. ISSN 0028-0836. doi: 10.1038/nature09460. URL http://www.nature.com/doifinder/10.1038/nature09460.

E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted l1 minimization, 2007.

E. Cerami, E. Demir, N. Schultz, B. S. Taylor, and C. Sander. Automated network analysis identifies core pathways in glioblastoma. *PloS*

one, 5(2):e8918, Jan. 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0008918. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2820542&tool=pmcentrez&rendertype=abstract.

M. Chidgey and C. Dawson. Desmosomes: a role in cancer? *British journal of cancer*, 96 (12):1783–1787, 2007.

R. Ciampi, J. Knauf, R. Kerler, M. Gandhi, Z. Zhu, M. Nikiforova, H. Rabes, J. Fagin, Y. Nikiforov, et al. Oncogenic akap9-braf fusion is a novel mechanism of mapk pathway activation in thyroid cancer. *J Clin Invest*, 115(1):94–101, 2005.

L. A. D. Cooper, J. Kong, D. A. Gutman, F. Wang, J. Gao, C. Appin, S. Cholleti, T. Pan, A. Sharma, L. Scarpace, T. Mikkelsen, T. Kurc, C. S. Moreno, D. J. Brat, and J. H. Saltz. Integrated morphologic analysis for the identification and characterization of disease subtypes. *Journal of the American Medical Informatics Association*, 19(2):317–323, 2012. doi: 10.1136/amiajnl-2011-000700. URL http://jamia.bmj.com/content/19/2/317.abstract.

V. Djonov, R. Ball, S. Graf, A. Mottaz, A. Arnold, K. Flanders, U. Studer, and V. Merz. Transforming growth factor-β3 is expressed in nondividing basal epithelial cells in normal human prostate and benign prostatic hyperplasia, and is no longer detectable in prostate carcinoma. *The Prostate*, 31(2):103–109, 1997.

M. J. Donovan, S. Hamann, M. Clayton, F. M. Khan, M. Sapir, V. Bayer-Zubek, G. Fernandez, R. Mesa-Tejada, M. Teverovskiy, V. E. Reuter, P. T. Scardino, and C. Cordon-Cardo. Systems pathology approach for the prediction of prostate cancer progression after radical prostatectomy. *Journal of Clinical Oncology*, 26(24):3923–3929, August 20, 2008. doi: 10.1200/JCO.2007.15.3155. URL http://jco.ascopubs.org/content/26/24/3923.abstract.

R. Elliott and G. Blobe. Role of transforming growth factor beta in human cancer. *Journal of Clinical Oncology*, 23(9):2078–2093, 2005.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent, 2009.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010a. URL `http://www.jstatsoft.org/v33/i01/`.

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso, 2010b. URL `http://www.citebase.org/abstract?id=oai:arXiv.org:1001.0736`.

C. B. Gambacorti-Passerini, R. H. Gunby, R. Piazza, A. Galietta, R. Rostagno, and L. Scapozza. Molecular mechanisms of resistance to imatinib in philadelphia-chromosome-positive leukaemias. *The Lancet Oncology*, 4(2):75 – 85, 2003. ISSN 1470-2045. doi: 10.1016/S1470-2045(03)00979-3. URL `http://www.sciencedirect.com/science/article/pii/S1470204503009793`.

C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y.-E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M.-H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–8, 2007. URL `http://www.ncbi.nlm.nih.gov/pubmed/17344846`.

B. Győrffy, A. Lanczky, A. Eklund, C. Denkert, J. Budczies, Q. Li, and Z. Szallasi. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast cancer research and treatment*, 123(3):725–731, 2010.

P. Havel. Role of adipose tissue in body-weight regulation: mechanisms regulating leptin production and energy balance. In *PROCEEDINGS-NUTRITION SOCIETY OF LONDON*, volume 59, pages 359–371. Cambridge Univ Press, 2000.

K. Horwitz and W. McGuire. Estrogen control of progesterone receptor in human breast cancer. correlation with nuclear processing of estrogen receptor. *Journal of Biological Chemistry*, 253(7):2223, 1978.

M. Hu, W. Qiu, Y. Wang, D. Hill, B. Ring, S. Scully, B. Bolon, M. DeRose, R. Luethy, W. Simonet, et al. Fgf-18, a novel member of the fibroblast growth factor family, stimulates hepatic and intestinal proliferation. *Molecular and cellular biology*, 18(10):6063–6074, 1998.

S. Jones, X. Zhang, D. W. Parsons, J. C.-H. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, H. Kamiyama, A. Jimeno, S.-M. Hong, B. Fu, M.-T. Lin, E. S. Calhoun, M. Kamiyama, K. Walter, T. Nikolskaya, Y. Nikolsky, J. Hartigan, D. R. Smith, M. Hidalgo, S. D. Leach, A. P. Klein, E. M. Jaffee, M. Goggins, A. Maitra, C. Iacobuzio-Donahue, J. R. Eshleman, S. E. Kern, R. H. Hruban, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu, and K. W. Kinzler. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science (New York, N.Y.)*, 321(5897):1801–6, Sept. 2008. ISSN 1095-9203. doi: 10.1126/science.1164368. URL `http://www.ncbi.nlm.nih.gov/pubmed/18772397`.

G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(suppl 1):D428–D432, 2005. doi: 10.1093/nar/gki072. URL `http://nar.oxfordjournals.org/content/33/suppl_1/D428.abstract`.

G. Kops, B. Weaver, and D. Cleveland. On the road to cancer: aneuploidy and the mitotic checkpoint. *Nature Reviews Cancer*, 5(10):773–785, 2005.

O. Kranenburg. The kras oncogene: past, present, and future. *Biochimica et biophysica acta*, 1756(2):81, 2005.

V. Le Doussal, M. Tubiana-Hulin, S. Friedman, K. Hacene, F. Spyratos, and M. Brunet. Prognostic value of histologic grade nuclear components of Scarff-Bloom-Richardson (SBR). An improved score modification based on a multivariate analysis of 1262 invasive ductal breast carcinomas. *Cancer*, 64(9):1914–1921, Nov. 1989. ISSN 1097-0142. doi: 10.1002/1097-0142(19891101)64:9\%3C1914::AID-CNCR2820640926\%3E3.0.CO;2-G. URL http://dx.doi.org/10.1002/1097-0142(19891101)64:9%3C1914::AID-CNCR2820640926%3E3.0.CO;2-G.

K. Lee, Y. Liu, J. Mo, J. Zhang, Z. Dong, and S. Lu. Vav3 oncogene activates estrogen receptor and its overexpression may be involved in human breast cancer. *BMC cancer*, 8(1):158, 2008.

K. M. Mani, C. Lefebvre, K. Wang, W. K. Lim, K. Basso, R. Dalla-Favera, and A. Califano. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Molecular systems biology*, 4:169, Jan. 2008. ISSN 1744-4292. doi: 10.1038/msb.2008.2. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2267731&tool=pmcentrez&rendertype=abstract.

E. Mosca, R. Alfieri, I. Merelli, F. Viti, A. Calabria, and L. Milanesi. A multilevel data integration resource for breast cancer study. *BMC systems biology*, 4(1):76, 2010.

S. Mukherjee. *The Emperor of All Maladies: A Biography of Cancer*. Fourth Estate (GB), 2010. ISBN 0007250916. URL http://www.worldcat.org/isbn/0007250916.

T. C. G. A. R. Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, Sept. 2008. ISSN 0028-0836. doi: 10.1038/nature07385. URL http://dx.doi.org/10.1038/nature07385.

R. Nicholson, J. Gee, and M. Harper. Egfr and cancer prognosis. *European Journal of Cancer*, 37:9–15, 2001.

S. G. O'Brien, F. Guilhot, R. A. Larson, I. Gathmann, M. Baccarani, F. Cervantes, J. J. Cornelissen, T. Fischer, A. Hochhaus, T. Hughes, K. Lechner, J. L. Nielsen, P. Rousselot, J. Reiffers, G. Saglio, J. Shepherd, B. Simonsson, A. Gratwohl, J. M. Goldman,

H. Kantarjian, K. Taylor, G. Verhoef, A. E. Bolton, R. Capdeville, and B. J. Druker. Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. *New England Journal of Medicine*, 348(11):994–1004, 2003. doi: 10.1056/NEJMoa022457. URL http://www.nejm.org/doi/full/10.1056/NEJMoa022457.

D. Patey, R. Scarff, and M. H. M. School. *The Position of Histology in the Prognosis of Carcinoma of the Breast*. 1928. URL http://books.google.com/books?id=maL8GwAACAAJ.

H. Pereira, S. E. Pinder, D. M. Sibbering, M. H. Galea, C. W. Elston, R. W. Blamey, J. F. Robertson, and I. O. Ellis. Pathological prognostic factors in breast cancer. IV: Should you be a typer or a grader? A comparative study of two histological prognostic features in operable breast carcinoma. *Histopathology*, 27(3):219–226, Sept. 1995. ISSN 0309-0167. URL http://view.ncbi.nlm.nih.gov/pubmed/8522285.

L. A. Pray. Gleevec: The breakthrough in cancer treatment. *Nature Education*, 2008. URL http://www.nature.com/doifinder/10.1038/306239a0.

S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub. A molecular signature of metastasis in primary solid tumors. *Nature genetics*, 33(1):49–54, Jan. 2003. ISSN 1061-4036. doi: 10.1038/ng1060. URL http://www.ncbi.nlm.nih.gov/pubmed/12469122.

T. Shimokawa, Y. Furukawa, M. Sakai, M. Li, N. Miwa, Y. Lin, and Y. Nakamura. Involvement of the fgf18 gene in colorectal carcinogenesis, as a novel downstream target of the β-catenin/t-cell factor complex. *Cancer research*, 63(19):6116, 2003.

K. Shin, J. Lee, N. Guo, J. Kim, A. Lim, L. Qu, I. Mysorekar, and P. Beachy. Hedgehog/wnt feedback supports regenerative proliferation of epithelial stem cells in bladder. *Nature*, 472(7341):110–114, 2011.

M. Sideridou, R. Zakopoulou, K. Evangelou, M. Liontos, A. Kotsinas, E. Rampakakis, S. Gagos, K. Kahata, K. Grabusic, K. Gkouskou, I. P. Trougakos, E. Kolettas, A. G.

Georgakilas, S. Volarevic, A. G. Eliopoulos, M. Zannis-Hadjopoulos, A. Moustakas, and V. G. Gorgoulis. Cdc6 expression represses E-cadherin transcription and activates adjacent replication origins. *JOURNAL OF CELL BIOLOGY*, 195(7):1123–1140, DEC 26 2011. ISSN 0021-9525. doi: {10.1083/jcb.201108121}.

T. Sjöblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. V. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, and V. E. Velculescu. The Consensus Coding Sequences of Human Breast and Colorectal Cancers. *Science*, 314(5797):268–274, Oct. 2006. ISSN 1095-9203. doi: 10.1126/science.1133427. URL `http://dx.doi.org/10.1126/science.1133427`.

V. Spencer, S. Costes, J. Inman, R. Xu, J. Chen, M. Hendzel, and M. Bissell. Depletion of nuclear actin is a key mediator of quiescence in epithelial cells. *Journal of Cell Science*, 124(1):123–132, 2011.

B. Sriperumbudur and G. Lanckriet. On the convergence of the concave-convex procedure. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1759–1767. 2009.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005. doi: 10.1073/pnas.0506580102. URL `http://www.pnas.org/content/102/43/15545.abstract`.

B. S. Taylor, N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, B. S. Carver, V. K. Arora, P. Kaushik, E. Cerami, B. Reva, Y. Antipin, N. Mitsiades, T. Landers, I. Dolgalev, J. E. Major, M. Wilson, N. D. Socci, A. E. Lash, A. Heguy, J. a. Eastham, H. I. Scher, V. E. Reuter, P. T. Scardino, C. Sander, C. L. Sawyers, and W. L. Gerald. Integrative Genomic Profiling of Human Prostate Cancer. *Cancer cell*, 18(1):

11–22, July 2010. ISSN 1878-3686. doi: 10.1016/j.ccr.2010.05.026. URL http://www.ncbi.nlm.nih.gov/pubmed/20579941.

M. Teverovskiy, V. Kumar, J. Ma, A. Kotsianti, D. Verbel, A. Tabesh, H.-y. Pang, Y. Vengrenyuk, S. Fogarasi, and O. Saidi. *IMPROVED PREDICTION OF PROSTATE CANCER RECURRENCE BASED ON AN AUTOMATED TISSUE IMAGE ANALYSIS SYSTEM*, volume 2, pages 257–260. IEEE, 2004. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1398523.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):pp. 267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

M. J. van de Vijver, Y. D. He, L. J. van 't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002. doi: 10.1056/NEJMoa021967. URL http://www.nejm.org/doi/full/10.1056/NEJMoa021967.

F. Vandin, P. Clay, E. Upfal, and B. Raphael. Discovery of mutated subnetworks associated with clinical data in cancer. *Pac Symp Biocomput*, 2012.

R. Verhaak, K. Hoadley, E. Purdom, V. Wang, Y. Qi, M. Wilkerson, C. Miller, L. Ding, T. Golub, J. Mesirov, and Others. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1):98–110, 2010. ISSN 1535-6108. doi: 10.1016/j.ccr.2009.12.020.An. URL http://linkinghub.elsevier.com/retrieve/pii/S1535610809004322.

I. Virgolini, M. Raderer, A. Kurtaran, P. Angelberger, Q. Yang, M. Radosavljevic, M. Leimer, K. Kaserer, S. Li, G. Kornek, P. HÃŒbsch, B. Niederle, J. Pidlich, W. Scheithauer, and P. Valent. 123i-vasoactive intestinal peptide (vip) receptor scanning: Update

of imaging results in patients with adenocarcinomas and endocrine tumors of the gastrointestinal tract. *Nuclear Medicine and Biology*, 23(6):685 – 692, 1996. ISSN 0969-8051. doi: 10.1016/0969-8051(96)00066-2. URL http://www.sciencedirect.com/science/article/pii/0969805196000662. <ce:title>New Trends in Nuclear Oncology</ce:title>.

R. Weinberg. *The biology of cancer*, volume 255. Garland Science New York, 2007.

R. Wooster, G. Bignell, J. Lancaster, S. Swift, S. Seal, J. Mangion, N. Collins, S. Gregory, C. Gumbs, G. Micklem, et al. Identification of the breast cancer susceptibility gene brca2. *Nature*, 378(6559):789–792, 1995.

A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Comput.*, 15 (4):915–936, Apr. 2003. ISSN 0899-7667. doi: 10.1162/08997660360581958. URL http://dx.doi.org/10.1162/08997660360581958.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x. URL http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.