Genomic Analysis of Pathway Signaling in Glioblastoma and Other Cancers

by

Jason Windham Reeves

University Program in Genetics and Genomics
Duke University

Date:_____
Approved:

_____
Joseph R. Nevins, Supervisor

_____
John H. Sampson

_____
Christopher M. Counter

_____
Gregory E. Crawford

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy
in the University Program in Genetics and Genomics
in the Graduate School
of Duke University

2012

ABSTRACT

Genomic Analysis of Pathway Signaling in Glioblastoma and Other Cancers

by

Jason Windham Reeves

University Program in Genetics and Genomics
Duke University

Date:_____
Approved:

_____
Joseph R. Nevins, Supervisor

_____
John H. Sampson

_____
Christopher M. Counter

_____
Gregory E. Crawford

An abstract of a dissertation submitted in partial fulfillment of
the requirements for the degree Doctor of Philosophy in
the University Program in Genetics and Genomics
in the Graduate School
of Duke University

2012

Copyright by
Jason Windham Reeves
2012

Abstract

The disease process giving rise to cancer involves the consecutive accumulation of genetic or genomic alterations impacting the normal regulation of cellular functions. In cases of hereditary cancers, this process may be stepwise, with a shared initiating lesion leading to common subsequent alterations. However, in many non-hereditary forms of cancer the initiating and subsequent alterations giving rise to the tumor can vary substantially from individual to individual, and multiple molecularly distinct subsets of the disease can exist within histopathologically similar tumors. This molecular heterogeneity between patients hinders the ability to identify which alterations are responsible for tumor development and subsequent maintenance, and confounds the ability to effectively treat patients as response to a particular therapeutic intervention may be highly dependent on the molecular composition of the disease.

To further our understanding of the molecular alterations associated with tumorigenesis, we analyzed aggressive brain cancer, glioblastoma (GBM), samples for which multiple types of genome-wide information was available. We utilized a series of *in vitro* or clinically derived gene expression signatures by comparing gene expression of samples based on whether a particular cellular signaling pathway was known to be active or inactive. Using these signatures for cellular signaling deregulation, we examined the association between various genomic alterations and the relative activity

of each pathway, identifying alterations that were enriched within patients that harbored similar profiles of pathway activation. These analyses lead to the identification of numerous previously uncharacterized alterations in GBM, including the identification of a ubiquitin-like gene, UBL3, that was associated not only with pathway signaling, but was also associated with poor patient outcome, as well as response of GBM xenograft models to treatment with standard of care therapeutic agents.

Further, given that the challenges involved in analyzing clinical samples include development methods for timely analysis of genomic data, we have described a framework to utilize these genomic signatures in a prospective setting by incorporating a non-overlapping reference dataset of similar tumor samples. This methodology allows the examination of pathway signaling, as captured by the signature, to be run in real-time when only a single patient sample is analyzed, and has a high degree of fidelity to the results generated from retrospective analysis across multiple tumor types. Together these studies have provided a novel framework for identification of significant genomic alterations that impact pathway signaling, as well as moving providing the mechanisms to analyze genomic signatures in a robust manner that accounts for the challenges associated with the prospective clinical setting.

Dedication

This work is dedicated to my parents, Jim and Kathy Reeves, and to my loving husband, Jason Clark.

Contents

Abstract iv

List of Tables x

List of Figures xi

Acknowledgements xiv

1. Introduction 1

1.1 Molecular heterogeneity and signaling deregulation in cancer 1

1.2 Tools and paradigms for clinical genomic analysis 5

1.2.1 Investigational algorithms for clinical genomics 7

1.2.2 Statistical validation of genomic analyses 13

1.2.3 Practical considerations for clinical application 14

1.3 Glioblastoma 17

1.3.1 Clinical presentation of glioblastoma and other diffusely infiltrative gliomas 17

1.3.2 Molecular characteristics of glioblastoma 19

1.3.3 Genomic classifications of gliomas 21

1.3.4 Surgical and therapeutic interventions for treatment of glioblastoma 25

1.4 Research rationale and summary 29

2. A Pathway Association Framework for Identification of Significant Genome Alterations in Glioblastoma 32

2.1 Summary 32

2.2 Background 33

2.3 Results 35

2.3.1 Expression signatures reflect the distinctions among GBM subtypes 35

2.3.2 Integrating pathway data with genomic alterations. 43

2.3.3 Pathway association analysis identifies known and putative drivers in GBM. 49

2.3.4 Patient outcome is stratified by pathway signaling 55

2.3.5 UBL3 is associated with IFNγ signaling, patient survival and response to therapeutic intervention. 61

2.4 Discussion 73

2.5 Materials and methods 79

3. Utilization of Cell Line Derived Genomic Signatures to Predict Oncogenic Pathway Activity in Human Tumors and Prospective Clinical Studies 85

3.1 Summary 85

3.2 Background 86

3.3 Results 89

3.3.1 Development of genomic signatures derived from experimental perturbation of pathway activity 89

3.3.2 A general strategy to evaluate expression signatures in prospective clinical studies 96

3.4 Discussion 110

3.5 Materials and Methods 114

4. Concluding Remarks 119

4.1 Extending genomic signatures of pathway signaling 121

4.1.1 Deeper integration and network analysis 121

4.1.2 Cytomegalovirus signaling in GBM 123

4.1.3 Factor analysis as a statistical extension to the signature framework 125

4.2 Additional practical concerns for prospective genomic analysis 129

Appendix A 132

Appendix B 134

References 137

Biography 160

List of Tables

Table 1: Copy number alterations significantly associated with pathway activity in all GBM expression datasets 50

Table 2: Multivariate Cox Proportional Hazards analysis of E2F1 and prognostic markers in GBM patients 59

Table 3: Association between candidate expression and xenograft treatment response .. 70

Table 4: Survival and expression characteristics of GBM xenograft models 71

Table 5: Parameters for pathway signature development 132

List of Figures

Figure 1: Scheme for the development and assessment of a pathway activation signature. 12

Figure 2: Utilizing pathway signatures for a genetic association study. 37

Figure 3: Pathway activation in GBM samples. 39

Figure 4: Overlap between pathway signature, GBM subtype and highly differentially expressed GBM genes 41

Figure 5: Robustness of pathway signatures when limited to GBM-related genes 42

Figure 6: Linking driver mutations with pathway activation. 43

Figure 7: Association between copy number alterations and pathway activity. 45

Figure 8: Location of GISTIC and Pathway Associated Regions of Interest 46

Figure 9: Distance between GISTIC ROIs and pathway associated loci 48

Figure 10: Association between E2F1 signaling and copy number alterations 53

Figure 11: RFC4 expression in normal brain and glioblastoma tissue samples 54

Figure 12: Patient outcome during aggressive therapeutic intervention is stratified by E2F1, IFNγ, and Ras pathway activity 56

Figure 13: E2F1 is a poor prognostic marker in the non-aggressive treatment setting 58

Figure 14: E2F1 pathway signaling reflects proliferative signaling in GBM tumors 60

Figure 15: Association between copy number alterations and IFNγ signaling. 62

Figure 16: Comparison of alterations identified on chromosome 13 by GISTIC or pathway association analysis 63

Figure 17: Expression of candidate genes in normal tissues 65

Figure 18: Effect of copy number alteration on mRNA expression of UBL3 and HMGB1 relative to normal brain 66

Figure 19: UBL3 loss associates with poor patient survival 68

Figure 20: Analysis of overall survival of glioma patients based on alteration of genes flanking UBL3 in Chr13q12.3 69

Figure 21: Xenograft response to standard of care intervention 72

Figure 22: Principal component analysis of training and test data before and after normalization using shift-scale. 91

Figure 23: Strategies for applying BinReg methodology to predict E2F1 pathway status in ovarian cancer samples. 93

Figure 24: Prediction of RB1 status with an E2F1 and a Src signature. 95

Figure 25: Framework for prospective analysis of genomic signatures 97

Figure 26: Effect of size of reference dataset on accuracy of predictions. 99

Figure 27: Prospective analysis of GBM and Ovarian E2F1 pathway activity 101

Figure 28: Permutation testing of prospective analysis 102

Figure 29: Batch effects on prospective simulation of pathway predictions. 103

Figure 30: Effect of independent GBM reference datasets on the concordance of simulated prospective predictions. 105

Figure 31: Effect of using a single independent GBM reference datasets on the concordance of simulated prospective predictions. 106

Figure 32: Effect of ER status in batch-independent breast cancer reference datasets on the concordance of simulated prospective predictions. 108

Figure 33: Pathway biased reference datasets compromise prospective predictions 109

Figure 34: Impact of reference array platform on prospective simulation 113

Figure 35: Expression normalization and validation of associations 133

Figure 36: Pathway signaling identifies key microRNA network hubs 134

Figure 37: CMV-encoded microRNAs associate with pathway signaling 135

Figure 38: CMV microRNA Expression-based clusters of GBM 136

Acknowledgements

I could not be more grateful for the guidance, the mentorship, and the perspective that my advisor, Joseph Nevins, has provided. He, and John Sampson, laid the groundwork for the projects described in this dissertation and the feedback and advice they have given me has been invaluable. I also owe a huge debt of gratitude for the advice and suggestions of the other current and former members of my committee, Jen Tsan Chi, Christopher Counter, Gregory Crawford.

The support, guidance, and enthusiasm of the members of the Nevins' lab were without a doubt a key component to my success and growth as a researcher at Duke. Dawn Chasse, Rachel Rempel, Jenny Freiedman, Mike Gatza, Steve Angus and So Young Kim have provided a sounding board for experiments and shared in my excitement along the way. Erich Huang, Jeff Chang, and Joe Lucas have shaped my understanding of computational biology. I cannot thank enough the graduate students that I overlapped with, Eric Jiang, Heather LaBreche, and William Kim, for their guidance and perspective on research and graduate school. Similarly, there are numerous members of the Brain Tumor Center without whom this project would have not been possible. These projects represent not only my work, but collaborative efforts from enough Oscar Perrson, Steve Keir, David Reardon, and Divya Ajay, and their assistance during this research has been invaluable.

The community within the Duke graduate school has also been an integral part of my success. The administrators of my program, Mohamed Noor, Beth Sullivan, Joseph Heitman, and Doug Marchuk have provided unwavering support throughout my career, and for this support I am continually grateful. It also goes without saying that I would not have been able to navigate my graduate studies without the aide of Andrea Lanahan and Leslie Mavengere.

The fellow students, researchers and friends that I have been lucky enough to meet and interact with at Duke have been a great blessing during this project. There is no way to express how grateful I am to have had the pleasure to know Kat Mitchel, Jeff Bandy, Liz Ballou, Ginger Hunter, Amie Rodgers, Aaron Sarazan, Martha Alonzo, Justin Johnson, and Kira and Brian Roberts. They have shared in my greatest triumphs, pushed me to be a better researcher and person, and humbled me with their friendship.

Finally, none of this work would have been possible without the steadfast love and support of my parents, Jim and Kathy Reeves, my brothers, Jonathan and Jordan, and especially my husband, Jason Clark. Their constant encouragement, patience, and support have been a key to my continued pursuit of science. I could not have done any of this without them.

1. Introduction

1.1 Molecular heterogeneity and signaling deregulation in cancer

Cancer can be viewed primarily as a disease which subverts normal cellular processes to create a favorable environment for continual cellular growth and division. This picture of the disease, however, is an incomplete description of the systematic and context specific processes that occur in order to promote malignant transformation. The formation of a tumor and its subsequent progression and metastasis requires that the malignant cells acquire several key characteristics which allow them to escape tightly regulated cellular processes that constrain proliferation, growth arrest and programmed cell death. These characteristics, which progressively build upon each other, provide the mechanisms for these cells to exit from the normal homeostatic environment and regulatory signaling pathways and transition to a state that favors sustained tumor growth and subsequent invasion into normal tissue (1).

In order for tumor cells to survive and produce a malignant lesion they must acquire the ability to divide indefinitely while avoiding apoptotic signaling, sustain growth signaling pathways while ignoring growth suppressive cues, increase local angiogenesis, and invade or metastasize into surrounding tissues. However, these characteristics are not isolated events; they occur together to promote malignant progression. Additionally tumor cells must also acquire the ability to avoid immune surveillance and, in many cases, to deregulate the normal metabolic pathways within

the cell (2). These processes require a complex interplay not only between tumor cells, but also endothelial cells, fibroblasts, and immune cells. Even though these neighboring cells, making up the tumor stroma, are not malignant themselves, several cancers rely on stromal signaling to support their ongoing proliferation (3, 4), and mutations in stromal cells have also been found to enhance tumor formation and malignancy (5, 6). Moreover, this heterotypic cellular environment is found not only in primary tumor lesions, but is also present in subsequent metastatic lesions.

These tumor-promoting characteristics arise from the acquisition of alterations to the loci encoding normally functioning genes or the regulatory mechanisms governing their expression. These alterations can be hereditary, spontaneously induced, or driven by exposure to exogenous physical and chemical agents. Regardless of the mechanism by which an alteration is induced, they primarily include DNA damage in the form of mutations, epigenetic deregulation or copy number alteration, all of which functionally deregulate cell signaling at the level of mRNA, microRNA and protein expression and activation.

Many of the initial discoveries into the molecular lesions driving tumor formation came from the study of virally encoded genes, as the lytic production of viral particles required the ability to overcome many of the same hurdles that are deregulated in malignant cells, including activation of cellular proliferation machinery and avoidance of apoptotic signaling pathways. Studying the human genes which were

homologous to virally encoded genes identified key genes associated with human malignancies, such as Ras and Src (7, 8). These studies also lead to the understanding that tumor-promoting alterations could occur when these genes are overexpressed or constitutively activated that is sufficient to promote malignant transformation. Conversely, seminal studies of the hereditary cancer retinoblastoma demonstrated that alterations could occur in genes that required alteration across both copies of the gene, and served to suppress tumor formation (9). This two-hit hypothesis proposed that a tumor-suppressor gene would first acquire an inactivating mutation in one copy of the gene, which then was followed by loss of heterozygosity by deletion or gene conversion of the other copy of the gene. While TP53 was later characterized to repress tumor formation as well (10-12), owing to deletions and dominant negative mutations associated with colorectal cancer, this scheme functionally segregated genes that participated in the tumorigenic process into tumor suppressors and proto-oncogenes. Since these studies, our understanding of the complex interplay of alterations has broadened, demonstrating that regulators of numerous key pathways, from biogenesis to inflammatory response, play roles in tumor development when disrupted or deregulated.

These studies also lead to the observation that cancer represented a progressive disease, where an initial alteration of a single gene could initiate the tumorigenic process, but that aggressive tumor development required additional subsequent

alterations. However, in many cancers there are multiple different ways in which tumorigenesis can take place, and these models can be a poor reflection of the tumorigenic process if they are inflexible to the inherently stochastic process of the cumulative acquisition of genetic alterations. While a single mechanism for progression is clearly the case for some types of cancer, such familial colorectal cancers or chronic mylogenous leukemia (13, 14), further research into tumor development has indicated that many cancers are more molecularly diverse than initially assumed, and that the alterations driving the cancer are restricted based on cell and tissue of origin.

As the methods for investigating the collective cancer genome have become widely available and cost effective, it has become increasingly evident that in a variety of cancers there is a large diversity in the scope and coordination of alterations that are found in the patient population. There has been a growing body of evidence demonstrating that there is substantial inter-patient heterogeneity at the molecular level, defining distinct subpopulations of tumors that reflect the diversity present in the molecular etiology of the disease. This has perhaps been most appreciated in the setting of breast cancer where there appear to be four or more distinct subsets of the disease, depending on the methodology employed and patient population analyzed (15, 16). However, this has also been demonstrated in other types of tumors (17-19), further highlighting the stochastic and sporadic developmental process underlying tumorigenesis.

1.2 Tools and paradigms for clinical genomic analysis

One of the challenges to studying the development and progression of a disease with a complex and stochastic underlying mechanism, such as cancer, is the ability to detect biologically relevant characteristics promoting disease progression and subsequently validate the effect of these alterations. The availability of the complete sequence of the human genome has provided access to a wealth of information, and was proposed to provide the key to understanding the disease process at a much more granular level (20, 21). Coupled with the development of microarrays (22), these resources have provided researchers access to a rapidly expanding set of candidate disease-related genes as well as the means to test them concurrently.

Regardless of platform used, microarrays rely on the principle that nucleic acid sequences that are affixed to a substrate will preferentially hybridize to complimentary sequences when combined *in vitro*. While they offer a much denser information content than non-multiplexed assays, microarrays represent an extension of the technologies that have been well established in the study of nucleic acids (23, 24). The combination of having access to annotations spanning the entire genome, and methods by which to test those hundreds of thousands of genes simultaneously has lead to an explosion of studies examining the genomic complement of mRNA or microRNA expression, copy number status, and methylation patterns associated with various disease and cellular states. However, with this wealth of information, researchers have been challenged to analyze

and interpret this data using appropriate statistical methods so that the results of such studies would be robust and reproducible.

Methods used during low-throughput analyses were no longer appropriate to employ without significant adjustments in genomic studies, as most of these studies consider at most several hundred samples while at the same time analyzing hundreds of thousands of probes. This highly multiplexed data allows researchers to test far more hypotheses than the number of samples found in a study can statistically support (25), because the rate of false positives associated with analyses increases proportionally with the number of tests considered. Best practices for how to define, test and validate candidate genes and models built from genomic data were developed to safeguard against over-fitting the data (26). To guard against the increased error rates, researchers have incorporated techniques originally pioneered in the fields of machine learning, applied physics, and classical statistics.

The source of the data and the hypothesis being tested determine the framework under which a hypothesis can be tested. The analysis can be run in a supervised manner, with distinct classes of samples being identified prior to analysis and informing the models derived, or can be unsupervised, where the goal of the analytical framework is to identify novel characteristics that are emergent from the data itself without researcher guidance. Primarily the difference between these approaches depends on whether there is a robust pre-determined phenotype that can be used to segregate or order samples

prior to analysis. Both frameworks can be implemented for many of the methods described below, and both require validation regardless of the training set upon which the models, genes or patients of interest are built or identified.

1.2.1 Investigational algorithms for clinical genomics

Researchers studying genome-wide data have utilized a variety of approaches to explore the complex data available in genomic analysis, each that offer unique strengths and caveats to consider based on the hypothesis being tested. In clinical genomics, most of these algorithms are implemented to identify clusters of similar patient samples or genomic features; to quantify correlations and relationships within a dataset; or to identify prognostic or predictive features within a dataset (19, 27, 28). To move this complex information into a more manageable analytical space, many genomic analysis methods focus on reducing the number of features being tested in order to identify significant characteristics of genomic data that may be obscured by the volume of data assayed. Other methods rely on analysis of depicting covariance between samples or probesets analyzed, assuming that the structure of covariance is important in developing robust models of a disease state.

The clinical and cancer genomics community has relied heavily on clustering algorithms that identify groups of samples or genomic variables that are consistently and quantitatively correlated with each other. One of the most commonly used

clustering algorithms in clinical genomics has been agglomerative unsupervised hierarchical clustering (USHC) (29). This algorithm does not define distinct clusters on its own, but rather presents a matrix view of samples ordered by how close or far they are relative to other samples in the datasets, and which can be calculated using a variety of distance metrics. Researchers can then heuristically or using various thresholds for clusters, identify feature sets within the study to further investigate.

On its own, USHC does not directly imply or define what the most important clusters of samples are. However, various methods have been developed to identify robust and meaningful structure within the hierarchical relationships identified. These methods primarily rely on permutation testing or bootstrap analysis wherein different subsets of the patient or genomic data is repeatedly tested to identify the most robust and reproducible structures within the data (30-32). Another commonly used method of cluster identification is K-means clustering (33), which requires the researcher to defines how many clusters they anticipate identifying (*K*). The algorithm calculates the best fit for samples into *K* clusters based on how close samples fall to the estimated mean or centroid value. In contrast to USHC, one of the direct outputs of K-means analysis is a calculation of cluster membership, either presented as the probability a sample falls within a cluster or the best-fit cluster for a sample.

These methods are important for identifying emergent properties within multidimensional datasets, however, post-hoc analysis of a cluster is usually required to

interpret what the structure identified within the data represents. This often leads researchers to use annotation based approaches to understanding which genes are enriched within a particular cluster. Such analyses can be implemented to compare a list of genes to previously annotated ontologies, comparing the probability that those genes would be found together by chance (34-36). Conversely, groups of samples can be compared to identify the whether a set of genes is enriched or depleted within one group of samples relative to other groups of samples using Gene Set Enrichment Analysis (GSEA) (37). These methods most often define the likelihood that group is enriched, and whether that is statistically significant based on how the analysis was run.

While annotation based approaches have the ability to identify biological signaling that reflects known biological events within a disease or experimental setting, this class of analyses has the caveat that the enrichment of annotations is directly dependent on the accuracy of an annotated list within the test setting. Network analysis can further extend annotation based analyses further, or refine the connections predefined by an annotated relationship (38-42). This approach relies on many techniques developed in graph theory and aims to examine either annotated relationships to identify coordinately enriched nodes within a dataset, or build novel network arrangements based on co-expression and correlation of expression data. Fundamentally, this method can be seen as an extension of either the clustering

algorithms or the ontological analyses that provides a visual graph of how a network of biological signaling is constructed.

Most of the approaches described above offer qualitative assessments of the structure and variation within a particular set of genomic samples, and often these algorithms do not produce robust quantitative metrics that can be extended or employed in further analyses. To move beyond a descriptive understanding of genomic data, the broad class of algorithms used for dimensionality reduction allows for both concise visualization and quantification of multidimensional data, though the biological interpretation of the condensed data may not always be readily apparent. As a whole these analytical approaches includes regression analyses, feature selection, principal component (PCA) or singular value decomposition (SVD) analyses, self-ordered maps, and factor analyses (43). Perhaps the most common application of dimension-reduction techniques used in cancer genomics is PCA (44), as this method is useful for visualizing the underlying structure of a dataset being analyzed and identifying outliers. This approach identifies the orthogonally constrained multivariate axes that define the greatest variation among the data analyzed and allows for both quantitative and visual representation of genome-wide data without need for a direct mechanistic understand of what this variation captures.

Regression and factor analyses can improve upon a PCA by identifying variation within multidimensional spaces that is related to a particular phenotype of interest, or

by identifying variation that is not orthogonally constrained. These methods form the basis for building biologically relevant models which can later be applied to or projected onto additional datasets, and which affords significantly more flexibility than clustering based models which are, by definition, discrete. This ability, allows researchers to investigate the underpinnings of a defined biological state that is quantitatively captured, rather than qualitatively developed such as in annotation or cluster based approaches.

It is this class of analytical methods that our lab has leveraged to understand the inherent heterogeneity, and the clinical implications of that heterogeneity, that is present in a variety of cancers (45-49) (Figure 1). We have applied regression techniques to model key cellular signaling pathways that are commonly deregulated across cancer, and this approach provides a platform to predict how active or repressed a particular signaling pathway appears to be within a new biological sample. But perhaps most importantly, this methodology calculates a quantitative measurement of signaling deregulation that can then serve as a phenotype for further inquiry and discovery that qualitative assessments of pathway deregulation cannot offer. Even though these models have been built primarily from *in vitro* experiments, they have been shown to model deregulation which occurs across various cancer types and to robustly capture not only signaling deregulation but also clinical outcome and response to targeted therapeutic intervention (16).

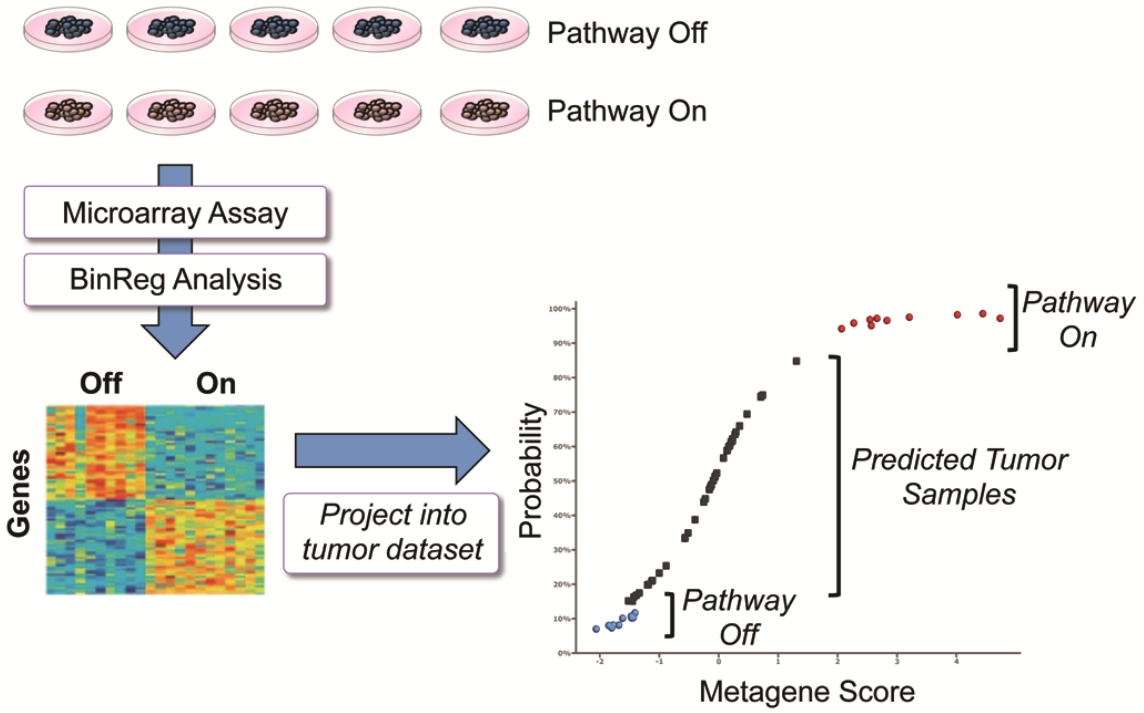

Figure 1: Scheme for the development and assessment of a pathway activation signature.

The process begins with the creation of a training dataset in which a given pathway is experimentally activated within a population of cell lines, RNA is prepared, and then used to generate whole genome gene expression data using DNA microarray analysis. This data is then employed with a Binary Regression methodology to develop a model reflected as a collection of differentially expressed genes with associated regression weights. The model can then be applied to a test dataset, such as a collection of tumor samples, to predict the status of the pathway, represented as probability values that lie between the extremes of the training data.

1.2.2 Statistical validation of genomic analyses

With all of these approaches, from dimension reduction to ontology analysis to cluster analysis, a key concern is the possibility of over-fitting data to a model that performs well in one dataset but fails to be reproducible outside of that dataset. This particular problem has been remarked on numerous times in the literature (50-52). Initially, leave-one- or leave-some-out methodologies were proposed to validate the utility of a genomic analysis. However, this method of validation can still be prone to over-fitting if the discovery dataset is not reflective of the general population being tested. Given this concern, holdout analysis should be used as initial assessment of robustness, though further analysis is necessary to robustly validate a model.

Ideally validation of a genomic model should entail analysis in at least one separate, non-overlapping dataset in which the original finding is reproduced. A validation dataset could consist of samples collected from the same study but which withheld and processed and analyzed completely separately. However, this framework for validation assumes that the patient population accrued at during the study is reflective of the entire patient population as a whole, which may not always be the case. More often, however, validation datasets consist of publicly available data acquired from studies which were developed to address unrelated questions in the same disease or cellular setting. With the growth of resources such as GEO (53), the cancer genome atlas (54), and oncomine (55), it is increasingly possible to use such publicly available

data to validate the findings of genomic analyses, and this should be considered to be a critical component to any analytical assessment of genomic data.

In addition to these methods of validation, one of the fundamental principles implemented in the analysis of high-dimension datasets such as microarray data is adjusting the analytical outputs to account for type-I and type-II error rates associated with testing numerous hypotheses concurrently. One of the simplest and most conservative methods to accomplish this is to adjust for family-wise error using a Bonferroni adjustment. While other methods for adjusting family-wise error have been proposed to improve the high type-II error rate of the Bonferroni adjustment, Bonferroni remains the most widely used adjustment for multiple testing. An alternative method for adjusting reported significance is to estimate the false-discovery rate, which focuses on controlling type-II rather than type-I error. The most common implementation of false-discovery rate adjustments is the Benjamini–Hochberg step-up procedure, though this has been adapted and extended based on the specific usage scenario that the researcher employs (25, 56).

1.2.3 Practical considerations for clinical application

While genomic techniques have been implemented to understand the landscape of alterations that exist for numerous diseases, many of the analytical models generated cannot be implemented in the clinical setting where individual samples must be dealt

with in a timely manner. Also, gene expression arrays have not found their way into the clinical setting though they are relatively cost-efficient. This is due in part to ongoing concerns about the effects of batch variability on the data analyzed (26, 57). While the research community is well aware of the problem of batch-specific limitations on gene expression arrays, these caveats are not limited to gene expression analysis, and other important concerns exist when analyzing samples using microarrays or other genomic platforms (58-61).

One of the other hurdles facing the clinical genomics community is the fact that most methods for classifying, clustering or otherwise analyzing patient samples rely on standardized or relativistic values of gene expression, and usually involve z-score transformation of the expression data. Such transformations, or calculations of fold-change, directly require the analysis of multiple samples, and even during the preprocessing steps some of these challenges exist as well (62, 63). However, the prospective nature of clinical trials, where samples are collected in a rolling and time-dependant manner, prevents such standardization because they are based on calculated population summary statistics as these cannot be assessed until the trial is finished. Moreover, centroid based gene expression signatures or classification schemes developed based on standardized expression data require similar manipulations of datasets to fit a sample to a centroid (64, 65). Even the recently reported 'single-sample' framework for GSEA analysis relies on standardized expression values, and as such is

only 'single-sample' with respect to the fact that each sample is scored for enrichment, rather than looking for enrichment differences across multiple groups of samples (66). The current methodological framework for GSEA provides no means to rank genes within a single-sample, leaving the researcher to perform such ranking prior to analysis.

As mentioned above, the other critical component to consider is batch-specific variation, which represents the technical differences between samples run on the same array platform but on different dates, with different allotments of reagents, by different technicians or at different locations. While batch related variation has been extensively dealt with in the genomics community (67-71), it is a primary concern when analyzing genomic data. Numerous methods have been developed to identify and address this source of noise in genomic datasets including approaches that implement empirical Bayes, support vector machine, or factor based frameworks to identify and systematically adjust batch differences. In the case of non-array type platforms, the techniques to adjust batch effects are still nascent, though there is a clear necessity to consider this type of variation even during discretely quantitative analysis of sequencing data (26, 72). Given these concerns, prospective analysis of samples is not feasible without development of novel methods to adjust for gene expression differences between batches while also being able to robustly score individual samples that will not be collected as part of a large cohort of samples to be analyzed at the same time.

1.3 Glioblastoma

1.3.1 Clinical presentation of glioblastoma and other diffusely infiltrative gliomas

Primary brain cancers, while presenting as distinct histological diseases, contain significant molecular heterogeneity within each classification of disease, like many of the cancers described previously (73). While brain cancers are not among the most common form of cancer, nor are they the most lethal as a whole, most malignant primary brain cancers are clinically intractable and patients diagnosed with these diseases generally have extremely poor prognoses. Given the molecular heterogeneity found within this class of diseases, it is important not only to understand the presentation of this disease, but also the molecular drivers of tumor formation and clinical treatment response.

The most commonly diagnosed form of malignant brain cancer is glioblastoma, which affects between two to five out of 100,000 people annually, and has a median overall survival of 12 to 16 months after diagnosis depending on the treatment a patient receives (74, 75). The majority of these tumors are initially diagnosed as World health organization (WHO) grade IV malignant glioblastoma (76), referred to as primary or *de novo* glioblastomas, which are distinct from secondary glioblastomas that result from the progression from a lower grade glioma. These tumors are a subset of glioma, tumors that derive from or resemble glial cells, which also include astrocytomas, oligodendrogliomas, and mixed oligoastrocytomas.

Most patients are diagnosed in their fifth or sixth decade of life, and there is a slight sex bias in disease incidence, with more men diagnosed with glioblastoma than women (75). Patients are most commonly diagnosed after MRI scan following a seizure, with tumors primarily present in the temporal or frontal lobe of the brain with significant edema associated with the tumor. Grading and definitive diagnosis are dependent upon histological examination of a suspected tumor mass from a stereotactic biopsy sample, however, preliminary assessment is based on a CT or MRI scan.

Glioblastomas, along with grade II and III gliomas, are diffusely infiltrative tumors that have a central core to the tumor that is surrounded by tumor cells invading through the surrounding non-neoplastic tissue, which preferentially occurs along white matter tracks in the brain. These tumors are also highly proliferative and are characterized by the presence of a core tumor associated with an extremely high degree of hypoxic tissue, tumor necrosis, neovascularization, and psuedopalisading cells surrounding necrotic regions of the tumor.

As glioblastoma tumors are diffusely infiltrative, recurrence of tumors occurs in nearly all patients due to the inability of surgical debulking of the tumor core to remove all tumor cells. Tumors may recur directly adjacent to the previous site of the tumor, or have been known to develop in previously unaffected areas of the brain, though this is less common. In contrast to other malignant tumors, patients with malignant gliomas rarely metastasize to other organ systems.

1.3.2 Molecular characteristics of glioblastoma

In addition to the complex clinical presentation associated with glioblastomas, these tumors contain diverse molecular alterations responsible for tumor formation (77). Many alterations that occur commonly in glioblastoma have been well characterized for decades, though that knowledge has not translated into improved outcome for these patients. The most common alterations include amplification or mutation of EGFR (78, 79), deletion of PTEN (80, 81), and deletion of the locus encoding $p16^{INK4A/ARF}$ (CDKN2A) (82, 83). Less frequent alterations, such as PDGFRA amplification or mutation, TP53 mutation or deletion, and RB1 loss have also been well documented (84-87). Since these initial alterations have been discovered there has been intense focus on understanding the prevalence of these and other alterations driving tumorigenesis in this disease setting. Indeed, the cancer genome atlas (TCGA), profiled GBM tumors for their initial pilot project, detecting alterations that existed at the sequence, copy number, epigenetic, and expression levels throughout the genome. This confirmed the presence of not only these alterations but also a host of less frequent alterations, many of which had been documented as well (54).

These alterations are responsible for promoting the aggressive proliferation, invasion, and vascularization associated with this disease. The premise underlying the majority of these studies was that by better understanding which lesions associated with glioblastoma, researchers would be able to target tumors harboring those specific

alterations and devising ways to subvert the malignant processes that a tumor relies on. However, these tumors don't just depend on mutations; they are subject to the same heterotypic cellular environment described earlier. Within these tumors, it has been widely demonstrated that there is a progenitor cell niche that is responsible for tumorigenesis as well as tumor recurrence after therapeutic intervention (88-91). While there is ongoing debate about the exact identity of these stem-like cancer cells, it is clear that this is an important subset of tumor cells to focus on during research and drug development.

While it was proposed that other characteristics of tumor development did not necessarily apply to brain tumors because of immune privilege afforded by the blood brain barrier (92), this assumption has been called into question given the number of immunological therapeutic approaches that appear to be efficacious (93, 94), and the expression of immune repressive signaling cascades in these tumors (95). Similarly, metabolic deregulation in gliomas has become an increasingly active area of research after mutations in isocitrate dehydrogenase genes (IDH1/2) were identified in secondary glioblastomas and low-grade gliomas (96-98).

There has also been a growing appreciation for the impact of deregulation that occurs across multiple levels of the genome, including the epigenetic regulation of expression as well as deregulation of microRNA signaling. Studies from the cancer genome atlas have identified a strong protective affect that one class of methylation

patterns confers to a subset of gliomas (99). Moreover, numerous microRNAs have been characterized to be either induced or repressed during glioblastoma tumorigenesis, which adds another means by which normal cellular signaling can be deregulated in this setting (100-102). Many of these microRNAs have been shown to affect key oncogenic or neurological development pathways, including regulation of the Myc, TP53, or EGFR pathways as well as the REST and co-REST signaling axis (103-106).

While these alterations are present glioblastomas, there clearly exists a continuum of cooperative alterations and tumor heterogeneity reminiscent of other aggressive cancer settings between glioblastomas tumors. Even within the same tumor, studies have demonstrated that spatial heterogeneity exists for MGMT promoter methylation (107), as well as RTK amplification status (108). It has also been demonstrated recently that mutations such as EGFRvIII can actively drive intratumor heterogeneity (109). The presence of this intratumor and interpatient variability directly implies that any hope of understanding the underpinnings of this disease depends on the molecular context of which alterations are present, and which signaling cascades within a tumor are actively deregulated to promote malignancy.

1.3.3 Genomic classifications of gliomas

Many of the studies enumerated above describe various alterations or signaling networks that are disrupted in glioma, though the majority of these studies focus only

on the context of a single alteration or a small set of alterations. To move beyond this analysis, where alterations are counted and enumerated, placing these genomic alterations in the context of each other and the population of patients affected by GBM provides a richer description of the disparate events that can drive tumorigenesis. Classifying tumors based on the coexpression of alterations may therefore lead to a better understanding of how to treat these tumors, and has become significantly more feasible due to the development of genome-scale assays. However, even at a more basic level, examining genomic alterations provides more concise and comprehensive information than simple histological examination which only detects the presence of a few key proteins or markers.

Within the glioma community, there has been an intense focus on identifying molecular distinctions that exist within pathologically indistinguishable tumors. Initially, molecular profiling of a tumor was demonstrated to better reflect patient outcome than histological grading (110, 111), supporting the idea that these genome-wide studies could provide contextual understanding not previously afforded. Since then there have been numerous studies investigating the molecular underpinnings of patient outcome and development of gliomas, with many studies including both high and low grade tumors to understand the impact of gene co-expression clusters or copy number alterations (19, 73, 112-114). There is also a growing body of evidence that

suggests that *de novo* glioblastomas are etiologically distinct from low grade tumors, which appear to develop from a distinct niche within the brain (97, 115-118).

While the results from these studies were somewhat discordant, it is clear that there are certain common themes among the prognostic glioma subgroups identified. Several studies noted that tumors with proliferative or mesenchymal characteristics had poorer outcome than other subgroups. Research into the mesenchymal class of tumors demonstrated that this affect was driven by activation of Stat3 signaling (95). Conversely, several these studies identified that expression of genes involved in neurogenesis was associated with better patient outcome. Phillips *et al.* noted that in a number of cases where they arrayed patient tumors both before and after recurrence, that tumors that were initially proliferative or proneural in their expression profile often transitioned upon recurrence to a mesenchymal classification (19).

Several more recent studies have focused on understanding glioblastoma as a distinct disease setting, with research primarily driven by the poor patient outcome. To date the TCGA has profiled hundreds of glioblastomas at the expression, copy number, sequence, and methylation level (54). Initial studies from this effort identified and corroborated many previously known alterations, and emphasized the perspective that the p53, RB-E2F, and Receptor Tyrosine Kinase pathways are central to GBM signaling. Later research from the TCGA identified four subgroups of glioblastoma using an unsupervised clustering analysis of gene expression (119), and noted that these groups

similar but not analogous to previously described prognostic subgroups of gliomas (19, 112). Under this framework, the TCGA again identified a mesenchymal and proneural group, as well as a classical representation of glioblastoma and a neural group. Further analysis of these groups identified key alterations associating with each subtype, and demonstrated that the mesenchymal and classical subgroups of the disease that responded better to aggressive intervention than tumors from other subgroups. Subgroup classification schemes have also been developed using other genomic data sources, including protein or microRNA expression (100, 120), though these groupings are not completely concordant with gene expression derived classifications.

To date however, beyond the prognostic implications mentioned above, few clinically actionable insights have been gained during these studies. Recent research has refined some of these studies to identify a panel of expression biomarkers that can be used to predict clinical outcome (121, 122), but targeting therapeutic interventions to a specific subgroup has thus far not been implemented. The subgroup definitions characterized by the cancer genome atlas and by Phillips *et al.* have become the most widely adopted classifications for gliomas, and glioblastoma in particular. While they have brought about a better understanding of the coincidence and potential cooperative alterations that exist, these studies conform to the idea that this disease can and should be understood as built by discrete and distinguishable entities.

This framework creates a simplistic rationale for diagnostic purposes, but such discrete delineations are not clearly reflected across all levels of genomic alteration within the same panel of tumors studied. The fact that few of the mutations or copy number alterations identified to be enriched within a given subgroup are exclusive to that subgroup directly argues that this disease exists as a continuum of disease. Additionally, some tumors have been identified in these studies have no clear driving alteration associated with them. Therefore, even though these studies have identified much of the genomic heterogeneity present in this disease, there is a clear need to better understand the drivers of these diverse phenotypes and to improve our understanding of patient outcome.

1.3.4 Surgical and therapeutic interventions for treatment of glioblastoma

While genomics has yet to directly inform the therapeutic intervention used in this disease setting, numerous advances have paved the way to begin approaching GBM treatment in a more personalized manner. The challenges to treating patients diagnosed with GBM are not exclusive to this disease setting and similar to many late stage malignancies glioblastoma tumors are both extremely aggressive as well as refractory to most interventions, with nearly all tumors recurring after surgical and therapeutic treatment. Despite the advances made in understanding the molecular biology of this disease, the most effective treatment methods developed to date involve surgical

removal of the tumor and radiation therapy. Still, even drastic resection techniques, such as removal of an entire hemisphere of the cerebellum harboring the tumor, have proven unable to fully cure patients diagnosed with this disease (117). However, despite nearly universal recurrence, the extent of resection has been demonstrated to be a clear determinant of patient outcome and overall survival, even if it does not offer a cure (28, 123-128).

Even when surgical resection is not possible nearly all patients receive radiation therapy, as early studies demonstrated that administration of 50-60 Gy of whole-brain radiation provided a clear benefit to patient survival (129, 130). While neurological deficits can results from radiation, radiotherapy is an important arm in treatment of gliomas, and is often employed when the tumor is deemed inoperable. However, the stem-like cell niche of glioblastomas has been demonstrated to be largely radiation resistant (89, 131, 132). Thus, like surgery, radiation therapy has extended patient overall and disease-free survival; however, it does not prevent recurrence nor cure patients of this disease.

Until recently chemotherapeutic agents were used to little effect. The most commonly used therapies include cytotoxic and DNA damaging agents, such as carmustin (BCNU), PVC (procarbazine, vincristine, and lomustine (CCNU)), and temozolomide (133, 134). Temozolomide has been the most successful general treatment of patients with malignant gliomas, and patients treated with concurrent radiation and

temozolomide had an increased median overall survival time of approximately four to six months compared to patients that received radiation alone (135), and has since become the standard of care for this disease setting. Accurate measurements of progression free survival have been harder to quantify, as it has been demonstrated that a significant proportion of patients undergo psuedoprogression while receiving this treatment (136, 137). Similar to radiotherapy, several studies have proposed that stem cell niche of glioblastomas is largely unaffected by temozolomide treatment (88, 91, 138), and that poor patient outcome was related to a 'self-renewal' or stem-like phenotype (139).

Temozolomide is converted to its active form active form, 3-methyl-(triazen-1-yl)imidazole-4-carboxamide (MTIC), during systemic circulation at physiological pH and acts by methylating guanine residues at O-6 or alkyating N-7. Cytotoxicity to tumor cells is induced by the inability to repair this damage. The methylated guanine adduct is the substrate for the DNA repair enzyme Methylguanine methyltransferase (MGMT), while the alkyated adduct is repaired by O-6-alkylguanine-DNA alkyltransferase (AGT or AGAT). Guanine crosslinkages, which can be induced by BCNU, are also repaired by MGMT. Methylation of the MGMT promoter was associated with increased patient response to BCNU or temozolomide with or without radiation (140-143). Subsequent studies have demonstrated that MGMT promoter methylation is a significant prognostic

indicator that should be considered during the treatment of glioblastoma (28, 92, 127, 144-148).

Another significant development in the treatment of glioblastoma has been the use of the VEGFR-targeting monoclonal antibody bevacizumab (avastin) in recurrent glioblastomas (149, 150). Similar to treatment with temozolomide, there are concerns about the ability to track tumor recurrence during avastin treatment due to the phenomenon of psuedoresponse, which most likely reflects reduced edema surrounding the tumor and normalization of the blood brain barrier (151). This has, however, been posited to be one of the mechanisms for avastin's efficacy, as reducing often improves quality of life for patients, though recurrences for patients treated with avastin are frequently characterized by extremely diffuse tumors (152). Concurrent avastin and temozolomide treatment at initial diagnosis was observed to increase progression free survival, but not overall survival (144).

Targeted therapeutic intervention has also been an area of keen interest, given the prevalence of specific alterations in glioblastoma and the successes seen in other cancers. EGFR-specific small molecule inhibitors such as gefinitib or erlotinib that bind to the kinase domain of EGFR have had mixed results in clinical trials (153-157), and other studies have demonstrated that glioma-specific mutations in the extracellular domain or concurrent PTEN mutation affects the efficacy of these inhibitors (158, 159)

Similarly, initial clinical trials with lapatinib and cetuximab, which target the extracellular domain, have not had impressive results to date (160, 161).

Targeting the appropriate subset of glioblastoma patients is a key area of concern when using any of these agents, as demonstrated by the effect of MGMT. Selectively targeting an appropriate subset of patients has also shown success with immunological approaches to target the EGFRvIII mutation (93). Other targeted therapeutics or immunological approaches are actively being tested in glioblastoma (134, 162), though EGFR remains one of the most attractive broad-spectrum target owing to the frequency at which alterations are found in this disease. However, identifying additional molecular determinants governing which patients respond the commonly used cytotoxic agents or new targets for small-molecule based therapeutics is a key strategy to improving the outcome of patients diagnosed with malignant gliomas.

1.4 Research rationale and summary

The ability of genomic technologies to expand our understanding of complex disease states is not in question. However, there are clear areas where our current understanding of malignant gliomas can and should be improved. The underlying mutations and other alterations that have been described in this cancer have primarily been identified based on the frequency, predicted functionality, and amplitude of an alteration. While these studies have lead to understanding of this disease at a fairly

atomized level, too little emphasis has been placed on the biological context in which an alteration occurs. Though studies have focused on alterations occurring in discrete subsets of these tumors, even this description appears incomplete, as some subtypes have no clear driving alterations. As the genomic complexity of cancer becomes more thoroughly characterized, it is imperative that we do not ignore the potential role that less frequent or less obvious alterations play during tumor development and malignant progression. The research described in chapter two documents our approach to annotating genomic alterations as they co-occur with signaling deregulation in malignant gliomas. We note that this method not only identifies many well characterized genes that have been commonly described in this setting, such as PTEN, EGFR and RB1, but also identifies recently characterized alterations that have not been implicated in glioma development before. Moreover, we expand this research to investigate the clinical impact these alterations have, identifying an ubiquitin-like protein (UBL3) on chromosome 13 that associates with patient outcome and xenograft response to standard of care intervention.

Then, research presented in chapter three focuses on the clear gaps in our ability to move the genomic techniques we utilize from the discovery phase, where samples can be analyzed retrospectively without consequence, to a translational setting where timely analysis of genomic data is essential. We address the concern that pathway signatures, while useful for annotating and studying the disease process, are incompatible with the

prospective clinical setting by developing a framework to implement accurate and rapid analysis of genomic signatures based on data obtained from a single patient. To overcome batch and standardization related issues, we propose the use of a predefined reference dataset, and demonstrate that such datasets provide the necessary backbone to predict an individual sample with a high degree of fidelity.

2. A Pathway Association Framework for Identification of Significant Genome Alterations in Glioblastoma

2.1 Summary

The complexity of the genomic alterations present in most cancers, including glioblastoma, challenges the ability to identify events that drive the oncogenic process and disease phenotype. While numerous studies have catalogued the genetic complexity of this disease, alterations are identified primarily based on their frequency, rather than their functional impact. To better understand driving alterations in glioblastoma, we implemented an association study to identify loci linked with oncogenic phenotypes captured using a series of gene expression signatures reflecting key signaling pathways. We found that several of the pathways analyzed were differentially enriched across the subgroups of glioblastoma, were associated of with the patient survival, and were linked to mutations in known pathway regulators. We extended this framework to identify driving copy number alterations of in the cancer genome atlas (n = 305). We validated these associations at the expression level in the cancer genome atlas, as well as in samples from previously published glioblastoma studies (n = 480) and a panel of *in vitro* and *in vivo* models (n = 49). Among the candidate alterations identified was chromosome 13q12.3, which was inversely associated with IFNγ signaling ($P = 2.64 \times 10^{-6}$). We demonstrated that loss of UBL3, specifically, was associated with poor patient outcome (P = 0.002), which we validated in a panel of xenograft lines. In these lines, low

expression of UBL3 corresponded to poor response to treatment with either temozolomide ($P = 0.015$) or avastin ($P = 0.034$).

2.2 Background

Glioblastoma (GBM), like most human cancers, is characterized by substantial heterogeneity, both with respect to clinical features as well as molecular characteristics. This heterogeneity confounds not only the understanding of the molecular mechanisms giving rise to GBM, but also impedes the development of therapeutic strategies that match the characteristics of the individual patient's tumor. The power afforded by genome-scale analyses of gene expression, DNA sequence alterations, and chromosomal alterations has begun to provide a more detailed understanding of the complexity of the disease. Analysis of global gene expression using DNA microarrays has demonstrated the ability of genomic scale investigations to differentiate glioma grade (111), oligodendroglial and astrocytic tumors (110), and *de novo* and progressive GBMs (73, 116). Expression analyses have also shown the ability to delineate subsets of patients with different prognoses and identify genes distinguishing patients with a good prognosis from those with a poor prognosis (19, 112, 122, 163-165). Similarly, studies of chromosomal alterations, mutations, protein expression, and microRNA expression have shown the capacity to capture key information regarding tumor heterogeneity and alterations driving gliomagenesis (97, 100, 102, 120, 166-168).

A key challenge posed by the enormous complexity of data now being developed through these genomic studies is the ability to distinguish which low frequency events are functionally important and which represent noise. Recent studies within the Cancer Genome Atlas (TCGA) and other large-scale projects have used copy number alterations to identify GBM subtype specific alterations (166), epigenetic expression regulation across genome and cooperative alterations that appear to be maintained together (54, 169, 170). Further, studies of high-grade gliomas spontaneously induced in mice have recently suggested that initial alterations in the TP53, PTEN, and RB1 genes can generate the heterogeneity of copy number alterations, as well as patterns of genome-scale gene expression, found in the human disease state (171). This may provide an important experimental model, but such studies have not addressed what function the range of subsequent alterations play in maintaining the tumor phenotype or how they promote the range of phenotypes observed.

To address this issue, and identify potential drivers of GBM tumorigenesis, we have sought to understand the copy number landscape in GBM not only as a function of the rate at which a particular alteration occurs, but also by analyzing whether an alteration co-occurs with the activation of key signaling pathways that are deregulated in most cancers. For this analysis, pathway activity was quantified using a suite of gene expression-based genomic signatures derived from *in vitro* experiments that have been shown to predict patient outcome and sensitivity to therapeutics targeting these

pathways (16, 45). Our analytical framework for integrating genomic signatures with copy number data draws from previous work in breast cancer where a genomic signature lead to the identification of CSN5 as a modulator of the wound-healing phenotype through coordination with Myc (172). The work we describe here builds from that study, employing the concept of genetic association analysis as the framework for discovery of genome alterations driving disease states. The results from our study demonstrate the ability to use of expression signatures of oncogenic pathway activation as relevant quantifiable phenotypes for an association framework, and the capacity for this approach to identify novel alterations in the complex disease setting of glioblastoma.

2.3 Results

2.3.1 Expression signatures reflect the distinctions among GBM subtypes

Genetic association studies seek to identify genetic variants that exhibit a statistically significant linkage with a given phenotype across a population of individuals. In the context of human cancers, genetic variants can be represented by germline polymorphisms as well as somatic gene mutations or chromosomal alterations, either of which can result in a gain or loss of function. Similarly, phenotypes can be represented not only as observable traits but also any quantifiable measure of relevance to the disease state.

Many recent studies have emphasized the view that the consequence of gene mutation and genome alteration can be best understood as the deregulation of the cell signaling pathways in which these genes function (54, 173). Recognizing the opportunity to use pathway activation as a relevant cancer phenotype, together with the wealth of integrated genomic data in GBM, we have carried out an analysis aimed at identifying genomic loci that, when altered, coincide with deregulation of relevant oncogenic signaling pathways (Figure 2). Our approach to generating measures of pathway activity makes use of expression signatures derived from defined experimental perturbation of a given signaling activity. These signatures offer the advantage of providing a measure of pathway activity irrespective of how the pathway may have been activated (16, 45, 174). The signatures can then be projected into additional samples, including tumors, resulting in the probability of pathway activation for those samples and can serve as a quantitative trait across a population of tumors (45).

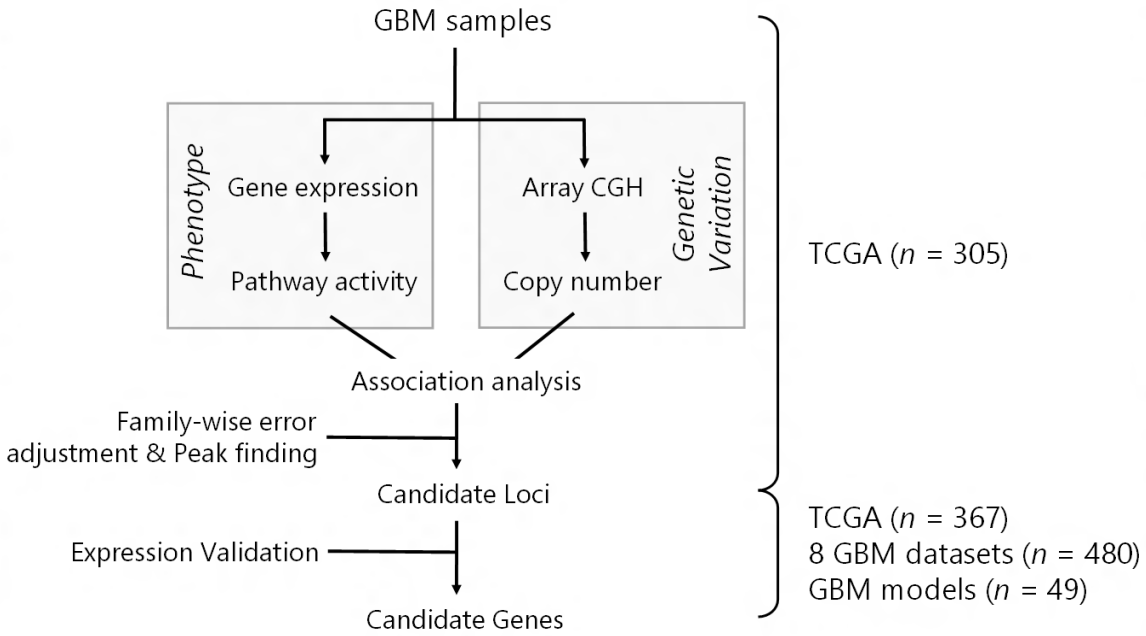

Figure 2: Utilizing pathway signatures for a genetic association study.

Scheme for identification of novel candidate genes associated with pathway signaling utilizing integrated genomic data from the TCGA glioblastoma project to link pathway activity with copy number alterations. Family-wise error adjustment and peak finding are implemented to identify candidate loci for association and validation of expression level associations across multiple GBM datasets is used to identify candidate genes.

As a starting point in the analysis, we recapitulated the gene expression-derived GBM subtypes described by the TCGA as a framework to evaluate the pathway signatures (Figure 3A) (119). We then used the signatures to predict the deregulation of pathway activity associated with these four subtypes. As seen in Figure 3B, distinct patterns of pathway activity were evident within these subtypes ($P < 0.01$, Kruskal-Wallis), which reflected previous characterizations of each group in many cases. Examples include activation of β-Catenin, E2F1 and Myc within the Proneural subtype, consistent with the proliferative and stem cell-like characteristics of this subtype. Similarly many mitogenic and inflammatory response pathways, such as Ras, Stat3, interferon (IFN) and receptor tyrosine kinase (RTK) signaling, were highly active in the Mesenchymal subtype. Classical samples showed activation of the HER2 pathway and active P53 signaling, while Neural tumors were characterized by elevated Akt and Src signaling.

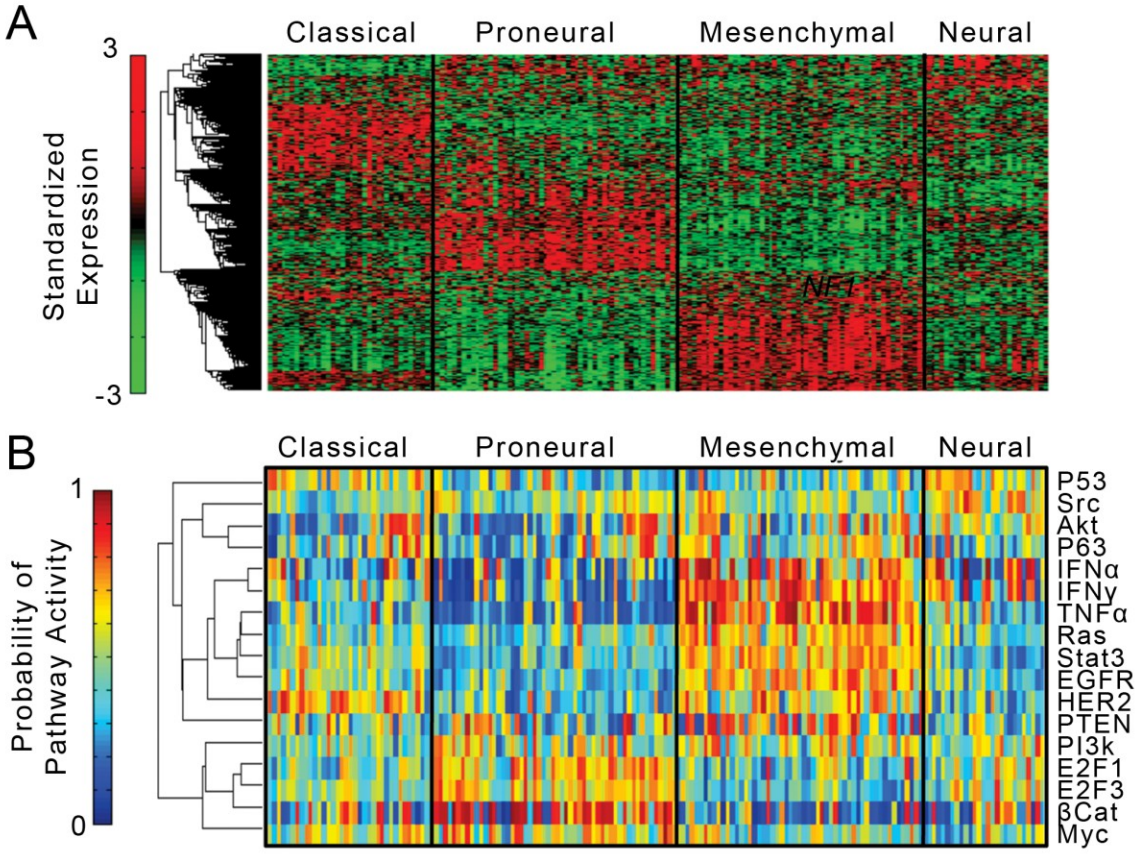

Figure 3: Pathway activation in GBM samples.

A) Unsupervised hierarchical clustering (USHC) of standardized gene expression characterizing the core TCGA subtypes (n = 176). High gene expression is shown in red and low gene expression is shown in green. B) Unidirectional unsupervised hierarchical clustering of core TCGA GBM samples across cell signaling pathways. High pathway activity is shown in red and low pathway activity is shown in blue. Patient samples ordered as in A.

To further determine whether these signatures were an appropriate phenotypic marker for pathway signaling in GBM, we examined the impact of limiting the binary regression to a predefined set of GBM-related expression probesets (Figure 4). We first filtered the Affymetrix probesets for each training file to only the genes that were previously defined as consistently and highly differentially expressed in GBM tumors by the TCGA (119). We then rebuilt the pathway signatures on this limited probeset and predicted the pathway activity of the GBM tumors with these re-derived signatures. Under this analysis, we found that in general many of the signatures had a strong degree of correlation when generated on either the whole expression dataset or the GBM-related expression subset of probes, though the PTEN and p63 predictions were poorly correlated between the original and filtered signatures ($R < 0.5$, Pearson correlation, Figure 5). This suggests that most of these signatures, while derived in and for a heterologous background, can still be used in the setting of GBM to obtain accurate predictions, though results from the PTEN and p63 signature may be less reliable than some of the other signatures.

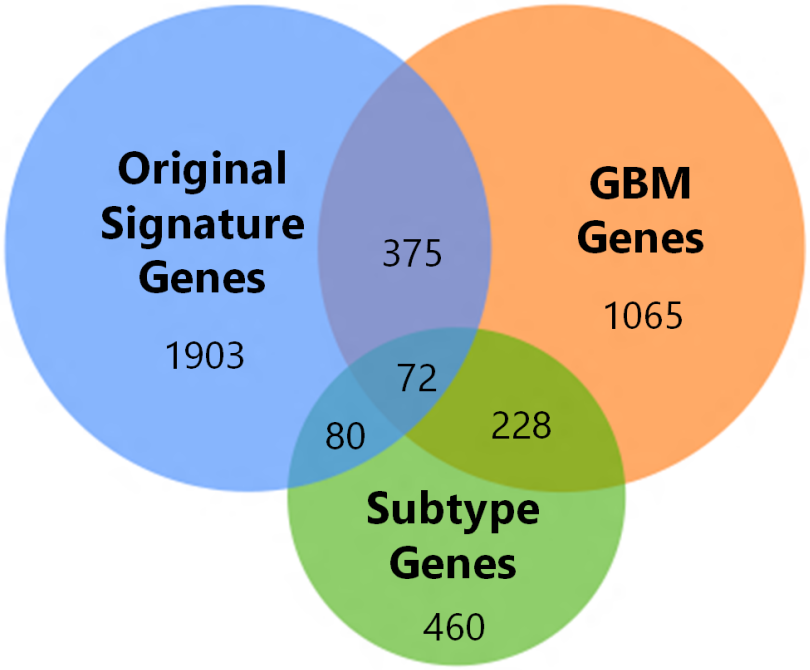

Figure 4: Overlap between pathway signature, GBM subtype and highly differentially expressed GBM genes

Venn diagram depicting genes identified either during pathway signature development (blue), or by Verhaak *et al.* as forming the basis for the GBM subtypes (green) or genes which were identified as being consistently and highly differentially expressed across 3 array platforms by Verhaak *et al.* (orange). The highly differentially expressed genes (orange) were the only genes upon which the filtered pathway signatures in Figure 5 were trained.

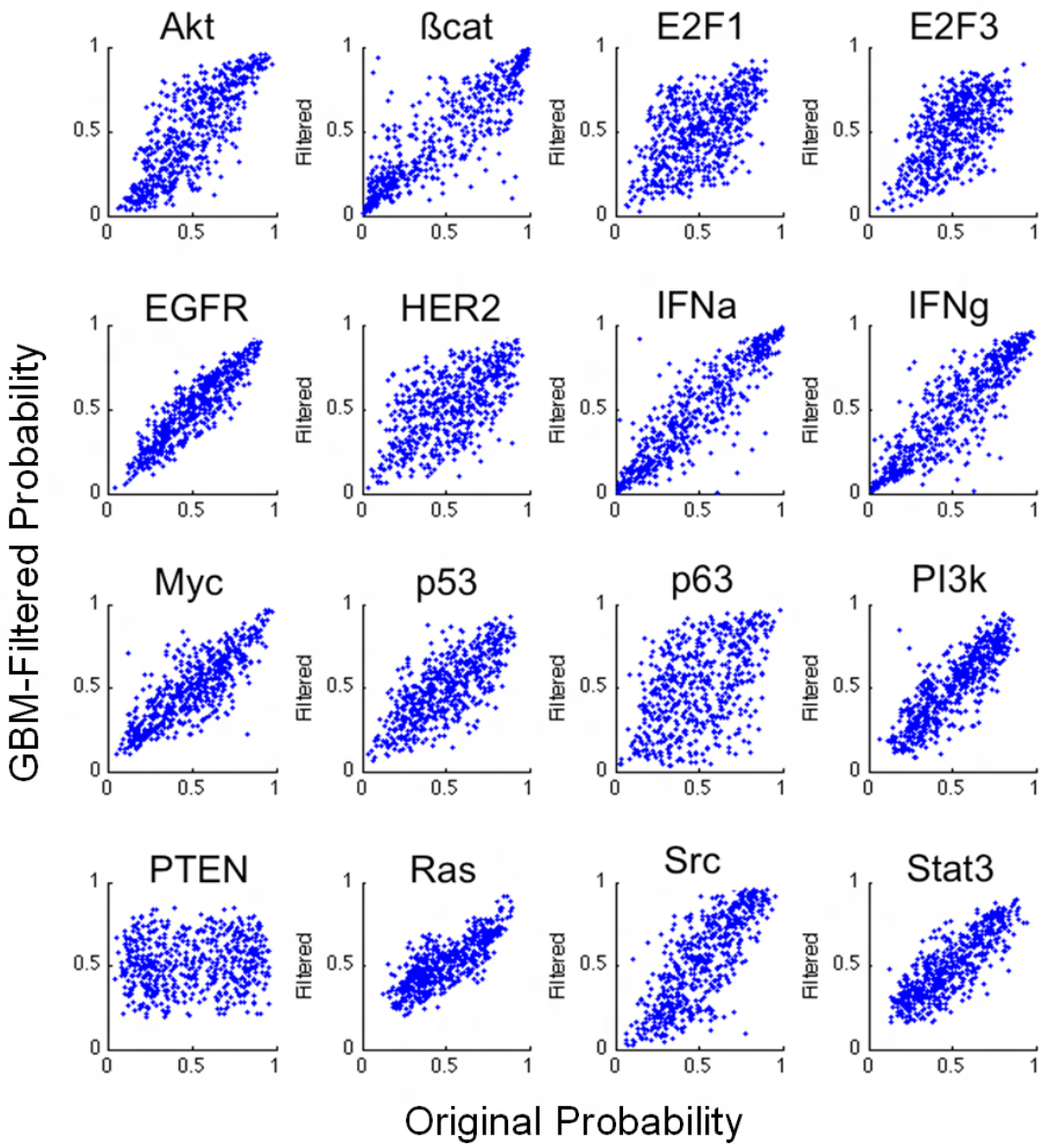

Figure 5: Robustness of pathway signatures when limited to GBM-related genes

Comparison of the predictions from the original pathway signatures and predictions generated using pathway signatures that were built using only the previously identified GBM-related genes. Pathway signatures were run on TCGA GBM samples (n = 356).

2.3.2 Integrating pathway data with genomic alterations.

To further explore the use of these signatures in an association framework, we tested the ability to link oncogenic mutations with activation or repression of relevant pathways (Figure 6). Mutations in TP53 were associated with low predicted p53 pathway activity ($P = 0.027$, Wilcoxon Rank-Sum), consistent with TP53 mutations being largely dominant negative. Likewise, mutations in negative regulators of pathway activity that include NF1 ($P = 0.001$, Ras pathway), RB1 ($P = 0.009$, E2F1), and PTEN ($P = 0.004$, PI3K) were each enriched in samples with high pathway activity. While EGFR activity only approached significance when comparing EGFR mutant and wild-type samples ($P = 0.06$), high activity trended towards enrichment in the tumor samples carrying a mutation in EGFR.

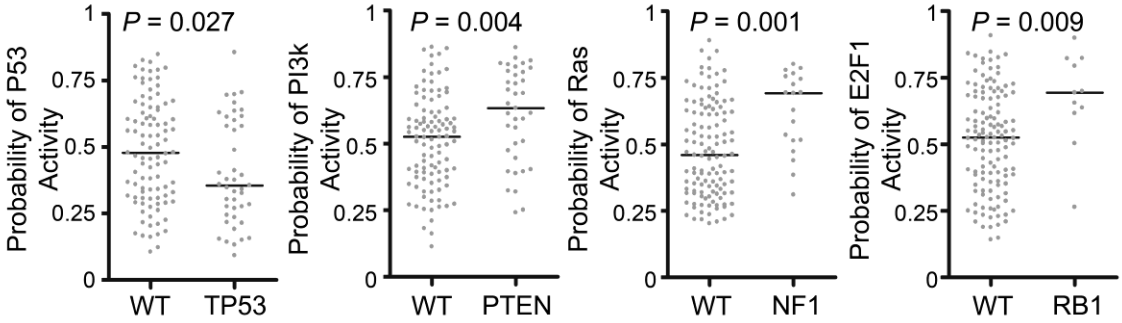

Figure 6: Linking driver mutations with pathway activation.

Comparison of pathway activity of patients with or without validated non-synonymous coding mutations in key pathway regulators. Wilcoxon Rank-Sum *P*-values shown. Bars indicate median pathway activity.

Given the ability to link pathway activity with mutations in known driver genes, we extended this analysis to compare each of the pathways to the landscape of copy number alterations across TCGA GBM samples (n = 305, Figure 7A). The association analysis was directional, providing the ability to differentiate between putative activators and repressors of cell signaling. Activators were expected to have significant direct associations with pathway activation, while negative regulators would exhibit an inverse association with pathway activation. Many of the associations reached statistical significance after accounting for family-wise error when testing across all genomic loci ($P < 4.48 \times 10^{-6}$, Spearman correlation) (Figure 7B). We then identified global and local minima of significance, identifying 74 1.5 Mbp regions that approached or reached significance on 13 of the autosomes (Table 1, Figure 8).

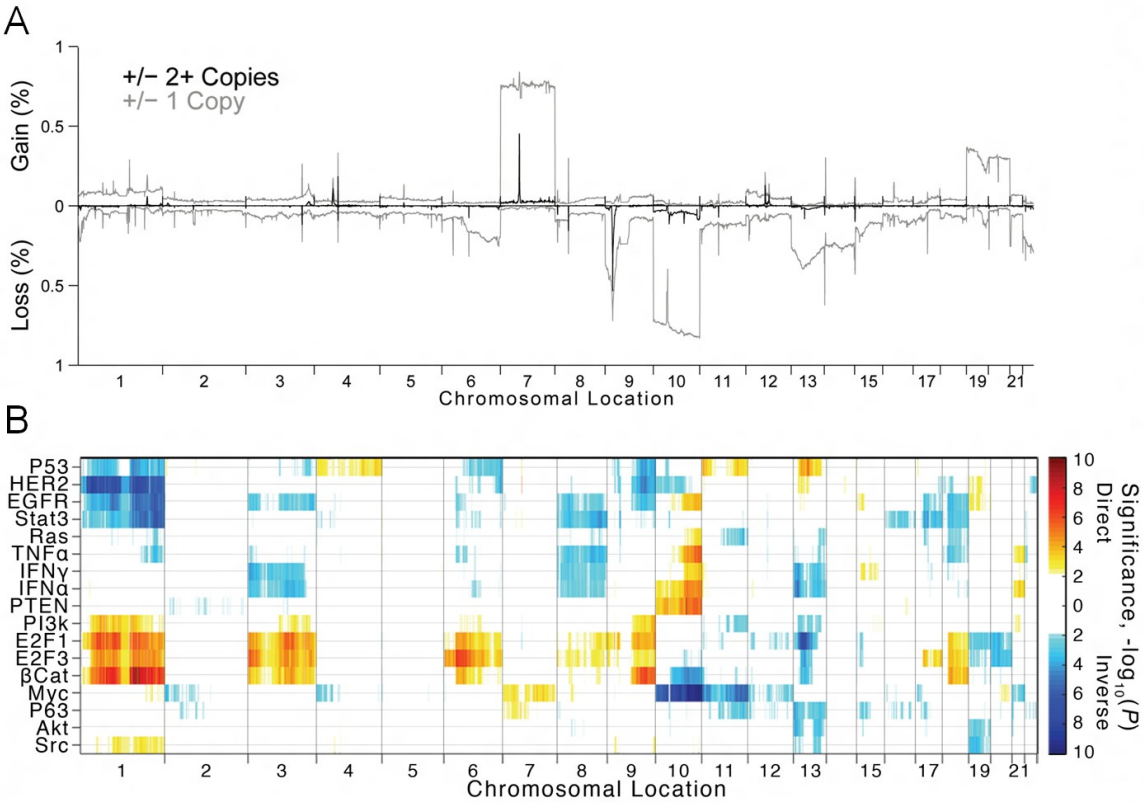

Figure 7: Association between copy number alterations and pathway activity.

A) Landscape of copy number alterations present in GBM tumors. The proportion of patients altered at a particular locus is depicted for (top) gains or (losses) and the amplitude of the alteration. B) Spearman correlation *P*-values for association between copy number alterations and individual pathways across the autosomes. Direct correlations are shown in red, and inverse correlations are shown in blue.

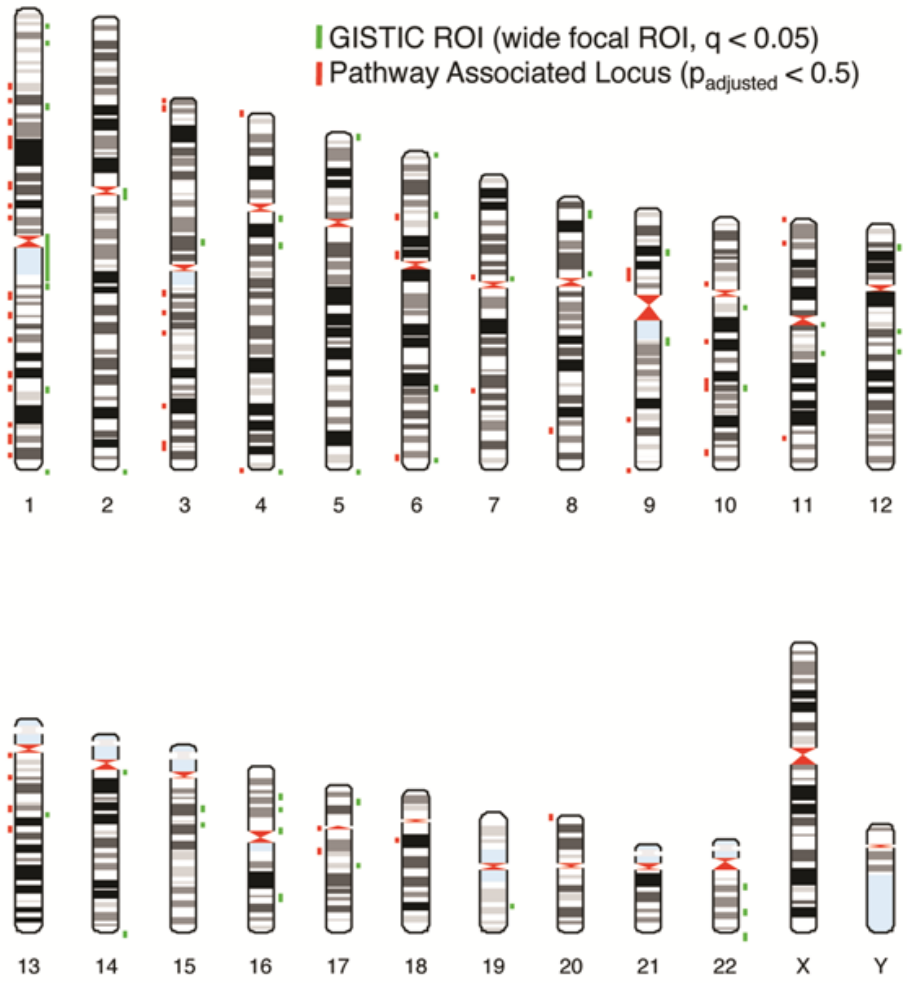

Figure 8: Location of GISTIC and Pathway Associated Regions of Interest

Chromosome ideogram displaying the location of significant focal regions of interests (ROI) identified by GISTIC ($q < 0.05$, green, right side) or pathway associated loci ($P_{adjusted} < 0.5$, red, left side) for all chromosomes.

We also examined the extent to which pathway association analysis provided an opportunity for discovery beyond the commonly used GISTIC algorithm (Figure 8), which identifies genomic regions that are altered in a population of samples more frequently and with a higher amplitude than expected by chance (175, 176). To assess the differences in the results from these two methods, we calculated the distance between each pathway associated locus and the nearest significant GISTIC focal region of interest (ROI, $q < 0.05$). While there was a subset of pathway-associated loci that coincided with GISTIC ROIs, such as the loci containing known alterations such as EGFR, PTEN and PARK2, we found that more than half of all pathway associated loci were located more than 10 megabase pairs (Mbp) from the nearest GISTIC ROI (Figure 9). These distinctions suggest that the pathway association analysis, in principle, can identify novel alterations that play a role in tumorigenesis that would otherwise remain obscured because they are not the most frequent or have the highest amplitude compared to neighboring loci.

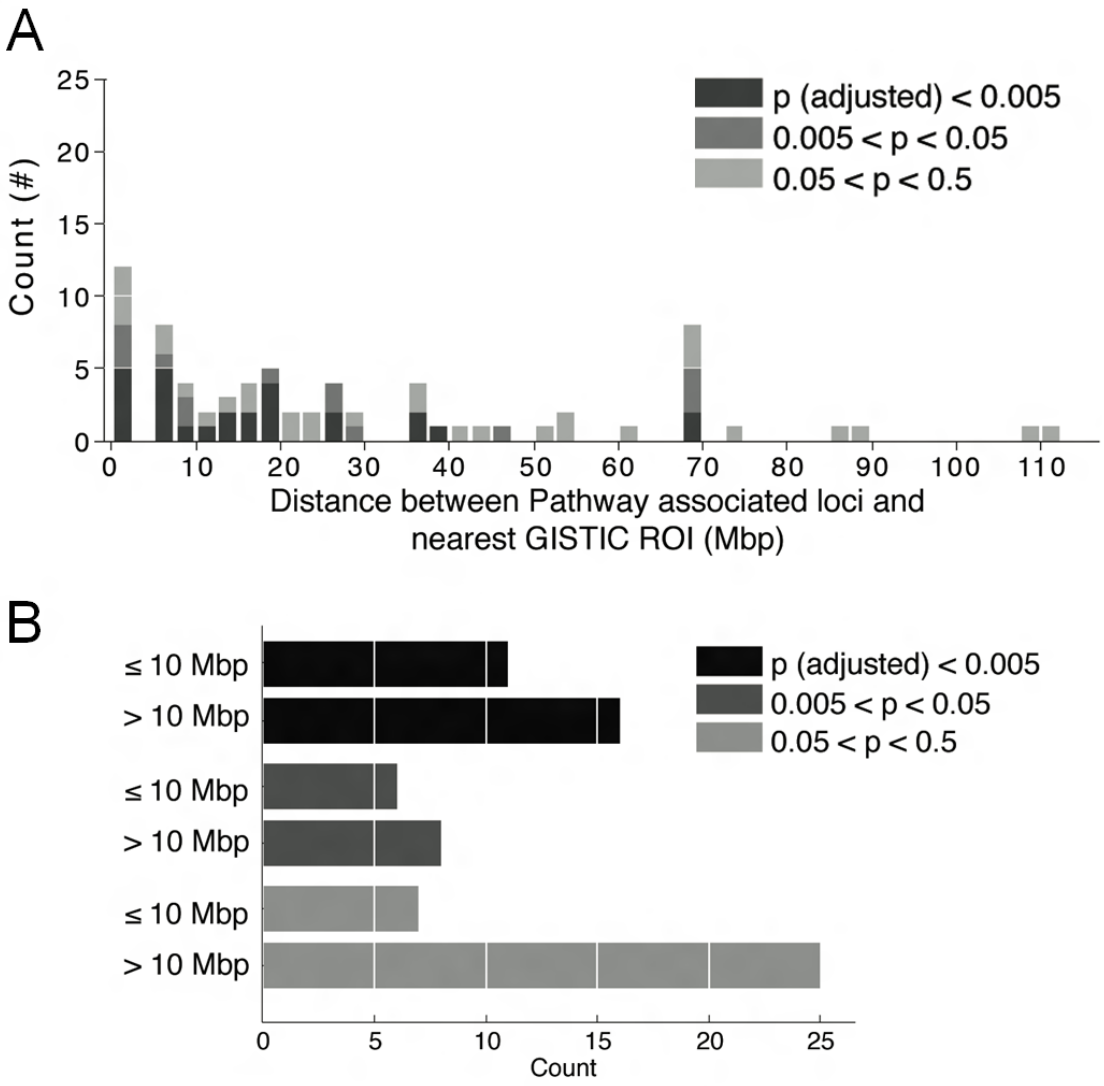

Figure 9: Distance between GISTIC ROIs and pathway associated loci

Determination of the distance between a particular pathway associated locus and the nearest GISTIC ROI. Pathway associated loci which are on chromosomes with no GISTIC ROI withheld from this analysis, and the significance at which a pathway is associated with a given locus is depicted by the color of the bars shown. A) Histogram of the distance between a given pathway associated locus and the nearest GISTIC ROI, binned every 2.5 Mbps. B) Histogram depicting whether the nearest GISTIC ROI is closer or further than 10 Mbp from a given pathway associated locus

2.3.3 Pathway association analysis identifies known and putative drivers in GBM.

In addition to regulators of pathway signaling in GBM, we anticipated that some of the regions associated with pathway activity would include genes that were present within the pathway signatures, as these are the most likely candidate drivers of the pathway associations. This was the case for PTEN ($P = 2.07 \times 10^{-11}$, PTEN signature), EGFR ($P = 1.92 \times 10^{-6}$, HER2) and a suite of IFN response genes neighboring PTEN, IFIT1/2/3 ($P = 1.23 \times 10^{-6}$, IFNα). In contrast to these loci, most of the regions identified did not contain genes that were also present in the expression signature. To identify which genes within a candidate region were driving the association, we analyzed the expression of these candidate genes anticipating that a copy number change relevant for control of a pathway would coincide with a change in expression of candidate drivers (Appendix A). In addition to analyzing expression from TCGA samples ($n = 367$), we also ran an analysis of an additional dataset consisting of GBM samples ($n = 480$) aggregated from 6 previously published studies (19, 112, 114, 139, 177, 178), confirming association at the gene expression level after accounting for family-wise error due to multiple testing ($P < 2.85 \times 10^{-5}$, Spearman correlation). We further validated nominal expression-level association of candidate genes in a panel of GBM xenograft and cell culture lines that we arrayed in replicate ($n = 49$). These results are summarized in Table 1.

Table 1: Copy number alterations significantly associated with pathway activity in all GBM expression datasets

			Copy Number Association		Expression Association		
Locus	Pathway	Direction	*P*	Gene Symbol	*P*, TCGA	*P*, Validation	*P*, GBM Model
1, Mbp 154.6	BCAT	Direct	1.58E-10	ILF2	*P* < 1E-20	2.16E-85	3.68E-07
				CKS1B	*P* < 1E-20	2.52E-84	9.22E-04
				C1orf77	*P* < 1E-20	6.34E-35	3.94E-02
				FDPS	*P* < 1E-20	3.64E-32	4.79E-02
				UBAP2L	3.84E-11	4.51E-33	1.39E-04
1, Mbp 196.1	BCAT	Direct	1.10E-08	ASPM	*P* < 1E-20	6.98E-128	3.27E-04
1, Mbp 69.6	BCAT	Direct	8.76E-08	DEPDC1	*P* < 1E-20	3.36E-62	3.01E-04
				LRRC40	*P* < 1E-20	3.59E-28	1.51E-03
1, Mbp 95.1	BCAT	Direct	1.91E-08	DNTTIP2	4.57E-09	1.12E-23	7.07E-03
3, Mbp 126.0	BCAT	Direct	8.40E-06	MCM2	*P* < 1E-20	1.48E-129	1.98E-05
				SNX4	2.96E-16	6.70E-32	4.22E-03
9, Mbp 140.2	BCAT	Direct	1.10E-07	TUBB2C	*P* < 1E-20	3.16E-15	3.71E-03
				NOTCH1	1.29E-07	6.02E-23	2.81E-02
1, Mbp 154.6	E2F1	Direct	1.97E-07	ILF2	3.15E-17	4.09E-33	2.86E-04
				CKS1B	8.52E-15	1.45E-34	2.91E-02
1, Mbp 94.1	E2F1	Direct	2.43E-07	MTF2	8.01E-13	3.00E-18	5.68E-06
6, Mbp 56.6	E2F1	Direct	6.53E-06	PRIM2	1.65E-18	1.03E-22	2.87E-03
9, Mbp 140.2	E2F1	Direct	5.26E-09	TUBB2C	5.82E-15	6.39E-19	1.64E-03
				NOTCH1	8.43E-06	2.93E-13	1.01E-02
9, Mbp 33.7	E2F1	Direct	2.30E-05	FANCG	7.39E-19	1.77E-31	6.65E-03
				DCTN3	4.71E-11	3.20E-11	1.79E-02
1, Mbp 154.6	E2F3	Direct	3.76E-07	C1orf77	3.85E-13	2.29E-18	2.24E-03
				ILF2	4.34E-11	3.30E-20	9.37E-05
				UBAP2L	2.65E-07	1.64E-06	4.20E-03
1, Mbp 202.6	E2F3	Direct	3.60E-06	IPO9	1.68E-12	1.22E-12	8.44E-04
				RABIF	9.61E-12	1.25E-10	3.32E-03
				TMEM183A	1.85E-06	4.16E-09	4.97E-04
				KDM5B	5.44E-06	2.26E-05	4.28E-02
3, Mbp 185.5	E2F3	Direct	8.04E-06	RFC4	1.45E-18	4.75E-34	1.62E-03
				TRA2B	3.79E-13	1.67E-18	4.05E-03
				TBCCD1	3.45E-08	4.35E-07	2.27E-02
6, Mbp 35.1	E2F3	Direct	3.21E-08	FANCE	1.87E-14	6.28E-14	1.01E-02
9, Mbp 140.2	E2F3	Direct	1.16E-06	TUBB2C	2.56E-18	1.57E-26	1.02E-02
				NOTCH1	1.64E-12	1.11E-13	1.42E-03
10, Mbp 125.6	EGFR	Direct	3.46E-05	CHST15	1.28E-11	3.88E-25	2.32E-03
1, Mbp 163.6	EGFR	Inverse	2.12E-07	PBX1	9.34E-16	3.17E-07	1.19E-03
3, Mbp 187.5	EGFR	Inverse	1.21E-05	MASP1	2.25E-09	1.10E-09	1.45E-02
9, Mbp 140.2	EGFR	Inverse	3.86E-05	TUBB2C	1.51E-16	3.49E-17	1.16E-02
				NELF	4.65E-06	9.52E-13	7.14E-03

Table 1 (continued)

Locus	Pathway	Direction	Copy Number Association *P*	Gene Symbol	Expression Association *P*, TCGA	*P*, Validation	*P*, GBM Models
1, Mbp 164.6	HER2	Inverse	6.71E-09	MGST3	3.61E-06	4.34E-07	4.24E-02
1, Mbp 41.6	HER2	Inverse	2.04E-09	CTPS	3.21E-14	3.54E-06	2.85E-05
10, Mbp 91.6	IFNA	Direct	1.23E-06	IFIT3	2.34E-38	4.68E-53	2.09E-07
				IFIT2	2.86E-24	9.10E-20	2.66E-04
				IFIT1	2.93E-24	6.08E-26	6.40E-07
				FAS	4.15E-08	4.21E-06	7.16E-03
13, Mbp 31.1	IFNA	Inverse	2.96E-07	HMGB1	7.64E-18	5.75E-13	2.67E-02
				UBL3	8.98E-06	1.95E-09	7.71E-03
13, Mbp 31.1	IFNG	Inverse	2.64E-06	HMGB1	2.45E-20	2.52E-16	5.24E-03
				UBL3	5.32E-09	6.74E-18	4.77E-03
10, Mbp 126.6	MYC	Inverse	1.81E-11	LHPP	5.73E-12	9.37E-30	7.06E-04
11, Mbp 117.2	MYC	Inverse	7.70E-07	SIDT2	2.95E-10	3.47E-15	4.37E-03
				BACE1	7.54E-06	6.15E-07	2.02E-03
11, Mbp 12.7	MYC	Inverse	7.65E-06	DKK3	$P < 1E-20$	5.09E-25	8.10E-04
11, Mbp 0.2	P53	Direct	6.75E-06	SIRT3	$P < 1E-20$	1.06E-36	8.26E-08
				TMEM80	2.14E-09	4.70E-06	2.08E-03
				PTDSS2	1.03E-08	1.64E-09	3.61E-04
1, Mbp 153.1	P53	Inverse	2.17E-06	ILF2	$P < 1E-20$	1.75E-29	7.41E-03
				UBAP2L	1.80E-07	2.32E-14	5.72E-04
1, Mbp 112.1	PI3K	Direct	3.11E-05	WDR77	9.19E-06	3.83E-19	8.81E-04
1, Mbp 153.6	PI3K	Direct	4.08E-05	CKS1B	$P < 1E-20$	4.98E-27	1.28E-03
				ILF2	5.00E-11	5.44E-36	5.48E-03
				C1orf77	7.58E-08	2.05E-15	2.38E-02
10, Mbp 89.6	PTEN	Direct	2.07E-11	PTEN	9.64E-130	1.09E-201	$P < 0.01$
1, Mbp 177.1	STAT3	Inverse	5.02E-07	FAM5B	7.99E-07	9.21E-07	6.65E-05
17, Mbp 35.5	STAT3	Inverse	1.60E-05	PCGF2	$P < 1E-20$	1.95E-09	2.20E-02
8, Mbp 125.1	STAT3	Inverse	9.95E-06	MTSS1	1.60E-11	1.20E-14	6.93E-05
				SQLE	1.48E-08	4.79E-09	4.65E-03
9, Mbp 37.2	STAT3	Inverse	2.61E-05	ZBTB5	$P < 1E-20$	6.97E-08	2.17E-03
				RNF38	5.01E-06	4.26E-09	8.79E-04
10, Mbp 125.6	TNFA	Direct	2.57E-07	CHST15	1.47E-06	6.09E-16	8.42E-03

To use one analysis as an example, we focused on candidate genes associated with the E2F1 pathway (Figure 10). By combining copy number alterations with expression change to identify candidate genes driving the association within these regions, we positively identified NOTCH1 as a regulator of E2F1 ($P = 5.26x10^{-9}$), as well as E2F3 ($P = 1.16x10^{-6}$) and ß-catenin ($P = 1.10x10^{-7}$), which has been shown to occur through LEF1 (179). This is also consistent with recent proteomic analysis of GBM that demonstrated that the cleaved form of NOTCH1 associated with phosphorylated RB, indicative of active E2F signaling (120). Examples of previously uncharacterized GBM pathway regulators include a suite of genes that putatively regulate the activity of the E2F pathways. Two of these genes, CKS1B ($P = 1.97x10^{-7}$, E2F1 pathway) and KDM5B ($P = 3.60x10^{-6}$, E2F3), are known to act as upstream regulators of the E2F pathway (180, 181). In addition to KDM5B, MTF2 ($P = 2.43x10^{-7}$, E2F1) was also identified, and both are involved in embryonic stem cell pluripotency and epigenetic regulation (182, 183), perhaps reflective of the proneural association with E2F1. We also identified a Fanconi Anemia gene, FANCE, which was associated with E2F3 activation ($P = 3.21x10^{-8}$). Finally, we found that ASPM was significantly associated with β-catenin signaling ($P = 1.10x10^{-8}$), consistent with the reported role for ASPM in GBM tumorigenesis and WNT signaling in the developing brain (184, 185).

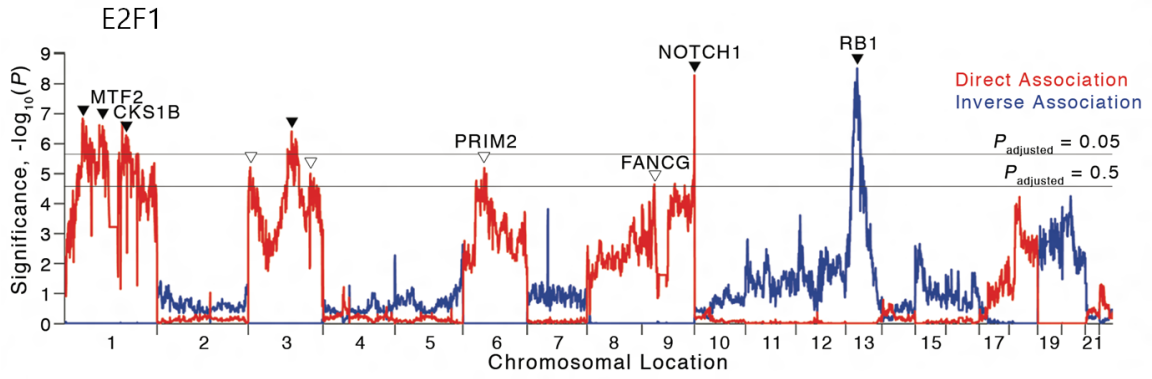

Figure 10: Association between E2F1 signaling and copy number alterations

Association between copy number alteration and E2F1 pathway activity. Thresholds for statistical significance are shown at a family-wise error adjusted *P*-value of 0.5 (-log10(*P*) = 4.35) and 0.05 (-log10(*P*) = 5.35). Candidate loci are indicated with an arrow (adjusted *P* < 0.05, black; adjusted *P* < 0.5 and > 0.05, white). Genes that were known to drive E2F1 or had significantly associated expression values with E2F1 activity across three independent datasets are labeled at the corresponding locus. Both direct (red) and inverse (blue) associations are shown.

Additionally, we found that several known pathway regulators approached significance in the copy number association analysis while being significantly associated with pathway activity at the gene expression level, and vice versa. Loci that approach copy number association but which had strong evidence of association across the three expression datasets include another Fanconi Anemia gene, FANCG ($P = 2.30 \times 10^{-5}$, E2F1), as well as SIRT3 ($P = 6.75 \times 10^{-6}$, P53 pathway) and DKK3 ($P = 7.65 \times 10^{-6}$, Myc). SIRT3 is a known modulator of apoptosis that has been shown to be deregulated at the expression level in GBM (186). Similarly, DKK3, a dikkopft family member, is known to repress Myc expression (187). Two replication complex associated genes, RFC4 ($P = 8.04 \times 10^{-6}$,

E2F3) and PRIM2 ($P = 6.53x10^{-6}$, E2F1), also approached significance. Other studies have found that RFC4 is overexpressed in other cancers (188, 189), which we also observed in both the TCGA GBM and REMBRANDT glioma databases (Figure 11). Finally, copy number of RB1, the canonical E2F repressor, was inversely associated with E2F1 activity ($P = 3.11x10^{-9}$) and approached significance at the expression level in the TCGA dataset ($P = 8.63x10^{-5}$).

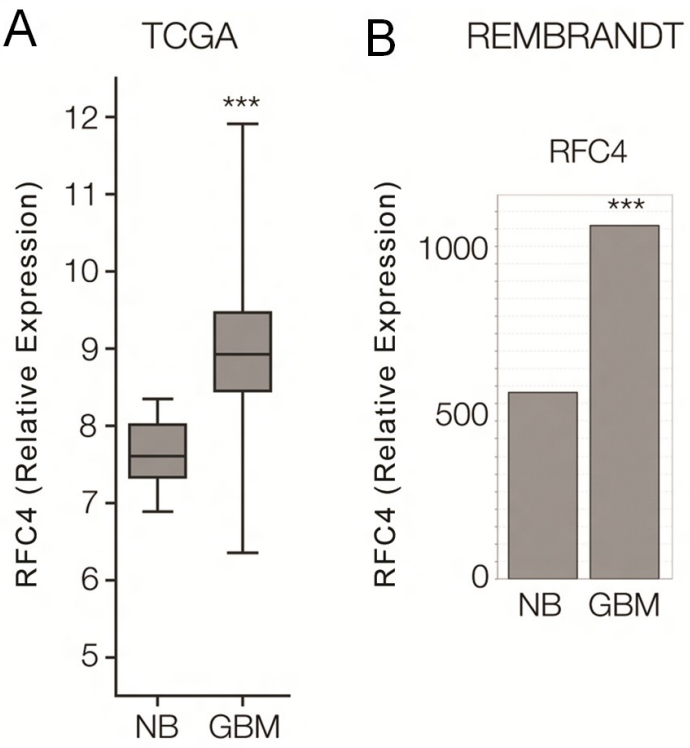

Figure 11: RFC4 expression in normal brain and glioblastoma tissue samples

Relative expression level of RFC4 measured in either normal brain (NB) samples or glioblastoma samples (GBM). A) TCGA samples analyzed based on log_2 expression values. B) REMBRANDT database samples. *** $P < 0.001$, Wilcoxon Rank-Sum Test

2.3.4 Patient outcome is stratified by pathway signaling

While the identification of known and likely regulators of pathway activity in GBM may lead to a more complete understanding of the alterations that drive tumorigenesis, this framework may also identify clinically relevant alterations that were previously uncharacterized. Therefore, we also examined the extent to which pathway activity was associated with clinical outcome of patients, focusing on patients that received aggressive therapeutic intervention (> 3 cycles of targeted therapeutics or chemotherapy, or concurrent chemotherapy and radiation). We compared differences in overall survival by stratifying patients based on whether they had high (Probability ≥ 0.5) or low (Probability < 0.5) pathway activity.

Three pathways reached significance after accounting for multiple testing ($P < 2.94 \times 10^{-3}$, log-rank; Figure 12) including the E2F1 ($P = 1.41 \times 10^{-3}$, log-rank) pathway, where high pathway activity coincided with longer patient survival. Additionally, elevated activity of the IFNγ ($P = 2.25 \times 10^{-3}$) and Ras ($P = 5.7 \times 10^{-4}$) pathways was associated with poor patient outcome. Identification of the Ras and IFNγ pathways was consistent with the generally poor prognosis of Mesenchymal tumors (19). Previously described links between patient outcome and two additional pathways characteristic of the Mesenchymal subtype, STAT3 ($P = 6.83 \times 10^{-3}$) and TNFα ($P = 3.64 \times 10^{-3}$), approached but did not reach significance (95, 190, 191).

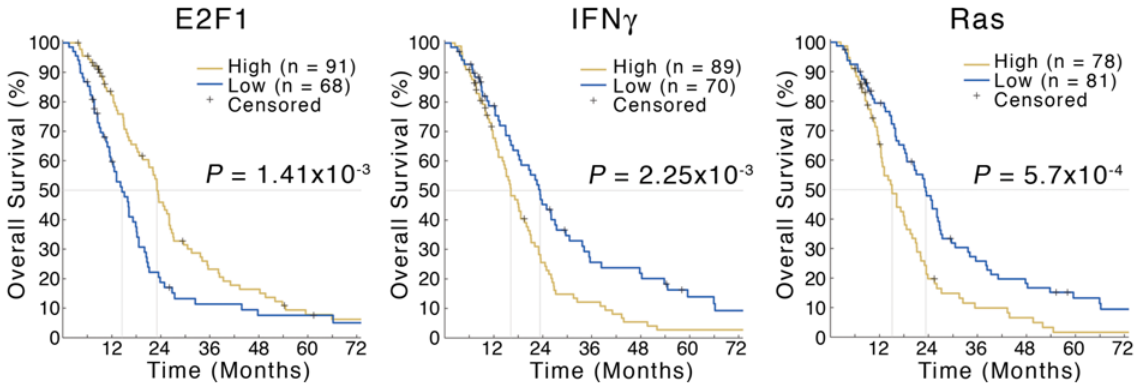

Figure 12: Patient outcome during aggressive therapeutic intervention is stratified by E2F1, IFNγ, and Ras pathway activity

Overall survival of TCGA patients receiving aggressive therapeutic intervention (>3 cycles of chemotherapy or targeted therapy, or concurrent chemotherapy and radiation treatment; $n = 164$). Patient samples are group as having either high (Probability ≥ 0.5, beige) or low (Probability < 0.5, blue) activity for E2F1, IFNγ, or Ras. Nominal log-rank *P*-values are shown.

While the identification of pathways associated with mesenchymal signaling was anticipated, E2F1 acting as a beneficial prognostic marker was surprising because proliferative signatures and genes have been demonstrated to be associated with poor patient outcome (19, 112, 121, 122). However, these studies compared all patients, regardless of intervention. This identification could reflect the apoptotic control that E2F1 exerts on cellular signaling, or the role that proliferation may play during induced mitotic catastrophe. To further understand this relationship between E2F1 signaling and patient outcome, we also analyzed patients which did not receive aggressive

intervention. In this setting, in contrast to the aggressive setting, high E2F1 activity coincides with poor patient prognosis, consistent with previous findings (Figure 13).

Moreover, under multivariate analysis accounting for other known prognostic indicators this relationship is still noted, with E2F1 having a strong prognostic effect that was dependant on treatment. High E2F1 activity was both a good prognostic marker in the aggressive treatment setting ($P = 0.014$, HR = 0.78 (0.61 – 0.95)) while being a poor prognostic marker in the nonaggressive setting ($P = 0.002$, HR = 1.70 (1.21 – 2.39)) (Table 2). Considering that some of these studies were conducted prior to or concurrent with the landmark clinical trial investigating Temozolomide treatment in glioblastoma (135), and that none of these studies recorded or stratified patients based on intervention, this could explain some of these findings as temozolomide treatment is by far the most common treatment recorded by the TCGA.

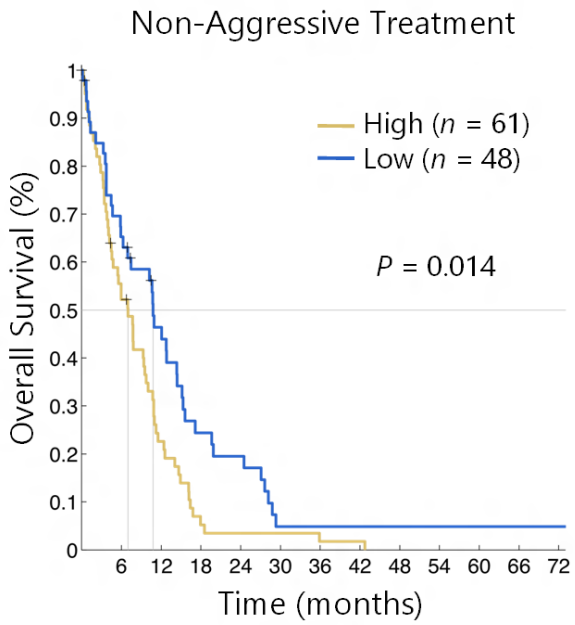

Figure 13: E2F1 is a poor prognostic marker in the non-aggressive treatment setting

Overall survival of TCGA patients receiving non-aggressive therapeutic intervention (<4 cycles of chemotherapy or targeted therapy, and non-concurrent chemotherapy and radiation treatment; $n = 109$). Patient samples are group as having either high (Probability ≥ 0.5, beige) or low (Probability < 0.5, blue) activity for E2F1. The nominal log-rank *P*-value is shown comparing survival between the two groups.

Table 2: Multivariate Cox Proportional Hazards analysis of E2F1 and prognostic markers in GBM patients

	Cox *P*	HR (95% CI)
Age at Diagnosis	**0.001**	1.02 (1.01 – 1.03)
KPS	**0.004**	0.98 (0.97 – 0.99)
MGMT Promoter Methylation†	0.810	0.96 (0.66 – 1.39)
E2F1 Activity **	**0.014**	0.78 (0.61 – 0.95)
Nonaggressive Treatment†	**0.004**	1.67 (1.18 – 2.36)
E2F1 : Nonaggressive	**0.002**	1.70 (1.21 – 2.39)

** E2F1 pathway scores logit transformed prior to cox PH analysis
† variables entered as logical (true/false) values for analysis

To determine whether E2F1 signaling captures the proliferative capacity of these tumors, similar to the other prognostic signatures generated by previous studies, we compared E2F1 pathway signaling to the mean expression of each of the signatures developed for predicting patient outcome, determining the concordance between these signatures (Figure 14). Under this analysis we also found that the E2F1 shows a strong degree of correlation with the Proliferative subtype and the H2CA glioma cluster when applying those signatures to the TCGA GBM database, as well as HJURP and CHAF1B, genes previously noted for their association with proliferation (19, 112, 121, 122). Therefore, we conclude that the signatures, while not identical, capture similar characteristics of GBM tumors, and the activity of E2F1, while initially perplexing, is consistent with the proliferative signatures developed in previously published studies.

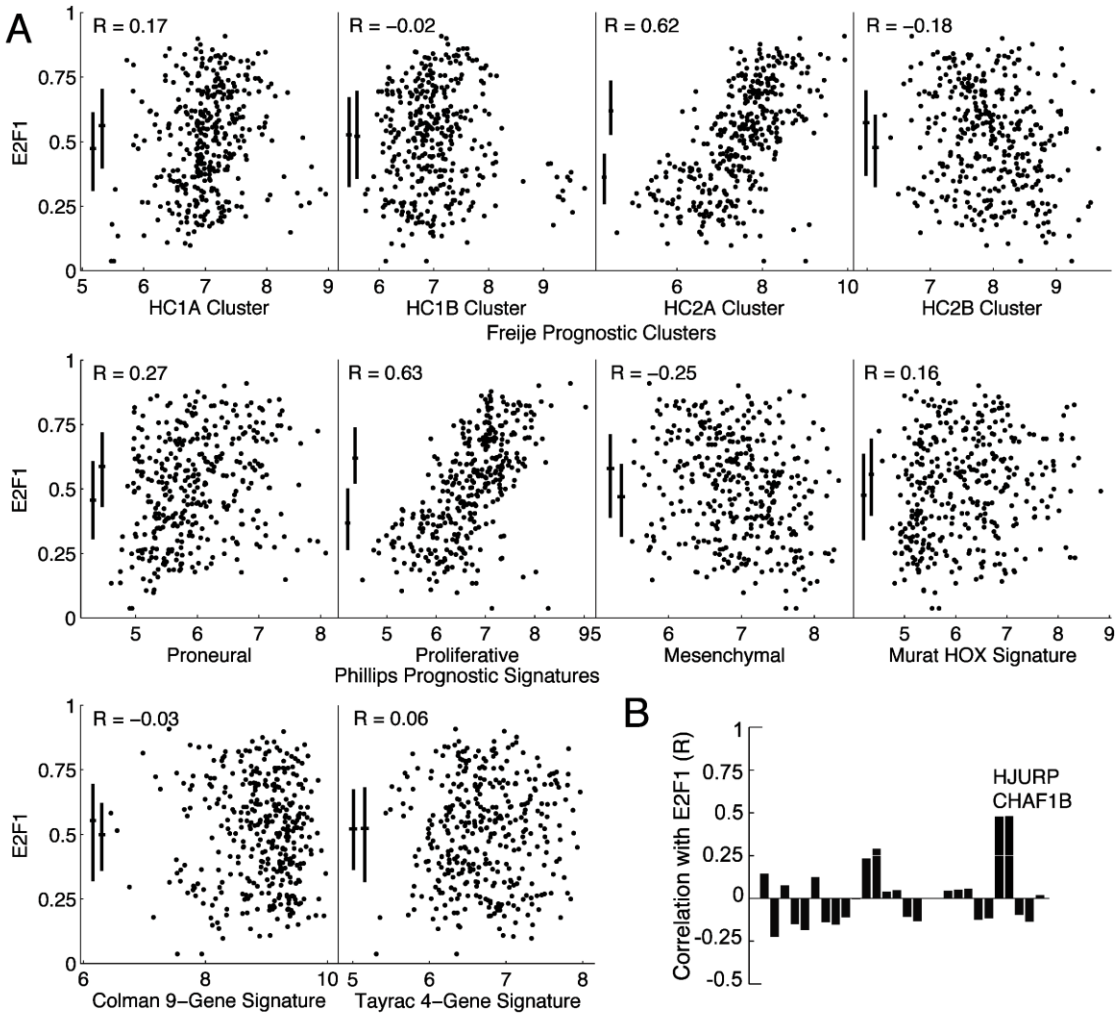

Figure 14: E2F1 pathway signaling reflects proliferative signaling in GBM tumors

TCGA GBM samples were analyzed using either the E2F1 pathway signature or the prognostic signatures developed by Phillips *et al.,* Freije *et al.,* Murat *et al.,* Colman *et al.* or deTay Rac *et al.* Pearson correlation coefficient shown for between each signature or gene and the predicted E2F1 activity of a patient. A) Prognostic signatures were analyzed by comparing calculating the centroids of the signature genes listed for each prognostic signature. Bars on the left of each graph show the mean (dot) and inter-quartile range (line) for the E2F1 pathway activity of samples with prognostic signatures either below (left) or above (right) the mean. B) Single-gene prognosis predictors were also analyzed for the concordance of expression between each gene and E2F1 pathway signaling.

2.3.5 UBL3 is associated with IFNγ signaling, patient survival and response to therapeutic intervention.

After adjusting for family-wise error, there were seven candidate regions associated with the pathways that stratified patient outcome. Of these regions, six were associated with E2F1 signaling, one with IFNγ signaling, while no alterations were significantly associated with Ras signaling.

One locus of interest that was identified was chromosome 13q12.3, which is 18 megabases upstream of the RB1 locus. This region was inversely associated with both IFNα ($P = 2.96 \times 10^{-7}$) and IFNγ ($P = 2.64 \times 10^{-6}$, Figure 15) pathways, but was not identified as a focal ROI by GISTIC (mean q = 0.023, Figure 16). 13q12.3 was altered in 29.8% (91/305) of patients, with most alterations being losses. Two of the genes within this region, UBL3 and HMGB1, were significantly associated with IFNα/γ signaling at the expression level across all three of the expression datasets we analyzed (Table 1).

HMGB1 is a well-characterized regulator of response to ischemic stroke and inflammation of in the brain and a regulator of glioma migration and survival (192-195), which makes it an attractive candidate for association with IFNγ and patient outcome. In contrast, UBL3 is a poorly characterized membrane anchored ubiquitin-like gene that has been implicated as a downstream member of MITF signaling in melanoma (196, 197), though its function is not known.

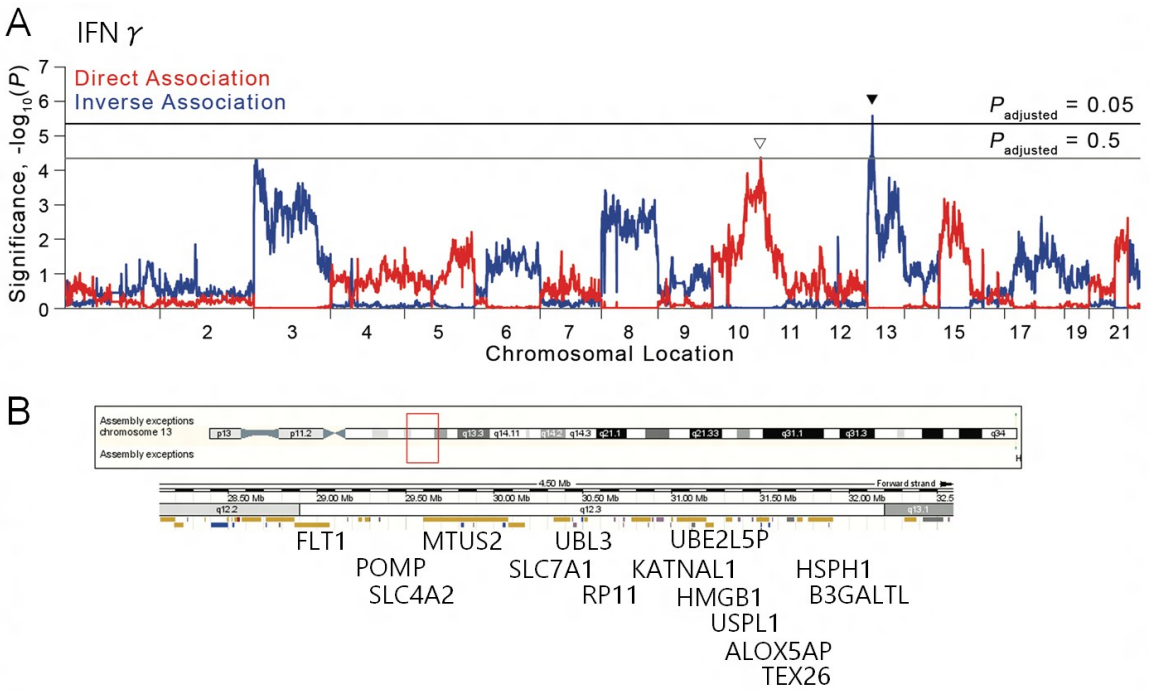

Figure 15: Association between copy number alterations and IFNγ signaling.

A) Association plot between copy number alteration and IFNγ pathway activity. Thresholds for statistical significance are shown at a family-wise error adjusted *P*-value of 0.5 ($-\log_{10}(P) = 4.35$) and 0.05 ($-\log_{10}(P) = 5.35$). Candidate loci are indicated with an arrow (adjusted $P < 0.05$, black; adjusted $P < 0.5$ and > 0.05, white). Both direct (red) and inverse (blue) associations are shown. B) Diagram depicting the protein-coding genes located within 13q12.3, which was identified as being significantly associated with IFNγ pathway activity at $P < 0.05$ after family-wise error adjustment.

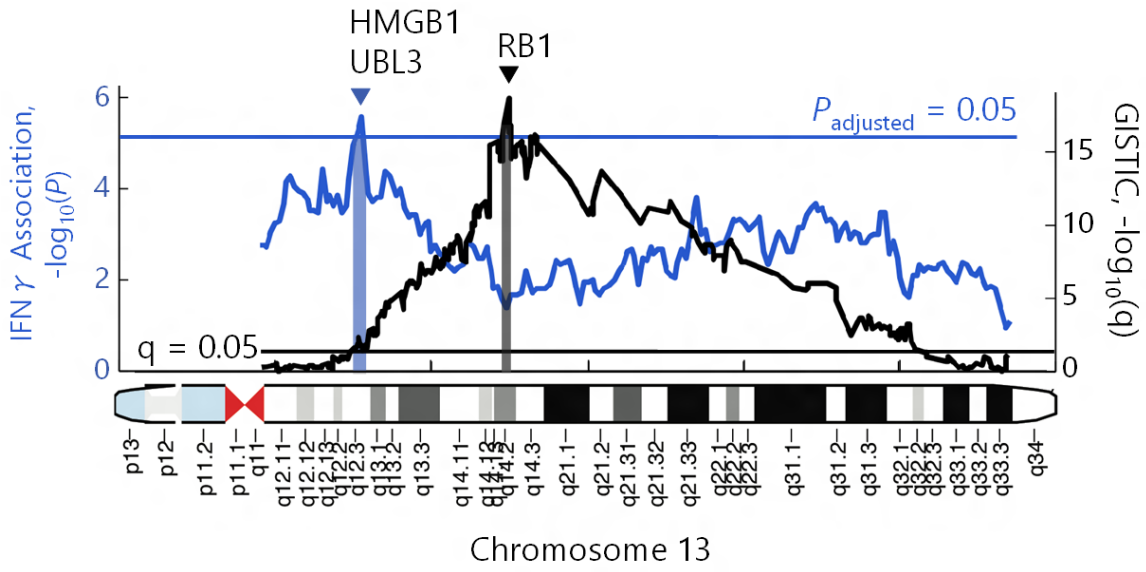

Figure 16: Comparison of alterations identified on chromosome 13 by GISTIC or pathway association analysis

Plot of chromosome 13 associations identified using either the pathway association analysis of inverse associations with IFNγ (blue, significance plotted on left axis), or using GISTIC 2.0 (black, significance plotted on the right axis). Thresholds for significance are indicated as solid lines showing either significance at a $P < 0.05$ after family-wise error adjustment for the pathway association analysis (blue), or $q < 0.05$ for the GISTIC analysis (black). Candidate genes within the loci identified by either pathway association or GISTIC focal ROI are displayed above the plot.

To better understand these genes, and identify which of these two might be important in GBM tumorigenesis and patient survival, we analyzed the expression of these genes across various tissues throughout the body. In this analysis, we demonstrated that HMGB1 is overexpressed in the hematopoietic lineage and UBL3 is overexpressed in the brain (Figure 17). This particular finding supports a specific role for UBL3 in the normal functioning of the brain, and a perhaps broader role for HMGB1 throughout the body, and in immune signaling.

Considering that alteration of this locus primarily occurs as a loss of one or more copies, this makes UBL3 an interesting candidate, as does the fact that MITF signaling modulates temozolomide response in melanoma (198). Therefore to better understand whether loss of this locus affected overall expression of these genes, we analyzed expression of UBL3 and HMGB1 in normal brain samples, GBM samples without loss at 13q12.3 and GBM samples with 1 or more copy loss at 13q12.3 (Figure 18). Under this analysis, we found that expression of UBL3 was significantly lower in GBM samples with at least one copy loss of 13q12.3 relative to both normal brain samples and other GBM samples without loss at this locus. In contrast HMGB1 expression was not significantly different between normal brain samples and GBM tumors will loss of this locus. There was, however, a significant decrease observed in GBM tumors with loss of this locus compared to tumors that retained both copies of 13q12.3. Together, these findings suggest that UBL3 may indeed play a role in the tumorigenic process.

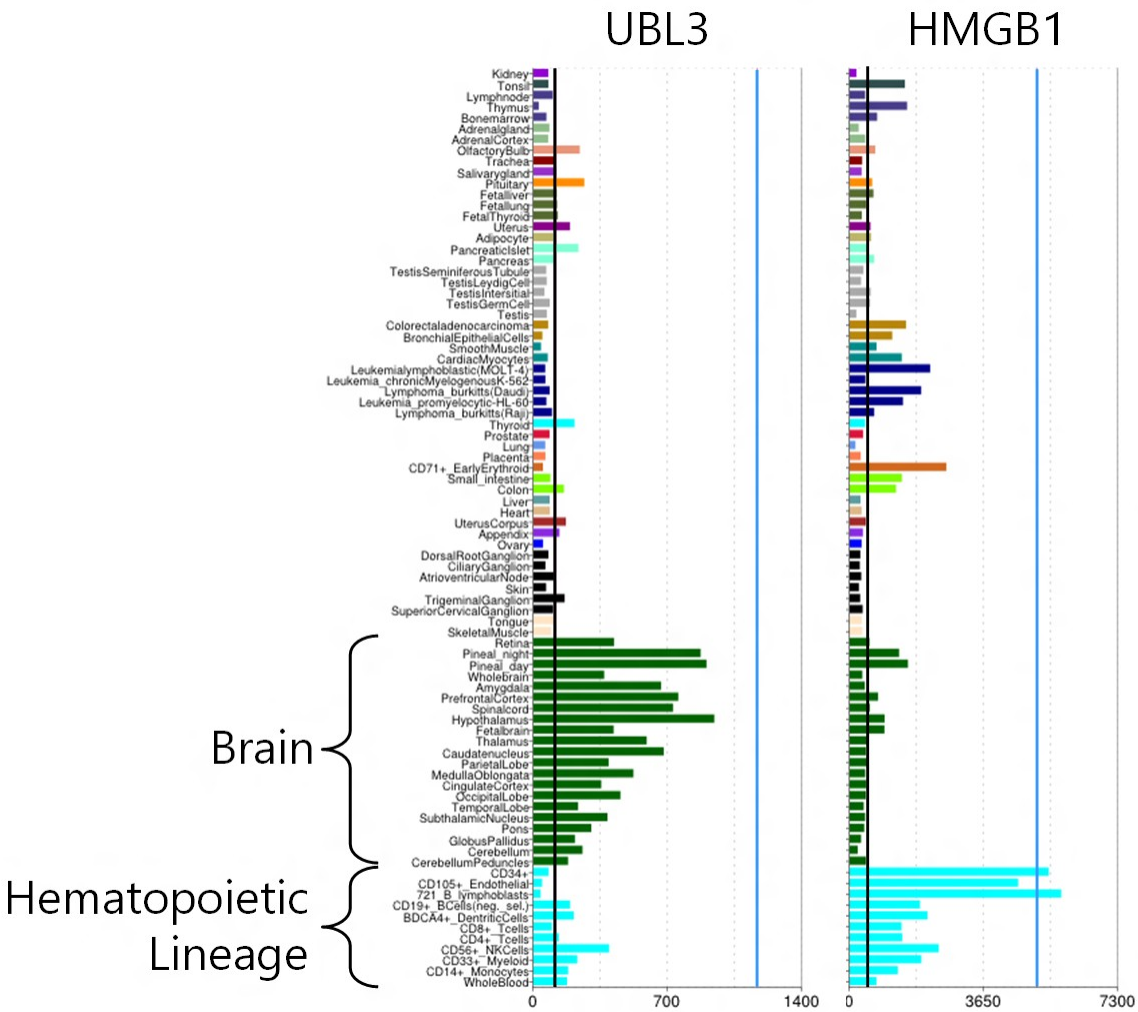

Figure 17: Expression of candidate genes in normal tissues

Expression of UBL3 (left) or HMGB1 (right) mRNA measured across different normal tissue specimens color coded by gross site of collection. Samples from various regions within the brain are shown in green, while tissue consisting of the hematopeitic lineage is indicated in cyan. Median (black line) and 10X median (blue line) expression for either UBL3 or HMGB1 is indicated on the graph.

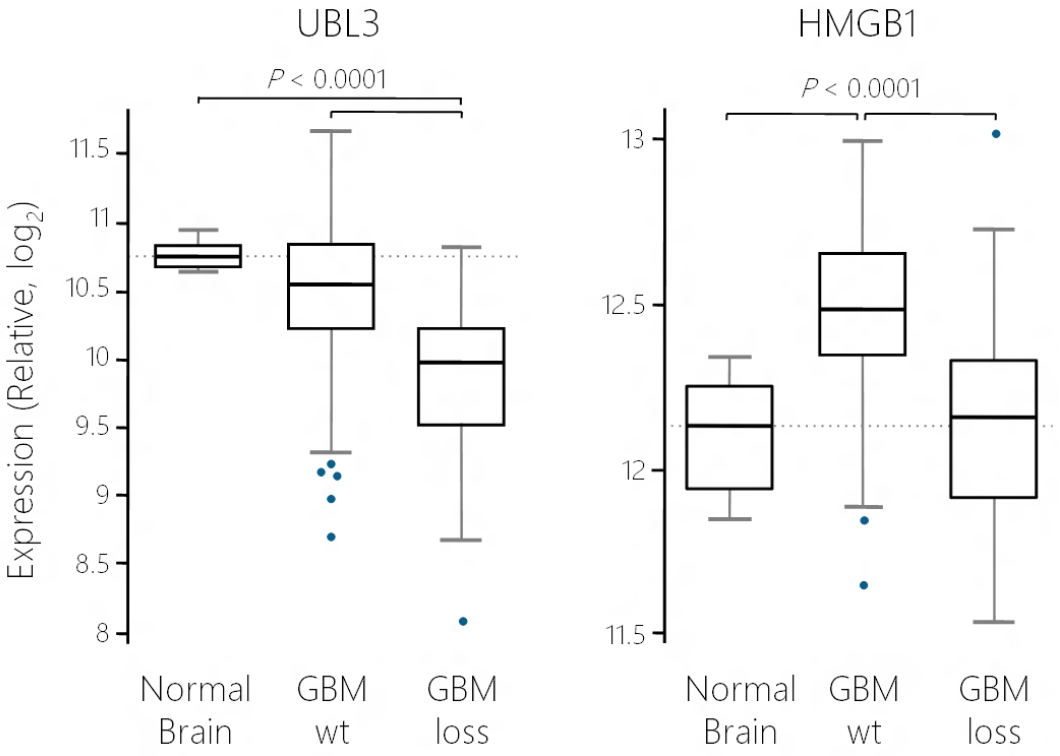

Figure 18: Effect of copy number alteration on mRNA expression of UBL3 and HMGB1 relative to normal brain

Relative expression levels of UBL3 (left) or HMGB1 (right) mRNA in TCGA samples of either normal brain, GBM tumor samples without alteration at 13q12.3, or GBM samples with at least 1 copy loss of 13q12.3. *P*-values are shown for Wilcoxon Rank-Sum test when significant at $P < 0.05$. Mean expression level for each gene in normal brain samples is shown (dotted line). Plot shows the mean (center line), interquartile range (box), and 2x the interquartile range (extended bars), as well as outlier samples (blue points).

As we noted previously, IFNγ pathway activity was associated with survival differences in aggressively treated patients. To determine whether either of these candidates associated with patient outcome, we examined the overall survival of glioma patients in the REMBRANT database using copy number status that was ascertained on

the Affymetrix SNP6 arrays, which provide better resolution than the aCGH data analyzed in the TCGA. In this dataset HMGB1 failed to associate with patient survival ($P = 0.942$, log-rank, Figure 20), but alteration of UBL3 significantly stratified patient survival ($P = 0.002$, log-rank; Figure 19). Furthermore, other genes both upstream and downstream of UBL3 failed to significantly stratify patient outcome ($P > 0.05$, log-rank, Figure 20).

Because loss of UBL3, indicative of high IFNγ signaling, was associated with poor patient outcome these results lend further support for a role for UBL3 in patient outcome. Moreover, these results demonstrate that UBL3, rather than loss of Chromosome 13q12.3 in general acts as a prognostic marker in this setting. However, while UBL3 loss at the expression level also associated with patient outcome, it was also associated with tumor grade, showing a significant reduction in expression in GBM patients compared to normal brain ($P < 0.001$, Wilcoxon Rank-Sum Test). Reduced expression was not observed in lower grade astrocytomas or oligodendrogliomas ($P > 0.05$), and suggests that UBL3 loss could represent a GBM specific marker that was associated with grade, rather than a prognostic marker for this setting.

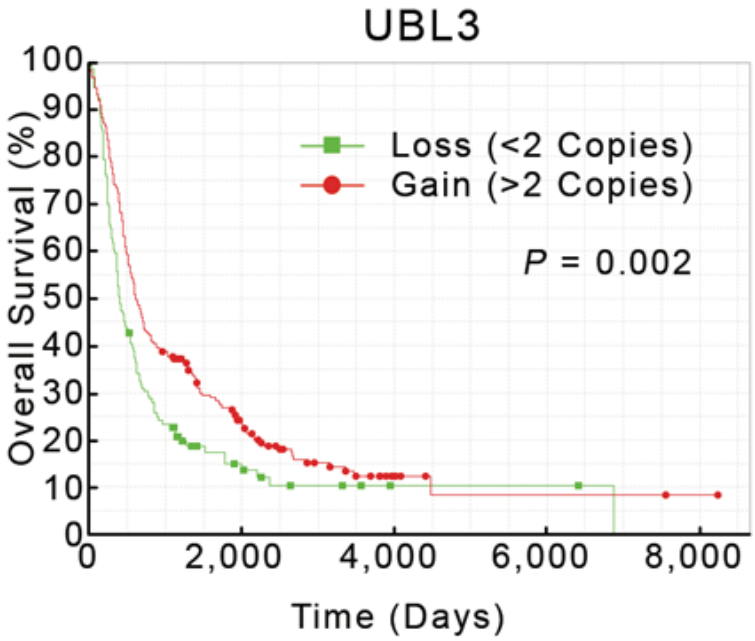

Figure 19: UBL3 loss associates with poor patient survival

Overall survival of REMBRANDT glioma patients (n = 351) stratified by loss (< 2 copies, green) or gain (≥ 2 copies, red) of UBL3 determined by the mean of all reporters for UBL3 per patient. The log-rank P-value is shown.

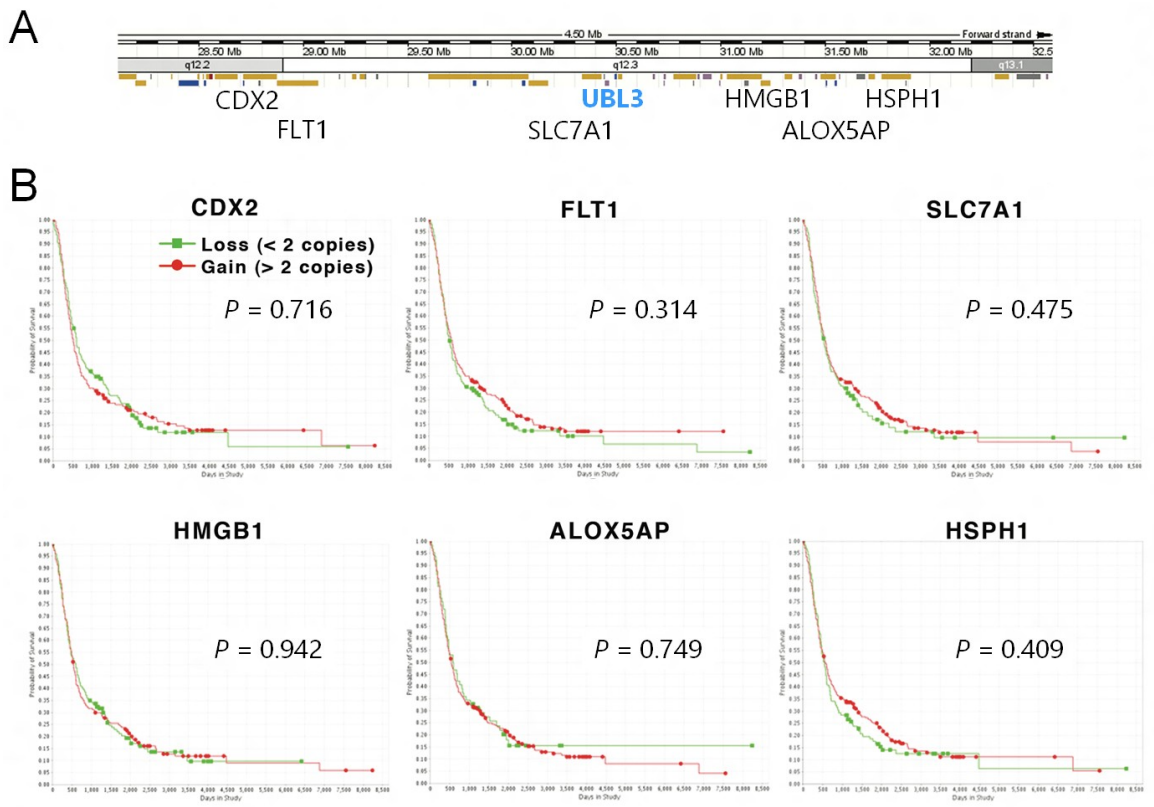

Figure 20: Analysis of overall survival of glioma patients based on alteration of genes flanking UBL3 in Chr13q12.3

A) Schematic of chromosome 13q12.3 highlighting UBL3 (blue) and three genes upstream and downstream of UBL3 for which SNP6 data was available (black). B) Overall survival of REMBRANDT glioma patients (n = 351) stratified by loss (< 2 copies) or gain (≥ 2 copies) of the genes indicated in A. Alteration status of the genes was determined by the mean of all reporters per patient for either gene. The log-rank P-value is shown.

To validate these findings and to clarify the role of UBL3 in patient outcome, we examined response of GBM xenograft lines treated with standard of care agents. Xenografts were treated with either Temozolomide (n = 16 lines, 10 replicates per line), or Avastin (n = 17 lines, 10 replicates per line) and analyzed for the extent to which

either treatment prolonged median survival relative to saline-treated control mice (Table 3 & 4). First, we ensured that these models reflected known modifiers of treatment response by confirming that high expression of MGMT in untreated xenografts corresponded to poor response to temozolomide ($P = 0.031$, $\rho = -0.539$, Spearman correlation; Table 3), consistent with the role of MGMT promoter methylation in GBM treatment response (140).

We then compared expression of UBL3 and HMGB1 to xenograft response to either intervention. While we found that HMGB1 was not significantly associated with response to treatment ($P > 0.05$, Table 3), expression of UBL3 was significantly associated with xenograft response to treatment for both Temozolomide ($P = 0.015$, $\rho = 0.593$; Figure 21) and Avastin ($P = 0.034$, $\rho = 0.515$; Figure 21). We also note that low expression of UBL3 was associated with a poor outcome in the treated mice, consistent with our findings in both the TCGA and REMBRANDT databases. Taken together, these results strongly suggest that loss of UBL3 is a GBM specific event that affects expression of this gene in a way that impacts patient response to therapeutic intervention.

Table 3: Association between candidate expression and xenograft treatment response

		MGMT	HMGB1	UBL3	NOTCH1
Temozolomide	*P* (Spearman):	**0.031**	0.692	**0.015**	0.520
	ρ (Spearman):	**-0.539**	0.107	**0.593**	0.174
Avastin	*P* (Spearman):	0.070	0.450	**0.034**	**0.037**
	ρ (Spearman):	-0.450	-0.196	**0.515**	**0.509**

Table 4: Survival and expression characteristics of GBM xenograft models

	Increase in Median Survival (%, Treatment - Control)**		Log$_2$ Expression Level (RMA)			
Line ID	Avastin	Temozolomide	MGMT	HMGB1	UBL3	NOTCH1
08-695	55	76	4.2775	12.227	8.9177	9.0475
08-499	129	449	4.3956	12.634	9.6571	9.3911
245-PR	45	-4	4.3872	12.564	8.2558	8.828
x43	11	4	6.3031	12.495	6.7941	5.0077
x693	42	91	4.138	12.662	7.9509	9.534
x54	7	136	4.3983	12.308	8.3526	7.3732
X2224	46	39	6.1343	12.244	8.372	8.8135
08-537	35	13	4.5725	12.409	7.7926	9.0183
x263	200	75	4.6609	12.19	9.131	8.5764
x245	73	118	4.2256	12.316	8.5229	8.6276
08-308	60	N/A	4.6714	12.123	8.121	7.3268
x317	36	N/A	4.3453	12.577	7.3831	7.1631
x456	31	39	4.714	12.531	7.7619	8.4086
X2339	N/A	97	5.2129	12.414	7.8637	6.4878
x717	21	74	5.2583	12.512	9.268	7.8928
X2159	39	129	4.2793	12.806	8.9718	8.1172
x475	21	53	5.8262	12.228	8.2605	8.6872
x320	145	145	4.2381	12.549	9.0161	10.345

** Increase in median survival is calculated as the percentage increase in median survival of xenograft mice receiving treatment (Avastin or Temozolomide, $n = 10$) compared to vehicle control treated mice ($n = 10$)

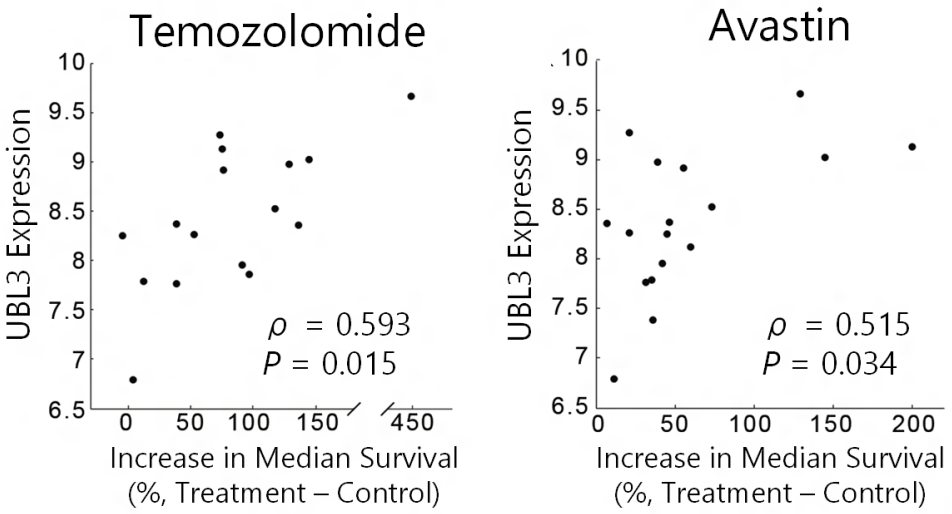

Figure 21: Xenograft response to standard of care intervention

Response of GBM xenograft lines treated with either Temozolomide (n = 16 xenograft lines, left) or Avastin (n = 17 xenograft lines, right) compared to control treatment. Response is measured as the percentage increase in median survival for a xenograft line between treatment (n = 10) and control (n = 10) settings. Correlation coefficients and P-values are shown for the Spearman correlation between response and UBL3 RNA expression level as measured by RMA normalized Affymetrix U133A arrays.

Finally, we note that another candidate gene associated with a survival-related pathway, NOTCH1, was also correlated with xenograft Avastin response (P = 0.037, ϱ = 0.508; Table 3). Notably, low expression of NOTCH1 was associated with poor response to Avastin, which was consistent with the role NOTCH1 signaling plays during CD133+ cell differentiation into CD133+/CD144+ cells that subsequently give rise to VEGFR2 producing tumor endothelial cells in GBM (199).

2.4 Discussion

Perhaps the greatest challenge in oncology is facing the reality of the extreme complexity and heterogeneity of the disease process. With few exceptions, most cancers represent a myriad of diseases resulting from a very large array of genome alterations including mutations, copy number alterations, aberrant epigenetic markers or chromosome rearrangements. Although there are instances of genes that are frequently altered in particular cancers, such as EGFR or CDKN2A in GBM, it is becoming increasingly clear that events that occur in a fraction of tumors represent a challenge to understanding underlying mechanisms of tumorigenesis and ultimately limit the effective treatment of individual patients (140, 158, 191, 200). Paramount in this challenge is the ability to identify which of these events are functionally important and how they alter tumor signaling and response to intervention.

Others have argued that an analysis of pathways that are deregulated as a result of the alterations represents a logical approach to addressing the challenge of identifying the significant events (54, 173). In essence, this approach aims to focus on the functional consequence of pathway activity regardless of the nature of the initiating event. The analysis of pathways in the context of complex genomic datasets has generally taken two approaches. First, pathway activity can be inferred by the alteration of a candidate driver, such as a truncating mutation in an upstream repressor or amplification of a growth factor receptor (e.g. EGFR and Her2). Second, various gene annotation tools,

such as GSEA (37), have been applied to identify the enrichment of genes that cluster within a known pathway. Both approaches have strengths and have extended our understanding of many types of tumors, but they also have significant limitations, especially when the function of an alteration is not known or appreciated *a priori*.

In contrast, our approach relied on the quantitative analysis of pathway activity defined by *in vitro* experimentation to form the basis for the identification of functionally significant genome alterations. While there could be some concern that these signatures, which were developed in a heterologous cellular background, might not be appropriate to use in the context of GBM, we have presented several pieces of compelling evidence that these pathway signatures do indeed perform well in this setting. Many of the pathway signatures were clearly enriched within distinct subtypes of GBM, and these results were consistent with previous ontological analysis of these subtypes. However, beyond this analysis, when we limited the gene expression data to only GBM-related genes as defined by another research group analyzing this dataset, we found that most of the signatures were able to produce relatively concordant results. This directly implies that the signatures that we have developed, while meant for one cellular context, can extend well beyond that initial target.

These results are consistent with previous findings from our lab that described during the initial development of these signatures, and demonstrated a broad utility across multiple cancer types (45). The identification of pathway signatures that

associated with patient outcome is also encouraging. While the E2F1 signature appeared to have a duel role in GBM, the predictions of this signature were concordant with previously developed glioma survival signatures. Finally, the numerous alterations that we identified that are known to regulate pathway activity directly confirm that these signatures are performing as expected in this setting.

We believe this framework provides an approach to identifying alterations that are low frequency, and thus difficult to distinguish from bystander events, but nevertheless represent functionally significant regulators of pathway activity. The identification of many well-established regulators of the E2F pathway, including several not previously appreciated to be altered in GBM, provides support for the logic and power of this approach. For instance, a role for alterations in CKS1B, KDM5B and RFC4, has been described to various degrees in other cancers but not previously been observed for GBM (180, 181, 188, 189). Likewise, the numerous other pathway regulators identified by this approach reflect the power of linking multiple measures of genomic alterations with their functional outcome, in this case measured by activity of a given pathway. It is likely that some of these signatures could be improved, to have better specificity and sensitivity, in this setting. There are a number of key signatures, such as the Ras pathway, which did not identify any significant copy number alterations. While NF1 loss was located in a peak region, it was well under even the less stringent threshold we placed, suggesting that either the Ras signature could be more accurate or

that incorporation of mutation data along with copy number alterations is necessary to fully model this signaling deregulation.

The identification of UBL3 highlights many of the strengths of this analytical approach. Not only is UBL3 a poorly understood gene that would generally be considered a poor target for further study using candidate gene approaches, it is also part of a locus that frequency-based analyses treated as insignificant. It was only when analyzing the association between copy number alterations and activation of signaling pathways that this locus was indentified as a significant candidate. Given that the association between both HMGB1 and UBL3 expression and IFNγ signaling was validated across three independent expression datasets of GBM tumors or GBM models demonstrates the robustness of the locus identified using the pathway association framework.

However, association analysis alone would still have failed to differentiate between HMGB1 and UBL3 had we not further explored the biological context of UBL3 and HMGB1 expression as well as the clinical significance of this region. Moreover, HMGB1 has been better characterized in glioma tumorigenesis, and thus was initially considered a much more promising candidate gene. However the weight of evidence supporting the role of UBL3 in glioma tumorigenisis comes from examining the expression of these genes in normal tissue, as well as the role of loss of either gene in an additional clinical database and across a diverse set of xenograft models. We were able

to demonstrate that UBL3 is specifically overexpressed in normal brain tissue and is selectively downregulated in GBM tumors that harbor a loss at this locus, consistent with loss of 13q12.3 being a GBM-related lesion. Moreover, UBL3 loss was associated both with patient outcome and response to therapeutic intervention with standard of care agents. These results, given the multiple datasets analyzed and the panel of *in vivo* models used, should accurately reflect the genomic and clinical diversity that is observed across patients, and lead to more robust results than any single xenograft model alone would have generated.

With further analysis and validation, UBL3 loss could become an additional biomarker of patient response to intervention. Targeting UBL3 is not likely to of particular use given that loss of the gene is associated with poor response in the settings that we analyzed. Moreover, there is very little currently known about the functional role that UBL3 plays during development and in the adult brain, though it is clearly overexpressed there. Until we known more about the specific biological impact of this alteration, it remains unclear how UBL3 directly impacts patient outcome when treated with aggressive therapeutic intervention.

The role that UBL3 plays is also an area of particular interest that further studies need to address. To date, there have been very few studies that have implied or directly tested the function of this gene, and none have been conducted in the setting of gliomas. Some studies have suggested that the role could be tied to NF1, and have found that

UBL3 expression was decreased significantly in hyperpigmented melanocytes of NF1+/- mice (201), which represent a pre-neoplastic model of melanoma. This is particularly compelling given that UBL3 loss was identified as being linked to expression of the IFNγ pathway, which we demonstrated to be indicative of the mesenchymal phenotype. If there is indeed a cooperative role between NF1 loss and UBL3 then *in vivo* analysis of human GBM tumors or genetically engineered mouse models of GBM bearing NF1 loss could provide further insight into the interplay between these two genes. Moreover, the relationship between UBL3 and MITF also speaks to a potential link that underlies both gliomas and melanomas, and is particularly striking considering these are the only two cancers for which Temozolomide is a common therapeutic agent (197, 198).

While we focus on copy number alterations and UBL3, the analytical association framework presented here can be extended to a various genomic datasets and sources. Modification of this approach could be used to identify additional types of drivers from distinct genomic sources such as epigenetic or microRNA regulators of pathway activity, or could integrate other network analysis approaches. There is no reason to presume that copy number aberrations are the only mechanistic means by which a signaling pathway is deregulated in GBM, as this would be inconsistent with many of the foundational studies that form the basis for our understanding of oncogenes and tumor suppressors. By integrating further genomic data sources, we will gain a more complete picture of the means by which a particular signaling pathway is repressed or activated

across these tumors. Moreover, understanding how the alterations linked to a given pathway co-occur, may provide further insight into the progressive steps involved with leveraging that particular pathway. These findings highlight the utility of applying a genetic association framework to identify novel drivers of tumor signaling, and demonstrate that these methods represent a complementary approach to frequency- and amplitude-based analyses of genomic alterations in cancer, though they are only one step to a fully integrated genomic understanding of this disease.

2.5 Materials and methods

Gene Expression Analysis and Genomic Signatures.

Affymetrix U133A or U133+2 .CEL files were downloaded for the TCGA (n = 386) and from 8 other datasets from previous publications (n = 482) (19, 112, 114, 139, 177, 178). TCGA and validation samples were normalized using both MAS5 and RMA in R using Bioconducter and MAS5 data was $\log_2$ transformed. Bayesian Factor Regression Modeling was used to remove batch discrepancies in MAS5 and RMA normalized datasets as previously described (16, 68, 69). Principal component analysis was used to confirm reduction in batch-effects prior to further analysis (Appendix A). Genomic signatures consisted of 16 previously described signatures (16, 45), along with a PTEN signature developed using data from Vivanco *et al.*, 2007 (GSE7562). A Bayesian Binary

Regression algorithm (BinReg 2.0) was used to generate signatures and score samples for each genomic signature as previously described (16).

Previously developed prognostic signatures were applied to batch-normalized RMA data by determining the centroid of all probesets representing genes included in a particular signature (19, 112, 121, 122, 139). These centroids were then directly compared against the probability of E2F1 pathway signaling, calculating the linear correlation (R, Pearson) between the two scores. Tissue-specific expression data was obtained through bioGSP[1]

Xenograft Chemotherapy Treatment.

Male and female athymic mice (nu/nu genotype, Balb/c background, 6 to 8 weeks old) were used for all studies. Animals were maintained in filter top cages in Thoren units (Thoren Caging Systems, Inc., Hazelton, PA). All animal procedures conformed to Institutional Animal Care and Use Committee and National Institute of Health guidelines.

Patient-derived human tumor xenografts were passaged in athymic mice and were excised from host mice under sterile conditions in a laminar flow containment hood. In preparation for transplantation, tumors were placed into a modified tissue press and homogenized. The resulting homogenate was then loaded into a repeating

[1] http://biogps.org/

Hamilton syringe dispenser. Tumor homogenate was injected sc into the right flank of the athymic mouse at an inoculation volume of 50 μl with a 19 gauge needle (202). Subcutaneous tumors were measured twice weekly with hand-held vernier calipers (Scientific Products, McGraw, IL). Tumor volumes, V, were calculated with the following formula: $[(\text{width})^2 \times (\text{length})]/2 = V\ (\text{mm}^3)$ (203).

Groups of mice were stratified by tumor volumes and were treated when the median tumor volumes were on average 250 mm^3 and were compared with control animals receiving vehicle control. Bevacizumab was administered at a dose of 5mg/kg IP twice weekly for 5 consecutive weeks and temozolomide was administered at a dose of 75mg/kg IP for 5 consecutive days. Mice were maintained until they reached five times greater than the measured volume at the start of treatment, or 120 days after start of treatment. The response of sc xenografts to treatment was assessed by delay in tumor growth defined as the percentage increase in median survival time for tumors in treated animals compared to control animals.

Xenograft and Cell Line Arrays.

Xenograft lines were grown to approximately 500 mm^3 in *balb/c* athymic mice, excised, and snap frozen as whole tumor samples. At least two replicates of each xenograft line were harvested. Tissue samples were homogenized in RLT buffer with 1% β-mercaptoethanol using a rotor-stator. Cell lines were previously established in either serum media or in stem-cell media. For each cell line, $1\text{-}2\times10^6$ cells were harvested and

homogenized in RLT buffer with 1% β-mercaptoethanol using QiaShredder spin columns. RNA was extracted using the Qiagen RNeasy kit, using an additional on column DNA digestion step according to the manufacturer's instructions. Samples were arrayed on Affymetrix U133A 2.0 chips. A total of 117 samples were generated, and were batch corrected as described above and mean centered across replicates before being analyzed. Expression data for xenograft and cell lines is available on GEO (GSE26344).

Clinical Analysis.

Analysis was performed on TCGA samples for which clinical data was available. Patient treatment status was classified as aggressive (n = 164) or nonaggressive (n = 112) intervention based on whether patients received either concurrent chemotherapy and radiation or more than 3 cycles of chemotherapy or targeted therapeutic intervention. Patients for which clinical, treatment or expression data was not available were excluded from analysis. Survival analysis was calculated using the log-rank test. Analysis of REMBRANDT samples was conducted using the REMBRANDT data portal, stratifying patients (n = 351) using the mean of all reporters for each candidate gene. Multivariate Cox proportional hazards calculations were run for all patients with available MGMT methylation, pathway prediction, KPS, Age at Diagnosis and treatment data (n = 157). Pathway scores were logit transformed prior to analysis in the Cox model.

Copy Number Analysis.

Segmented aCGH data was downloaded directly from the TCGA, samples without paired expression data were removed, and log2 ratio values were mean centered in 0.5 Mbp windows starting at the first unique segment available on each chromosome as was described previously by Bredel *et al.* (169). Data from the X and Y-chromosomes were excluded from analysis. Thresholds for gain and loss were set at a log2 ratio value of >0.2 and <-0.2 for ±1 copy and >0.8 and <-0.8 ±2+ copies, respectively. Copy number or gene expression associations with pathway status were calculated using the Spearman correlation between the probability of pathway activity and either the log2 ratio value for a given locus or the relative expression of a gene. Local minima of significance were calculated for each pathway, requiring an increase in significance of at least two orders of magnitude spanning at least three loci between each minimum. GISTIC (version 2.0) was run using the same thresholds for gain and loss using the gene pattern module (175, 176), identifying wide and focal alterations significant at $q < 0.05$.

Statistical Analysis.

Enrichment of pathway activity within core GBM tumor subtypes was calculated using the Kruskal-Wallis test, based on previously published subtype classifications (119). Differential pathway activity was tested in sequenced GBMs ($n = 140$) comparing tumors with validated non-synonymous mutations to non-mutated samples using the Wilcoxon Rank-Sum test. The spearman correlation was calculated between expression

of candidate genes and xenograft response to intervention using the Affymetrix probeset for each candidate with the highest mean expression across the xenograft models. Thresholds for significance were determined using a Bonferroni-adjustment for family-wise error, although *P*-values reported are nominal values.

3. Utilization of Cell Line Derived Genomic Signatures to Predict Oncogenic Pathway Activity in Human Tumors and Prospective Clinical Studies

3.1 Summary

In vitro experiments provide an opportunity to derive genomic expression signatures of key biological processes, including oncogenic pathway activation, that also have the potential to direct the use of investigational drugs that target such processes. However, the distinctions in the general characteristics of gene expression between in vitro and in vivo samples presents a substantial challenge in the utilization of these signatures in clinical studies. Here we describe methods for the effective utilization of a pathway signature, standardizing between the in vitro and in vivo samples, and using as an example an E2F1 pathway signature to predict the status of RB, a key regulator of E2F1. Further, anticipating the use of these methods in prospective clinical studies where one sample at a time must be analyzed, methods were tested to identify reference datasets for the purpose of standardizing the distinctions between *in vitro* and *in vivo* samples. We suggest that these methods provide a framework for the effective utilization of well-defined in vitro derived signatures in the context of prospective clinical studies where a goal would be to achieve an enrichment of those patients likely to respond to pathway-specific therapeutics.

3.2 Background

A major challenge in oncology is the ability to match the best therapy with the individual cancer patient, balancing relative benefit with risk, to achieve the most favorable outcome (204-206). The importance of selecting patients who will respond to a given therapeutic agent is perhaps best illustrated by the example of trastuzumab. In the absence of any selection, the overall response rate in breast cancer patients would be less than 10%. In contrast, for those patients selected on the basis of Her2 amplification, the overall response rate rises to 35-50% (207). Clearly, an assay for Her2 amplification is now viewed as a critical measure in the decision to use trastuzumab for the treatment of breast cancer. Likewise, other work has linked the efficacy of EGFR inhibitors such as erlotinib or gefitinib with the presence of activating mutations in the receptor (208-211). Conversely, the presence of K-Ras mutations signals resistance to EGFR monoclonal antibody therapy such as cetuximab (212).

While the use of gene mutation or gene amplification provides a powerful means to direct the use of targeted therapies such as trastuzumab, erlotinib, and cetuximab, these examples are limited and likely provide only one avenue to the goal of systematic strategies for directing therapy. Based in part on the observations from the recent cancer genome sequencing efforts, many have argued that a focus on pathways rather than single genes represents a better view of the consequence of these genome alterations

(173). In part, this derives from the realization that the consequence of these mutations is deregulation of signaling pathways that drive the cancer phenotype.

Although the use of various biochemical assays that measure the activation of a specific pathway, such as immunohistochemistry, fluorescent *in situ* hybridization (FISH) and sequence analyses, have made important contributions, these assays are often inadequate for many of the signaling pathways and are limited in the extent to which they can provide a global view of the contributions of pathways to cancer phenotypes. As an alternate approach, we have previously described the development of methodologies to develop 'signatures' of pathway activation (16, 45, 213). These are based on well-controlled experimental perturbation of a pathway regulatory activity coupled with the generation of genome-scale gene expression data that can then be used for the development of predictive statistical models. Because these signatures reflect not just the known role of specific genes but rather measure the consequence of perturbation of key regulatory activities, they have the potential to more broadly measure the events associated with a given pathway.

Importantly, because the signatures are based on a measure of gene expression that is common to any biological sample, the signatures have the potential to bridge various biological contexts. For instance, cell line based signatures have been shown to accurately predict the activity of cell signaling pathways often deregulated in human cancers and mouse models of cancer (45, 213, 214). Further work has shown that the

prediction of pathway activation using these signatures can serve to dissect the heterogeneity of cancer and at the same time to also predict sensitivity to therapeutic agents that target components of the pathway (16, 45, 215, 216). As such, the patterns of pathway activity that can be measured across populations of patient tumor samples then has the potential for predicting response to a given agent and thus serve as a mechanism to match those individual patients with the drug most likely to be of benefit. An example of such a strategy can be seen in an ongoing breast cancer trial of the Src inhibitor dasatinib (NCT00780676) in which patients are selected for inclusion in the study based on one of three genomic signatures (217).

In considering the next steps in the development and use of these signatures as predictive biomarkers that might eventually be applied in the context of prospective clinical trials, we recognize that there are several significant challenges that must be overcome. Foremost is the challenge of using expression signatures derived from *in vitro* sources to apply to *in vivo* tumor samples where global distinctions in expression characteristics can nullify the predictive capacity of the signature. Beyond this is the challenge posed by sample collection in the clinical setting, where patient samples are received at different time points, usually are obtained individually, and must be analyzed in a time dependent manner. We have addressed each of these issues with methods that we believe can lead to effective application of these signatures in retrospective and prospective clinical studies.

3.3 Results

3.3.1 Development of genomic signatures derived from experimental perturbation of pathway activity

We have previously described the development of genomic signatures reflecting oncogenic pathway activity based on the perturbation of activating components of these pathways (45, 218). A schematic representation of the methodology was depicted in Figure 1, where the training set used to derive the signature is a series of cell cultures that have been brought to quiescence to silence the pathway and were subsequently infected with an adenovirus expressing a gene of interest that leads to pathway activation or a control adenovirus containing only a GFP gene. These two sets of samples, representing pathway off and on states, is used to train and develop a predictive model. This model is projected into a test dataset derived from a series of tumor samples yielding probability scores that range between the two extremes defined by the model.

A critical aspect of the methodology used to apply the *in vitro* derived signatures to *in vivo* tumor samples is the recognition that the global gene expression profiles from these two contexts can differ substantially and compromise the utility of the signatures. An example illustrating the challenge is shown in Figure 22 in which principal component analysis is used to examine the global distinctions in expression data. In this example, gene expression data derived from a training set for an E2F1 pathway signature is displayed together with glioblastoma tumor data to be used for assessment

of the signature. It is evident from this analysis that the training data exhibits very distinct characteristics from the glioblastoma data (Figure 22). This broad distinction would make analysis of the two populations difficult and ultimately drive the results towards the global distinctions rather than the more subtle differences reflective of pathway activity. To adjust for these global differences, we have made use of several methods, one of which is shown in Figure 22, that involves a simple adjustment of the tumor data at the gene level to match the mean and variance of the training data – a process termed shift-scale normalization. Other methods have been described to achieve a similar goal including DWD and Combat (67, 70). We have also used methods in previous work in which a singular value decomposition and principal component development makes use of the combined data from cell culture samples and the target samples such as the tumor samples (45).

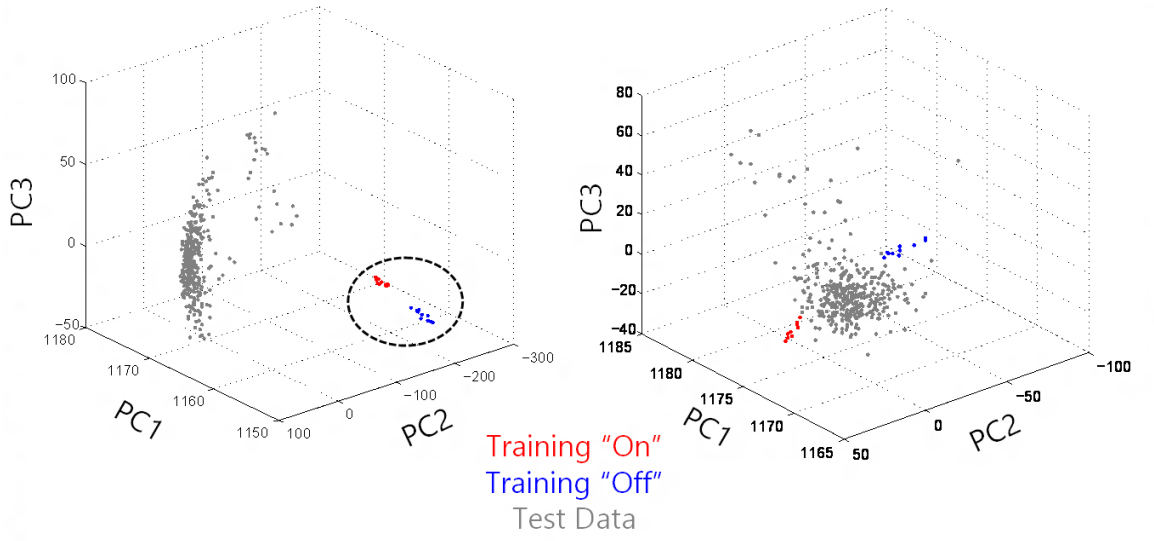

Figure 22: Principal component analysis of training and test data before and after normalization using shift-scale.

Principal component analysis of datasets from either test or training data before (left) or after (right) shift-scale normalization. Training samples are shown in red (E2F1 overexpression) and blue (GFP infected controls). Test cases from the TCGA GBM dataset are gray.

The impact of adjusting global expression differences is shown in Figure 23. For this analysis, we have made use of a recently developed implementation of the BinReg methodology within the GenePattern framework (219). This provides a platform that is user-friendly and still robust for customizing parameters. In this example, the training data is derived from activation of the E2F1 pathway in human mammary epithelial cell cultures. The test dataset derives from a collection of glioblastoma tumor samples

obtained from the Cancer Genome Atlas (TCGA) database[1] (220). The schematic on the left illustrates the basic framework employed when using what we refer to as the BinReg2 method. The global measure of gene expression in the tumor samples can be adjusted to match that of the cell line training data prior to decomposing the data by SVD. This adjustment, shift-scale normalization, uses a simple linear discriminate analysis to standardize the mean and variance of expression data at the feature level but various other methods could be employed as well. The training data is unaltered; it is the test data that is adjusted to correspond to the same general characteristics of the training data. The software then builds principal components from a SVD of cell line data alone, and predicts the probability of pathway activity in the tumor samples. The genes included in the model are illustrated in the heatmap shown in Figure 23B and the resulting predictions in the ovarian tumor dataset are shown in Figure 23C. The plot shown in Figure 23C depicts predicted probabilities for both the training data (blue and red symbols) as well as the test data (black symbols). When applied in this manner, the training samples define the extremes of pathway activation probabilities and the tumor samples then exhibit probabilities lying between these two extremes.

[1] http://www.tcga-data.nci.nih.gov/tcga/

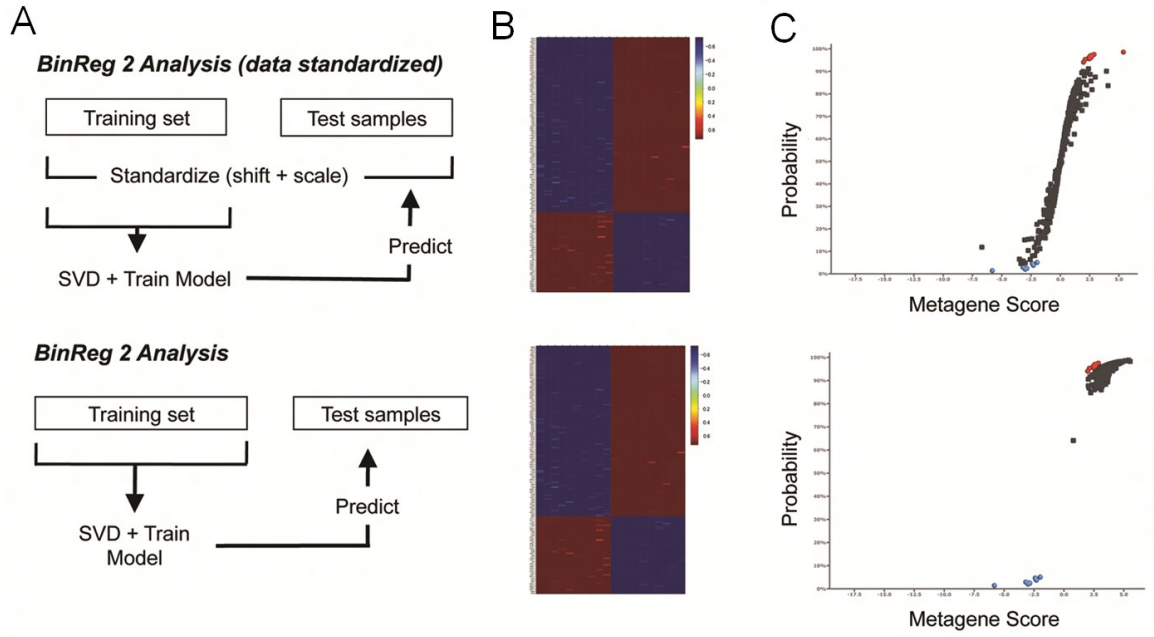

Figure 23: Strategies for applying BinReg methodology to predict E2F1 pathway status in ovarian cancer samples.

A) The schematic depicts the use of the BinReg2 methodology with (top) or without (bottom) use of the shift-scale methodology to adjust the mean and variance of the test data to match the training data. The training data alone is used for singular value decomposition (SVD) and principal component development and trained using only the information from the training dataset (pathway off and pathway on) to identify the genes constituting the model (panel B). C) The model is then used to predict the pathway status in the collection of ovarian tumor samples and the resulting probabilities of the tumor sample predictions (black symbols) are displayed along with the probabilities of the training data (blue and red symbols).

In sharp contrast to these results, an analysis in which the model is developed from only the cell line training data and without adjustment for the global gene expression differences yields a very different result. The genes in the model are the same but the predicted probabilities of the tumor samples are now compressed, not exhibiting the spread between the two sets of training samples as seen in the other examples and thus severely compromising the ability to discern differences in pathway activation across the test dataset.

Of course, an important question is the extent to which this analysis yields accurate predictions of pathway activation, reflecting the state of the *in vitro* activation achieved in cell cultures. We have assessed this in GBM and ovarian tumor samples based on an analysis of status of the RB1 tumor suppressor, an inhibitor of E2F1 activity. Given the clearly established role of RB1 in regulating E2F, one would anticipate that samples exhibiting loss of RB1 function would also exhibit elevated E2F1 pathway activity. As depicted in Figure 24, this is indeed the case with a significant distinction evident based on the prediction of E2F1 pathway activity in GBM ($P = 0.018$, Wilcoxon rank-sum; AUC = 0.634, ROC analysis) or ovarian tumors ($P < 0.0001$, Wilcoxon rank-sum; AUC = 0.782, ROC analysis).

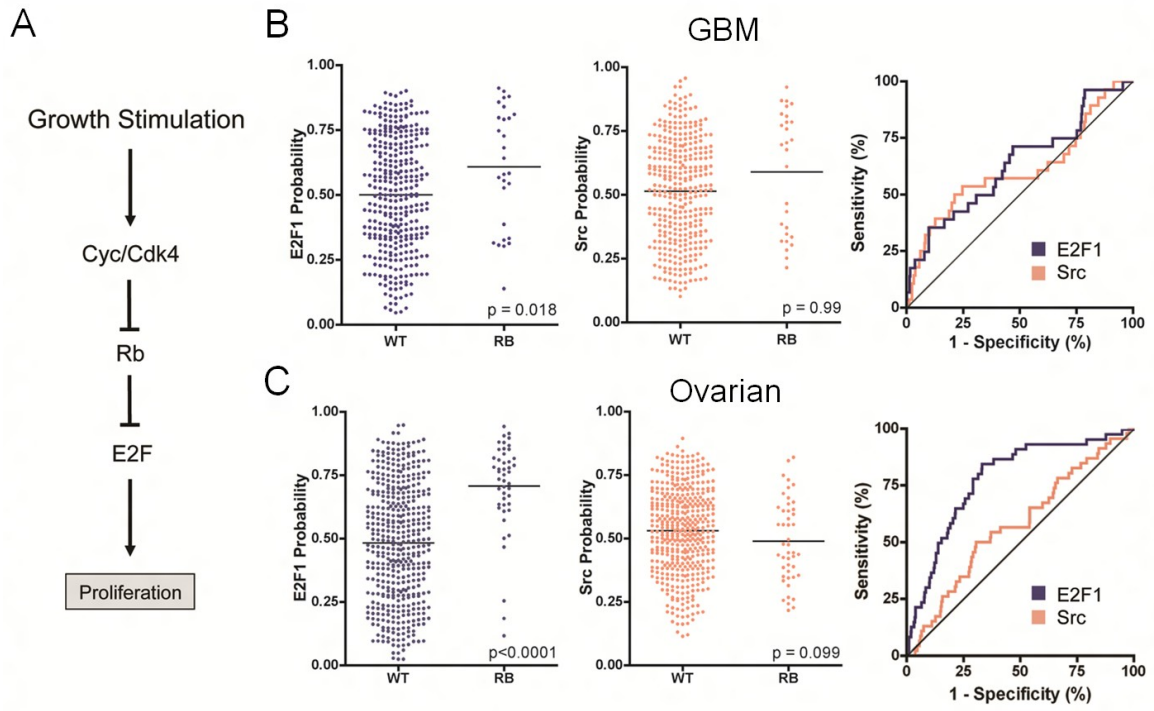

Figure 24: Prediction of RB1 **status with an E2F1 and a Src signature.**

A) The schematic depicts the organization of the RB1 pathway and the role of RB1 in controlling E2F transcription factor activity. B-C) Predictions of E2F1 and Src pathway activity in either GBM (B) or ovarian (C) TCGA datasets. Probabilities of E2F1 (left) or Src (middle) pathway activity are plotted as a function of RB1 gene status, along with and ROC curve (right) for both E2F1 (blue) and Src (red). *P*-values shown are for a Wilcoxon rank-sum Test.

Taken together, these analyses illustrate that in order for *in vitro* derived signatures to be applicable to human tumor samples, the process must properly adjust or account for variation in the gene expression characteristics from the two biologically distinct settings (cell lines vs. patient tumor samples).

We have also assessed the extent to which the predictions are specific by using a second pathway signature, one developed to measure activation of the Src pathway. As shown in Figure 24, applying this signature to the same datasets fails to discriminate between samples that were either wild type or RB1 null in GBM tumors ($P = 0.998$, Wilcoxon rank-sum; AUC = 0.608, ROC analysis) as well as in ovarian tumors ($P = 0.099$, Wilcoxon rank-sum; AUC = 0.574, ROC analysis), where the association with E2F1 predictions is more dramatic.

3.3.2 A general strategy to evaluate expression signatures in prospective clinical studies

The ability of expression signatures to predict pathway activity in tumor samples, together with past work that has demonstrated a link between prediction of pathway activation and sensitivity to pathway-specific therapeutics (45, 216), creates a potential opportunity to use these tools to guide the use of new investigational drugs. For instance, a pathway signature might be employed as an inclusion criterion to enrich for patients likely to respond to a pathway-specific drug. However, the requirement for data normalization as we have highlighted here, to accommodate the distinctions of the

expression data from *in vitro* training datasets and *in vivo* test datasets, presents a significant challenge for application of a signature in a prospective clinical study considering the fact that in a prospective study, one patient sample must be analyzed at a time. As such, this presents a unique context for carrying out the predictions, since the full data set will not be available for normalization.

To address this challenge, we have developed a strategy, as outlined in Figure 25, which makes use of three sources of information: the training dataset that derives from the cell lines identified as pathway off and pathway on, a reference dataset that provides the tumor- and analysis center-specific gene expression context, and the one test sample.

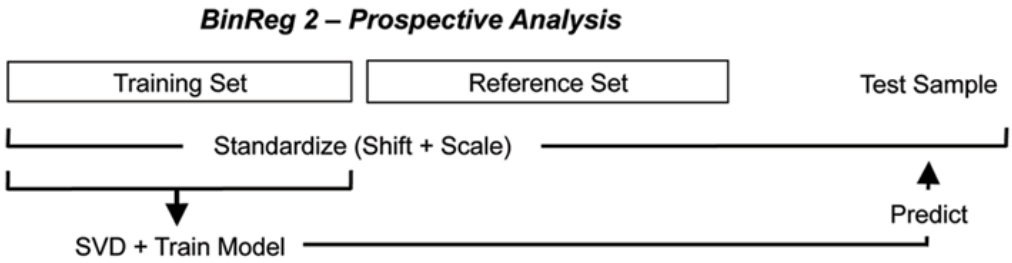

Figure 25: Framework for prospective analysis of genomic signatures

Schematic depiction of the framework for implementing genomic signatures in a prospective framework. A static reference dataset is included in the model to allow standardization of test samples that are independently preprocessed. The test sample is normalized using shift+scale standardization, and then the model is build on training samples and project back to the test sample.

To derive the reference data and test data, we used the TCGA glioblastoma dataset. An initial analysis used nested subsets of 10, 30 or 100 samples selected at random to serve as the reference dataset and another 50 randomly selected samples were used as the test data. Each of the 50 test samples was normalized independently, as would likely be the case in a clinical setting. We then ran 50 permutations of this analysis, randomly selecting nested reference sets and non-overlapping test samples. Using the BinReg2 methodology and adjusting the mean and variance within the reference set and the one test sample to match that of the training data, we evaluated the extent to which the size of the reference dataset affected the predictions when comparing results performed one sample at a time versus as a batch retrospective analysis. While larger reference datasets did improve the concordance of the simulated prospective predictions with the retrospective predictions ($P < 0.01$, Repeated measures ANOVA), the prospective analysis using a 10-sample reference sets still showed strong concordance with the retrospective analysis ($R^2 > 0.75$; $R^2 = 0.901 \pm 0.00615$, mean ± SEM) (Figure 26).

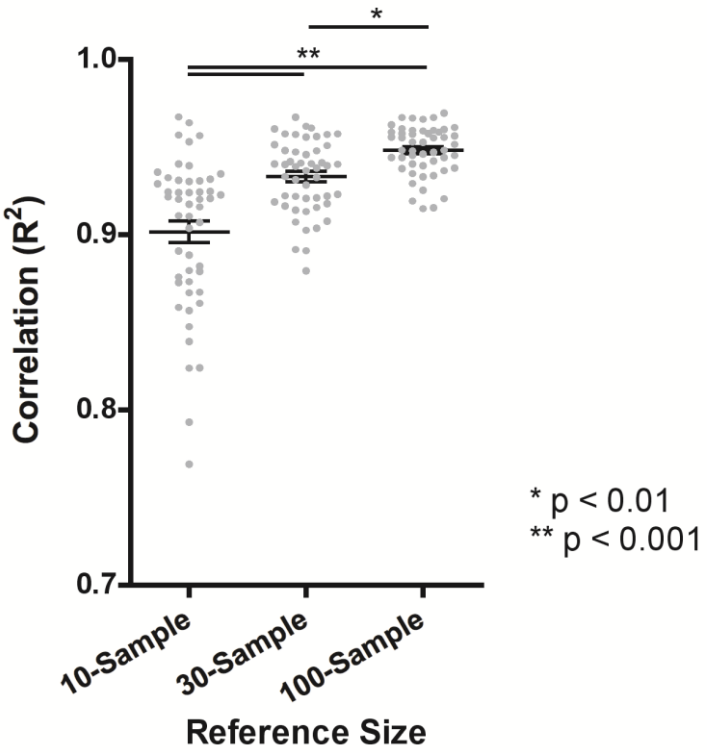

Figure 26: Effect of size of reference dataset on accuracy of predictions.

Concordance of pathway predictions for each of the samples in the test dataset plotted as a function of predictions carried out as a batch or in prospective simulation using a nested reference set of 10, 30 or 100 additional samples. R-squared values are shown for the Pearson correlation between the prospective simulations and batch predictions for 50 permutations of randomly sampled nested reference and non-overlapping test samples. Significant increases in R-squared values were determined using the repeated measures ANOVA.

To evaluate the ability of the prospective analysis framework to predict RB1 status, we used both the ovarian and GBM datasets. The reference dataset was again comprised of 10 randomly selected samples and the remaining samples were used as the test set for a prospective simulation using the E2F1 pathway signature. This analysis was also repeated 50 times, and assessing both the concordance of E2F1 pathway signature predictions and the significance of differences between the RB1 altered or wild type samples for both GBM and Ovarian samples. As was observed with the full batch analysis, we again observe a significant discrimination between those tumors that are wild type for RB1 versus those with an RB1 mutation or deletion in GBM ($P = 0.0159 \pm 0.00177$, Wilcoxon rank-sum, mean ± SEM) and ovarian datasets ($P = 6.62\text{x}10^{-8} \pm 1.86\text{x}10^{-8}$, Wilcoxon rank-sum, mean ± SEM) (Figure 27, Figure 28). We note that during this analysis RB1 altered samples had significantly higher E2F1 signaling in all of the permutations run on the ovarian samples and all but 2 of the permutations of the GBM data ($P > 0.05$), however in both of these cases the P-value approached significant at $P < 0.06$. A comparison of the consistency of the predictions of pathway activity for the samples, assessed either by batch or in the simulated prospective analysis, indicates a very close concordance for GBM and ovarian tumors with mean R^2 values of 0.8995 ± 0.00556 and 0.9194 ± 0.00316, respectively.

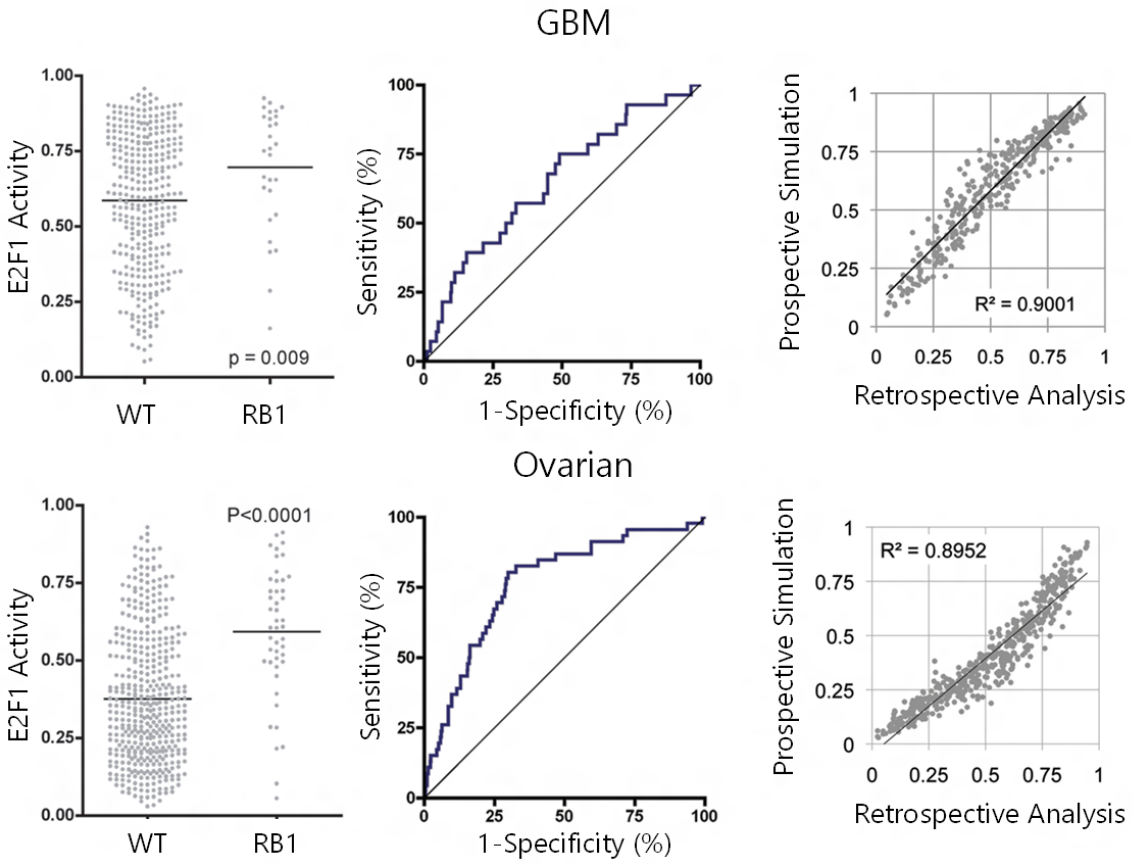

Figure 27: Prospective analysis of GBM and Ovarian E2F1 pathway activity

The predictions of GBM (top) or ovarian (bottom) TCGA samples, carried out one at a time, are plotted as a function of RB1 mutation status (left panel). An ROC analysis for the predictions is shown in the middle panel. The right panel depicts the concordance in predictions of E2F1 pathway activity in the batch set of samples or samples predicted one at a time. *P*-values shown are for a Wilcoxon rank-sum test, and R^2 Pearson correlation coefficients are shown.

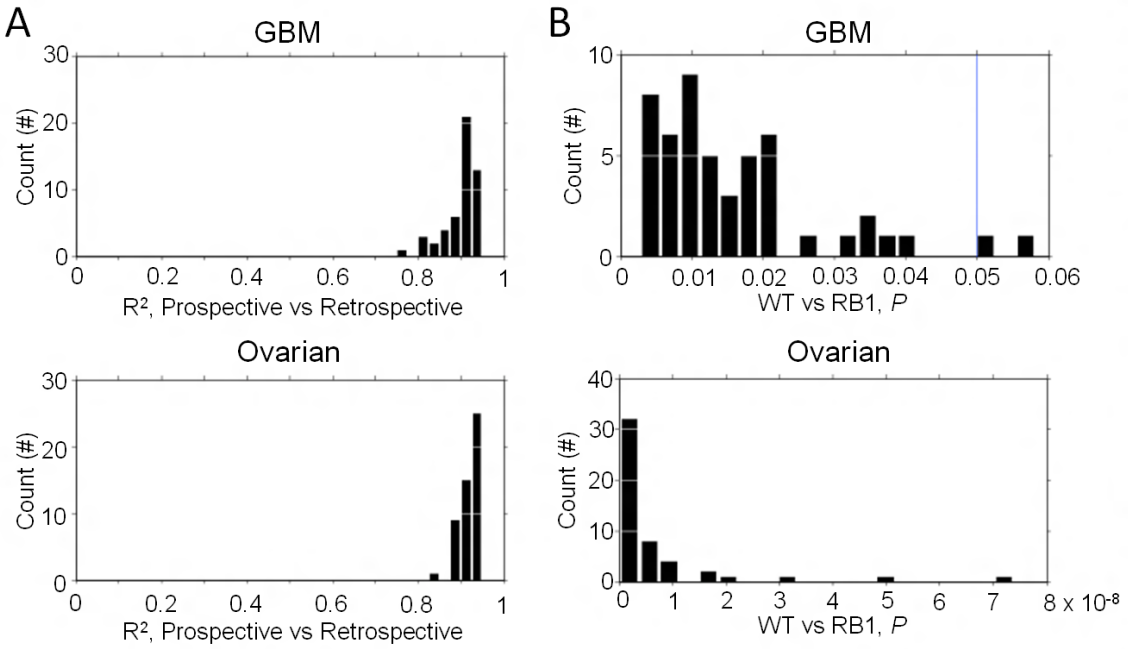

Figure 28: Permutation testing of prospective analysis

Permutation testing of the A) the concordance between retrospective analysis and prospective simulation and B) the significance of the difference in E2F1 pathway activity between RB1 altered and wildtype samples during prospective simulations. 50 permutations selecting a random 10-sample reference set for the remaining samples of GBM (top) or ovarian (bottom) tumors. Concordance is measured as the Pearson's correlation coefficient (R^2), and significance between RB1 altered and wildtype samples was determined using the Wilcoxon Rank-Sum test. The blue line indicates a *P*-value of 0.05.

Beyond the impact of sample size of the reference set, we also investigated whether the batch source of samples used in a reference dataset would affect the concordance of these results. We found that when using 10 reference samples from the one single batch in the TCGA set of samples, there was a noticeable shift away from concordance in the GBM predictions when compared to a reference that was a mix of the batches ($R^2 = 0.6704$, Figure 29).

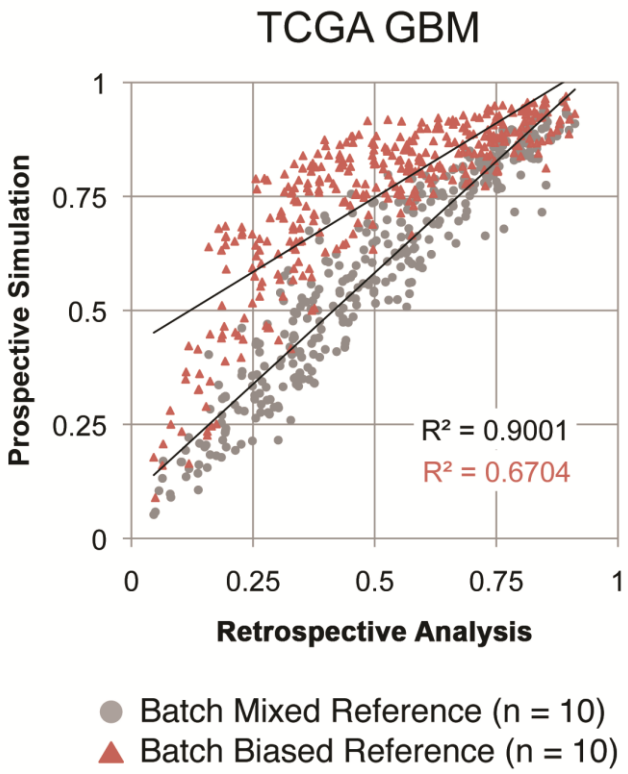

Figure 29: Batch effects on prospective simulation of pathway predictions.

Concordance of pathway predictions for each of the samples in the GBM test dataset plotted as a function of predictions carried out as a batch or one at a time using a reference set of 10 additional mixed-batch (gray circles) or single-batch (red triangles) samples. R-squared values are shown for the Pearson correlation between the prospective simulations and batch predictions.

This challenge, along with the fact that few clinical trials will be able to collect 30-100 samples for a reference dataset provides a significant hurdle to implementing these methods in a prospective setting. However, previous research has demonstrated that reference datasets do not necessarily need to be overlapping to produce concordance RMA values, and as such a distinct unrelated dataset could theoretically be used to address these concerns (62, 63). To overcome the concern about a batch-biased reference dataset and the challenge posed by needing to collect a large number samples, we investigated whether it was possibly to employ publicly available samples to serve as an appropriate reference dataset during prospective analysis. To accomplish this, we compared samples from 5 different GBM datasets arrayed on U133A standard arrays, independently preprocessing the samples in one of the datasets while using all four of the remaining datasets as a reference. This allowed us to determine whether a non-overlapping, but batch-mixed dataset could overcome the skewing induced by using a single batch as the reference (Figure 30). We also investigated whether a single non-overlapping dataset could serve as an appropriate reference. This analysis demonstrated that a single dataset with sufficient variation and a large enough number of samples could serve as an appropriate reference dataset, although a mixture of these datasets generally performed best (Figure 30Figure 31).

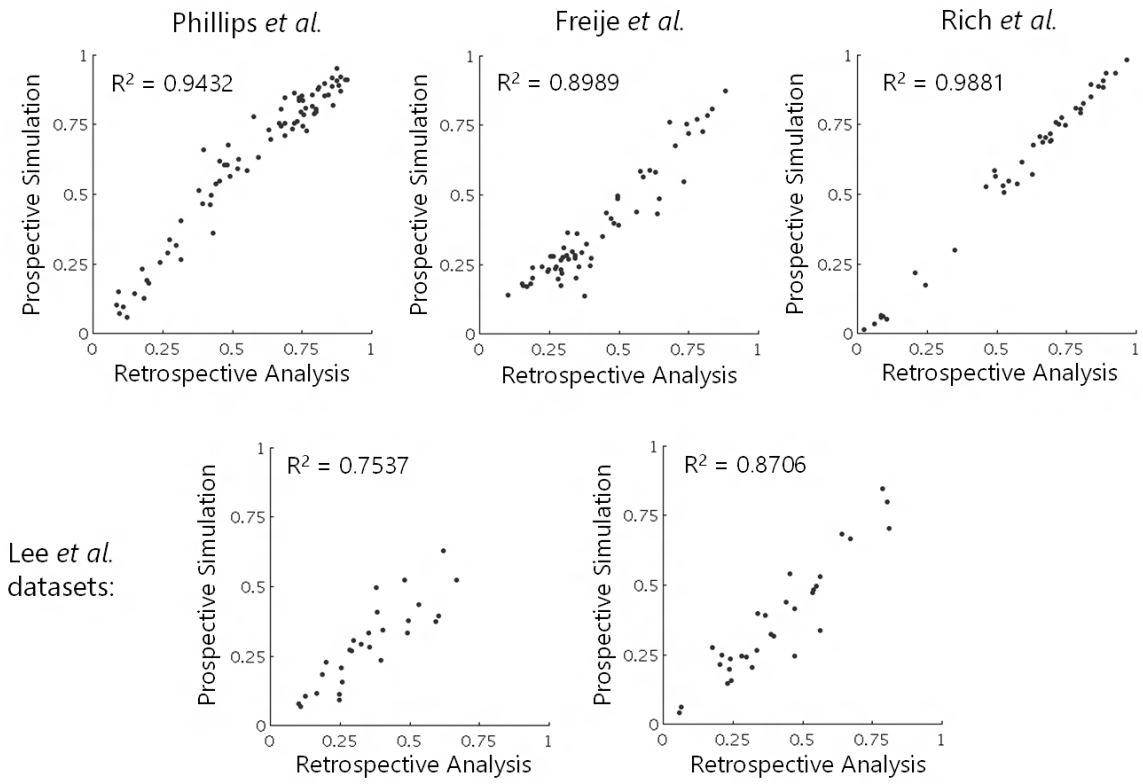

Figure 30: Effect of independent GBM reference datasets on the concordance of simulated prospective predictions.

Concordance of E2F1 pathway prediction for each of the samples in a given test dataset plotted as a function of predictions carried out during retrospective analysis or prospective simulations using a reference set built four unrelated GBM datasets processed together. R-squared values are shown for the Pearson correlation between prospective simulations and batch predictions.

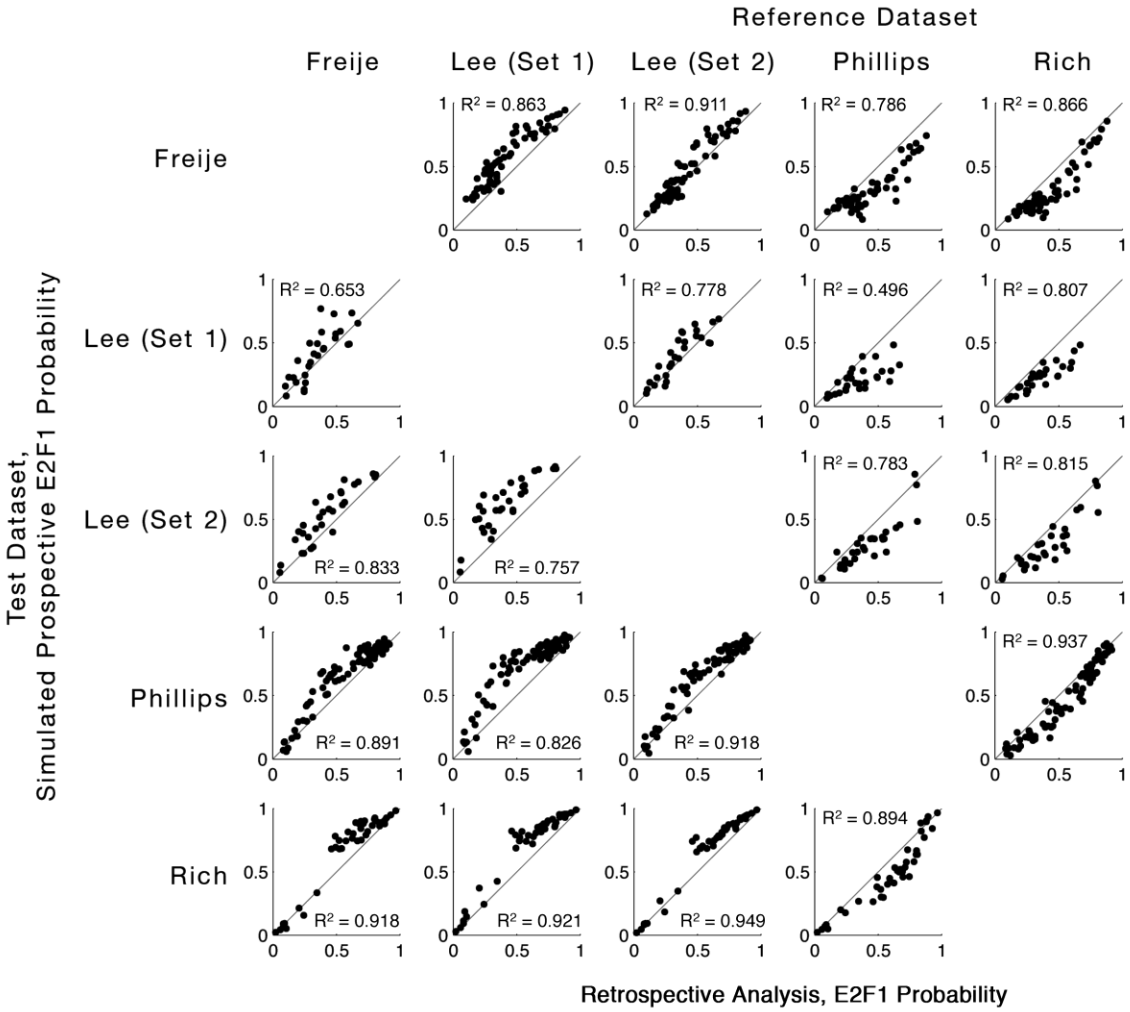

Figure 31: Effect of using a single independent GBM reference datasets on the concordance of simulated prospective predictions.

Concordance of E2F1 pathway prediction for each of the samples in a given test dataset plotted as a function of predictions carried out as a batch or during prospective simulations using a reference set built each of the four unrelated GBM datasets or all four datasets processed together. R^2 values are shown for the Pearson correlation between prospective simulations and batch predictions. The line of identity is shown in for reference purposes.

Finally, to determine if a bias in an important biological variable would result in a similar skewing of predicted pathway activity, we evaluated a series of breast cancer samples using either an independent ER+ or an ER- dataset as the source of the reference data. As shown in Figure 32, we found that there is a strong concordance between the predicted E2F1 pathway status when either an ER- reference or an ER+ reference dataset is used, regardless of which original dataset was the source for the reference ($R^2 = 0.987$ and 0.856). We note that when using the Ivshana *et al.* dataset as a reference, we do see predictions shift away from a linear relationship, with a number of the probabilities being significantly higher when standardized using the ER+ reference (Figure 32).

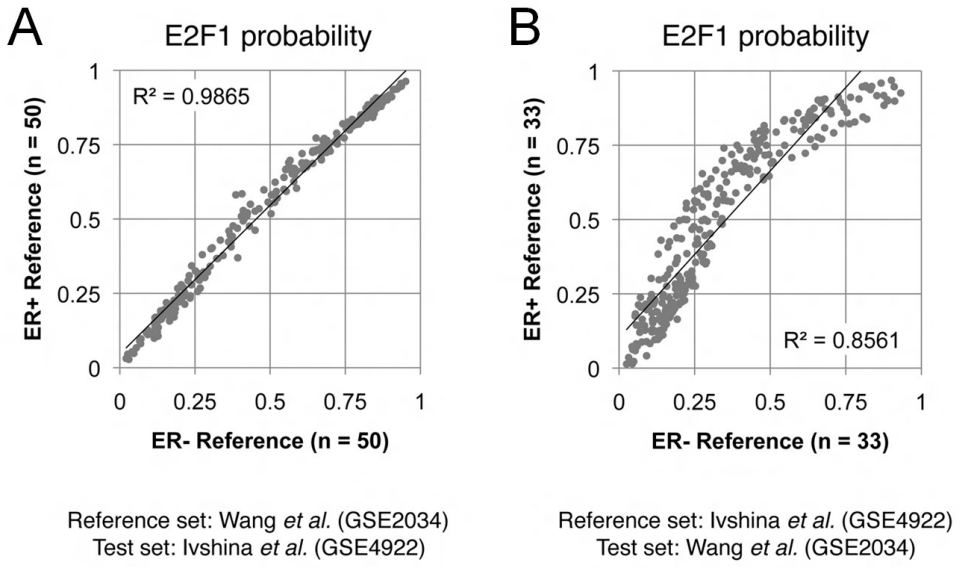

Figure 32: Effect of ER status in batch-independent breast cancer reference datasets on the concordance of simulated prospective predictions.

Concordance of pathway predictions for samples in a given test dataset plotted as a function of predictions carried during prospective simulations using a reference set of either only ER+ or only ER- breast cancer. A) Test dataset of GSE4922 was processed using a reference set of either ER+ or ER- samples from GSE2034. B) Test dataset of GSE2034 was processed using a reference set of either ER+ or ER- samples from GSE4922. R^2 values are shown for the Pearson correlation between predictions using ER+ or ER- reference sets.

The fact that this type of effect is not consistently observed when using a biologically biased reference dataset argues that the shift away from concordance is likely due to batch related effects rather than biological biases in the dataset. To further determine the likelihood that a biased reference set will influence the outcome of prospective analysis, we examined a reference dataset composed of only samples that were *a priori* known to have either low or high predicted pathway activity (Figure 33),

but were composed of samples from multiple batches. In this analysis we also found that there was significant deviation from concordance between retrospective analysis and prospective simulations (R^2 = 0.6055, 0.4934). However, it should be noted that during the permutation testing reported above, we did not encounter a single case where the correlation was as this weak, thus suggesting that obtaining a reference dataset that is extremely skewed would be an extremely rare case, and that larger accrual or incorporation of multiple historical datasets would overcome this concern.

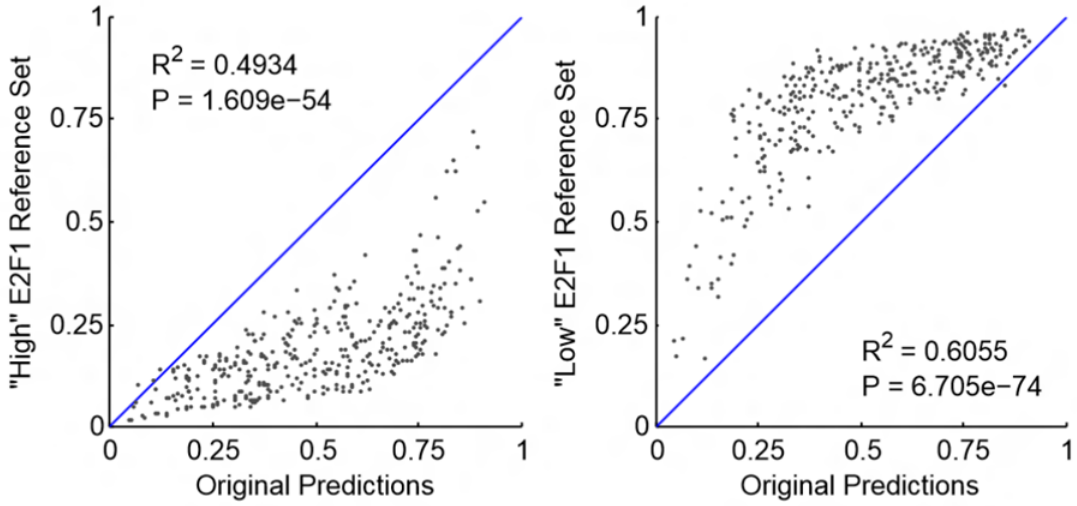

Figure 33: Pathway biased reference datasets compromise prospective predictions

Concordance of E2F1 pathway prediction between retrospective analysis (x-axis) and prospective simulation (y-axis) using either a reference dataset selected to contain only samples with E2F1 pathway activity > 0.75 (left) or < 0.25 (right). R^2 and *P*-values are shown for the Pearson correlation. The line of identity is shown in blue for reference purposes.

Based on these results, we can conclude that it is possible to use the framework outlined here for applying genomic signatures derived from cell line models in the context of a prospective clinical study. With an appropriately diverse reference dataset, we find that this framework for prospective analysis accurately reflects the application of these signatures in a retrospective batch analysis.

3.4 Discussion

The challenge of achieving the goal of personalized cancer therapy is daunting, driven in large part by the enormous complexity and heterogeneity of disease. It is well recognized that cancer is not one disease but many, each requiring a therapeutic regimen tailored to the characteristics of the disease process. To further magnify the challenge, it is likely that effective cancer treatment regimens will require not single agents but combinations of agents that can target the vulnerabilities presented by that particular cancer. Present day treatment guidelines for most cancer modalities have been derived from clinical trials that show benefit across a population of patients without regard to the inherent heterogeneity of the disease presented by these patients. As such, the benefit for any individual patient may not reflect this overall result and the benefit for an individual patient may be minimal. In essence, this is the challenge of personalized cancer medicine, the capacity to progress beyond studies of drug efficacy

in mixed populations to trial designs that allow the development of drugs that are matched with the individual patient.

Many have argued that the use of gene mutations or specific gene alterations such as copy number alterations will be the mechanism to most effectively achieve the goal of personalizing therapy. In other cases, gene signatures that involved complex statistical models have been evolved to simple PCR assays of single genes or a limited number of genes (221). While these forms of assay are conceptually simple and easy to interpret, we believe the more complex assays such as those described here will also play an important role in being able to more effectively address the complex nature of the biology confronted by cancer states.

A critical aspect of initiating prospective studies to evaluate the performance of the signatures is to provide the opportunity to define the conditions and procedures necessary for implementation in day-to-day clinical practice. These include the ability to confirm the feasibility of conducting quality control of the tissue specimen and microarray data (60), evaluate the technical variability of microarray data (222), recognize the necessary step of assessing the accuracy and precision of signatures in replicate tumor samples (223), to report genomic results in a timely manner, and to overcome the challenges discussed here when using *in vitro*-based expression signatures in the context of predicting a single patient sample. With respect to normalization procedures between disparate sources of gene expression data (cell lines vs. human

tumor data), we show here that if the appropriate measures are taken these methods have the capacity to adjust for these differences such that gene expression signatures derived from cell culture samples can be applied to the analysis of patient tumor samples.

Finally, we note that the ability to create a framework for prospective predictions using a cell line derived signature can be influenced by the nature of the samples used as the reference dataset. Perhaps not surprising, we found that substantial shifting occurs when using array data from disparate array platforms, such as when attempting to use standard U133A array data to serve as a reference for high-throughput U133A arrays (Figure 34). A more subtle effect is seen where a single batch of samples was used as a reference for normalization, leading to some skewing in the pathway predictions between the prospective simulation and the retrospective analysis. Finally, there was also some evidence for a modest potential bias as a function of a biological variable such as ER status (Figure 32).

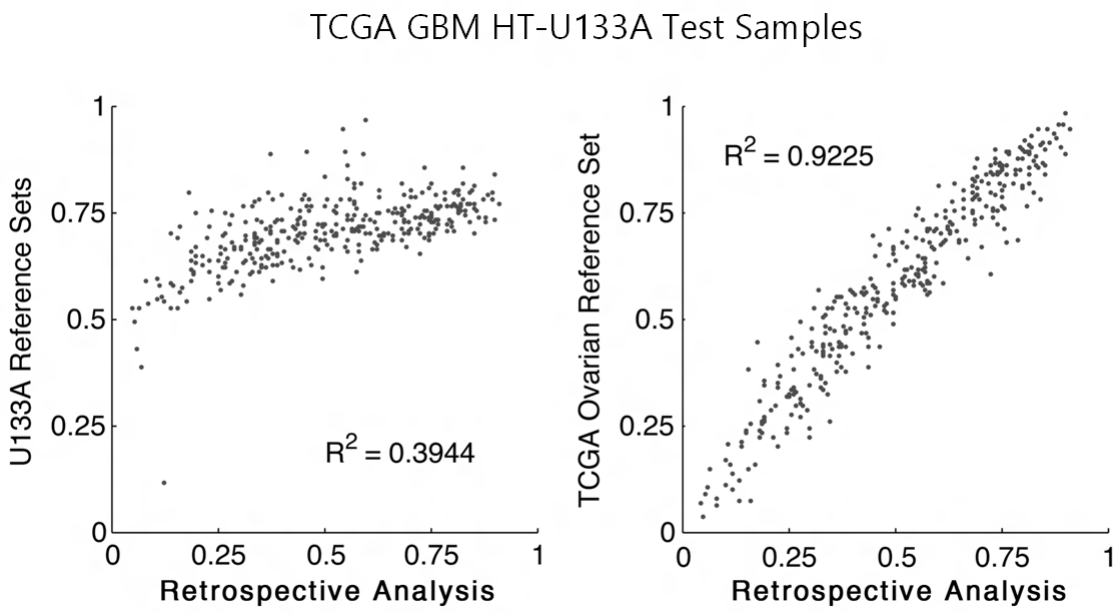

Figure 34: Impact of reference array platform on prospective simulation

HT-U133A GBM samples were used as the test samples for retrospective analysis or prospective simulation relying on a reference dataset that was either built from standard U133A arrays (left) or from HT-U133A arrays (right). Concordance of predictions is shown as the Pearson correlation coefficient (R^2).

Ideally, one would strive to utilize a set of samples as the reference dataset that were closely matched to those expected to be collected prospectively. For instance, this might involve an initial lead-in collection of pre-trial samples from equivalent patients to be enrolled in the trial, facilitated by the fact that even small reference sets can provide adequate information to produce robust results. But, it is also true that a reference dataset composed of multiple batches from unrelated and non-overlapping datasets can produce results that are highly concordant with a retrospective analysis. Given that

accrual during a clinical trial may challenge the ability to collect sufficient numbers of samples to generate a robust reference dataset, the ability to leverage the wealth of public gene expression data to generate more universal reference datasets may ultimately be the most robust approach. Importantly, these results provide a framework to take advantage of the power of *in vitro* experimental studies that can accurately define cellular states such as pathway activation that can then be translated into tumor analyses, including in the context of prospective clinical studies.

3.5 Materials and Methods

Pathway signature and GenePattern Implementation.

The development of genomic signatures predicting activation of various oncogenic signaling pathways, including the E2F1 pathway, has been described in previous work (45, 218). The parameters for development of an E2F1 and a Src pathway signature were described previously (16, 219). Briefly, in this analysis we used BinReg 2.0 version 2 (BinReg 2) (224, 225). The E2F1 signature was built using the default parameters of 150 genes with 2 metagenes, and the Src signature was built using (the default) 85 genes with 3 metagenes. All other parameters for analysis were also left to their default settings in the GenePattern module. Training data files were RMA normalized, and are provided as Supplemental Tables 1 and 2.

We have recently described a computational platform making use of the Broad Institute GenePattern framework that combines these in user-friendly interface, using the principals previously established (219). The BinReg statistical algorithm is one of the components of the gene expression signature analysis platform collectively termed SIGNATURE and available from a GenePattern implementation[2]

To allow this framework to prospectively predict pathway signatures for individual samples, the GenePattern module was modified to allow the user to load a Reference dataset (Supplemental Figure 1). The normalization reference should contain a dataset processed using the same normalization method (RMA or MAS5) as both the training dataset and the test dataset. The reference is then combined with one of the test samples, and normalization (quantile or shift + scale) proceeds on that combined dataset. This is repeated for each of the test samples, ultimately resulting in a test set where each sample was normalized independently.

TCGA test datasets.

Alteration status of RB1 was downloaded from cBioPortal[3] for ovarian cancer samples (n = 488). Data used to impute RB1 status for glioblastoma (GBM) Samples (n = 368) was downloaded from the TCGA web-portal[4]. RB1 was classified as altered for

---

[2] https://genepattern.uth.tmc.edu/gp/pages/login.jsf
[3] http://www. cBioportal.org/public-portal/index.do
[4] http://tcga-data.nci.nih.gov/tcga

GBM samples with a homozygous deletion, defined as a log2 ratios of < -0.8 on the SNP6 platform, or with a validated non-silent mutation in RB1. Level 1 Affymetrix HT-U133A expression .CEL files for ovarian and GBM samples were downloaded from the TCGA web-portal and RMA normalized using Bioconductor[5] in R[6], batching by tumor type. Only the first available .CEL file for a given patient was analyzed, all other replicate .CEL files were removed from analysis.

Reference datasets and prospective simulation.

We RMA normalized nested sets of 10, 30 or 100 random GBM samples from our test dataset to serve as a reference. A non-overlapping random subset of 50 samples was chosen for prospective analysis simulation. Each .CEL file for the 50 test samples was RMA normalized independently and then the 50 resulting files were merged into a single test dataset. Reference samples, test samples and signature files were loaded into the GenePattern module for prospective simulation, and run with inputs described previously. This analysis was repeated 50 times, randomly selecting the nested reference and non-overlapping test sets. Once the size of the reference was analyzed, we extended the analysis to the full GBM or ovarian datasets, RMA normalizing a random set of 10 samples to serve as the reference and independently processing and merging the

[5] http://www.bioconductor.org
[6] http://www.r-project.org

remaining samples as described above. This analysis was also repeated 50 times, randomly selecting 10 samples to serve as reference for the remaining samples.

To analyze the effects of a biased reference set on prospective simulations, we used 10 GBM samples from the same batch as a reference for the remaining GBM TCGA samples as described above. We also used Affymetrix U133A GBM expression data from four publications (19, 112, 177, 226), representing five distinct batches. Samples from each dataset were independently normalized and merged to create test datasets as described above. We RMA normalized data from the remaining datasets together as the normalization reference, loading each dataset and the accompanying reference datasets into Gene Pattern. Results were compared to the predictions generated when all five datasets were RMA normalized together and run using BinReg2 + Shift/Scale.

To determine the effect of using a biologically biased reference, we analyzed two previously published Affymetrix U133A breast cancer expression datasets (GSE2034 and GSE4922) for which estrogen receptor status was annotated (227, 228). We selected either 50 (GSE2034) or 33 (GSE4922) random ER+ and 50 (GSE2034) or 33 (GSE4922) random ER- tumors to serve as a normalization reference datasets. ER+ samples were RMA normalized together, and ER- samples were RMA normalized together. CEL files from the non-reference dataset were normalized individually, merged and loaded into Gene Pattern along the training files and either the ER+ or ER- reference set as described above.

Statistical Analysis.

Performance of the genomic signature was assessed using a Wilcoxon rank sum test and Receiver-Operator Characteristic (ROC) analysis. Concordance among retrospective and prospective applications of the model was evaluated using Pearson correlation coefficient, R^2. Values are presented as the mean ± standard error of the mean where appropriate.

4. Concluding Remarks

The past decade of research into the genomic underpinnings of malignant gliomas has demonstrated that glioblastomas, like many other cancers, are a heterogeneous collection of diseases that results from numerous potential alterations and combinations of those alterations (19, 54, 73, 112, 166, 169). This disease, by any definition, represents a complex setting where tumor development and malignant transformation can be viewed as having numerous potentially transforming alterations that are redundant to each other. The idea that these tumors would be dependant on the signaling of an initiating alteration has thus far not translated to an effective means to manage these tumors for long-term care. In several cases tumor recurrence has been demonstrated to occur by shifting the genomic resources used by the tumor either down the signaling cascade to drive the same signaling pathway, or to an analogous pathway that continues to promote tumor growth (19, 93).

The work described in this thesis aims to address some of the critical concerns currently facing the cancer genomics community in the setting of malignant gliomas. While we are increasingly aware of the diverse molecular subtypes that are present in these tumors, the assumption that this is the only means to investigate and search for drivers of tumorigenesis overlooks the fact subtype-related alterations are rarely exclusive to one type, and that using different genomic sources to develop subgroups does not completely recapitulate the previous definitions of this disease. Even in other

brain cancers, such as medulloblastomas, the process of tumor development, while heterogeneous, produces more molecularly distinct clusters of tumors and occurs in clinically distinct populations of patients (18). This underscores the need for other methodological approaches to understand tumor signaling in the setting of glioblastoma, and provides weight to looking at this disease as a continuum rather than discrete sets of tumors.

The research described here provides a general mechanism to identify candidate genes involved in the tumorigenic process of signaling deregulation as well as alterations associated with clinical outcome and response to therapeutic intervention. The pathway association framework we apply to glioblastoma could just as easily be extended to any of the numerous tumor settings that the TCGA or other large-scale genomic studies where there are paired genomic data sources. Moreover, this framework has proven its capacity not only to identify novel alterations associated with pathway deregulation, but to detect significant loci that other methods for studying copy number alterations fail to identify. It is precisely because we integrate pathway signaling with copy number alterations that we are able to identify and annotate these alterations, and this integration is a key component to the utility of this approach. However, it should be noted that follow-up validation efforts, such as *in vivo* or clinical analyses are still necessary to understand the role of these alterations, as the data present in the TCGA provides a starting place for further studies.

Additionally, this thesis has described a significant methodological extension of these genomic signatures by developing a means to apply these methods in the prospective clinical trial setting. While other research has focused on producing robust expression values for individual samples, we have demonstrated that even with robust values, the ability to normalize a training model with prospective samples still presents a barrier to implementing these methods. The research presented here has demonstrated that not only is it feasible to robustly predict the activity of a pathway signature in the prospective setting, but that the historical datasets already generated in many disease settings provide enough genomic variation to accomplish this without needing to use clinical samples from the trial itself. This in and of itself, represent a large step forward for the field, as numerous methodologies for classifying genomic samples exist beyond the pathway signatures that we use, and will likely also benefit from these types of approaches.

4.1 Extending genomic signatures of pathway signaling

4.1.1 Deeper integration and network analysis

One of the key areas that we can move these analyses forward is by expanding the contexts in which we utilize pathway signaling as a quantitative phenotypic measurement. While this thesis has focused on linking copy number alterations and mutations to pathway activity, we can similarly approach other key regulatory

infrastructures within the cell. We have already begun to investigate the role of microRNAs that are deregulated in GBM (Figure 36 in Appendix B), and the same analytical frameworks could reasonably be applied to protein expression and epigenetic alterations. However, one of the key features afforded by copy number alterations is the ability to identify which alterations are most likely important based on the linkage between adjacent loci. While there may be similar information that can be used in epigenetic analysis, the boundaries of chromatin modification are likely much more fluid and discrete than the chromosomal boundaries created during tumorigenesis. Therefore, it is likely that other analytical approaches will be necessary to further extend our understanding of signaling deregulation in this complex setting.

In the particular case of microRNAs, one of the key features that may lead to a better understanding of the role these short non-coding RNA sequences play is the ability to calculate the likelihood that they target a particular mRNA based on homology and sequence conservation (229). In fact we note that careful examination of microRNAs expressed in GBM tumors shows that pathway signaling can act directly as a filter to identify candidate microRNAs that appear to be exerting a stronger effect on their predicted downstream target compared to analyses of overexpression alone. Clearly this information, like the physical proximity when analyzing copy number alterations, provides a means to better understand the functional impact of microRNAs while

informing the next experimental steps to take to unravel the role these short non-coding RNAs play in the tumorigenic landscape.

While predicting target activation adds a key element to these analyses, full integration across datasets is clearly the ultimate goal of a project such as the TCGA. Combining the diverse genomic datasets will require more flexibility than our current model affords, which is a relatively simplistic implementation of the association framework. While we have clearly demonstrated the ability to identify numerous clear and well characterized pathway regulators, many of which are overlooked by other approaches, we have yet to fully encapsulate the diversity of genomic data that we have access to. To extend these models, a key area to consider is building a networked framework where we can imply or require additive, synergistic, or inhibitory affects between the different types of genomic data so that we can better understand the alterations that leverage pathway signaling in this disease setting and in others.

4.1.2 Cytomegalovirus signaling in GBM

One of the intriguing findings that occurred during the analysis of microRNA expression was the identification of significant associations between non-human microRNAs and pathway signaling. In addition to assaying human microRNAs, the Agilent array chip used by the TCGA also assays microRNAs encoded by Herpes Simplex Virus 1 (HSV1), Cytomegalovius (CMV), Ebstein-Barr Virus (EBV), and Kaposi

Sarcoma Virus (KSHV). While several of these viruses have been well characterized for their contribution to tumorigenesis in the context of other human cancers, in the setting of GBM it has been CMV that has received the most attention, with conflicting reports that detail a strong association of CMV with grade or the inability to identify viral genomic content in these tumors (230).

The fact that CMV microRNAs were associated with the activity of several signaling pathways (Figure 37 in Appendix B) was of interest during the analysis of TCGA microRNA platform. Especially intriguing was the fact that Stat3 signaling was among the strongest associations identified, consistent with recent work linking CMV signaling and Stat3 activity in glioblastomas (231). To better understand the relationship between these microRNAs and GBM characteristics, we investigate microRNAs encoded by CMV, to determine whether there were any clinically relevant findings associated with CMV transcripts. We were able to find a robust association of CMV-microRNA-derived clusters of patients with the previously described human mRNA subgroups, as well as a robust association between one of the microRNAs, hcmv-miR-UL148D and patient outcome (Figure 38 in Appendix B).

The results of these experiments are ongoing, but the initial associations that we have described, as well as apparent clinical affect of these microRNAs, provides a fertile ground for future studies. While the role of CMV has not been conclusively proven in GBM, and there are clearly conflicting reports detecting virally encoded genes in this

setting, evidence is accumulating that supports some interplay between reactivation of this virus and tumor signaling. The genomic analysis described here provides an initial support for a role of viral signaling in GBM deregulation. Clearly there is a need to further expand these studies, and while we have started to analyze patient samples using alternate methodologies, there is much work still left to be done to demonstrate that the viral microRNA signaling that we have detected in the TCGA reflects real biological results and not cross-hybridization of human sequences to these probesets.

4.1.3 Factor analysis as a statistical extension to the signature framework

The other clear means to extend these analytical approaches is to develop novel models of pathway signaling that are directly relevant to brain cancers. This could take the form of developing astrocytic or neural-stem cell based signatures of alterations that we do not have in our library of signatures to date, such as PDGFRA. This represents the traditional approach to pathway signatures, and would create unique, and biologically-well defined signatures to understand this disease further. However, given the heterogeneity of this tumor setting, the best *in vitro* model to use for this purpose is debatable, and the fact that we have demonstrated the utility of signatures derived in other heterologous backgrounds argues that in some cases we would not need to use a brain tumor specific background for experimentation. Still, there are likely key features of tumor signaling that we have yet to capture in this setting, and expanding our

repertoire of genomic approaches would provide further insight into the disease processes governing GBM tumorigenesis.

An alternative approach that could be used to expand our repertoire of pathway signaling models would be to use regression based analyses to develop glioma-specific signatures. Our lab has already implemented Bayesian Factor Regression Modeling (BFRM) approaches that use the currently available signatures to develop context-specific factor models (232). The idea here is that the signatures that we have developed, while important for capturing the dominant aspect of a signaling pathway *in vitro*, may provide the necessary information to start to understand more refined components within a pathway. Factor modeling, in this regard, provides a mechanism to develop multiple distinct quantitative models based on two data sources: the genes that are contained in the signature, and the expression data present in a context-specific dataset. The regression approach uses the signature genes as a seed to build new factor models which are not orthogonally constrained, but which capture significant variation within the target dataset related to the original seed of the signature genes. This process involves an iterative factor regression approach, and thus starts with the input seed of genes, and then evolves that list of genes into the ones that compose the final factor models.

Work described by Chang *et al.* has demonstrated that the BFRM approach can not only can evolve a signature into its underlying component parts from a statistical

perspective but that these components identify the expression signaling patterns related to the control of specific downstream signaling elements within a pathway (48, 233). This idea, that factors built upon an expression signature reflect the core signaling arms that make up a pathway, has the ability to refine the information that we already have available in the pathway signatures. It is a powerful extension of the pathway framework that adds further specificity to a pathway signature.

This method also provides a means to add context to a given signature that may be initially lacking. Work by Erich Huang in the lab has demonstrated that factor-based models built from a given signature can outperform the original signature when predicting drug response, in this case for EGFR and erlotinib as well as Src and dasatinib. While this would be expected, what was intriguing about these results is that distinct factors associated with drug response in different tumor settings, even though they were developed concurrently. This tumor-type specificity is an attractive feature of factor development, and could prove useful not only in the glioma setting, but also in other settings similar hypotheses about clinical response to targeted therapeutic intervention are currently a key area of research.

While we have demonstrated that most of the signatures appear to perform reasonably well in the reduced GBM-specific expression dataset, evolving factors that are related to the underlying expression patterns present in glioblastoma would provide a more thorough and insightful understanding of this disease setting than relying on the

original signatures alone. Moreover, further analytical approaches, such stochastic shotgun search provide us the ability to build combinatorial models integrating multiple different factors to better model a phenotype of interest (234). These methods may prove useful for identifying sets of factors that on there own contribute to response to therapeutic intervention, or disease progression, but which only account for a fraction of the variation found within that phenotype.

With all of the benefits mentioned, the primary drawback to this approach is that while we understand the biological context in which the signatures themselves are generated, it will not be directly apparent initially as to which facets of glioma signaling we capture by evolving factors in a glioma-specific dataset. While it can be expected that the details that a quantified by this method will relate to the original signature, as demonstrated above, the exact underlying meaning of this quantitative data will have be determined post-hoc. So, though we will undoubtedly capture important signaling information, we lose the direct contextual connection developed by manipulating the *in vitro* environment. However, given the benefits afforded by the statistical extension of these signatures, and the challenges in selecting the appropriate *in vitro* model to use for signature development, factor analysis offers a direct and amenable line of further inquiry into the signaling deregulation present in these tumors.

4.2 Additional practical concerns for prospective genomic analysis

As genomic technologies continue to mature, the prospect of creating a genomic biomarker or assay will continue to be a moving target. Not only are there technological hurdles for implementing microarray-based approaches in the clinic, but there are practical considerations that are as important as the scientific ones. It has been well noted that the cost of DNA sequencing is dropping at a rate at least equivalent to Moore's law, though these numbers do not always include processing time and data storage costs associated with this incredibly dense information source. It is extremely likely that within the next decade that microarrays will no longer be necessary, as genome-wide sequencing will be more cost effective, and will not require the *a priori* definition of targets to be assayed. Unless sequencing technologies hit a cost-barrier, which currently appears unlikely, then this transition is inevitable and in some research settings is already happening.

The obvious hurdle, or opportunity, that this creates is the need to transition data away from microarrays and towards sequencing based approaches. While transitioning to a sequencing based approach may become appealing from a purely monetary approach, if the switch to sequencing represents a clear break with microarray technologies then in many cases there will be a wealth of historical data that becomes wasted or unusable. We demonstrated that historical datasets were able to overcome

batch-related biases in the prospective modeling of pathway signaling. In a disease setting as rare as glioblastoma, such historical datasets provide more value than the cost-differential of switching to sequencing currently affords. The field of genomics should invest in preserving the utility of these datasets by determining how to robustly integrate gene expression microarrays with next generation sequencing data, and how to transition between the two technologies. To assume that this transition will not happen would be naïve, but to not address the technical hurdles or to let these resources, built from patient samples, public funding, and years of research, lay unused because of a technological transition, would be wasteful. The fact that we can and do integrate genomic technologies from disparate data sources argues that this integration will happen, but it will require a concerted effort to fundamentally understand how these two technologies relate to each other and how to overcome the biases inherent in both platforms.

Finally, the work presented here primarily relies upon tumor tissue obtained as a biopsy or at the time of surgery. In the case of GBM, in which surgery is nearly always performed if possible, it can be assumed that for a large fraction of patients this will be feasible. But for inoperable tumors, or tumors in which surgery is not routine, such as lower-grade gliomas, the ability to apply a particular signature will need to be examined in the setting of a surrogate biomarker rather than bulk tumor tissue. Previous work from our lab by LaBreche *et al.* has demonstrated that this is indeed feasible, and can

serve as a marker for disease onset (235). In conjunction with the work presented here we now have a means by which we can use a genomic signature as a biomarker of disease presence, and examine that biomarker in a prospective manner. These methods will need to be expanded, and it is still unclear whether other material rather than peripheral blood might serve as a better biomarker for disease activity. Other biomarkers, such as tumor-related microvessicles, cerebrospinal fluid or circulating tumor cells may prove a more useful source for understanding the underpinnings of a tumor while still providing a less- or non-invasive assay. Regardless of the tissue source, we are approaching a time in which we can reasonable expect to implement personalized genomic screening, and the methods and frameworks described here form the basis for such capabilities.

Appendix A

Table 5: Parameters for pathway signature development

Signature **	# Genes	# Metagenes	Apply Quantile Normalization
AKT	250	3	No
BCAT	85	2	Yes
E2F1	150	2	No
E2F3	250	2	No
EGFR	500	2	Yes
HER2	250	3	Yes
IFNA	100	4	Yes
IFNG	100	4	Yes
MYC	500	2	No
P53	250	3	No
P63	75	3	No
PI3K	250	3	Yes
PTEN	150	2	Yes
RAS	350	2	No
SRC	85	3	Yes
STAT3	125	3	No
TNFA	100	4	Yes

** All signatures are run with Shift+Scale normalization using BinReg version 2 with 1000 burn-ins, 5000 iterations, 3 skips and a 95% credible interval

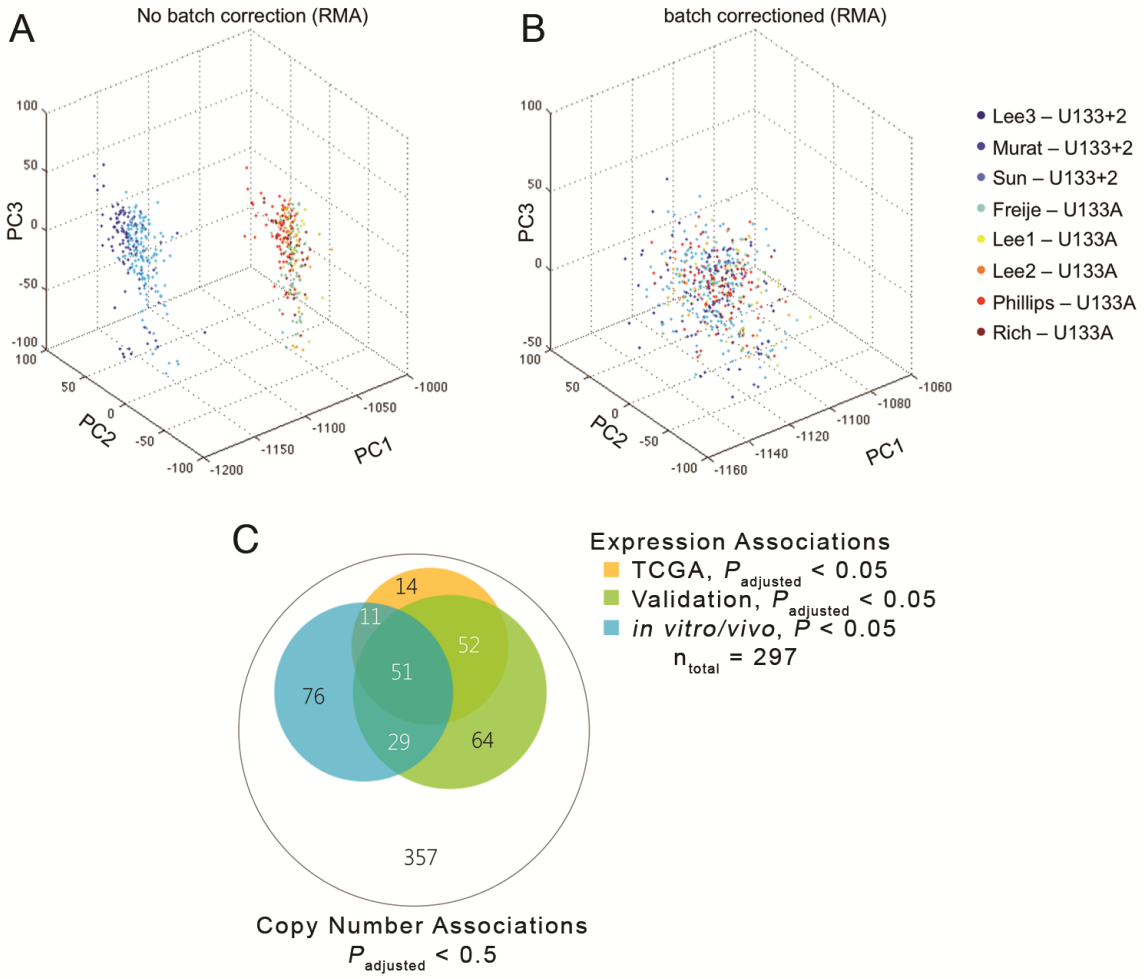

Figure 35: Expression normalization and validation of associations

Expression data from 8 datasets was analyzed to create a single dataset for validation of expression-level associations between candidate genes and associated pathway signaling. PCA of datasets A) before and B) after application of Bayesian Factor Regression Modeling to remove batch effects on expression data. Samples are color coded based on the dataset from which they are derived. C) Venn-diagram depicting the number of genes that are found within copy number associated loci, and which also show significant expression level associations in the TCGA database (n = 356, yellow), validation GBM clinical dataset (n = 480, green), or the GBM *in vitro* and *in vivo* model dataset (n = 49, blue).

Appendix B

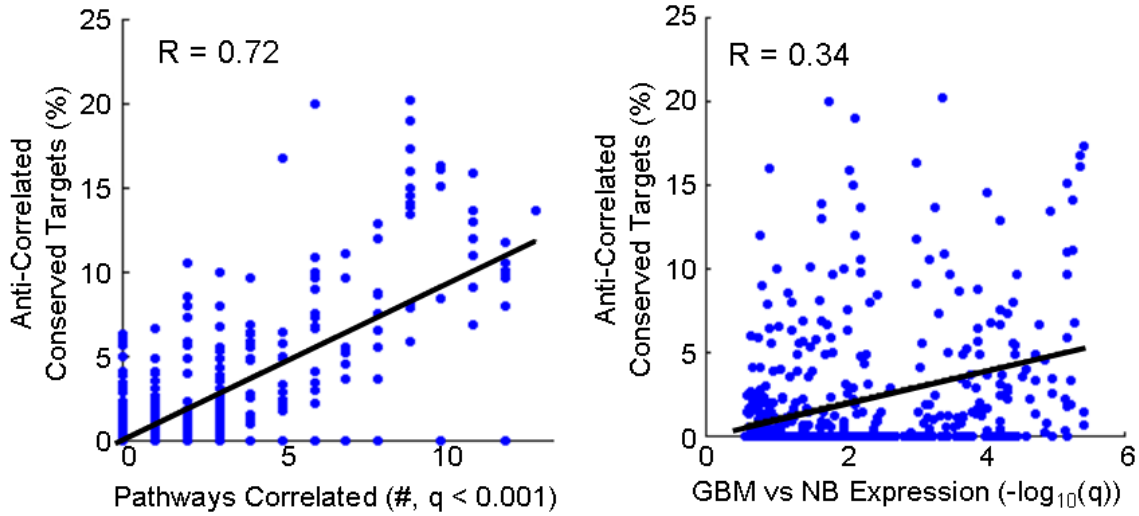

Figure 36: Pathway signaling identifies key microRNA network hubs

The expression of 536 microRNAs were correlated with pathway activity (left, Spearman correlation) or between normal brain and GBM samples (right, Wilcoxon rank-sum test). Significance was calculated using the FDR adjustment. microRNAs were then compared to the expression of predicted conserved targets to identify potential microRNA-mRNA repressive events which were calculated as the percentage of predicted conserved targets exhibiting significant anti-correlation ($\rho < -0.5$, Spearman). Under this analytical framework, the number of pathways that are strongly associated ($q < 0.001$) identifies a consistent relationship between microRNA-mRNA interactions, while overexpression or underexpression alone does not.

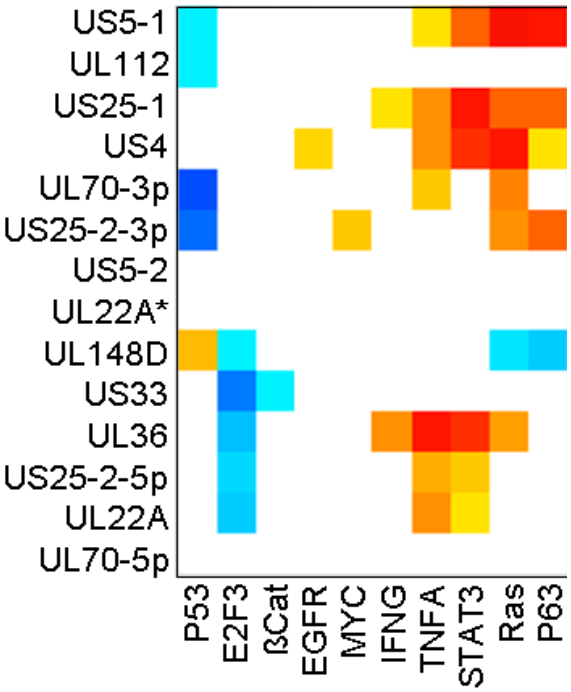

Figure 37: CMV-encoded microRNAs associate with pathway signaling

Correlation between microRNAs encoded by CMV and pathway activity. Significance of association is indicated by darker color with direct associations shown in red and inverse associations shown in blue for $q < 0.01$.

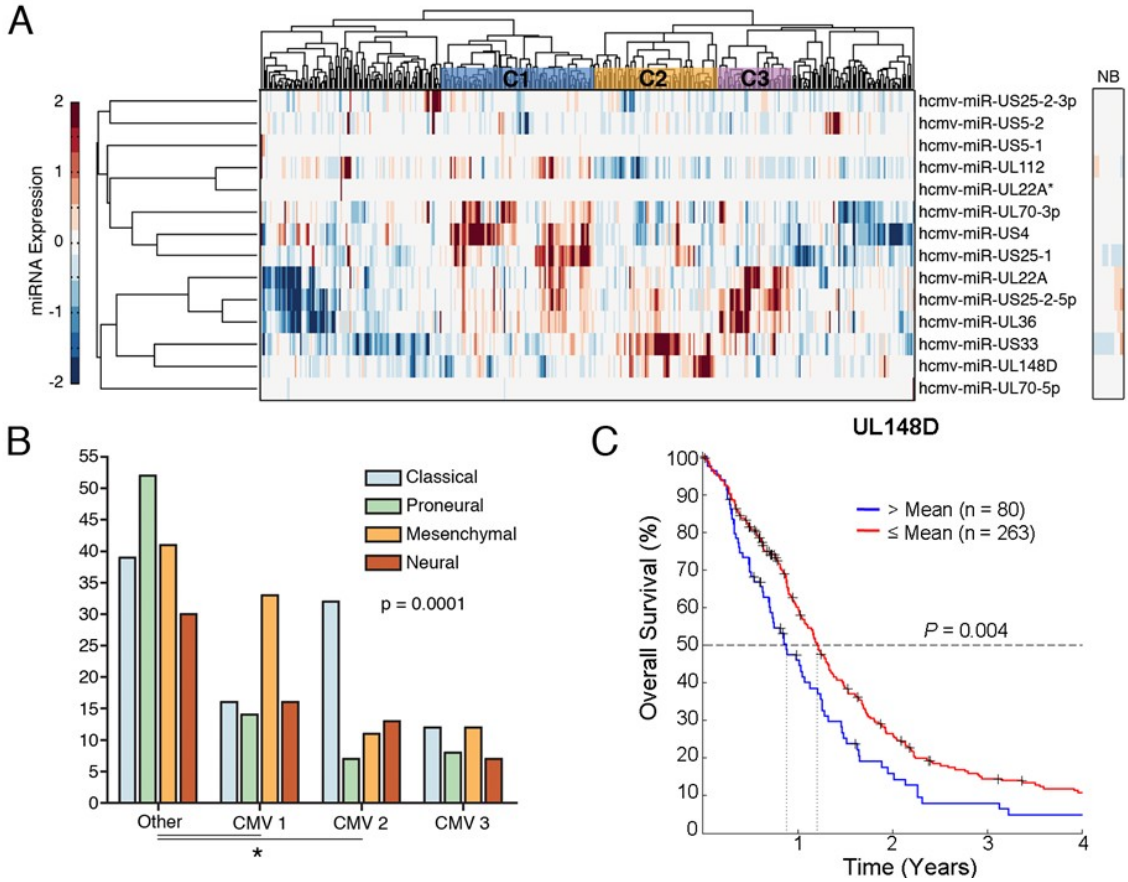

Figure 38: CMV microRNA Expression-based clusters of GBM

CMV microRNA expression in GBM tumors associates with distinct subgroups of GBM, and patient outcome. A) USHC of microRNA expression (left) and paired normal brain expression of microRNAs (right) for 14 microRNAs encoded by CMV. B) Enrichment of GBM subtypes in clusters identified in panel A. * $P < 0.05$, Fisher's Exact Test. C) Kaplan-Meier analysis of hcmv-miR-UL148D expression in patient samples. Log-rank *P*-value shown.

References

1. Hanahan D & Weinberg RA (2000) The Hallmarks of Cancer. *Cell* 100:57-70.

2. Hanahan D WR (2011) Hallmarks of cancer: the next generation. *Cell* 144(5):646-674.

3. Karnoub AE, *et al.* (2007) Mesenchymal stem cells within tumour stroma promote breast cancer metastasis. *Nature* 449(7162):557-563.

4. Wiseman BS & Werb Z (2002) Stromal effects on mammary gland development and breast cancer. *Science* 296(5570):1046-1049.

5. Heinrich MC, *et al.* (2003) PDGFRA activating mutations in gastrointestinal stromal tumors. *Science Signalling* 299(5607):708.

6. Hill R, Song Y, Cardiff RD, & Van Dyke T (2005) Selective evolution of stromal mesenchyme with p53 loss in response to epithelial tumorigenesis. *Cell* 123(6):1001-1011.

7. Courtneidge SA & Smith AE (1983) Polyoma virus transforming protein associates with the product of the c-src cellular gene. *Nature* 303:435-439.

8. Chang EH, Gonda MA, Ellis RW, Scolnick EM, & Lowy DR (1982) Human genome contains four genes homologous to transforming genes of Harvey and Kirsten murine sarcoma viruses. *Proceedings of the National Academy of Sciences* 79(16):4848-4852.

9. Knudson AG (1971) Mutation and cancer: statistical study of retinoblastoma. 68:820-823.

10. Baker SJ, *et al.* (1989) Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science* 244(4901):217-221.

11. Zakut-Houri R, *et al.* (1983) A single gene and a pseudogene for the cellular tumour antigen p53.

12. Oren M & Levine AJ (1983) Molecular cloning of a cDNA specific for the murine p53 cellular tumor antigen. *Proceedings of the National Academy of Sciences* 80(1):56-59.

13. Druker BJ, *et al.* (2001) Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with Philadelphia chromosome. *N. Engl. J. Med.* 344:1038-1042.

14. Aaltonen LA, *et al.* (1993) Clues to the pathogenesis of familial colorectal cancer. *Science* 260:751-752.

15. Perou CM, *et al.* (2000) Molecular portraits of human breast tumours. *Nature* 406(6797):747-752.

16. Gatza ML, *et al.* (2010) A pathway-based classification of human breast cancer. *Proc. Nat'l. Acad. Sci.* 107:6994-6999.

17. Dave SS, *et al.* (2006) Molecular diagnosis of Burkitt's lymphoma. *N. Engl. J. Med.* 354(23):2431-2442.

18. Thompson MC, *et al.* (2006) Genomics identifies medulloblastoma subgroups that are enriched for specific genetic alterations. *Journal of clinical oncology* 24(12):1924-1931.

19. Phillips HS, *et al.* (2006) Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis.[see comment]. *Cancer Cell* 9(3):157-173.

20. Lander ES, *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860-921.

21. Venter JC, *et al.* (2001) The sequence of the human genome. *Science* 291(5507):1304-1351.

22. Lipshutz RJ, Fodor SPA, Gingeras TR, & Lockhart DJ (1999) High density synthetic oligonucleotide arrays. *Nature genetics* 21(Suppl 1):20-24.

23. Southern E (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. 98:503-517.

24. Alwine JC, Kemp DJ, & Stark GR (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences* 74(12):5350-5354.

25. Storey JD & Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* 100(16):9440-9445.

26. Leek JT SR, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews* 11(10):733-739.

27. Tusher VG, Tibshirani R, & Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98(9):5116-5121.

28. Felsberg J RM, Loeser S, Fimmers R, Stummer W, Goeppert M, Steiger HJ, Friedensdorf B, Reifenberger G, Sabel MC. (2009) Prognostic significance of molecular markers and extent of resection in primary glioblastoma patients. *Clin Cancer Res* 15(21):6683-6693.

29. Eisen MB, Spellman PT, Brown PO, & Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* 95(25):14863-14868.

30. Monti S, Tamayo P, Mesirov J, & Golub T (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* 52(1):91-118.

31. Kerr MK & Churchill GA (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences* 98(16):8961-8965.

32. Allison DB, Cui X, Page GP, & Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* 7(1):55-65.

33. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, & Church GM (1999) Systematic determination of genetic network architecture. *Nature genetics* 22:281-285.

34. Chang JT & Nevins JR (2006) GATHER: a systems approach to interpreting genomic signatures. *Bioinformatics* 22:2926-2933.

35. Ogata H, *et al.* (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research* 27(1):29-34.

36. Thomas PD, *et al.* (2003) PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic acids research* 31(1):334-341.

37. Subramanian A, *et al.* (2005) 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.

38. Friedman N, Linial M, Nachman I, & Pe'er D (2000) Using Bayesian networks to analyze expression data. *Journal of computational biology* 7(3-4):601-620.

39. Faith JJ, *et al.* (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology* 5(1):e8.

40. Cerami E, Demir E, Schultz N, Taylor BS, & Sander C (2010) Automated network analysis identifies core pathways in glioblastoma. *PLoS one* 5(2):e8918.

41. Shannon P, *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 13(11):2498-2504.

42. Bredel M, *et al.* (2005) Functional network analysis reveals extended gliomagenesis pathway maps and three novel MYC-interacting genes in human gliomas. *Cancer research* 65(19):8679-8689.

43. Saeys Y, Inza I, & Larrañaga P (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19):2507-2517.

44. Wold S, Esbensen K, & Geladi P (1987) Principal component analysis. *Chemometrics and intelligent laboratory systems* 2(1):37-52.

45. Bild A, *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439:353-357.

46. Huang E, *et al.* (2003) Gene expression predictors of breast cancer outcomes. *Lancet* 361:1590-1596.

47. Huang E, *et al.* (2003) Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat Genet.* 34(2):226-230.

48. Chang JT, *et al.* (2009) A genomic strategy to elucidate modules of oncogenic pathway signaling networks. *Molecular Cell* 34:104-114.

49. Chen JL, *et al.* (2008) Genomic analysis of response to lactic acidosis and acidosis in human cancers. *PLoS Genetics* 4(12):e1000293.

50. Simon R (2005) Roadmap for developing and validating therapeutically relevant genomic classifiers. *Journal of Clinical Oncology* 23(29):7332-7341.

51. Nevins JR, *et al.* (2003) Towards integrated clinic-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Hum. Mol. Genet.* 12:R153-157.

52. Handl J, Knowles J, & Kell DB (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21(15):3201-3212.

53. Edgar R, Domrachev M, & Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* 30(1):207-210.

54. Network CGAR (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455(7216):1061-1068.

55. Rhodes DR, *et al.* (2004) ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia (New York, NY)* 6(1):1.

56. Benjamini Y & Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach tomultiple testing. *J. Royal Stat. Soc.* 57:289-300.

57. Marko NF, Quackenbush J, & Weil RJ (2011) Why is there a lack of consensus on molecular subgroups of glioblastoma? understanding the nature of biological and statistical variability in glioblastoma expression data. *PloS one* 6(7):e20826.

58. Freidlin B & Simon R (2005) Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* 11(21):7872-7878.

59. Freidlin B, Jiang W, & Simon R (2010) The cross-validated adaptive signature design. *Clinical Cancer Research* 16(2):691-698.

60. Owzar K, Barry WT, Jung SH, Sohn I, & George SL (2008) Statistical challenges in preprocessing in microarray experiments in cancer. *Clinical Cancer Research* 14(19):5959-5966.

61. Tinker AV, Boussioutas A, & Bowtell DDL (2006) The challenges of gene expression microarrays for the study of human cancer. *Cancer cell* 9(5):333-339.

62. McCall MN, Bolstad BM, & Irizarry RA (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics* 11(2):242-253.

63. Katz S, Irizarry RA, Lin X, Tripputi M, & Porter MW (2006) A summarization approach for Affymetrix GeneChip data using a reference training set from a large, biologically diverse database. *BMC bioinformatics* 7(1):464.

64. Alon U, *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences* 96(12):6745-6750.

65. Golub TR, *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science* 286(5439):531-537.

66. Barbie DA, *et al.* (2009) Systematic RNA Interference Reveals that Oncogenic KRAS-driven cancers require TBK1. *Nature* 462(7269):108-112.

67. Johnson WE, Li C, & Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8:118-127.

68. Lucas JE, *et al.* (2006) Sparse statistical modelling in gene expression genomics. *Bayesian Inference for Gene Expression and Proteomics*, ed M V (Cambridge University Press, Cambridge), pp 155-176.

69. Carvalho C, *et al.* (2008) High-dimensional sparse factor modelling: applications in gene expression genomics. *J. Am. Stat. Assoc.* 103:1438-1456.

70. Benito M, *et al.* (2004) Adjustment of systematic microarray data biases. *Bioinformatics* 20:105-114.

71. Chen C, *et al.* (2011) Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one* 6(2):e17238.

72. Auer PL & Doerge R (2010) Statistical design and analysis of RNA sequencing data. *Genetics* 185(2):405-416.

73. Godard S, *et al.* (2003) Classification of human astrocytic gliomas on the basis of gene expression: a correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes. *Cancer Res* 63(20):6613-6625.

74. Larjavaara S, *et al.* (2007) Incidence of gliomas by anatomic location. *Neuro-oncol* 9(3):319-325.

75. Horner M, *et al.* (2009) SEER Cancer Statistics Review, 1975-2006, National Cancer Institute. Bethesda, MD.

76. Louis DN, *et al.* (2007) The 2007 WHO classification of tumours of the central nervous system. *Acta neuropathologica* 114(2):97-109.

77. Ohgaki H & Kleihues P (2007) Genetic pathways to primary and secondary glioblastoma. *Am J Pathol* 170(5):1445-1453.

78. Libermann TA, *et al.* (1985) Amplification, enhanced expression and possible rearrangement of EGF receptor gene in primary human brain tumours of glial origin.

79. Ekstrand AJ, Sugawa N, James CD, & Collins VP (1992) Amplified and rearranged epidermal growth factor receptor genes in human glioblastomas reveal deletions of sequences encoding portions of the N-and/or C-terminal tails. *Proceedings of the National Academy of Sciences* 89(10):4309-4313.

80. Li J, *et al.* (1997) PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* 275(5308):1943-1947.

81. Tohma Y, *et al.* (1998) PTEN (MMAC1) mutations are frequent in primary glioblastomas (de novo) but not in secondary glioblastomas. *Journal of neuropathology and experimental neurology* 57(7):684.

82. Jen J, *et al.* (1994) Deletion of p16 and p15 genes in brain tumors. *Cancer Research* 54(24):6353-6358.

83. Schmidt EE, Ichimura K, Reifenberger G, & Collins VP (1994) CDKN2 (p16/MTS1) gene deletion or CDK4 amplification occurs in the majority of glioblastomas. *Cancer Research* 54(24):6321-6324.

84. James CD, *et al.* (1988) Clonal genomic alterations in glioma malignancy stages. *Cancer research* 48(19):5546-5551.

85. Henson JW, *et al.* (2004) The retinoblastoma gene is involved in malignant progression of astrocytomas. *Annals of neurology* 36(5):714-721.

86. Louis DN (1994) The p53 gene and protein in human brain tumors. *Journal of neuropathology and experimental neurology* 53(1):11.

87. Clarke I & Dirks P (2003) A human brain tumor-derived PDGFR-α deletion mutant is transforming. *Oncogene* 22(5):722-733.

88. Liu G, *et al.* (2006) Analysis of gene expression and chemoresistance of CD133+ cancer stem cells in glioblastoma. *Molecular cancer* 5(1):67.

89. Tamura K AM, Wakimoto H, Ando N, Nariai T, Yamamoto M, Ohno K. (2010) Accumulation of CD133-positive glioma cells after high-dose irradiation by Gamma Knife surgery plus external beam radiation. *J Neurosurg* 113(2):310-318.

90. Singh SK, *et al.* (2004) Identification of human brain tumour initiating cells. *nature* 432(7015):396-401.

91. Chen J LY, Yu TS, McKay RM, Burns DK, Kernie SG, Parada LF. (2012) A restricted cell population propagates glioblastoma growth after chemotherapy. *Nature* 488:522-526.

92. Weller M & Fontan A (1995) The failure of current immunotherapy for malignant glioma. Tumor-derived TGF-β, T-cell apoptosis, and the immuneprivilege of the brain. *Brain Reseach Reviews* 21(2):128-151.

93. Sampson JH HA, Archer GE, Aldape KD, Friedman AH, Friedman HS, Gilbert MR, Herndon JE 2nd, McLendon RE, Mitchell DA, Reardon DA, Sawaya R, Schmittling RJ, Shi W, Vredenburgh JJ, Bigner DD. (2010) Immunologic escape after prolonged progression-free survival with epidermal growth factor receptor variant III peptide vaccination in patients with newly diagnosed glioblastoma. *J Clin Oncol* 28(31):4722-4729.

94. Liau LM, *et al.* (2005) Dendritic cell vaccination in glioblastoma patients induces systemic and intracranial T-cell responses modulated by the local central nervous system tumor microenvironment. *Clinical Cancer Research* 11(15):5515-5525.

95. Carro MS LW, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H, Lasorella A, Aldape K, Califano A, Iavarone A. (2010) The transcriptional network for mesenchymal transformation of brain tumours. *Nature* 463(7279):318-325.

96. Dang L WD, Gross S, Bennett BD, Bittinger MA, Driggers EM, Fantin VR, Jang HG, Jin S, Keenan MC, Marks KM, Prins RM, Ward PS, Yen KE, Liau LM, Rabinowitz JD, Cantley LC, Thompson CB, Vander Heiden MG, Su SM (2009) Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* 465(966):739-744.

97. Parsons DW, *et al.* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science* 321(5897):1807-1812.

98. Nobusawa S WT, Kleihues P, Ohgaki H. (2009) IDH1 mutations as molecular signature and predictive factor of secondary glioblastomas. *Clin Cancer Res* 15(19):6002-6007.

99. Noushmehr H, *et al.* (2010) Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* 17(5):510-522.

100. Kim TM HW, Park R, Park PJ, Johnson MD (2011) A developmental taxonomy of glioblastoma defined and maintained by MicroRNAs. *Cancer Res* 71(9):3387-3399.

101. Wuchty S AD, Li A, Kotliarov Y, Walling J, Ahn S, Zhang A, Maric D, Anolik R, Zenklusen JC, Fine HA. (2011) Prediction of Associations between microRNAs and Gene Expression in Glioma Biology. *PLOS ONE* 6(2):e14681.

102. Sumazin P YX, Chiu HS, Chung WJ, Iyer A, Llobet-Navas D, Rajbhandari P, Bansal M, Guarnieri P, Silva J, Califano A. (2011) An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* 147(2):370-381.

103. Yoo AS SB, Chen L, Crabtree GR (2009) MicroRNA-mediated switching of chromatin-remodelling complexes in neural development. *Nature* 460:642-646.

104. Li Y, *et al.* (2009) MicroRNA-34a inhibits glioblastoma growth by targeting multiple oncogenes. *Cancer Res* 69(19):7569-7576.

105. Papagiannakopoulos T SA, Kosik KS (2008) MicroRNA-21 targets a network of key tumor-suppressive pathways in glioblastoma cells. *Cancer Res* 68(19):8164-8172.

106. Chan JA KA, Kosik KS. (2005) MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells. *Cancer Res* 65(14):6029-6033.

107. Hamilton MG RG, Magliocco A, McIntyre JB, Parney I, Easaw JC. (2011) Determination of the methylation status of MGMT in different regions within glioblastoma multiforme. *J Neurooncol* 102(2):255-260.

108. Szerlip NJ, *et al.* (2012) Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with

distinct growth factor response. *Proceedings of the National Academy of Sciences* 109(8):3041-3046.

109. Bonavia R, *et al.* (2010) Tumor heterogeneity is an active process maintained by a mutant EGFR-induced cytokine circuit in glioblastoma. *Genes & development* 24(16):1731-1745.

110. Kim S, *et al.* (2002) Identification of combination gene sets for glioma classification. *Mol Cancer Ther* 1(13):1229-1236.

111. Rickman DS, *et al.* (2001) Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res* 61(18):6885-6891.

112. Freije WA, *et al.* (2004) Gene expression profiling of gliomas strongly predicts survival. *Cancer Research* 64(18):6503-6510.

113. Mischel PS, *et al.* (2003) Identification of molecular subtypes of glioblastoma by gene expression profiling. *Oncogene* 22(15):2361-2373.

114. Lee Y, *et al.* (2008) Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. *BMC Med Genomics* 1:52.

115. Lai A KS, Pope WB, Tran A, Solis OE, Peale F, Forrest WF, Pujara K, Carrillo JA, Pandita A, Ellingson BM, Bowers CW, Soriano RH, Schmidt NO, Mohan S, Yong WH, Seshagiri S, Modrusan Z, Jiang Z, Aldape KD, Mischel PS, Liau LM, Escovedo CJ, Chen W, Nghiemphu PL, James CD, Prados MD, Westphal M, Lamszus K, Cloughesy T, Phillips HS. (2011) Evidence for sequenced molecular evolution of IDH1 mutant glioblastoma from a distinct cell of origin. *J Clin Oncol* 29(34):4482-4490.

116. Tso CL, *et al.* (2006) Distinct transcription profiles of primary and secondary glioblastoma subgroups. *Cancer Res* 66(1):159-167.

117. Maher EA, *et al.* (2006) Marked genomic differences characterize primary and secondary glioblastoma subtypes and identify two distinct molecular and clinical secondary glioblastoma entities. *Cancer research* 66(23):11502-11513.

118. Groenendijk FH TW, Dubbink HJ, Haarloo CR, Kouwenhoven MC, van den Bent MJ, Kros JM, Dinjens WN. (2011) MGMT promoter hypermethylation is a

frequent, early, and consistent event in astrocytoma progression, and not correlated with TP53 mutation. *J Neurooncol* 101(3):405-417.

119. Verhaak RGW, *et al.* (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17:98-110.

120. Brennan C MH, Hambardzumyan D, Ozawa T, Tandon A, Pedraza A, Holland E (2009) Glioblastoma subclasses can be defined by activity among signal transduction pathways and associated genomic alterations. *PLOS ONE* 4(11):e7752.

121. de Tayrac M AM, Saïkali S, Etcheverry A, Surbled C, Guénot F, Galibert MD, Hamlat A, Lesimple T, Quillien V, Menei P, Mosser J. (2011) A 4-gene signature associated with clinical outcome in high-grade gliomas. *Clin Cancer Res* 17(2):317-327.

122. Colman H ZL, Sulman EP, McDonald JM, Shooshtari NL, Rivera A, Popoff S, Nutt CL, Louis DN, Cairncross JG, Gilbert MR, Phillips HS, Mehta MP, Chakravarti A, Pelloski CE, Bhat K, Feuerstein BG, Jenkins RB, Aldape K (2010) A multigene predictor of outcome in glioblastoma. *Neuro Oncol* 12(1):49-57.

123. Lacroix M A-SD, Fourney DR, Gokaslan ZL, Shi W, DeMonte F, Lang FF, McCutcheon IE, Hassenbusch SJ, Holland E, Hess K, Michael C, Miller D, Sawaya R. (2001) A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. *J Neurosurg* 95(2):190-198.

124. Salvati M PA, Piccirilli M, Floriana Brunetto GM, D'Elia A, Artizzu S, Santoro F, Arcella A, Giangaspero F, Frati A, Simione L, Santoro A. (2012) Extent of tumor removal and molecular markers in cerebral glioblastoma: a combined prognostic factors study in a surgical series of 105 patients. *J Neurosurg* 117(2):204-211.

125. Kuhnt D BA, Ganslandt O, Bauer M, Buchfelder M, Nimsky C. (2011) Correlation of the extent of tumor volume resection and patient survival in surgery of glioblastoma multiforme with high-field intraoperative MRI guidance. *Neuro Oncol* 13(12):1339-1348.

126. Sanai N PM, McDermott MW, Parsa AT, Berger MS. (2011) An extent of resection threshold for newly diagnosed glioblastomas. *J Neurosurg* 115(1):3-8.

127. Cao VT JT, Jung S, Jin SG, Moon KS, Kim IY, Kang SS, Park CS, Lee KH, Chae HJ. (2009) The correlation and prognostic significance of MGMT promoter methylation and MGMT protein in glioblastomas. *Neurosurgery* 65(5):866-875.

128. Clarke JL IF, Sul J, Panageas K, Lassman AB, DeAngelis LM, Hormigo A, Nolan CP, Gavrilovic I, Karimi S, Abrey LE. (2009) Randomized phase II trial of chemoradiotherapy followed by either dose-dense or metronomic temozolomide for newly diagnosed glioblastoma. *J Clin Oncol* 27(23):3861-3867.

129. Walker MD, *et al.* (1980) Randomized comparisons of radiotherapy and nitrosoureas for the treatment of malignant glioma after surgery. *New England Journal of Medicine* 303(23):1323-1329.

130. Walker MD, *et al.* (1978) Evaluation of BCNU and/or radiotherapy in the treatment of anaplastic gliomas. *Journal of Neurosurgery* 49(3):333-343.

131. Bao S WQ, McLendon RE, Hao Y, Shi Q, Hjelmeland AB, Dewhirst MW, Bigner DD, Rich JN. (2006) Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. *Nature* 444:756-760.

132. Wang J WT, Lathia JD, Hjelmeland AB, Wang XF, White RR, Rich JN, Sullenger BA. (2010) Notch promotes radioresistance of glioma stem cells. *Stem Cells* 28(1):17-28.

133. Carpentier AF (2005) Neuro-oncology: the growing role of chemotherapy in glioma. *The Lancet Neurology* 4(1):4-5.

134. Reardon DA & Wen PY (2006) Therapeutic advances in the treatment of glioblastoma: rationale and potential role of targeted agents. *The oncologist* 11(2):152-164.

135. Stupp R, *et al.* (2005) Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.* 352:987-996.

136. Brandsma D SL, Taal W, Sminia P, van den Bent MJ (2008) Clinical features, mechanisms, and management of pseudoprogression in malignant gliomas. *Lancet Oncol* 9(5):453-461.

137. Brandes AA FE, Tosoni A, Blatt V, Pession A, Tallini G, Bertorelle R, Bartolini S, Calbucci F, Andreoli A, Frezza G, Leonardi M, Spagnolli F, Ermani M. (2008) MGMT promoter methylation status can predict the incidence and outcome of

pseudoprogression after concomitant radiochemotherapy in newly diagnosed glioblastoma patients. *J Clin Oncol* 26(13):2192-2197.

138. Eramo A, *et al.* (2006) Chemotherapy resistance of glioblastoma stem cells. *Cell Death & Differentiation* 13(7):1238-1241.

139. Murat A, *et al.* (2008) Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. *J Clin Oncol* 26(18):3015-3024.

140. Hegi ME, *et al.* (2005) MGMT gene silencing and benefit from temozolomide in glioblastoma. *New England Journal of Medicine* 352:997-1003, 2005 Mar 1010.

141. Esteller M, *et al.* (2000) Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *New England Journal of Medicine* 343(19):1350-1354.

142. Hegi ME DA, Godard S, Dietrich PY, Regli L, Ostermann S, Otten P, Van Melle G, de Tribolet N, Stupp R. (2004) Clinical trial substantiates the predictive value of O-6-methylguanine-DNA methyltransferase promoter methylation in glioblastoma patients treated with temozolomide. *Clin Cancer Res* 10(6):1871-1874.

143. Friedman HS, *et al.* (1998) DNA mismatch repair and O6-alkylguanine-DNA alkyltransferase analysis and response to Temodal in newly diagnosed malignant glioma. *Journal of Clinical Oncology* 16(12):3851-3857.

144. Lai A TA, Nghiemphu PL, Pope WB, Solis OE, Selch M, Filka E, Yong WH, Mischel PS, Liau LM, Phuphanich S, Black K, Peak S, Green RM, Spier CE, Kolevska T, Polikoff J, Fehrenbacher L, Elashoff R, Cloughesy T. (2011) Phase II study of bevacizumab plus temozolomide during and after radiation therapy for patients with newly diagnosed glioblastoma multiforme. *J Clin Oncol* 29(2):142-148.

145. Rivera AL PC, Gilbert MR, Colman H, De La Cruz C, Sulman EP, Bekele BN, Aldape KD. (2010) MGMT promoter methylation is predictive of response to radiotherapy and prognostic in the absence of adjuvant alkylating chemotherapy for glioblastoma. *Neuro Oncol* 12(2):116-121.

146. Karayan-Tapon L QV, Guilhot J, Wager M, Fromont G, Saikali S, Etcheverry A, Hamlat A, Loussouarn D, Campion L, Campone M, Vallette FM, Gratas-Rabbia-Ré C. (2010) Prognostic value of O6-methylguanine-DNA methyltransferase

status in glioblastoma patients, assessed by five different methods. *J Neurooncol* 97(3):311-322.

147. Stupp R HM, Mason WP, van den Bent MJ, Taphoorn MJ, Janzer RC, Ludwin SK, Allgeier A, Fisher B, Belanger K, Hau P, Brandes AA, Gijtenbeek J, Marosi C, Vecht CJ, Mokhtari K, Wesseling P, Villa S, Eisenhauer E, Gorlia T, Weller M, Lacombe D, Cairncross JG, Mirimanoff RO; European Organisation for Research and Treatment of Cancer Brain Tumour and Radiation Oncology Groups; National Cancer Institute of Canada Clinical Trials Group. (2009) Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol* 10(5):459-466.

148. Gorlia T vdBM, Hegi ME, Mirimanoff RO, Weller M, Cairncross JG, Eisenhauer E, Belanger K, Brandes AA, Allgeier A, Lacombe D, Stupp R. (2008) Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of EORTC and NCIC trial 26981-22981/CE.3. *Lancet Oncol* 9(1):29-38.

149. Vredenburgh JJ DA, Herndon JE 2nd, Marcello J, Reardon DA, Quinn JA, Rich JN, Sathornsumetee S, Gururangan S, Sampson J, Wagner M, Bailey L, Bigner DD, Friedman AH, Friedman HS. (2007) Bevacizumab plus irinotecan in recurrent glioblastoma multiforme. *J Clin Oncol* 25(30):4722-4729.

150. Friedman HS, *et al.* (2009) Bevacizumab alone and in combination with irinotecan in recurrent glioblastoma. *Journal of Clinical Oncology* 27(28):4733-4740.

151. Clarke JL & Chang S (2009) Pseudoprogression and pseudoresponse: challenges in brain tumor imaging. *Current neurology and neuroscience reports* 9(3):241-246.

152. Norden A, *et al.* (2008) Bevacizumab for recurrent malignant gliomas Efficacy, toxicity, and patterns of recurrence. *Neurology* 70(10):779-787.

153. van den Bent MJ BA, Rampling R, Kouwenhoven MC, Kros JM, Carpentier AF, Clement PM, Frenay M, Campone M, Baurain JF, Armand JP, Taphoorn MJ, Tosoni A, Kletzl H, Klughammer B, Lacombe D, Gorlia T. (2009) Randomized phase II trial of erlotinib versus temozolomide or carmustine in recurrent glioblastoma: EORTC brain tumor group study 26034. *J Clin Oncol* 27(8):1268-1274.

154. Brown PD KS, Sarkaria JN, Wu W, Jaeckle KA, Uhm JH, Geoffroy FJ, Arusell R, Kitange G, Jenkins RB, Kugler JW, Morton RF, Rowland KM Jr, Mischel P, Yong

WH, Scheithauer BW, Schiff D, Giannini C, Buckner JC; North Central Cancer Treatment Group Study N0177 (2008) Phase I/II Trial of Erlotinib and Temozolomide With Radiation Therapy in the Treatment of Newly Diagnosed Glioblastoma Multiforme: North Central Cancer Treatment Group Study N0177. *J Clin Oncol* 26(34):5603-5609.

155. Prados MD CS, Butowski N, DeBoer R, Parvataneni R, Carliner H, Kabuubi P, Ayers-Ringler J, Rabbitt J, Page M, Fedoroff A, Sneed PK, Berger MS, McDermott MW, Parsa AT, Vandenberg S, James CD, Lamborn KR, Stokoe D, Haas-Kogan DA. (2009) Phase II study of erlotinib plus temozolomide during and after radiation therapy in patients with newly diagnosed glioblastoma multiforme or gliosarcoma. *J Clin Oncol* 27(4):579-584.

156. Uhm JH BK, Wu W, Giannini C, Krauss JC, Buckner JC, James CD, Scheithauer BW, Behrens RJ, Flynn PJ, Schaefer PL, Dakhill SR, Jaeckle KA. (2011) Phase II evaluation of gefitinib in patients with newly diagnosed Grade 4 astrocytoma: Mayo/North Central Cancer Treatment Group Study N0074. *Int J Radiat Oncol Biol Phys* 80(2):347-353.

157. Rich JN RD, Peery T, Dowell JM, Quinn JA, Penne KL, Wikstrand CJ, Van Duyn LB, Dancey JE, McLendon RE, Kao JC, Stenzel TT, Ahmed Rasheed BK, Tourt-Uhlig SE, Herndon JE 2nd, Vredenburgh JJ, Sampson JH, Friedman AH, Bigner DD, Friedman HS. (2004) Phase II trial of gefitinib in recurrent glioblastoma. *J Clin Oncol* 22(1):133-142.

158. Mellinghoff IK, *et al.* (2005) Molecular determinants of the response of glioblastomas to EGFR kinase inhibitors. *New England Journal of Medicine* 353(19):2012-2024.

159. Vivanco I RH, Rohle D, Campos C, Grommes C, Nghiemphu PL, Kubek S, Oldrini B, Chheda MG, Yannuzzi N, Tao H, Zhu S, Iwanami A, Kuga D, Dang J, Pedraza A, Brennan CW, Heguy A, Liau LM, Lieberman F, Yung WK, Gilbert MR, Reardon DA, Drappatz J, Wen PY, Lamborn KR, Chang SM, Prados MD, Fine HA, Horvath S, Wu N, Lassman AB, DeAngelis LM, Yong WH, Kuhn JG, Mischel PS, Mehta MP, Cloughesy TF, Mellinghoff IK. (2012) Differential sensitivity of glioma- versus lung cancer-specific EGFR mutations to EGFR kinase inhibitors. *Cancer Discov* 2(5):458-471.

160. Thiessen B SC, Tsao M, Kamel-Reid S, Schaiquevich P, Mason W, Easaw J, Belanger K, Forsyth P, McIntosh L, Eisenhauer E (2010) A phase I/II trial of GW572016 (lapatinib) in recurrent glioblastoma multiforme: clinical outcomes,

pharmacokinetics and molecular correlation. *Cancer Chemother Pharmacol* 65(2):353-361.

161. Neyns B, *et al.* (2009) Stratified phase II trial of cetuximab in patients with recurrent high-grade glioma. *Annals of Oncology* 20(9):1596-1603.

162. Van Meir EG, *et al.* (2010) Exciting New Advances in Neuro,ÄêOncology: The Avenue to a Cure for Malignant Glioma. *CA: a cancer journal for clinicians* 60(3):166-193.

163. Yamanaka R, *et al.* (2006) Identification of expressed genes characterizing long-term survival in malignant glioma patients. *Oncogene* 25(44):5994-6002.

164. Marko NF, Toms SA, Barnett GH, & Weil R (2008) Genomic expression patterns distinguish long-term from short-term glioblastoma survivors: a preliminary feasibility study. *Genomics* 91(5):395-406.

165. Gravendeel LA, *et al.* (2009) Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer Res* 69(23):9065-9072.

166. Nigro JM, *et al.* (2005) Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma. *Cancer Research* 65(5):1678-1686.

167. Lo KC, *et al.* (2007) Candidate glioblastoma development gene identification using concordance between copy number abnormalities and gene expression level changes. *Genes Chromosomes Cancer* 46(10):875-894.

168. Ruano Y, *et al.* (2006) Identification of novel candidate target genes in amplicons of Glioblastoma multiforme tumors detected by expression and CGH microarray profiling. *Mol Cancer* 5:39.

169. Bredel M SD, Harsh GR, Bredel C, Chandler JP, Renfrow JJ, Yadav AK, Vogel H, Scheck AC, Tibshirani R, Sikic BI (2009) A network model of a cooperative genetic landscape in brain tumors. *JAMA* 302(3):261-275.

170. Kotliarov Y KS, Charong N, Li A, Walling J, Aquilanti E, Ahn S, Steed ME, Su Q, Center A, Zenklusen JC, Fine HA (2009) Correlation analysis between single-nucleotide polymorphism and expression arrays in gliomas identifies potentially relevant target genes. *Cancer Res* 69(4):1596-1603.

171. Chow LM ER, Zhu X, Rankin S, Qu C, Zhang J, Broniscer A, Ellison DW, Baker SJ (2011) Cooperativity within and among Pten, p53, and Rb pathways induces high-grade astrocytoma in adult brain. *Cancer Cell* 19(3):305-316.

172. Adler AS, *et al.* (2006) Genetic regulators of large-scale transcriptional signatures in cancer. *Nat. Genet.* 38:421-430.

173. Jones S, *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 321(5897):1801-1806.

174. Vivanco I, *et al.* (2007) Identification of the JNK signaling pathway as a functional target of the tumor suppressor PTEN. *Cancer Cell* 11(6):555-569.

175. Beroukhim R GG, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S, Du J, Kau T, Thomas RK, Shah K, Soto H, Perner S, Prensner J, Debiasi RM, Demichelis F, Hatton C, Rubin MA, Garraway LA, Nelson SF, Liau L, Mischel PS, Cloughesy TF, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A* 104(50):20007-20012.

176. Mermel CH SS, Hill B, Meyerson ML, Beroukhim R, Getz G (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 12(4):R41.

177. Rich JN, *et al.* (2005) Gene expression profiling and genetic markers in glioblastoma survival. *Cancer Res.* 65(10):4051-4058.

178. Sun LJ, *et al.* (2006) Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell* 9(4):287-300.

179. Zhou F ZL, Gong K, Lu G, Sheng B, Wang A, Zhao N, Zhang X, Gong Y (2008) LEF-1 activates the transcription of E2F1. *Biochem Biophys Res Commun* 365(1):149-153.

180. Westbrook L MM, Kern FG, Estes NR 2nd, Ramanathan HN, Thottassery JV (2007) Cks1 regulates cdk1 expression: a novel role during mitotic entry in breast cancer cells. *Cancer Res* 67(23):11393-11401.

181. Hayami S YM, Veerakumarasivam A, Unoki M, Iwai Y, Tsunoda T, Field HI, Kelly JD, Neal DE, Yamaue H, Ponder BA, Nakamura Y, Hamamoto R (2010) Overexpression of the JmjC histone demethylase KDM5B in human

carcinogenesis: involvement in the proliferation of cancer cells through the E2F/RB pathway. *Mol Cancer* 9(59).

182. Xie L PC, Wang W, Bashar A, Varlamova O, Shadle S, Impey S. (2011) KDM5B regulates embryonic stem cell self-renewal and represses cryptic intragenic transcription. *EMBO* 30(8):1473-1484.

183. Zhang Z JA, Sun CW, Li C, Chang CW, Joo HY, Dai Q, Mysliwiec MR, Wu LC, Guo Y, Yang W, Liu K, Pawlik KM, Erdjument-Bromage H, Tempst P, Lee Y, Min J, Townes TM, Wang H. (2011) PRC2 complexes with JARID2, MTF2, and esPRC2p48 in ES cells to modulate ES cell pluripotency and somatic cell reprogramming. *Stem Cells* 29(2):229-240.

184. Horvath S, *et al.* (2006) Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc Natl Acad Sci U S A* 103(46):17402-17407.

185. Buchman JJ DO, Tsai LH (2011) ASPM regulates Wnt signaling pathway activity in the developing brain. *Genes Dev* 25(18):1909-1914.

186. Kim HS PK, Muldoon-Jacobs K, Bisht KS, Aykin-Burns N, Pennington JD, van der Meer R, Nguyen P, Savage J, Owens KM, Vassilopoulos A, Ozden O, Park SH, Singh KK, Abdulkadir SA, Spitz DR, Deng CX, Gius D (2010) SIRT3 is a mitochondria-localized tumor suppressor required for maintenance of mitochondrial integrity and metabolism during stress. *Cancer Cell* 17(1):41-52.

187. Yue W SQ, Dacic S, Landreneau RJ, Siegfried JM, Yu J, Zhang L (2008) Downregulation of Dkk3 activates beta-catenin/TCF-4 signaling in lung cancer. *Carcinogenesis* 29(1):84-92.

188. Smith IM MS, Liu C, Chang SS, Begum S, Dhara M, Westra W, Sidranksy D, Califano JA (2009) Novel integrative methods for gene discovery associated with head and neck squamous cell carcinoma development. *Arch Otolaryngol Head Neck Surg* 135(5):487-495.

189. Narayan G BV, Chaganti S, Arias-Pulido H, Nandula SV, Rao PH, Gissmann L, Dürst M, Schneider A, Pothuri B, Mansukhani M, Basso K, Chaganti RS, Murty VV. (2007) Gene dosage alterations revealed by cDNA microarray analysis in cervical cancer: identification of candidate amplified and overexpressed genes. *Genes Chromosom Cancer* 46(4):373-384.

190. Bredel M BC, Juric D, Duran GE, Yu RX, Harsh GR, Vogel H, Recht LD, Scheck AC, Sikic BI. (2006) Tumor necrosis factor-alpha-induced protein 3 as a putative regulator of nuclear factor-kappaB-mediated resistance to O6-alkylating agents in human glioblastomas. *J Clin Oncol* 24(2):274-287.

191. Bredel M SD, Yadav AK, Alvarez AA, Renfrow JJ, Chandler JP, Yu IL, Carro MS, Dai F, Tagge MJ, Ferrarese R, Bredel C, Phillips HS, Lukac PJ, Robe PA, Weyerbrock A, Vogel H, Dubner S, Mobley B, He X, Scheck AC, Sikic BI, Aldape KD, Chakravarti A, Harsh GR 4th (2011) NFKBIA deletion in glioblastomas. *The New England journal of medicine* 364(7):627-637.

192. Bassi R, *et al.* (2008) HMGB1 as an autocrine stimulus in human T98G glioblastoma cells: role in cell growth and migration. *Journal of neuro-oncology* 87(1):23-33.

193. Curtin JF, *et al.* (2009) HMGB1 mediates endogenous TLR2 activation and brain tumor regression. *PLoS medicine* 6(1):e1000010.

194. Yu M, *et al.* (2006) HMGB1 signals through toll-like receptor (TLR) 4 and TLR2. *Shock* 26(2):174.

195. van Beijnum JR, Buurman WA, & Griffioen AW (2008) Convergence and amplification of toll-like receptor (TLR) and receptor for advanced glycation end products (RAGE) signaling pathways via high mobility group B1 (HMGB1). *Angiogenesis* 11(1):91-99.

196. Chadwick B, Kidd T, Sgouros J, Ish-Horowicz D, & Frischauf AM (1999) Cloning, mapping and expression of< i> UBL3</i>, a novel ubiquitin-like gene. *Gene* 233(1):189-195.

197. Hoek KS, *et al.* (2008) Novel MITF targets identified using a two-step DNA microarray strategy. *Pigment cell & melanoma research* 21(6):665-676.

198. Garraway LA, *et al.* (2005) Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* 436(7047):117-122.

199. Wang R CK, Wilshire J, Kowalik U, Hovinga KE, Geber A, Fligelman B, Leversha M, Brennan C, Tabar V (2010) Glioblastoma stem-like cells give rise to tumour endothelium. *Nature* 468(7325):829-833.

200. Karapetis CS, *et al.* (2008) K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N. Engl. J. Med.* 359:1757-1765.

201. Boucneau J, *et al.* (2005) Gene expression profiling of cultured human NF1 heterozygous (NF1+/–) melanocytes reveals downregulation of a transcriptional cis-regulatory network mediating activation of the melanocyte-specific dopachrome tautomerase (DCT) gene. *Pigment cell research* 18(4):285-299.

202. Friedman HS CO, Skapek SX, Ludeman SM, Elion GB, Schold SC Jr, Jacobsen PF, Muhlbaier LH, Bigner DD (1988) Experimental chemotherapy of human medulloblastoma cell lines and transplantable xenografts with bifunctional alkylating agents. *Cancer Res* 48(15):4189-4195.

203. Friedman HS SSJ, Bigner DD (1986) Chemotherapty of Subcutaneous and Intracranial Human Medulloblastoma Xenografts in Athymic Nude Mice. *Cancer Res* 46:224-228.

204. Herbst RS, *et al.* (2006) Clinical Cancer Advances 2005; major research advances in cancer treatment, prevention, and screening - a report from the American Society of Clinical Oncology. *J. Clin. Oncol.* 24:190-205.

205. Breathnach OS, *et al.* (2001) Twenty-two years of phase III trials for patients with advanced non-small-cell lung cancer: sobering results. *J. Clin. Oncol.* 19(6):1734-1742.

206. Minna JD, Gazdar AF, Sprang SR, & Herz J (2004) Cancer. A bull's eye for targeted lung cancer therapy. *Science* 304(5676):1458-1461.

207. Vogel CL, *et al.* (2002) Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J. Clin. Oncol.* 20:719-726.

208. Pao W, *et al.* (2004) EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proceedings of the National Academy of Sciences of the United States of America* 101(36):13306-13311.

209. Paez JG, *et al.* (2004) EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 304:1497-1500.

210. Lynch TJ, *et al.* (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *The New England journal of medicine* 350(21):2129-2139.

211. Janne PA, *et al.* (2004) Outcomes of patients with advanced non-small cell lung cancer treated with gefitinib (ZD1839, "Iressa") on an expanded access study. *Lung Cancer* 44(2):221-230.

212. Karapetis CS, *et al.* (2008) K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N. Engl. J. Med.* 359:1757-1765.

213. Huang E, *et al.* (2003) Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat Genet* 34(2):226-230.

214. Liu Z, *et al.* (2008) Singular value decomposition-based regression identifies activation of endogenous signaling pathway *in vivo*. *Genome Biology* 9(12):R180.181 - R180.111.

215. Bild AH, *et al.* (2009) An integration of complementary strategies for gene-expression analysis to reveal novel therapeutic opportunities for breast cancer. *Breast Cancer Res* 11(4):R55.

216. Loboda A, *et al.* (2010) A gene expression signature of RAS pathway dependence predicts response to P13K and RAS pathway inhibitors and expands the population of RAS pathway activated tumors. *BMC Medical Genomics* 3(26):1775-8794.

217. Moulder S, *et al.* (2010) Development of candidate genomic markers to select breast cancer petients for dasatinib therapy. *Mol. Cancer Ther.* 9:1120-1127.

218. Huang ES, Black EP, Dressman H, West M, & Nevins JR (2003) Gene expression phenotypes of oncogenic signaling pathways. *Cell Cycle* 2(5):415-417.

219. Chang JT, *et al.* (2011) SIGNATURE: a workbench for gene expression signature analysis. *BMC Bioinformatics* 12:443.

220. Network TCGAR (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455(7216):1061-1068.

221. Parker JS MM, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS. (2009) Supervised risk

predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 27(8):1160-1167.

222. Consortium M (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gen expression measurements. *Nat. Biotechnol.* 24:1151-1161.

223. Barry WT, *et al.* (2010) Intra-tumor heterogeneity and precision of microarray-based predictors of breast cancer biology and clinical outcome. *J. Clin. Oncol.* in press.

224. West M, *et al.* (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A* 98(20):11462-11467.

225. Spang R, *et al.* (2002) Prediction and uncertainty in the analysis of gene expression profiles. *In Silico Biol* 2(3):369-381.

226. Lee Y, *et al.* (2008) Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. *BMC Med. Genomics* 1:52.

227. Wang Y, *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365(9460):671-679.

228. Ivshina AV, *et al.* (2006) Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.* 66:10292-10301.

229. Lewis BP, Burge CB, & Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120(1):15-20.

230. CS C (2011) Evolving evidence implicates cytomegalovirus as a promoter of malignant glioma pathogenesis. *Herpesviridae* 1(2):10.

231. Soroceanu L ML, Bezrookove V, Harkins L, Martinez R, Greene M, Soteropoulos P, Cobbs CS. (2011) Human cytomegalovirus US28 found in glioblastoma promotes an invasive and angiogenic phenotype. *Cancer Res* 71(21):6643-6653.

232. Wang Q, Carvalho C, Lucase JE, & West M (2007) BFRM: Bayesian factor regression modelling. *Bulletin of the International Society for Bayesian Analysis* 14:4-5.

233. Chang JC, *et al.* (2007) Decomposing cellular signaling pathways into functional units: a genomic strategy. *Manuscript submitted*.

234. Hans C, Dobra A, & West M (2007) Shotgun Stochastic Search for Large P Regression. *J. Am. Stat. Assoc.* 102:507-516.

235. LaBreche HG, Nevins JR, & Huang E (2011) Integrating Factor Analysis and a Transgenic Mouse Model to Reveal a Peripheral Blood Predictor of Breast Tumors. *BMC medical genomics* 4(1):61.

Biography

Jason Windham Reeves was born in Rochester, NY on October 27th, 1982. He grew up and was raised in Wilmington, NC, and attended college at the Georgia Institute of Technology, receiving a bachelor of science in Applied Biology in 2004. Jason worked at Emory University in Dr. Gray Crouse's lab studying mismatch repair in *S. cerevisiae* until he joined Duke University's Program in Genetics and Genomics in 2006. At Duke, Jason joined Dr. Joseph Nevins' lab in 2008 studying the role of signaling deregulation in glioblastoma. In addition to his research, he worked as a fellow with the brain cancer research foundation Accelerate Brain Cancer Cure (ABC^2) in 2012.

Publications

Reeves JW, Persson O, Lucas JE, Gatza ML, Keir S, Reardon D, Kanaly C, Sampson J, Nevins J. A Pathway Association Framework for Identification of Significant Genome Alterations in Glioblastoma. *Manuscript Under Review*

Reeves JW, Chang JT, Lucas JE, Barry WT, Nevins J. Utilization of Cell Line Derived Genomic Signatures to Predict Pathway Activity in Human Tumors and Prospective Clinical Studies. *Manuscript Under Review*

Mravinac B, Sullivan LL, Reeves JW, Yan CM, Kopf KS, Farr CJ, Schueler MG, Sullivan BA. Histone modifications within the human X centromere region. PLoS One. 2009 Aug 12; 4(8): e6602

Kow YW, Bao G, Reeves JW, Jinks-Robertson S, Crouse GF. Oligonucleotide transformation of yeast reveals mismatch repair complexes to be differentially active on DNA replication strands. Proc Natl Acad Sci USA. 2007; 104(27): 11352–11357

Nicholson A, Fabbri RM, Reeves JW, Crouse GF. The effects of mismatch repair and RAD1 genes on interchromosomal crossover recombination in *Saccharomyces cerevisiae*. Genetics. 2006 Jun; 173(2): 647-59

Williams TM, Fabbri RM, Reeves JW, Crouse GF. A new reversion assay for measuring all possible base pair substitutions in *Saccharomyces cerevisiae*. Genetics. 2005 Jul; 170(3): 1423-6.

McDermid KJ, Gregoritza MC, Reeves JW, Freshwater DW. Morphological and Genetic Variation in the Endemic Seagrass *Halophila hawaiiana* (Hydrocharitaceae) in the Hawaiian Archipelago. Pacific Science. 2003; 57(2): 199-209