A STUDY OF BREAST CANCER HETEROGENEITY AND MOLECULAR MECHANISMS OF METASTASIS

By

Daniel Patrick Hollern

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Cell and Molecular Biology – Doctor of Philosophy

2015

ABSTRACT

A STUDY OF BREAST CANCER
HETEROGENEITY AND MOLECULAR MECHANISMS OF METASTASIS

By

Daniel Patrick Hollern

The biggest challenges clinicians face during treatment of breast cancer are tumor heterogeneity and tumor metastasis. With breast cancer tumor heterogeneity, the problem is that the genomic variability within tumors and between patients limits the efficacy of breast cancer therapy. Directed therapies for specific types of breast cancer improved breast cancer survival times, yet due to the molecular complexity of breast cancer, treatment is still inadequate; with tumors initially regressing only to reoccur and become resistant to therapy. Many reoccurring tumors manifest as distant metastasis. It is these metastases that lead to breast cancer lethality. To simplify the molecular complexity of breast cancer, researchers have taken advantage of mouse models where different cancer-causing events found in human breast cancer are used to initiate mammary tumors in mice. However, the degree to which mouse models are reflective of the heterogeneity of human breast cancer needed to be demonstrated. If mouse models with relationships to individual types of human breast cancer could be identified, such a finding would represent a major breakthrough and enhance the research of mechanisms and treatments for drivers of breast cancer progression using mouse models.

To test the hypothesis that genomic similarities exist between mouse models of breast cancer and human breast cancer, I characterized MMTV-Myc initiated tumors that had demonstrated histological heterogeneity. Using bioinformatic analysis of tumor gene expression data from MMTV-Myc mouse mammary tumors and human breast cancer samples, molecular similarities and mouse human counterparts were identified. As a result, I hypothesized that

molecular similarities between mouse and human breast cancer are widespread. To this end, I generated and analyzed a database of gene expression data from over 1000 mouse mammary tumors and over 1000 human breast tumors. I detected relationships between individual mouse model tumors and specific types of human breast cancer through gene expression patterns. This was extended to predict which signaling pathways were activated in both human breast cancer and the mouse models. This novel approach established relationships between individual mouse mammary tumors and human breast cancer, identifying shared pathways that may contribute to tumor progression in mouse and human breast cancer.

Using this database as a predictive resource, I developed the hypothesis that the E2F transcription factors regulate breast cancer metastasis. Using a genetic test in the MMTV-PyMT mouse model, I show that E2F1 and E2F2 are critical for progression through multiple stages of metastasis. Using predictive informatics and gene expression analysis, I show that multiple pro-metastatic features are impacted with E2F loss, including tumor angiogenesis and activation of the pro-metastatic hypoxia response gene expression program. As part of uncovering E2F1's role in tumor metastasis, I uncover new regulators of metastasis: Adm and Fgf13.

Collectively, the work in this dissertation demonstrates that integrating gene expression analysis, bioinformatics, mouse models and multiple experimental techniques provide the unique capacity to study the complex molecular differences and mechanisms across the spectrum of human breast cancer. Importantly, these strategies have allowed us to credential mouse models for relevance to human breast cancer and identify mechanistic features of breast cancer metastasis.

This work is dedicated to:
My wife, Alexandria.
My parents: Patrick and Michelle.
My grandparents: Dale, Isabelle, Ken, and Mary Lou.
My brothers: Richard and Joseph.
My sisters: Rachael, Amanda, Colleen, and Shannon.

ACKNOWLEDGEMENTS

The work presented within this dissertation could not have been possible without the support of the amazing people in my life. As such, I would like to acknowledge their efforts as part of my appreciation for them.

First of all, my family has been a blessing with motivation and support. Thank you to my wife, Alexandria, who has allowed me to pursue my ambition and dream of being a scientist. The sacrifices you have made are truly selfless and the patience you have shown for my busy schedule is incredible. Thank you to my parents Patrick and Michelle, who have invested their entire lives into me. I would not have made it this far in my education and career without you. Thank you to my grandparents, brothers, and sisters for the lifetime of love, support, and friendship.

At Michigan State, I have been privileged to work with people who have made a tremendous impact in my life and in my research. In particular, I would like to acknowledge my mentor Eran Andrechek. Thank you for giving me the opportunity to work with you. I appreciate all of the effort and time that you put into my training. You have been the most influential person I have met during my education. I firmly believe you have given me the skills and guidance I needed to pursue this career.

A special thank you to my committee: Dr. David Arnosti, Dr. Chengfeng Yang, Dr. Hua Xiao, and Dr. Christina Chan for the letters of recommendation, time spent listening to presentations and the valuable feedback that was instrumental in my research. Similarly, I'd like to acknowledge members of the cancer research network, especially Dr. Sandra Haslam, Dr. Michelle Fluck, and Dr. Kathy Gallo for their letters and guidance. Finally, I would like to thank

Dr. Susan Conrad and Dr. Kathy Meek and the Cell and Molecular Biology program for funding and providing me with the opportunity to earn my degree at Michigan State University.

Thank you to Lauren Aitch and the Aitch foundation for funding my research. I am incredibly honored to work with your foundation and will be forever grateful for the experiences and opportunities you have unlocked for me.

Thank you to members of the Andrechek lab. Thank you for all of the shared reagents, protocols, advice, good times and teamwork. I especially want to thank Jordan Honeysett for managing the breeding of the mice, genotyping and his help on my project, he was a big reason I was able to meet the MCB review effectively and quickly. I also want to thank Jon Rennhack for developing CRISPR protocol. This allowed me to work on other projects while they perfected the protocol. I also need to thank Sean Misek and Lisa Gullish for their help with my research. Thank you to Sophia Lunt for career advice and allowing me to collaborate on her Molecular Cell paper. It has been a joy to work with each person in the lab and I will miss all of you when I move to North Carolina.

Finally, thank you to Michigan State University. Attending this university was my dream and coming here has been a life changing experience beyond description. This is truly a special place.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO SYMBOLS OR ABBREVIATIONS

Adm- adrenomedullin

Areg- amphiregulin

Coro1C- Coronin, Actin Binding Protein, 1C

CRISPR- clustered regularly interspaced short palindromic repeats

DMBA- 7,12-dimethylbenz[a]anthracene

DMEM- Dulbecco's modified Eagle medium

Egfr- epidermal growth factor receptor

EMT- epithelial to mesenchymal transition

ENCODE- Encyclopedia of DNA Elements

ER- estrogen receptor

Fgf 7- fibroblast growth factor 7

Fgf13- fibroblast growth factor 13

GATHER- gene annotation tool to help explain relationships

GEM- genetically engineered mice

GSEA- gene set enrichment analysis

Her-2- human epidermal growth factor receptor, 2

Hbegf- heparin-binding EGF-like growth factor

HER2- human epidermal growth factor receptor 2

Hspb1- heat shock protein beta 1

L1Cam- L1 cell adhesion molecule

Lama 5- laminin 5

MMTV – mouse mammary tumor virus

PCR- polymerase chain reaction

PR- progesterone receptor

PCA- principle components analysis

Plaur- plasminogen activator, urokinase receptor

PyMT- polyoma middle T

qRT- quantitative reverse transcriptase

SAM- significance analysis of microarrays

SVD- singular value decomposition

TAG-large T antigen

TCA - the citric acid cycle

Tead1- TEA Domain Family Member 1 (SV40 Transcriptional Enhancer Factor)

TCGA-The Cancer Genome Atlas

Tgm2- transglutaminase 2

TNF- tumor necrosis factor

Tgfb- transforming growth factor, beta

TUNEL- Terminal deoxynucleotidyltransferase-mediated dUTP-biotin nick end labeling

TRANSFAC- transcription factor database

Vegfa- vascular endothelial growth factor

WGCNA- weighted correlation network analysis

INTRODUCTION

GENE EXPRESSION MICROARRAYS

The advent of high throughput technologies has become an integral tool in understanding biological systems. Indeed, the use of gene expression technology and sophisticated data analysis methods has provided a global view of the genetic changes that occur with the development and progression of cancer. Amongst the various gene expression analysis tools, two of the most frequently used platforms are the Agilent and Affymetrix microarrays.

Agilent gene chips can measure the expression of tens of thousands of messenger ribonucleic acid (mRNA) transcripts. Agilent offers both one color and two color arrays. However, since two-color arrays were utilized to generate the data for my work and for brevity, two color arrays will be the focus of this brief overview of Agilent array gene expression profiling. Current Agilent gene chips are built by printing oligonucleotide molecules on the glass surface of a chip [1, 2]. Typically Agilent oligonucleotide probes are 60 nucleotides (also referred to as 60-mer) in length. These printed molecules are designed to construct unique complimentary sequences to serve as annealing probes for tens of thousands of individual gene products. Further, each probe is printed according to specific coordinates, so that the signal detected during scanning can be assigned to the probe for each specific gene. Preparing mRNA for array hybridization requires several steps[3]. First, mRNA is converted to complementary deoxyribonucleic acid (cDNA) using a reverse transcriptase enzyme. Following this step, complementary RNA (cRNA) is transcribed using a RNA polymerase enzyme. To label cRNA with fluorescent dyes, two separate tubes are prepared for each sample. In one tube, cyanine 3-labeled cytidine triphosphate (cy3, a green fluorescent dye) is added so that it can be incorporated to the cRNA. In a separate tube, cyanine 5-labeled cytidine triphosphate (cy5, a red fluorescent dye) is added and incorporated into the cRNA. At this point, cRNA is purified to

remove unincorporated dye-labeled nucleotides which could cause background signal on the array each tube of labeled and the cRNA is assessed for concentration and quality. Next, the cy3 and cy5 labeled cRNAs are pooled and simultaneously hybridized onto the gene chip. After hybridization, the chip will be scanned using a confocal laser scanner on two channels : at 635nm wavelength to capture the cy5 signal and at 532nm wavelength to capture the cy3 signal. The intensity of fluorescence is relative to the abundance of the original mRNA transcript. Along with quantitation, intensity readings are corrected for background signal and noise. Next intensity readings for each channel are normalized. Normalization corrects for technological biases that cause non-biological perturbations to data. Although a variety of approaches are available here, Lowess normalization, which corrects for differences in how the cy3 and cy5 dyes impact cRNA hybridization and therefore intensity readings [4-7] , was the method utilized in the Agilent data used in this dissertation. Finally, to obtain actual expression values for each sample the log (2) ratio between the cy5 and cy3 channels are calculated. Frequently used software for normalization and intensity calculations of Agilent chips includes LIMMA [8] and other commercial products like Agilent's Feature Extraction Software.

Although somewhat similar in the approach there are key differences in how Affymetrix gene expression arrays work. For example, Affymetrix analysis differs from Agilent at the *in vitro* transcription step. Instead of dye labelled nucleotides, Affymetrix protocols utilize biotin-labelled nucleotides for incorporation into the cRNA molecules[9]. Taking advantage of streptavidin's binding affinity for biotin, cRNA hybridization to oligonucleotide probes is detected by staining with a fluorescent dye that is coupled to streptavidin. The Affymetrix oligonucleotide probes are synthesized on silicon wafers by photolithography. Affymetrix oligonucleotide probes are 25-mer and there are multiple pairs of probes (16-20), perfect match

and mismatched probes, for each gene. The perfect match probe contains sequence exactly complimentary to unique regions of the 3' end of its corresponding gene and measures the expression of the gene. The mismatch probe differs by a single base at the center of the probe and as a result the binding of the corresponding gene transcript is disrupted. This allows the background and nonspecific hybridization signal to be calculated for the perfect match oligonucleotide [10]. The mismatch probes assist with normalization and calculation of intensity values. One method that utilizes mismatch probes is Mas5 normalization. At the most basic level, Mas5 normalization adjusts each array independently for background and non-specific hybridization by subtracting the signal of the mismatch probes from the signal of the perfect match probe to obtain the expression value for each probe [11]. Another common normalization method for Affymetrix gene chips is robust multi-array analysis (RMA). While this method ignores the mismatch probes it is still able to remove background and deal with probe specific affinity effectively [12]. This method is meant to be employed for multiple chips all from the same batch (samples processed together under the same methods, settings, same facility, etc…). There are basic three steps of RMA normalization: background adjustment, quantile normalization, and median polish summarization. After background is adjusted, intensity values for perfect match probes are log transformed (log (2)). Next is quantile normalization; where the distribution of the intensity values amongst chips is measured and is corrected so that each chip has an equal distribution of intensity values [13]. Finally, median summarization identifies and removes outliers [14]. Conveniently, RMA normalization and Mas5 normalization to obtain gene expression levels can be done utilizing Affymetrix Expression Console Software [15] .

BIOINFORMATIC METHODS

With differences between microarray platforms, as well as differences in technical settings during array analysis, computational methods had to be developed that would remove this artificial variance between microarray studies (also referred to as batch effects) to allow for datasets to be combined. To remove, measure, and visualize these batch effects, one common method is principle components analysis [16]. One of the initial challenges in microarray data analysis and adjusting for batch effects is the high dimensionality of the data that comes with having tens of thousands of probes on an array and often a large number of samples in a dataset. At the most basic level, principal component analysis reduces this dimensionality of large datasets by calculating vectors (or principle components) that describe variability in the dataset [17]. To identify and describe each principle component, a mathematical method known as singular value decomposition is often employed. Singular value decomposition works to break down high dimensional expression data using linear reduction. In this way, singular linear vectors are calculated that simplify and summarize the both samples and genes [18]. A simplified way to think about this is that singular value decomposition identifies genes that correlate with one another across samples and describes them as a linear vector, these vectors are often referred to as "eigengenes". Similarly, samples that correlate with one another can also be described as mathematical vectors, and these are often referred to "eigenarrays". As mentioned, most principle components analysis methods rely on singular value decomposition. With this approach, the vector that describes the greatest amount of the variability in the data is the first principle component. The remaining maximal variance not described by the first principle component will be described with a second vector (the second principle component). This trend continues, with left over variance being described by subsequent vectors until the dataset's

variance is completely described. Applying this, a 2008 publication by Ringner used principle components analysis to summarize a 105 sample, 8,534 probe microarray dataset. He found that only 104 principle components were needed to describe the variance in the entire dataset [16]; thus illustrating how this method reduces dimensionality into more manageable units.

As to how principle component analysis assists with identifying and removing batch effects, consider that differences in chip type, protocol differences amongst labs, and cRNA/cDNA synthesis can cause large disparities in the scaling and variance between separate gene expression datasets. As a result, the perturbations of batch effects on the data is likely to be captured within the initial first several principle components. This makes mapping samples to their position on even the first three principle components a useful approach for visualizing and predicting batch effects between studies. Indeed, labs working to identify and correct batch effects frequently relied on singular value decomposition and principle components analysis [18-20]. While good as starting point, new batch adjustment correction approaches were developed that have been shown to outperform the singular value decomposition / principle components analysis method [21]. However, principle components analysis is still frequently used for viewing combined datasets to predict the presence of batch effects.

Developing robust methods for mediating batch effects continues to be a major goal within the field of bioinformatics. Amongst the wide variety of methods, Distance Weighted Discrimination (DWD), Combatting' Batch Effects When Combining Batches of Gene Expression Microarray Data (COMBAT), and Bayesian Factor Regression Modelling (BFRM) have all proven to be reliable tools for removing technical artifacts from microarray data.

Distance Weighted Discrimination was demonstrated to correct for technical biases in microarray data [21]and is based on an algorithm known as Support Vector Machines. The

Support Vector Machines algorithm is a supervised approach that builds a hyperplane in infinite-dimensional space [22]. Perhaps an easier way to think about this, is the algorithm identifies a line or plane in space that features the maximal difference/separation between to gene expression datasets. Distance weighted discrimination takes advantage of this strategy to remove batch effects by first projecting the different batches on the hyperplane (also referred to as the distance weighted discrimination plane). Next, then the mean is calculated for all genes within each batch separately. Finally, the distance weighted discrimination plane is subtracted out for the samples in each separate batch and multiplied by the projected mean for each gene. Essentially, this evens the scale of gene expression values between two datasets and removes the distance between each dataset so that Support Vector Machine could no longer identify differences between batches. One weakness of Distance Weighted Discrimination is that it does not work well when only a few samples are featured in a batch [23].

In the case of dealing with smaller datasets, an appropriate option would be to use COMBAT. COMBAT can deal with both small and large batch sizes and utilizes an Empirical Bayes approach to removing batch effects [24]. Importantly, this method assumes that batch effects are having the same impact on gene expression values. In the context of COMBAT, Empirical Bayes is being used to estimate degree of batch effects between two datasets based on the distribution of the data within each batch. As part of COMBAT a location and scale adjustment method is employed, where genes across each batch are analyzed and this information is used to establish the batch effect parameter describing the mean and variance for each gene. Together, distribution and batch estimates are then used to establish the correction the data for batch effects. Thus in overly simplified terms, COMBAT compares the variance of genes expression values between each batch to then establish model for adjusting gene

expression values in each batch so that there is similar degree of variance in the expression values across all batches.

While COMBAT and DWD can be used for correcting gene expression data for any microarray platform, the BFRM method was built specifically to work with Affymetrix datasets. BFRM takes advantage of probes for housekeeping genes. These 60-100 probe sets are present on Affymetrix chips and are used as controls because they will have no biological or hybridization variation across samples. As a result, this information can be utilized for measuring and correcting batch effects across microarray studies. With BFRM, a principal components analysis strategy is used to measure the variance between batches on the basis of these housekeeping genes and establishes the model for removing this variance between each batch [25]. Importantly, after employing BFRM, COMBAT, or distance weighted discrimination it is a good idea to use principal components analysis to test and measure batch correction. Once batch effects have been removed, samples in the dataset are ready to be tested for biologically significant relationships.

One of the most common ways to analyze the relationships amongst samples is by way of unsupervised hierarchical clustering. Unsupervised hierarchical clustering relies on a machine learning strategy referred to as unsupervised learning. The strategy of unsupervised learning dictates that patterns in the data are assembled without the input any information beyond the raw data [26]. Hierarchical clustering treats each data point, whether it be a gene or sample, as a single cluster. The most similar pair of clusters are merged are sequentially merged until all points have been merged into a single remaining cluster and are typically represented as a dendrogram [27]. Putting these concepts together in the context of gene expression analysis, this means that clusters of samples, for example, are assembled without any user input on sample

type or group. Instead, samples are arranged merely on the basis of having similar gene expression profiles. Similarly, when genes are clustered, the ordering of genes into various clusters is dependent on genes sharing similar patterns of covariance across samples in the dataset. In this way, unsupervised hierarchical clustering provides a non-biased approach for measuring and detecting similarities and differences amongst samples in a gene expression dataset.

In addition to measuring the similarities in gene expression patterns, it is also of great interest to identify genes that are altered in expression between groups of samples. The identification of these altered genes and their degree of variation is often referred to as a fold change analysis. Fold change analysis is especially useful for identifying genes that change between two biological states. However, when dealing with multiple samples in each biological condition, it is important to measure the statistical significance of the gene expression changes. To achieve this, an approach and software referred to as Significance Analysis of Microarrays or SAM was developed [28]. This method relies on t-tests and calculates how much the expression of each gene changes in relationship to the standard deviation of repeated measurements. For genes with changes greater than a user defined threshold, SAM measures the false discovery rate. The false discovery rate relates to the percentage of genes identified by chance. In this way, the uniformity of the gene expression changes can be assessed.

Once a list of significantly altered genes is obtained, additional discovery and prediction tools can be utilized that go beyond single gene analysis. For example, to explore the relationship of differentially expressed genes to categories relating to cellular functions can be achieved using a gene ontology analysis [29]. The current gene ontology system was developed with the motivation to establish a uniform system for annotating genes according broad categories such as

the biological process, molecular function and cellular component the gene most significantly associates with. There are a variety of free online tools for conducting gene ontology analysis including DAVID [30] and GATHER [31]. One of the weaknesses of gene ontology analysis is that the categories for gene association are broad and thus limits making more specific predictions about the gene expression changes.

Despite the limits of gene ontology analysis, GATHER has several additional predictive tools that can assist in making mechanistic predictions based on gene expression changes. For example, GATHER allows users to test lists of altered genes for significant overlap with KEGG pathways. KEGG pathways are a collection of major cell signaling pathways that are annotated for which gene products are participants and serves as an excellent prediction tool for querying gene lists to predict alteration of specific pathways [32]. In addition to KEGG pathways, GATHER also offers a tool to query the TRANSFAC database. The TRANSFAC database allows users to access and predict the presence of transcription factor binding sites for their gene of interest [33]. Importantly, GATHER contains measures of statistical significance for gene list queries for overlap with specific KEGG pathways, transcription factor binding sites, and association with gene ontologies. This statistical measure is reflected in the Bayes score. This score reflects the degree of relationship of a particular annotation with the list of genes, where the higher the Bayes score, the stronger the likelihood that the annotation corresponds to the list of genes being queried than other genes in the genome. Together, GATHER provides comprehensive tool for investigating gene lists to predict what the alteration of the listed genes may represent at a more functional level.

Another useful tool for investigating and understanding gene expression changes between two groups of samples is Gene Set Enrichment Analysis (GSEA) [34]. GSEA utilizes a database

of wide variety of gene sets derived from scientific investigations. For example there are gene sets that specify the genes that are up or down regulated in response to pathway activation in specific cells (i.e. genes upregulated in response to AKT activation). There are other gene sets that have identified genes that are up and downregulated in specific cell types. For example, a gene set that identifies the genes up-regulated in comparison of CD4 positive T cells versus myeloid cells. There also gene sets that define the genes that correspond to very specific processes, such as tumor angiogenesis, metastasis, and hypoxia. For statistical analysis, genes are ordered on the basis of fold change between two user defined groups, thus establishing a ranked list. Then, gene sets are mapped on to the ranked list. Each time a gene in the gene set is encountered it drives up the enrichment score, therefore, the higher the rank of the gene (meaning the higher the fold change), the greater the degree that the enrichment scores will increase. In addition, a false discovery rate is also included. Thus, gene sets that have a high degree overlap with the genes that are consistently the most up or down regulated genes are those that will have the highest enrichment score and have the greatest statistical significance. All in all, this collection provides a comprehensive tool for making very specific bioinformatic predictions and measuring the statistical significance of those predictions.

Each of the previous methods have focused on computationally comparing groups of samples. However, it is also desirable to make predictions on each sample individually. One of the more powerful bioinformatic methods to be developed is the gene signature approach to activation of major cell-signaling pathways on individual samples that was developed by the Nevins lab [35]. To establish pathway activation gene expression profiles, key cell signaling molecules are overexpressed in plates of human mammary epithelial cells using adenovirus. As a control, adenovirus is used to overexpress GFP. By comparing gene expression profiles between

the pathway activated samples and the GFP control cells, a transcriptional signature of pathway activation is obtained and is used to establish a gene signature. This transcriptional response of pathway activation is used to build a model that can be used to query other samples and predict whether or not a cell signaling pathway is active or not.

Building the model for predicting pathway activation integrates multiple statistical methods [35]. Specifically, for each training dataset signature genes are identified using singular value decomposition and principle component analysis [36]. With this the user is offered the opportunity to adjust the parameters of the model they are about to build for signature setup and pathway activation prediction. For example, the user can choose how many genes and metagenes the signature model should contain. A metagene is derived using SVD and is a group of genes that show a constant pattern of expression in relation to a discernable phenotype [35-37]. Thus, this approach allows the genes that define pathway activation to be identified. The reason a user should adjust the number of metagenes in the model is to enhance discrimination of the GFP overexpressing cells from the pathway activated cells. During model assembly, metagene scores are calculated are calculated for genes and training samples alike. Metagene scores reflect the ability to differentiate or predict the pathway on state from the pathway off state. Thus, to calculate the probability of pathway activity in a sample, the sample is mapped to the metagene signature (based on the same genes and metagenes in the signature model). Based on the metagene score for the sample that was mapped to the model, a probability of pathway activation is assigned using probit binary regression. At this step, probit regression predicts the probability of a binary outcome: zero, where the pathway is off and one, where the pathway is on. As mentioned, this is based on metagene scores. As a result, there is a correlation between the metagene score and probability that the pathway is activated. Samples that map to high metagene

scores have a high probability of pathway activation and samples with a low metagene score have a low probability of pathway activation [35]. This is done on training data on its own, on the training data during leave one out cross validation (more on this later) and similarly on non-training data samples that are being tested for probability of pathway activation.

To go into more detail, the process of mapping samples to the metagene score is achieved by Bayesian fitting of probit binary regression models. More specifically, the Bayesian analysis applied in this scenario is known as iterative Markov chain Monte Carlo (MCMC) simulation methods[36]. This approach uses multiple simulations (or iterations) to assess probability of a particular outcome out of a multiplicity of choices. Inherent to Bayesian approaches, application of MCMC works to generate probabilities based on prior knowledge [36, 38, 39]. The prior knowledge in this case is that we know which samples in the training data the pathway is activated in. Therefore, we also know the metagene structure that defines pathway activation. Therefore, the MCMC model is informed as to the conditions where pathway activation is more likely. Using MCMC, we can simulate thousands of iterations of fitting the regression models to the metagene signature to both predict probability of pathway activation and calculate the degree of certainty for each of these predictions (since it reports the range of the probabilities calculated over the multiple simulations)[36, 40].

With the application of predictive bioinformatic methods, an important issue to consider is overfitting. The term overfitting refers to a category of technical issues that can be experienced with computational modeling [41, 42]. The reason overfitting is problematic is the risk of generating non-predictive models that do not replicate or validate; or a simpler way to put it: overfitted models do not describe real life. Overfitting error stems from datasets that contain measurements for an abundant number of features, but contain disproportionately fewer

replicates (ie, the training dataset size is small). Generating models on datasets containing high dimensional measurements on relatively few cases runs the risk of invalid feature selection. As a result the model is describes random error. As a result, the assembled model only works on the training data and fails to accurately predict or validate when applied to separate, but similar data scenarios. Using a more concrete example, such as model assembly from gene expression microarray datasets, the concept of overfitting can be made clear.

The risk for overfitting is elevated in microarray datasets, where gene expression for tens of thousands of genes are measured, but often times a small number of samples are measured. In particular, problems can manifest from "noisy" genes that have large, but irrelevant deviation. Some examples where this can arise from are probe sets that have poor affinity (an issue that can be addressed by mismatch probes), or transcripts that have higher tendencies to form secondary structures, or transcripts with a shorter half-life. In the end, this results in false positives in the identification of truly differentially expressed genes [43, 44] and increases the risk of overfitting in classification and signature generation methods [45]. Providing an example for how measuring random error can generate invalid models, take for example an analysis using a training dataset of only two samples: one sample where the Egfr pathway is activated and one sample where it is not. One might be tempted to think that due to the differences between these two sample types, a model to describe activation of the Egfr pathway can be assembled. However, due to the high number of genes measured for each sample and the limited number of samples, there is a high likelihood that random selection of genes (many of which could be variable due to noise and probe set efficiency) could generate model that describes the variance between the Egfr activated and non-activated samples. However, due to overfitting the genes selected would not be likely to be biologically significant to Egfr signaling. In this scenario, the

model would be unique to the training dataset only and would therefore fail during predictive analysis and validation on other datasets. These types of overfitting issues can be encountered in a number of modeling approaches including regression, fold change analysis, and gene signature approaches to modeling specific features. While overfitting tends to occur more frequently with supervised analysis [46] highly variable probe sets can cause overfitting in unsupervised approaches as well, such as in unsupervised hierarchical clustering. However, for each of these methods, there are a number of strategies that can be employed to reduce the frequency and magnitude of overfitting.

One strategy is to increase the number of samples used to inform model assembly. Take for example, using a regression strategy to identify genes that correlate with metastatic outcome; it is easy to find genes that by chance correlate with metastatic outcome with limited observations. If you have two samples, one with no metastasis, and one with ten metastasis, even random selection of genes find an incredibly large number genes that correlate with metastasis given those two data points. However, if you had a thousand samples with metastatic annotations, random selection would be unlikely to identify genes that correlate with metastatic outcome and instead the subset of genes that did correlate with increasing metastatic outcome would have a higher likelihood of being biologically significant. Increasing sample size is also especially useful for reducing overfitting while developing gene signatures and conducting fold change analysis. By increasing the sample size we reduce the likelihood that randomly selected genes can distinguish two known groups decreases. In many cases increasing the number samples gives a better indication of the false discovery rate (which is measured during SAM analysis) for each gene. As a result, false discovery rate is an important measure to pay attention to during fold change analysis as genes with a lower false discovery may have a better chance of

being biologically significant (and not noise). Similarly, increasing sample size reduces the false discovery of genes that define a given feature and this increases the likelihood that gene signature models are predictive outside of the training dataset. Likewise, for unsupervised hierarchical clustering, having a higher sample size is important to reduce overfitting. With low sample numbers, there is a better chance that "noisy" probe sets to be selected and inform cluster assembly. This could lead to incorrect predictions about genes that distinguish sample types or incorrect indication of sample to sample relationships.

Sometimes it is not economically feasible to include additional samples. As a result a number of statistical approaches have been developed to address specific issues with overfitting. For example to deal with noisy probe sets that cause overfitting issues, Talloen et al. developed a method for filtering out these non-informative "noisy" genes [44]. Their approach makes use of the multiple probes for the same target mRNA on an Affymetrix gene chip. By utilizing these repeated measures, they obtain a signal to noise ratio for each probe set on the chip. They combine this information to inform a factor analysis (similar to principle component analysis) that summarizes the variance of each gene based on its probe sets across microarrays. Importantly, to test their approach they used a dataset with spike-in transcripts for specific probesets. Since, spike-ins provide a known amount of RNA, they can measured degree of hybridization between the spike-ins and the control probes to calculate hybridization efficiency. In addition, they can detect coordinate changes in signal across the probets with increases in RNA concentration across arrays. A probe set was deemed informative if the probes showed the same decrease or increase in intensity readings with the changed in RNA concentration. Those probes that did not meet this criteria were called non informative as the signal did not exceed the noise for this probeset, supporting exclusion. They use this information to establish parameters

for a Bayesian model for identifying non-informative genes. As a result, this method may reduce overfitting and prove useful as a pre-processing tool that can be employed prior to other bioinformatic applications such as clustering, regression, signature development.

Importantly, the method described by Nevins and colleagues [35-37, 47] for developing gene signatures to predict pathway activation has built in steps that reduce and assess overfitting. The first approach that reduces overfitting is the use of SVD to select metagenes that consistently discriminate pathway on from pathway off conditions. As reported, this approach eliminates non-discriminatory genes that often contain noise [36]; therefore limiting a major source to overfitting. To assess possible overfitting, Bayesian fitting of the binary regression models and the metagene signature is featured. This allows for metagene signature to be tested for its ability to classify training data samples correctly and measure the degree of certainty by which it does so [37]. An additional overfitting assessment feature is leave one out cross validation [48]. In leave one out cross validation, one sample in the training dataset is left out, the metagene model is regenerated from the remaining samples and the new model is used to predict remaining training dataset[35, 47]. This is performed continuously, until all samples have been left out and classified. By performing leave one out cross validation, the error rate of the model can be determined. As a result, this gives an indication as to how strong the training dataset is and its error rate during generation of a predictor (where models with high error rates would reveal possible overfitting).

Together, the previously mentioned methods and strategies highlight a few amongst many of the approaches that have been developed to deal with overfitting. While these methods offer promise to reduce false discovery, there are additional steps following application of predictive models to test for overfitting. One tactic is to test the model on an independent dataset

containing the appropriate corresponding sample types. In the context of gene signatures and fold change analysis, this would determine the degree to which model is correct by identifying the proportion of genes that were consistently detected. Those that passed criteria on both datasets could establish a consensus gene set carrying more biological significance. Similarly, in the context of unsupervised hierarchical clustering, testing on an independent dataset would determine whether or not the same relationships amongst sample types are upheld and whether or not specific genes are critical to the detected relationships. An additional control for overfitting in unsupervised hierarchical clustering is using principle components analysis. This would test whether or not the detected relationships were specific to the hierarchical clustering approach, or hold up to an additional test.

While many of the approaches previously discussed can reduce the likelihood for overfitting, the best way to deal with overfitting is validation. While validation goes beyond testing multiple independent datasets, there are some ways to validate using bioinformatics approaches. For example, if you generate a signature that predicts the probability of Egfr activation in tumor samples, the signature could be bioinformatically validated on gene expression data for tumor samples where Egfr signaling was inhibited. If the samples where Egfr was inhibited show consistently low probability of pathway activation, there is evidence that the gene signature is validated. The other validation approach relies of biochemical means. For example, the Egfr activation signature could be tested on a dataset of tumor samples with a corresponding tissue bank. Examining to the samples with the highest and lowest probability of Egfr pathway activation, the signature could be validated by doing a western blot for the active version of phosphorylated Egfr.

BREAST CANCER

The use of microarray technology and sophisticated bioinformatic methods has enhanced our understanding of breast cancer tremendously. Breast cancer refers to the transformation of cells in the breast to a state of abnormal cellular behavior resulting in the formation of malignant tumors. The breast consists of milk producing lobules, and ducts that serve as connective tubes for transporting the milk to the nipple. The remainder of the breast consists of adipose tissue, connective tissue (for example, blood vessels, extracellular matrix, and other individual cell types like immune cells) and lymphatic tissue. The cells that give rise to breast cancer are those cells of the lobules and cells of the ducts. Importantly, the incidence ductal carcinoma is higher than lobular carcinoma [49]. Overall, a recent epidemiological report shows breast cancer affects one in eight women, with more than 230,000 new cases each year [50]. The molecular basis of these tumors are known for some of the heritable cancers, with BRCA1 and BRCA2 mutations occurring in less than 10% of breast cancers [51]. Other cancers are a result of gene amplification and overexpression, one of the most well studied being the amplification of HER2 [52] [53]. In addition, to high incidence, breast cancer is second leading cause of cancer death among women with nearly 40,000 deaths each year[50]. This high mortality rate is largely due to tumor heterogeneity and tumor metastasis to distant organs.

TUMOR HETEROGENEITY

In fact, one of the hallmarks of breast cancer is tumor heterogeneity. This refers to the detail that there are many different tumor types across the individual cases of breast cancer. Among the variable features is tumor histology (the histological type of the tumor refers to the morphological and cytological patterns evident within a tumor such as lobular carcinoma versus ductal carcinoma), genetic events (for example, the presence of Her 2 amplification), and

hormone receptor status (estrogen receptor, or ER, status). Each of these differences impact the overall genomic context and phenotype of the tumor. As such, not every tumor can be treated the same way. Traditionally, breast cancer heterogeneity is largely based upon immunohistochemistry for the clinically relevant markers HER-2, ER, and progesterone receptor or PR. To be sure, tumor heterogeneity has provided major challenges for breast cancer therapy, with some success being realized in treating tumors according to their unique features. For example, without pre-selecting breast cancer patients for the Her-2 biomarker, only a small percentage of patients benefit from therapy with Herceptin (a monoclonal antibody targeting Her-2). On the other hand, when Herceptin is assigned exclusively to Her-2 positive patients, a significant improvement in the clinical response is realized [54]. Importantly, diagnostic gene expression assays have been developed to predict clinical outcomes and tailor therapies for breast cancer patients. Such tools have been able to predict the benefit from chemotherapy and have improved life expectancy for patients with estrogen receptor positive breast cancer [55, 56]. While some tumors present molecular alterations that present opportunities for therapies targeting those alterations(for example, Herceptin for Her-2 positive tumors and Tamoxifen for ER positive tumors), there are other tumor types (the triple negative breast cancers, which do not express Her-2, ER, or progesterone receptor) that are limited to surgery, radiation therapy, and general chemotherapy. As a result, one of the major goals of researchers is to expand targeted therapies to triple negative breast cancers and to improve targeted therapy approaches for ER and Her-2 positive patients. This requires a better understanding of the tumor heterogeneity at the genomic level.

Disentangling the heterogeneity of breast cancer has been largely significantly by the use of microarray technology. One of the ground breaking works using this technology, established

the classification of breast tumors into their molecular subtypes based on unique gene expression profiles [57]. As a result of this work, tumors are now described according to their "intrinsic subtype": basal, luminal A, luminal B, her-2 positive, and normal-like breast group. In more recent work, an additional subtype was discovered, the claudin low intrinsic subtype [58]. Relating these intrinsic subtypes back to their clinically relevant markers, the luminal A and luminal B tumors are ER positive, the Her-2 subtype is Her-2 +, while the claudin low and basal subtypes are triple negative [58-60]. Further dissecting triple negative breast cancer, it was found that claudin low tumors have gene expression features resembling mesenchymal cells, while basal breast cancer retains a more epithelial cellular identity [58]. In terms of overall survival, luminal A tumors carry the best prognosis with longer survival times than the other intrinsic subtypes [58, 60]. Importantly, these intrinsic subtypes of breast cancer now serve as the fundamental basis by which researchers classify tumor heterogeneity.

Since the development and identification of the intrinsic subtypes of breast cancer, researchers have expanded on this work to further define tumor heterogeneity. Among these, an important study detailing the pathway activation profiles within the intrinsic subtypes, demonstrated molecular complexity beyond the six subtypes of breast cancer [47]. Specifically, this work identified subgroups within the intrinsic subtypes, totaling up to 17 subtypes of breast cancer on the basis of predicted pathway activation profiles. Importantly, pathway classification and separation of luminal tumors identified a subgroup of tumors that correspond to better overall survival. Similarly, this analysis separated out a subtype of basal breast cancer that had lower activity of the Src signaling pathway that better overall survival than two other subsets of basal breast cancers.

It is also important to note, that this expanded view of tumor heterogeneity was upheld in a large genetic and molecular profiling effort known as The Cancer Genome Atlas (TCGA) project [61]. In this study, breast tumors were analyzed by DNA copy number arrays, DNA methylation, exome sequencing, messenger RNA arrays, microRNA sequencing and reverse-phase protein arrays. In agreement with an expanded number of breast cancer subtypes, sequence and copy number analysis showed that event profiles were largely variable within intrinsic subtypes. Illustrating the utility of gene expression profiling, gene expression analysis was able to capture all of the heterogeneity within the tumors. Expanding upon this work, researchers found a significant correlation between pathway activity and the corresponding copy number alteration[62]. For example, a signature for Her-2 signaling correlated with actual Her-2 amplification events. In addition, this work revealed new potential therapeutic opportunities for a subset of luminal tumors. Taken together, these gene expression profiling studies are important because they not only unravel and characterize tumor heterogeneity but also may serve as a launch pad for new personalized therapeutic strategies.

METASTASIS

In addition to tumor heterogeneity, gene expression analysis has been a powerful tool for investigations into the mechanistic features of breast cancer metastasis. Metastasis is a multi-step process by which tumor cells leave the primary tumor and colonize the local and regional tissues as well as distant organs in the body [63]. First, a metastasizing tumor cell acquires invasive characteristics. The acquisition of metastatic phenotypes is still not completely understood, as multiple factors have been demonstrated to be involved at the early initiation steps of metastasis. However, once this shift to increased invasive potential does occur, tumor cells can begin to

invade local tissue, or enter the lymphatic system where they typical spread to local or regional lymph nodes, or enter the circulation to spread to distant organs. Thus, one of the first rate limiting steps of metastasis to distant organs is the recruitment and development of tumor vasculature. This process by which the tumor cell develops a blood supply is referred to as angiogenesis. A tumor cell will enter the bloodstream in a process known as intravasation. Once in the blood stream a tumor cell must evade a variety of factors that could cause cell death. Finally, as a tumor cell becomes trapped in a capillary bed, tumor cells must extravasate out of the vasculature and into the organ to begin colonization of the distant organ. The most common distant organs for breast cancer metastases to colonize are the bone, liver, brain and lungs [64] ; though metastasis does occur at other organs [65]. While this is a very simplified overview of the steps of metastasis, the mechanism at the molecular and cellular interaction level is incredibly complex; with gene expression changes, molecular alterations, and coordination with companion cells occurring to enable tumor cells to transition throughout each of these steps.

At the clinical level there is a great deal of interest in the metastatic status of the tumor, mainly due to the implications that metastasis has on overall prognosis. In fact, there is staging criteria that in large part is a measure of the degree to which tumor cells have spread[66]. To review this staging criteria, stage 0 breast cancer is also referred to as either ductal carcinoma in situ or lobular carcinoma in situ depending on the location of the tumor. Ductal carcinoma in situ means that cancer cells began to grow abnormally and fill the duct. Similarly, in the case of lobular carcinoma in situ, cancer cells remain localized within the lobule. At stage 0, the cancer cells have not invaded surrounding tissues and as such surgery has a high success rate of removing all of the tumor cells. As a result, nearly 99% breast cancer patients with Stage 0 breast cancer survive for at least five years after being diagnosed [67]. Stage 1 breast cancer refers to

tumors that are either two centimeters or smaller (stage 1A) or where no very small (0.2 millimeters) or no tumor is found however tumor cells are found in local lymph nodes. Similar to stage 0, this type of breast cancer is more readily managed by surgery and carries a five year survival rate of nearly 99% [67]. Stage 2A breast cancer describes tumors that become larger than two centimeters but are still less than five centimeters with no spread to the lymph nodes. Or, the tumor may be smaller than two centimeters with cancer cells found in less than three axillary lymph nodes. For stage 2B, tumor are larger than 2 centimeters but not larger than 5 centimeters and cancers cells are found in the lymph nodes. Another condition that is characterized as 2B is tumors that are larger than 5 centimeters but have not spread to the lymph nodes. Patients with stage 2 types of breast cancer have a five year survival rate of 93% [67]. Stage 3 breast cancer has several subcategories as well. Conditions that meet stage 3A criteria include tumors of any size where cancer cells are found in four to nine axillary lymph nodes. Stage 3B cancer describes cancers where cells have invaded the chest wall and spread to nine axillary lymph nodes. For stage 3C breast cancer, cancer cells have spread to more than ten axillary lymph nodes. Altogether, 72% of the patients with stage 3 breast cancer survive for five years from the time of diagnosis [67]. Stage 4 breast cancer refers to cancers that have spread to distant organs. This stage carries the worse overall prognosis, with 5 year survival rates plummeting down to 22% [67]. As a result, there is great deal of interest in understanding the process of metastasis so that strategies to limit metastatic potential and treat tumors that have formed in distant organs can be realized.

As mentioned earlier, what actually causes tumor cells to become metastatic is still not completely clear, as multiple factors have been demonstrated to be involved at the early initiation steps of metastasis. This includes molecular alterations and gene expression changes to facilitate

invasion and cell migration [68]. There are also the molecular changes that occur for production and secretion of pro-angiogenic factors to recruit blood vessels [69], expression of matrix metalloproteases to digest the extracellular matrix [70], and epigenetic and global genomic changes for a transition to a mesenchymal cell state [69, 71]. Other factors include clonal interaction [72], acquisition of mutations [73], and a large variety influences from the tumor microenvironment including response to hypoxia [74] and immune cell [75, 76]. In addition, there are clear differences amongst the intrinsic subtypes for metastatic potential [77]. More specifically, luminal A tumors exhibited the lowest incidence of metastasis and that were Her-2 positive had the highest incidence.

Similarly, there are a variety of factors that have been identified to be critical for the later steps of metastasis where tumor cells invade and colonize a distant organ. For example, cancer cells have the ability to transition the distant organ microenvironment to a vulnerable, tumor fostering state [78]. Other processes involve recruitment of pro-tumor immune cells [79] and interaction with companion cells, such as the interaction of circulating tumor cells (CTCs) with platelets [80]. Taken together, these multiple factors demonstrate metastasis as a highly coordinated process with a complex molecular and cellular mechanism. Further, while we know of some of these factors involved in metastasis, our understanding is far from complete.

Due to the complexity of metastatic progression at the molecular level, powerful tools are needed to identify the gene expression changes associated with metastatic ability. As such the use of gene expression analysis has been a vital tool in uncovering the molecular participants associated with multiple facets of metastatic progression.. For example, gene expression analysis has allowed for the identification of genes that predict the potential for metastasis [81], genes

whose expression is correlated with the epithelial to mesenchymal transition [82], genes that predict metastasis for specific tumor types [83] and genes associated with organ specific colonization [84, 85]. Perhaps one of the most useful utilizations of tumor gene expression data is with the development of the online Kaplan-Meier-Plotter tool [86]. This tool makes use of the analysis of over 4,000 human breast cancer patient tumors that have been analyzed on microarray. These patients have been observed for a number of clinical observations, such as ER status, intrinsic subtype, overall survival, and metastatic status. By stratifying patient tumors on the basis of high or low expression of a given gene, a Kaplan Meier analysis allows for differences in the time it took for a patient to develop metastasis to be detected. Thus, for example, if a high expression of gene separates out tumors where patients experienced earlier metastasis, that gene might be predicted to function in metastasis.

HUMAN BREAST CANCER CELL LINES AND MOLECULAR BIOLOGY

Verifying these predictions have largely relied on the use of human breast cancer cell lines [87]. Typically, a potential metastatic regulator can be tested for functionality by transfecting or viral injection human breast cancer cells with constructs for expression of short-hairpin RNAs that can diminish expression of a target gene [88]. One weakness of this approach is that it does not completely eliminate the expression of such genes. This leaves for the possibility of false negatives in hypothesis testing especially for genes that are regulated far more at the level of protein stability. An alternative approach without this weakness is an option known as CRISPR (clustered regularly interspaced short palindromic repeats) and has the ability to eliminate the expression of genes in cells [89]. Briefly, CRISPR takes advantage of plasmids expressing a Cas9 nuclease and an expression guide RNA that the user can design to direct Cas9

to a target gene. Cas9 will cleave DNA normally 3 nucleotides upstream from sequence of DNA known as a PAM (protospacer adjacent motif) motif. The PAM motif is typically any nucleotide followed by two guanine nucleotide. This will generate double-strand breaks at target sites and DNA repair process leads to insertions, deletions or substitutions at target sites. As a result, this technology has the ability to completely knock out the gene of interest.

Regardless of the technology used to alter gene expression in human breast cancer cell lines, validation experiments for metastasis typically utilize both *in vitro* and *in vivo* strategies. Examples of *in vitro* investigations with relevance to metastasis include wound healing assays to measure cell migration ability and transwell invasion assays. Transwell invasion assays have the flexibility to be adjusted to test multiple characteristics relevant to metastasis including invasion through the extracellular matrix, migration toward specific chemokines, and the ability to migrate through endothelial cells. While in-vitro assays are an excellent screening tool with a high level of user control, it is also preferential to test cancer cells ability to metastasize *in vivo.* Testing human breast cancer cell lines *in vivo* requires the use of immuno-compromised mice. One way of testing metastatic ability involves injection of cancer cells into the mammary gland, allowing a tumor to develop, and then monitoring metastasis at end stage. An alternative strategy is to inject tumor cells directly into the bloodstream by way or tail vein or retro-orbital injection and observing metastasis at a set time point. In this way early rate limiting steps are bypassed and the later colonization steps can be tested. These *in vivo* strategies also offer the advantage for enhanced detection methods, for example bioluminescent or fluorescently tagged cancer cells can be monitored using advanced imaging systems such as IVIS which allow metastatic progression to be monitored while the mouse is still alive. One of the major weaknesses of the using human cell lines in immuno-compromised mice is that the immune

system has been shown to be a major player in tumor progression and metastasis [90]. In light of this, the use of genetically engineered mouse models of cancer offers the advantage and the opportunity to study tumor progression in an immuno-competent system.

MOUSE MAMMARY TUMOR MODELS

The study of mouse mammary tumor models began several decades ago with the study of inbred strains of mice presenting an elevated incidence of mammary tumor formation. This study led to the discovery of the mouse mammary tumor virus (MMTV) [91]. After the discovery of the virus, methods were then established to harness its promoter activity as a means for discovering mammary transforming oncogenes [92]. The ability of the MMTV promoter to target oncogene expression to the mammary gland, along with other mammary specific promoters, has enabled the study of a variety of genetic pathways in mouse models for effects on mammary tumorigenesis [93]. In addition to overexpression of specific genes, gene deletions in a mouse model have also helped demonstrate the impact of heritable gene mutations, such as BRCA1, BRCA2 and p53 mutations, to breast cancer development [94-96]. By combining oncogene overexpression and gene deletions, mouse models have become a powerful tool for detecting and testing pathway interactions in the course of mammary tumor development [97-99]. Moreover, more recent models with inducible oncogene expression has allowed the dependency on specific oncogenes to be examined [100-103]Taken together, previous work in mouse models has been integral to advancing our understanding of breast cancer to its current state. Further, by integrating the advances in technology with bioinformatic methods, the capacity of mouse models to dissect key features of human breast cancer has the potential for new directions and possibly development of new therapeutic strategies.

Given that extensive heterogeneity has been noted in human breast cancer through genomic methods, it is important to examine the heterogeneity of the mouse model systems. Given that mouse models are usually induced by the overexpression of a critical oncogene, one may have expected a lack of tumor heterogeneity. However, the integration of mouse model studies with bioinformatic analysis of mouse mammary tumor gene expression data has demonstrated significant heterogeneity in mouse models, analogous to human breast cancer. While some models, such as MMTV-Neu model display characteristic gene expression profiles with minimal heterogeneity, other models have been shown to induce tumors with wide ranging genomic and histological heterogeneity [104]. While no single mammary tumor model has been able to capture the full-spectrum of heterogeneity of human breast cancer, bioinformatic methods provide a means of defining the similarities and differences between mouse models and subtypes of human breast cancer. For example, just as human breast cancers have been classified into intrinsic subtypes, mouse models have also been characterized in this manner [105]. In this way, similarities between human breast cancer subtypes and mouse mammary tumor models were identified and validated with immunohistochemistry for specific markers. However, by combining mouse and human mammary tumor gene expression data, analysis of the gene expression profiles distinguish mouse and human mammary tumors; perhaps due to the limited number of tumor samples for mouse mammary tumors within each mouse model.

Even though the full spectrum of human breast cancer heterogeneity was not present in previous mouse model studies, continued bioinformatic analysis of heterogeneity within a mouse model presents the opportunity to detect relationships to human breast cancer. For example, differences in Myc expression and stability have led to divergent histological and genomic tumor types [106]. Another example of using bioinformatic analysis to dissect heterogeneity and

establish relationships to human breast cancer comes from the MMTV-Met model [107]. Dissecting heterogeneity to establish relationships between mouse models and human breast cancer can be important when testing for the clinical significance of specific observations within a mouse model. For example, in the study of MMTV-Met tumors, mouse mammary tumor gene expression data was used to construct a gene expression signature of Met receptor tyrosine kinase signaling. By applying this signature to probe human breast cancer gene expression data, samples were stratified as being either positive or negative for the Met signature. Importantly, this showed that most Met positive samples were of the basal type, correlating with a poor prognosis. As an additional demonstration, the importance of predicted pathway activation has been tested in the examination of the E2F transcription factors in the MMTV-Myc mouse model [108]. The E2Fs are broadly classed into transcriptional activators (E2F1-3) and transcriptional repressors (E2F4-8) and have been classically studied for their role in cell cycle progression [109]. In the mouse model study, E2F2 was predicted to be active in Myc tumors. Testing this hypothesis, a knockout of E2F2 was found to reduce Myc's proliferative effects on the mammary gland and delayed tumor onset. In demonstrating the relevance of the E2F effects in human breast cancer, a gene signature was applied to predict E2F2 signaling across a dataset featuring various human breast cancer subtypes. It was found that in Her-2 positive and basal breast cancers, E2F2 signaling was predicted to be significantly altered. While no differences in Luminal B breast cancers existed, in Luminal A breast cancers in E2F2 levels were predicted to be significantly lower. Importantly, these results carried clinical significance as low E2F2 levels were correlated with an increase in relapse free survival time in human breast cancer.

Bioinformatics is important for establishing relationships between mouse models and human breast cancer, but these methods can also help identify mechanistic features of tumor

metastasis. One model that is commonly utilized is the Polyoma Virus Middle T (PyMT) model. This model is characterized by a high incidence of pulmonary metastasis [110], making it an efficient platform for studies of metastasis. Integration of bioinformatic approaches to analyze the PyMT model have led to the identification of molecules that promote metastasis. For example, the importance of the interaction between mammary tumors and stromal adipocytes were detailed in a genetic test in the MMTV-PyMT model [111]. In this experiment adiponectin, a protein secreted by adipocytes, was ablated to determine its effects on mammary tumorigenesis. The knockout of this adipocyte factor led to delays in mammary tumor onset, with delays related to impaired angiogenesis. To further describe these tumors, gene expression profiles of late stage tumors from adinopectin deficient mice were analyzed. These tumors were found to contain gene expression profiles descriptive of aggressive tumor phenotypes. Not surprisingly, later studies in human breast cancer found that low levels of this adipocyte factor correlated with increased breast cancer mortality [112].

Not only has integrating bioinformatics with mouse models assisted the identification of molecular participants in metastasis, but it can also help illustrate the mechanistic relevance to human breast cancer. For example, a genetic test of the role of the Snf1-Kinase, Hunk, in the Myc model revealed that Hunk is required for the metastasis of Myc induced tumors [113]. A Hunk signature from the resulting tumors was used to predict whether human breast tumors had similarities to Myc in the presence or absence of Hunk. In agreement with the observations of metastasis in the Myc model, human breast tumors that were predicted to be wild type for Hunk carried a significantly higher incidence of metastasis. After identification of these effects in the mouse model, and demonstration of its clinical relevance, further studies demonstrated the mechanism behind these metastasis effects in human breast cancer[114]. In a different model,

mammary specific inactivation of the tumor suppressor protein PTEN, coupled with Her-2 overexpression, revealed a metastasis suppressor function for PTEN[115]. By analyzing gene expression data from the resulting tumors, tumors were noted to share molecular features with luminal types of human breast cancer. These examples show the utility of integrating mouse model studies and bioinformatic methods as a means to uncover mechanistic features of human breast cancer metastasis. Going forward, similar studies should provide more details about the metastasis mechanism and which features are deregulated in specific subtypes of human breast cancer.

RATIONALE FOR DISSERTATION

Collectively, the previous technology and data described displays the power of gene expression analysis for dealing with the complexity of breast cancer and highlights the critical areas of need for breast cancer research. One area of need concerns dissecting tumor heterogeneity. Here the major problem is that the genomic variability within tumors and between patients limits the efficacy of breast cancer therapy. To date, some success in understanding this variability has been achieved, with breast cancer cancers being organized into different subtypes on the basis of both key markers and gene expression profiles that contribute to tumor progression. Directed therapies for specific types of breast cancer have improved breast cancer survival times, yet due to the molecular complexity of breast cancer, treatment is still inadequate; with tumors initially regressing only to reoccur and become resistant to therapy. The second major factor impacting clinical outcome is breast cancer metastasis. As previously, discussed metastasis to distant organs has clear implications on breast cancer survival times. Identification of additional regulators of this process may uncover new therapeutic opportunities. As a result,

my dissertation research focused on addressing these challenges to breast cancer survival: tumor heterogeneity and tumor metastasis.

Clearly, understanding the genomic and histological heterogeneity of human breast cancer is an essential goal in order to improve diagnostic tests and for successful, targeted treatment of breast cancer patients. Pre-clinical mouse models need to be credentialed for their ability to model human breast cancer and the heterogeneity that is a hallmark feature of human breast cancer. However, the degree to which mouse models are reflective of the heterogeneity of human breast cancer has yet to be demonstrated with gene expression studies on a large scale. If mouse models with human breast cancer-like complexity and relationships to individual types of human breast cancer could be identified, such a finding would represent a major breakthrough and enhance the research of mechanisms and treatments for drivers of breast cancer progression using mouse models. To meet these needs, I developed a research plan that would integrate bioinformatic analysis, classic molecular techniques, and genetic tests in mouse models.

The goal was to first build off of the findings in an initial analysis in a MMTV-Myc model with demonstrated histological and genomic heterogeneity. The advantage inherent to this study as compared to others, was that our lab had generated gene expression data for a large number of samples, had a clear understanding of histological heterogeneity, and even knew of some of the mutations that occurred in this mouse model. We hypothesized that this would allow for more statistical power in detecting relationships between tumors in this mouse model to human breast cancer. Further, we could make more precise conclusions about the similarities and differences between tumor types by integrating mutation status, pathway activation predictions, and the histological annotations into the comparisons. In support of our hypothesis we found, MMTV-Myc mice with tumors of an EMT-like histology develop similar gene expression

patterns and signaling pathway utilization as a subtype of human claudin-low breast cancer. Refining the understanding of human tumor heterogeneity, our results point out a clear division in human claudin-low tumors based on Myc pathway activation and target genes. Moving forward, this work highlighted the possibility that a similar analysis for other mouse models was needed.

As a result, I decided to assemble an expansive database of gene expression data of all of the publicly available mouse mammary tumors spanning 23 major mouse models. With this, there were a number of different ways the information of this expansive database could be harnessed. We could test the degree of heterogeneity within mouse models, make comparisons between models, make a number of bioinformatic predictions to springboard additional hypotheses, and test for relationships between mouse models and human breast cancer at both the level of gene expression and predicted signal pathway activity. One the major outcomes of this was the identification of mouse models that do and do not have similarities to human breast cancer at the level of gene expression. Thus, this date highlights the importance of fully characterizing mouse tumor biology at molecular, histological and genomic levels before a valid comparison to human breast cancer may be drawn. In addition, it provided a major bioinformatic resource to research community with a large number of predictions regarding important molecular players in tumor progression in each mouse model. Using the predictions from this work, I was able to develop additional projects and meet the other major objective for my dissertation: identifying additional molecular regulators of metastasis.

Using bioinformatic predictions from the database, I identified an elevation in predicted activity for E2F1 in the MMTV-PyMT mouse model. This mouse model is widely known for the enhanced metastatic capacity of the tumors that give rise. This directed a further interrogation of

the activator E2F in human breast cancer to test for an association patient metastasis data. With these predictions, I hypothesized that E2F transcription factors regulate breast cancer metastasis. By genetically testing these predictions, I found that loss of E2F1 and E2F2 severely limited the metastatic potential of tumor cells. Further investigation showed an association with defects at multiple steps of the metastatic cascade. Since the E2F are transcription factors, it is logical to believe that regulation of metastatic ability occurs by controlling expression of multiple genes that promote metastasis. As such, the focus was to expand on our discovery that the E2Fs regulate metastasis and highlight the mechanism by identifying E2F target genes integrating a number of computational strategies and assays.

Moving back to utilizing the gene expression database of mouse mammary tumor models, one of the weaknesses for many tumor samples was the lack of a histological annotation. For tumors where histological annotations existed, we are able to be more specific regarding which tumor types from a mouse model resemble a specific subset of human breast cancer. As such, there was a need to provide annotations for nearly a thousand tumor samples in the database without a histological annotation. To meet this need, I have begun to develop gene expression signatures that can accurately predict tumor histology. The outcome of this work has several payoffs. The first is that we are able to predict and annotate tumor histology using gene expression data. The second, it will allows for the genomic markers of specific tumor histologies that arise in mouse to be tested for their relevance in specific human breast cancer subtypes and other cancer types with similar histologies.

Taken together, the big picture of the work that I have completed during my dissertation demonstrates the absolute complexity of human breast cancer and how existing strategies and technologies can be used to solve the problems this complexity imparts on our ability to

understand breast cancer heterogeneity and metastasis. Importantly in doing this work, there were major findings that have furthered our understanding of mouse models of breast cancer, tumor heterogeneity, and tumor metastasis. These major findings are described in more detail in chapters that follow.

CHAPTER 1:

A MOUSE MODEL WITH T58A MUTATIONS IN MYC REDUCES THE DEPENDENCE ON KRAS MUTATIONS AND HAS SIMILARITIES TO CLAUDIN LOW HUMAN BREAST CANCER

ABSTRACT

Expression of c-Myc is highly prevalent in human breast cancer and stability of the oncoprotein is regulated through Ras regulated phosphorylation at serine 62 and threonine 58. Previous studies have illustrated the importance of accumulation of KRas mutations in Myc mediated tumor formation. To examine Myc dependence upon Ras mutations we have generated MMTV regulated Myc and Myc T58A transgenic mice. Expression of the more stable T58A Myc allele resulted in a reduction in KRas activating mutations. However, in a low level expression T58A Myc transgenic, the majority of the tumors were squamous or epithelial to mesenchymal (EMT) in nature and accumulated KRas mutations at a higher frequency. Interestingly, we show that these mice develop similar gene expression patterns and signaling pathway utilization as a subtype of human claudin low breast cancer. Indeed, our results demonstrate a clear division in human claudin low tumors based on Myc pathway activation and target genes. Together, our results demonstrate that Myc expression and stability has critical effects on molecular heterogeneity in mouse mammary tumors that parallel subtypes of human breast cancer.

INTRODUCTION

Human breast cancer is a collection of remarkably diverse neoplasms. The combination and context of the events that initiate transformation lead to varied regulation of cellular processes and give rise to tumor heterogeneity. This is reflected in gene expression, allowing the categorization of tumors into molecular subtypes[57, 59, 116]. In addition, these molecular subtypes of breast cancer differ in terms of differentiation, response to therapy, and overall survival [58]. For some subtypes, associations have been made with specific oncogene activity. For example, the basal molecular subtype has been associated with high expression of the Myc

oncogene [117]. *Myc* is amplified in 15% of human breast cancers [118] and the degree of *Myc* amplification has been shown to correlate with mRNA levels and protein levels of Myc [119]. Moreover, when large cohorts of human breast cancers were examined for Myc pathway activation, nearly 25% of the tumors showed a high probability of Myc signal activation [47, 104], correlating with previous reports [120, 121].

The half-life of Myc is brief and the stability of Myc is regulated by phosphorylation events at key serine and tyrosine sites [122, 123]. The first phosphorylation event is mediated by Ras signaling through Erk, leading to phosphorylation at Ser62. This stabilizes Myc and allows subsequent phosphorylation by GSK3β at Thr58, targeting Myc for ubiquitination. This GSK3β phosphorylation event may be blocked through Ras activation of PI3K / AKT, extending the half-life of Myc. The active Ser62 phosphorylated Myc initiates transcription by forming heterodimers with the Max transcription factor [124]. Once activated, Myc has been demonstrated to regulate a wide range of transcriptional targets [125]. Thus, Myc initiated changes in gene expression have a variety of effects resulting in the formation of breast cancer.

Myc effects on mammary development and cancer has been examined through mouse models. Myc was initially overexpressed in the mouse mammary gland using the MMTV promoter, resulting in the formation of adenocarcinomas [126]. The importance of Ras in these tumors was demonstrated by the synergistic reduction in tumor latency observed by interbreeding MMTV-Myc and MMTV-Ras transgenic mice [127]. Additionally, in a conditional Myc model, KRas mutations were detected with Myc withdrawal, correlating with tumor progression and recurrence [101]. Subsequently, a synergy with KRas was detected in Myc initiated tumors where Ras was seen to be a more dominant oncogene [103]. In recent work, a knock-in approach to overexpress Myc at low levels demonstrated that decreased stability of

Myc by S62A mutation abrogated Myc's transformation ability. In contrast, a T58A mutation enhanced the formation of tumors [128]. However, the majority of mice in this study developed brain tumors, preventing examination of T58A Myc in breast cancer. Taken together, these mouse models illustrate the role of Myc in mediating breast cancer and the importance of the Ras pathway in this process.

In previous work, we demonstrated that Myc expression affected tumor heterogeneity at the genomic and histological level[106]. In light of these results, we hypothesized that a stabilizing T58A mutation in Myc would decrease the dependence upon activating mutations in KRas for tumorigenesis. To test this hypothesis, Myc and Myc T58A transgenic mice were generated. We noted a decrease in activating mutations in KRas in Myc T58A strains with high levels of Myc expression. Through comparative gene expression analysis, we established similarities between the EMT type T58A Myc tumors and human claudin low breast cancer.

RESULTS

MOUSE MODEL CHARACTERIZATION

Mammary tumors initiated by Myc in transgenic mice result in the accumulation of activating mutations in KRas. We hypothesized that tumor formation in transgenic mice with a stabilizing T58A version of Myc would not be dependent upon accumulation of activating mutations in KRas. To test this hypothesis we generated Myc transgenic mice with the constructs shown in FIGURE 1.1A. For the wild type Myc transgenics, a total of 4 lines were generated and T58A resulted in an additional 11 lines. An RNase protection assay was completed in virgin and lactating mice to assess transgene expression (FIGURE 1.1B). This revealed that two wild type Myc lines (WT13 and WT21) were expressing the transgene at a high level in the virgin mammary gland with a significant increase in expression levels in the

lactating samples. In the T58A Myc genotype, two lines were noted to have comparable levels of transgene expression (TA14 and TA41). In addition, we retained TA39 transgenic mice to assay for effects of Myc expressed at lower levels. We also assessed mammary glands for both HA-Myc (FIGURE 1.1C and D) and total Myc protein (FIGURE S 1.1A and 1C) levels. Western blot results show that the T58A mutation stabilized Myc in the mammary gland. This resulted in higher protein levels of Myc in the mammary gland given similar levels of mRNA when comparing the T58A to the wild type transgenic lines. In addition, we quantified total Myc levels in the mammary glands of non-transgenic FVB mice to assess the degree of protein overexpression in the T58A transgenic lines, demonstrating at least 3.5 fold more total Myc protein in the transgenic mice compared to non-transgenic FVB mice (FIGURE S 1.1E-F).

To determine the initial effect of the transgene expression on mammary gland development in these lines, wholemounts of the mammary glands were assessed at 8 weeks of development. In comparison to the wild type control (FIGURE 1.1 E), we assessed the WT13, WT21, TA14, TA41 and TA39 lines (FIGURE 1.1 F-J respectively). This revealed that the lines with elevated transcript levels had abnormal sidebud formation in comparison to the wild type control. The most prominent alterations were noted in the two strains with high levels of Myc expression, WT13 and TA14, noted with arrowheads relative to the control in FIGURE 1.1E, F and H. In addition, the low expressing TA39 transgene had minimal phenotypic effects on mammary development.

To determine the effects of the T58A mutation on tumor latency, females from the five lines were maintained in a constant breeding program. We observed a large number of mice for each strain including; WT13 n=70, WT21 n=49, TA14 n=60, TA41 n=52 and TA39 n=62. Surprisingly, there was no difference in tumor latency between the high level of Myc WT

expression and the high expressing MycT58A strain (FIGURE 1.2A). However, we did note a significant increase in tumor latency for the low expressing Myc T58A TA39 strain ($p<0.001$ relative to WT and $p<0.0001$ relative to the other T58A strains). Measuring the number of tumors in each mouse revealed no difference between the wild type MMTV-Myc and high expressing T58A Myc transgenics (FIGURE 1.2B). Consistent with the tumor latency effects we noted that there was a reduction in tumor burden in the TA39 line with the majority of mice harboring a single tumor (FIGURE 1.2B) ($p =0.0002$ compared to WT and 0.0005 compared to T58A high).

REDUCED KRAS MUTATION IN T58A TUMORS

To ascertain whether T58A mutations reduced the frequency of activating mutations in KRas, we sequenced for KRas activating mutations at codons 12, 13 and 61. An example of a wild type sequence trace for codons 12 and 13 is shown in FIGURE 1.3A. In contrast, a sequence trace harboring a heterozygous mutation in codon 12 is shown (FIGURE 1.3B). Of note is that heterozygosity has been preserved in 59 of the 61 tumors containing activating mutations in KRas. We then compared the KRas activation in the five transgenic lines. KRas mutations were observed in 22.4% and 25.6% in the two wild type Myc lines (n=24 mutant KRas alleles for 107 tumors in WT13 and 10 mutations for 39 tumors in WT21) (FIGURE 1.3C). A T58A mutation in Myc significantly lowered the percentage of tumors containing an activating mutation with rates of 14.6% and 16.1% in the T58A high expressing lines (n=13 mutations in 89 tumors for TA14 and 11 mutations in 68 tumors for TA41) ($p=0.042$ by Fisher's Two Tailed t-test). Interestingly, there was an increase in the number of mutations observed in the low expressing TA39 line relative to the high expressing lines with 22.2 % of tumor bearing mutations (n=27 tumors). Together this data suggests that the T58A mutation reduces the

dependence upon accumulation of KRas mutations. However, the wild type transgenic line mutation rate does not concur with previously published data for Ras mutations in the wild type strain where 44% of MMTV-Myc tumors were found to have Ras mutations [101]. To investigate this, we examined 21 of the original MMTV-Myc strain tumors and found 11 activating mutations in KRas, equal to 52% of tumors with mutations. To examine the difference between our MMTV-Myc WT13 line (22.4% mutation rate) and the original MMTV-Myc (52% mutation rate), we measured transgene levels through QRT-PCR. This revealed that Myc expression was nearly fourfold higher in lactating mammary gland of the WT13 line compared to original MMTV-Myc mice. (FIGURE S 1.2A). Thus, we find a reduction in KRas mutation rates (FIGURE S 1.2B) from 52% to 22.4% with a fourfold elevation of Myc expression in the mammary epithelium. Interestingly, in the TA14 line, Myc was found to be expressed at similar levels to the original MMTV-Myc line and had the expected reduction in KRas mutation frequency (FIGURE S 1.2B).

To examine whether the generation of an activating mutation in KRas was associated with alterations to latency, transgenic mice with and without activating mutations were compared. For wild type MMTV-Myc lines, we observed a significant reduction in latency for mice with tumors containing a KRas mutation relative to mice with tumors without a mutation ($p=0.0384$) (FIGURE 1.3D). Interestingly, the latency acceleration due to KRas mutations was not observed in the T58A lines with high levels of Myc expression (FIGURE 1.3E). To compare these mouse mutations to human breast cancer, the COSMIC database was searched for breast cancers with mutations in KRas. Interestingly, the mouse tumors and human breast cancers shared the same pattern of KRas mutations with a range of mutation types in the mice (FIGURE S 1.3 and TABLE 1.1).

GENOMIC CHARACTERIZATION OF T58A TUMORS

We previously demonstrated that there was significant heterogeneity in the development of tumors in a subset of Myc and Myc T58A lines (WT13, WT21, TA14 and TA41) [106]. To assess the KRas mutation status in the heterogeneous tumors, we examined the KRas mutation rate in the various histological subtypes of tumors, grouped into WT, T58A High and T58A Low expressing strains (FIGURE 1.3F) and individually (FIGURE S 1.4). This revealed that the vast majority of tumor types, and all transgenic lines, had tumors with activating mutations in KRas (FIGURE 1.3F). However, we noted an elevation in the frequency of activating mutations in KRas in the Epithelial to Mesenchymal Transition (EMT) subtype of tumors. Moreover, the majority of the tumors in the low expressing TA39 T58A transgenics were either squamous or EMT in nature.

We then assayed gene expression for the low expressing T58A tumors in relation to the gene expression data from the other Myc samples [106]. After batch effects were removed (FIGURE S 1.5), unsupervised hierarchical clustering showed that differences in tumor histology are reflected in gene expression. It is important to note that the number of tumors analyzed by microarray are not proportional to the number of tumors in each class. However, the addition of new gene expression data from TA39 transgenic refined prior clustering analysis and allowed the separation of EMT tumors and squamous tumors (FIGURE 1.4A and 1.4B). In addition, we observed that the majority of tumors clustered into the subtypes predicted by their histological classifications (FIGURE 1.4B). Cell signaling pathway signatures were applied to the entire dataset with the samples maintained in the order established through unsupervised clustering (FIGURE 1.4C). This revealed patterns of signaling pathway activation associated with the various histological subtypes. Importantly, the distinction between the EMT and Squamous

subtypes noted in unsupervised clustering was maintained and we noted significant differences between these subtypes in RhoA, AKT, p63, E2F1 and Beta- Catenin pathways (FIGURE 1.4C and FIGURE S 1.6). Importantly, we predicted a low probability of Myc signaling in tumors of the EMT histology compared to microacinar and papillary tumors. To validate the genomic data, tumors from the various subtypes and transgenic lines were assessed for levels of Myc. This revealed that the EMT tumors had significantly lower levels of Myc expression (FIGURE 1.4D, FIGURE S 1.7). Quantification of this result supported these findings (FIGURE 1.4E, FIGURE S 1.8). Confirming the accuracy of our gene signature approach, probability values for Myc activation show a positive correlation with western blot quantification of total and exogenous Myc levels in corresponding tumor samples (FIGURE S 1.9).

As noted above, when the Myc transgenic tumors were examined for activation of the Myc pathway, we observed variable levels of Myc activation in the various subtypes (FIGURE 1.4C). We hypothesized that this difference in predicted Myc signaling would result in differential activation of Myc target genes. To test this hypothesis we analyzed expression levels of Myc target genes using previously published ChIP-chip data to identify Myc target genes[125]. In addition, by using Significance Analysis of Microarrays (SAM)[28] we identified genes that were upregulated in human mammary epithelial cells (HMECs) that overexpress Myc relative to controls [37]. A Venn Diagram reveals that of the 118 genes identified, only 31 were Myc targets (FIGURE 1.5A). Subsequently, we compared gene expression between samples from each histological cluster. In this analysis, we found 1,958 significantly upregulated genes, of which 110 were Myc targets. Together, these results indicate that there is differential expression of Myc transcriptional targets in histological subtypes (FIGURE 1.5A). To visualize the difference in activation of Myc target genes we extracted Myc targets as identified by Chip-

Chip from the Myc tumor gene expression data for SAM analysis. By minimizing the false discovery rate and ranking genes on fold change, we identified the top 70 differentially regulated genes within each histological cluster. By using these 280 genes in unsupervised hierarchical clustering, we illustrate that tumors extracted from the histological clusters defined in FIGURE 1.4 differentially activate unique Myc target genes (FIGURE 1.5B).

COMPARISONS TO HUMAN BREAST CANCER

To compare these mouse models to human breast cancer we combined our mouse gene expression data with human breast cancer gene expression data after removing batch and platform effects (FIGURE S 1.10). Through unsupervised hierarchical clustering, we compared Myc mediated mouse mammary tumors and subtypes of human breast cancer that were annotated with the intrinsic classification [58, 59]. This revealed that Myc tumors with microacinar histology clustered with a subset of human luminal B tumors (FIGURE 1.6A). In addition, we observed that EMT subtype of Myc mouse tumors clustered with a subset of human claudin low tumors. These EMT type tumors with similarities to claudin low tumors were found in both the WT and T58A strains (FIGURE S 1.11), although the TA39 strain did have a greater percentage of tumors with the EMT histological pattern (FIGURE 1.3F). Importantly, this clustering approach split the claudin low tumor subtype into two distinct groups. When we assessed Myc targets for gene expression in the listed genes we observed a significant decrease in expression levels of Myc targets in the human breast cancer claudin low samples clustered with the EMT samples relative to the other main cluster of claudin low human breast cancers (Fisher's Exact $p<0.001$). A comparison of marker expression levels between EMT-type tumors and other types of Myc-induced tumors shows that the EMT subtype largely parallels human claudin low tumors (FIGURE S 1.12). For example, EMT mouse tumors show low expression

of cell-cell adhesion genes characteristic of human claudin low tumors (FIGURE 1.6B, FIGURE S 1.12). Like claudin low tumors, Myc induced EMT tumors show high expression of the mesenchymal marker vimentin, as well as the angiogenesis marker VEGFC (FIGURE 1.6B), as well as other markers of claudin low tumors (FIGURE S 1.12). Importantly, we note a difference in Myc target expression levels in these claudin low tumors, split into two clusters. In agreement with this result, these two clusters of claudin low tumors also significantly differed in predicted Myc signal pathway activation (FIGURE 1.6C). As a result, we define these two subclasses of human claudin low tumors on the basis of Myc activity: the Myc-low claudin low tumors and the Myc-high claudin low tumors.

In addition, previous work has shown that claudin low tumors have stem cell-like characteristics [129]. Given the gene expression similarities between the mouse EMT-type tumors and the human Myc-low claudin low tumors, we hypothesized that the mouse EMT-type of tumors also have stem cell-like characteristics. To test this hypothesis we utilized gene set enrichment analysis (GSEA) and employed two gene sets derived from previously published gene expression data from mammary stem cells[130]. By comparing EMT-type with the remaining histologies of Myc-induced tumors, we found that the EMT-type of tumors are significantly enriched ($p=0.021$) for expression of genes that are upregulated in mammary stem cells (FIGURE S 1.13A). Likewise, EMT-type Myc induced tumors are significantly enriched ($p<0.0001$) with low expression of genes that are downregulated in mammary stem cells (FIGURE S 1.13B). Together, these results show that the EMT-type of Myc tumors match human claudin low tumors for expression of claudin low markers and features of stem cells.

In comparing the EMT-type Myc initiated tumors to the Myc-low claudin low human breast cancers , we also noted significant similarities at the pathway activation level (FIGURE

1.6C). AKT, β-catenin, E2F1, Myc and p110 pathways are predicted to have low activity in EMT and the Myc-low claudin low tumors while TNF-alpha signaling is similarly activated in these two tumors types. These same pathways are predicted to have significantly different activity between the Myc-low claudin low tumors from the claudin low tumors that did not cluster with the mouse tumors (FIGURE 1.6 C, FIGURE S 1.14). In a similar experiment for microacinar tumors, we revealed that similar activation patterns of the β-catenin and Stat3 pathways occur between the mouse and largely luminal B human breast cancers within the cluster (FIGURE 1.6D). In addition, these pathways distinguish the microacinar-like luminal B tumors from other luminal B tumors where β-catenin is predicted to a significantly lower activity and Stat3 is predicted to have a significantly higher activity (FIGURE 1.6D, FIGURE S 1.15). Taken together, these data reveal that there are striking similarities between subtypes of Myc induced mouse tumors and subsets of primary human breast cancers.

Given the molecular similarities between the EMT-type of mouse mammary tumors and human claudin low tumors, we sought to determine whether the results for KRas mutations in the EMT-type of mouse mammary tumors also extended to human claudin low breast cancer. KRas mutations occur in 3% of breast cancers contained within the COSMIC database. However, the molecular subtype of the primary tumors with these mutations is not reported, precluding an examination of Ras mutation status in subtypes. Alternatively, we predicted the probability of Ras signal pathway activation in human breast tumors. We find that nearly 80% of the claudin low tumors predict an elevation in Ras signaling activity (FIGURE S 1.16). This may indicate that KRas mutations occur more frequently in claudin low tumors, similar to EMT tumors in the MMTV-Myc mouse model. To further establish the significance of KRas mutations to the human claudin low tumors and the EMT-type of MMTV-Myc initiated tumors they are tightly

correlated with, we used GSEA and a signature for KRas addiction[131]. A comparison between KRas mutant EMT-type tumors and all other KRas mutant mouse mammary tumors showed that while MMTV-Myc EMT-type of tumors develop KRas mutations, they have a low probability of dependence upon this mutation for tumorigenesis (FIGURE S 1.17A). Likewise, claudin low tumors, compared to human basal tumors which also feature high levels of predicted Ras signaling, also predict low probability of KRas addiction (FIGURE S 1.17B). This is in agreement with the previously published work that shows that KRas mutant tumors with a mesenchymal phenotype are not dependent upon KRas signaling during tumorigenesis[131]. Together, these data show a clear link between the Ras mutation findings in the mouse model EMT tumors and the claudin-low subtype.

DISCUSSION

Here we have characterized mammary gland development in T58A transgenic mice and observed mammary gland effects that correlate with elevated Myc expression. The increased sidebud formation we see in mice with high Myc expression is similar to mammary gland effects in other studies featuring elevated Myc [101, 132]. Although no mammary effects were observed in the TA39 line at these early time points, all strains developed mammary tumors. Surprisingly, the tumor onset in the high expressing WT and T58A strains was virtually identical with a delay observed in the TA39 strain. We had anticipated that the increased Myc stability would result in more rapid tumor onset. A possible explanation for the similarities in tumor onset between the T58A and WT strains is that signaling conditions during the time when the MMTV promoter is most active, facilitate phosphorylation of S62 in the WT Myc strains, mimicking a T58A mutation under these circumstances.

Our hypothesis that an increase in Myc stability through T58A mutation would result in a reduction in dependence upon KRas mutations was validated. We observed a 1.5 fold reduction in the number of KRas activating mutations in the T58A lines relative to the wild type Myc transgenic lines. In addition, we observed that tumors with the lowest levels of Myc protein, tumors with EMT histology, also had the highest frequency of KRas mutations. Moreover, there was a selective advantage in the wild type mice that accumulated KRas mutations that was absent in the T58A strain. When our wild type Myc transgenic mice are compared to previous studies with Myc mouse models, we note that our lines have far fewer KRas mutations, but this can be attributed to the lower levels of Myc in the previously characterized Myc strain [126]. However, the original Myc mice have equivalent Myc transcription to the T58A strains, demonstrating a far more impressive reduction in the requirement for activating KRas mutations with the stable T58A form of Myc. Yet, by comparison to WT strains, it is not readily apparent that T58A tumors express high levels of Myc. This suggests the possibility that K-Ras mutation incidence may extend beyond Myc levels. This hypothesis is reinforced by the distinct pathology of Myc induced tumors; for example tumors from T58A strains exhibit primarily a papillary histology.

Amongst the histological clusters, we noted diverse probability of activation of the Myc signaling pathway. We hypothesized that this difference would result in differential activation of Myc target genes in the different subtypes. Interestingly, we note that tumors with differences in histology and Myc stability also have differences in Myc target utilization. This suggests that differences in Myc expression and stability influence divergent tumor histologies by differential activation of direct Myc targets. We also noted variability in Myc pathway activation in human breast cancer samples. While HER2 and luminal subtypes show sporadic activation of Myc

signaling, we see that basal and a subset of claudin low tumors are predicted to have high activation of the Myc signaling pathway. In agreement with this, we noted high expression of a cluster of genes made up of mainly Myc targets in this subtype. Illustrating the validity of our approach, other studies have established have established that Myc is highly expressed in basal tumors [117]. Together, this data suggests that Myc may play a role influencing tumor heterogeneity both in mouse and human breast cancers.

Mouse models have been compared to human breast cancer through histology, signaling pathway activation and through genomic means [105]. In addition, recent work has identified a p53 mutant mouse model with similarities to the claudin low subtype [129]. However, our work is unique in that we have stratified human claudin low tumors into two subclasses based on Myc expression levels. With these results, we report remarkable similarities between the EMT-type of Myc-induced tumors and human claudin low tumors. Not only do we identify a correspondence in gene expression and signal pathway activation between EMT-type tumors and Myc-low claudin low tumors, but we also show that EMT tumors match claudin low tumors for stem cell characteristics and expression patterns of markers for claudin low tumors. Along with expression patterns for markers cell-cell adhesion, we show EMT-type tumors match human claudin low tumors for high expression of mesenchymal markers. This mesenchymal identity may be critical to the understanding of the significance of KRas mutations or increased Ras signaling in both the mouse model EMT-type tumors and the human claudin low tumors. Previous work has shown that KRas mutant epithelial tumors are addicted to KRas, while KRas mutant mesenchymal tumors are not dependent upon KRas for tumor proliferation [131]. Using a gene signature for KRas derived from this study, we used GSEA to predict whether the claudin low tumors and KRas mutant EMT-type tumors were independent of KRas. Consistent with their

mesenchymal identities, both the human claudin low tumors and the mouse model tumors are not predicted to be "KRas addicted". It is important to note that the mouse model EMT-tumors that clustered with the Myc-low claudin low tumors were found in all strains (FIGURE S 1.11). This suggests that the presence of EMT, and not a T58A mutation, drives the similarities between mouse model EMT tumors and claudin low tumors. Furthermore, with the human claudin low tumors split on the basis of Myc signaling, the low levels of Myc, along with the presence EMT, found in the mouse model EMT tumors likely drive the similarities to Myc-low human claudin low subclass. Taken together, these data suggest that the EMT subtypes of the Myc transgenic mice are a model for human claudin low breast cancer.

Overall, our results show the importance of Myc stability and expression in mammary gland development and tumorigenesis. We show that differences in oncogene expression and stability can significantly alter molecular features of mammary tumors. The reduced dependency upon KRas mutations in T58A strains, as well as differential activation of Myc target genes, makes it clear that the context of oncogene expression plays a critical role in molecular features of the tumor. Here we have identified Myc induced mouse mammary tumors with remarkable similarities to human luminal B tumors and claudin low tumors. Particularly, the similarities in gene expression, Myc target utilization, and signaling pathway activation between EMT tumors and claudin low tumors support the use of this model as a means for experimental study of this molecular subtype of human breast cancer.

METHODS

ANIMAL WORK

All animal work has been conducted according to national guidelines. All mice are from the FVB background. The transgenic mice were generated using the MMTV promoter as

previously described[106]. Comparison to the existing MMTV-Myc model [126] was conducted on mice obtained from the Mouse Models of Human Cancer Consortium. Upon genotyping, the transgenic mice were maintained in a constant breeding program to accelerate tumor formation. Tumors were detected through weekly palpation and tumor growth was routinely measured.

RNA AND MICROARRAY

Preparation of RNA samples from flash frozen tumors was done using the Qiagen RNeasy kit after roto-stator homogenization. RNA from 17 Myc induced tumors was submitted to the Duke Microarray Core facility for gene expression analysis using Mouse 430A 2.0 Affymetrix arrays.

Mutations in KRas were assessed after RT-PCR using Qiagen One-Step RT-PCR with the following primers; 5' GGAGAGAGGCCTGCTGAA and 3' TCTTCTTCCCATCTTTGC TCA. PCR products were gel purified and the samples were sequenced with a nested primer with the following sequence; 5' TAGAAGGCATCGTCAACA C 3'.

WESTERN BLOT ANALYSIS

Western blot analysis was conducted using antibodies for Grb2 (Cell Signaling 3972), C-Myc (Ab Cam ab32072), and HA-Myc (ABM G036) and HRP-conjugated anti-mouse (BD Biosciences 554002) and anti-rabbit (Ab Cam ab97051) were used for detection of the specified protein. HA-Myc and C-Myc levels were standardized to Grb2 using image-J software.

COMPUTATIONAL METHODS

Microarray gene expression data has been submitted to the Gene Expression Omnibus (GEO) as GSE30805. New gene expression data was combined with previously published mouse gene expression data from GSE15904. In a separate experiment, this combined dataset was analyzed with previously published human breast cancer gene expression datasets GSE6532,

GSE14020, GSE2034, GSE2603, and GSE4922. Annotation of human breast cancer samples with intrinsic subtypes used previously described methods [57, 58, 105]. All datasets were RMA or MAS5 normalized using Affymetrix Expression Console software depending on the application. When combining datasets, Bayesian factor regression methods were used to remove batch and platform effects [25, 47]. Principle components analysis and plots were generated using Matlab to assess batch and platform removal effects. Unsupervised hierarchical clustering was completed with Cluster 3.0 and results were visualized using Java Tree View. Significance analysis of microarrays (SAM) was conducted [28]. The gene lists for Venn diagram using ChIP-Chip Myc target data from [125] and other lists were created by using results from SAM comparing HMECs that overexpress Myc (GSE3151) and by comparing histological types of Myc induced tumors. Pathway predictions were conducted as previously described [37, 47, 106, 133]. Specific settings for unsupervised hierarchical clustering, gene signature application and input /output files are available at https://www.msu.edu/~andrech1/. All statistical tests and Kaplan-Meier plots were performed using GraphPad Prism 4 software.

CHAPTER 2:

A GENOMIC ANALYSIS OF MOUSE MODELS OF BREAST CANCER REVEALS MOLECULAR FEATURES OF MOUSE MODELS AND RELATIONSHIPS TO HUMAN BREAST CANCER

ABSTRACT

Genomic variability limits the efficacy of breast cancer therapy. To simplify the study of the molecular complexity of breast cancer, researchers have used mouse mammary tumor models. However, the degree to which mouse models model human breast cancer and are reflective of the human heterogeneity has yet to be demonstrated with gene expression studies on a large scale. To this end, we have built a database consisting of 1,172 mouse mammary tumor samples from 26 different major oncogenic mouse mammary tumor models. In this dataset we identified heterogeneity within mouse models and noted a surprising amount of interrelatedness between models, despite differences in the tumor initiating oncogene. Making comparisons between models, we identified differentially expressed genes with alteration correlating with initiating events in each model. Using annotation tools, we identified transcription factors with a high likelihood of activity within these models. Gene signatures predicted activation of major cell signaling pathways in each model, predictions that correlated with previous genetic studies. Finally, we noted relationships between mouse models and human breast cancer at both the level of gene expression and predicted signal pathway activity. Importantly, we identified individual mouse models that recapitulate human breast cancer heterogeneity at the level of gene expression. This work underscores the importance of fully characterizing mouse tumor biology at molecular, histological and genomic levels before a valid comparison to human breast cancer may be drawn and provides an important bioinformatic resource.

INTRODUCTION

Breast cancer is a heterogeneous disease with significant mortality associated with metastatic progression. Classification subdivides human breast cancer into six categories including Luminal A, Luminal B, HER2+, Basal, Claudin-low and normal-like [58]. Recent

work suggests additional subclasses exist within each intrinsic subtype including three basal subtypes with striking differences in overall survival [47]. Further, the TCGA and ENCODE projects show remarkable variability in genetic alterations beyond gene expression both across and within subtypes of human breast cancer. Together these genomic analyses demonstrate the complex nature of human breast cancer.

To more readily study mechanisms leading to breast cancer, research has turned to the mouse as a model. Mouse models of breast cancer have employed various methods of initiation, including MMTV infection, chemical mutagenesis and genetically engineered mice (GEM). This pioneering work identified and tested the role of many oncogenes in breast cancer. With the insertion of MMTV into the genome, numerous key oncogenes were uncovered [91, 92]. The later development of MMTV driven transgenics allowed for development of spontaneous models. With the identification of HER2 amplification in human breast cancer [52, 53], the observation that MMTV driven expression of the activated rat form of HER2 (NeuNT) resulted in breast cancer reinforced the importance of HER2 as a driving oncogene [134]. More recently, models have been refined to include tissue specific activation resulting in gene amplification, analogous to human HER2+ breast cancer [135], as well as temporal control where transgene expression can be activated or inactivated [101].

Individual mouse models have been used to model aspects of human breast cancer and the selection of the appropriate model to compare to human breast cancer has been directed by phenotype or known genetic events. For instance, the MMTV-PyMT model is widely used to examine metastasis [136] while P53 knockout mammary epithelium transplanted into wild type hosts results in tumors with various genetic mutations [129]. Another aspect is the histological subtype associated with various tumors in GEM models and the metastatic ability can be altered

with background [137]. Indeed, similarities between mouse models such as Neu and Wnt as well as their human counterparts have been previously noted [138, 139]. Importantly, in both human breast cancer and in many GEM models, there is significant histological heterogeneity [106, 107, 140]. These attributes illustrate the importance and utility of mouse models to examine breast cancer.

With the number and variety of GEM models, it is important to consider how accurately these various systems model human breast cancer. Initial studies using intrinsic clustering revealed similarities between mouse models and human breast cancer, albeit in a limited numbers of samples[105]. Yet, a more detailed characterization of a larger number of p53 null tumors revealed a variety of subtypes with strong similarities to human breast cancer [129], revealing the importance of examining a large number of samples to capture tumor heterogeneity and variability. Further, expanding the number of Myc induced tumors revealed that a subpopulation of Myc induced tumors had similarities to Claudin-low human breast cancer [141]. Taken together, recent comparative studies[108, 129, 140-143] highlighted a clear need for a comprehensive examination of the genomic features of mouse models of breast cancer and their relation to human breast cancer. To this end, we assembled an expansive dataset of mouse models of breast cancer. This dataset reveals the genomic heterogeneity of mouse models and offers a predictive resource for essential cell signaling pathways. Importantly, all comparisons between all models are made available with our report. These data demonstrate the similarities and differences of the various subtypes of mouse models to the key subtypes of human breast cancer and underscore the necessity for an informed choice of the appropriate mouse model for studying specific types of human breast cancer.

RESULTS

DATABASE ASSEMBLY

We assembled a database contained 1,172 samples from mouse mammary tumor models, cell types, and normal mammary gland. The major mouse models and descriptions are listed in TABLE 2.1. Within a number of these models, variants exist with different alleles, promoters, and genetic backgrounds. In assembling the database, we measured the non-biological variance between gene expression studies and batch correction with principle components analysis (PCA) (FIGURE S 2.1 A-D). PCA demonstrated that normalization successfully removed artificial variance between datasets (FIGURE S 2.1 B, D). As a control, we confirmed batch correction utilizing Neu-initiated tumors spanning the Affymetrix and Agilent platforms from several studies. Prior to normalization (FIGURE S 2.1 E) PCA demonstrated that Neu tumors varied by platform. After correction, Neu tumors clustered together in PCA, demonstrating that artifactual variance has been removed (FIGURE S 2.1 F). With platform and batch effects eliminated, we began to explore relationships in the mouse model database.

GENE EXPRESSION HETEROGENEITY IN MOUSE MODELS

Using unsupervised hierarchical clustering, we examined mouse mammary tumors initiated by various oncogenes. Unsupervised hierarchical clustering generated four major clusters (FIGURE 2.1A). We observed remarkable variability in gene expression profiles, including within model heterogeneity. For example, Myc initiated tumors span each of the major clusters in the dendrogram. In contrast, some models show uniformity in gene expression from tumor to tumor, including Ras initiated tumors that ordered into a single cluster. Interestingly there was significant interrelatedness between tumor models initiated with different oncogenes. Annotations for individual tumors revealed that similarities in tumor histology correlated with

relationships in gene expression profiles. For example, MMTV-Myc, MMTV-Met and a subset of DMBA induced tumors of the adenosquamous histology shared gene expression profiles. These data reveal mouse models with various levels of heterogeneity and illustrate some of the tumor phenotypes that drive relationships between different mouse models. We used SAM to identify differentially regulated genes that define tumors within each cluster (Additional file 6, https://www.msu.edu/~andrech1/BCR_Supplemental/BCR_Supplemental.html).

To describe each gene lists for possible functional gene ontologies we used GATHER (Additional file 6, https://www.msu.edu/~andrech1/BCR_Supplemental/BCR_Supplemental.html). For instance, FIGURE 2.1B shows the gene ontologies for the upregulated genes in the blue cluster in FIGURE 1A. Ontological categories included genes involved in biological processes and metabolism. To refine these results, tumors from each cluster were examined with GSEA (Additional file 7, https://www.msu.edu/~andrech1/BCR_Supplemental/BCR_Supplemental.html). Focusing on tumors in the black cluster, GSEA showed enrichment for gene sets separating mesenchymal cells from luminal cells (FIGURE 2.1C, FIGURE S 2.2 A), including low expression of Zeb1 target genes (FIGURE S 2.2 B). Gene lists that define mammary stem cells demonstrated that this cluster also had a gene expression profile enriched for mammary stem cell-like features (FIGURE S 2.2 C, D). In agreement, the majority of EMT like tumors were observed in the black cluster (FIGURE 2.1A, FIGURE S 2.3). GSEA also demonstrated that tumors from the other clusters had gene expression profiles consistent with luminal cells (FIGURE S 2.4). For example, tumors within the blue cluster correlated with gene signatures for luminal progenitor cells and the orange cluster had similarities in gene expression to mature luminal cells. Together, these results define the characteristics of the tumors contained in the major clusters.

FOLD CHANGE ANALYSIS

Given that unique initiating events in the tumor models should cause characteristic responses associated with the tumor initiating event, we used SAM to identify genes significantly altered within each model compared to all other models (Additional file 2, https://www.msu.edu/~andrech1/BCR_Supplemental/BCR_Supplemental.html).
Fold change differences were also calculated between the tumors within a model and normal mammary glands in the corresponding genetic background (Additional file 3, https://www.msu.edu/~andrech1/BCR_Supplemental/BCR_Supplemental.html). As an example, we determined fold change gene expression differences for Neu initiated tumors (FIGURE 2.2A). Collectively, SAM analysis provided a collection of genes that are differentially expressed in each model.

To identify possible transcription factors that could be active in mediating these gene expression changes, we annotated fold change results for each model using TRANSFAC (Additional file 2, 3, https://www.msu.edu/~andrech1/BCR_Supplemental/BCR_Supplemental.html). For example, genes regulated by Neu (FIGURE 2.2A), we predicted that a significant number of genes had predicted binding sites for the Krox family of transcription factors (FIGURE 2.2B). The complete results for the transcription factor binding predictions are included in the additional data for each of the models.

We also annotated fold change differences between each model using gene ontologies (Additional files 2, 3, https://www.msu.edu/~andrech1/BCR_Supplemental/BCR_Supplemental.html). As an example of the utility of the method, we examined the similarities and differences in gene ontologies in the Neu and Tag models (FIGURE 2.2C). Both Neu and TAG tumors featured biological processes, metabolism, and nucleic acid-related metabolism as major ontological

categories. Key differences included Neu tumors with genes related to transport, ion transport, and biosynthesis, categories not found with TAG gene expression changes. TAG tumors had major ontologies representing genes involved in cell cycle, cell organization, cytoskeleton organization and biogenesis, and cell organization and biogenesis. To expand ontology results we compared each model to all other models and separately to normal mammary gland using GSEA (Additional file 11, https://www.msu.edu/~andrech1/BCR_Supplemental/BCR_Supplemental.html). This analysis predicted unique features for all models including specific information on metabolism, microenvironment, metastasis, and possible pathway activation (FIGURE 2.3). For example, TAG tumors had down regulation of genes significantly enriched for the TCA cycle (FIGURE 2.3A). Wnt tumors were predicted to have upregulation of tumor angiogenesis (FIGURE 2.3B). Not surprisingly, PyMT tumors show enrichment for gene sets that predict metastasis (FIGURE 2.3C). Finally, GSEA results predict that p53 mutant tumors may have increased TNF signaling activity (FIGURE 2.3D). Together, these results provide a catalogue of possible important features corresponding to the transcriptional outcomes of an initiating oncogene event.

PATHWAY ANALYSIS

To expand the predictive analysis, we utilized a gene signature approach to predict pathway activation across mouse mammary tumors. The pathway prediction relationships between the various models were organized with unsupervised hierarchical clustering (FIGURE 2.4). Using this approach, we noted a large degree of heterogeneity within models. Myc tumors showed extensive variation in pathway activation profiles, spanning the spectrum of clusters. To understand heterogeneity and pathway activity within each model, we show pathway predictions (Additional file 12, https://www.msu.edu/~andrech1/BCR_Supplemental/BCR_Supplemental.html). For

example, in PyMT induced tumors, there is a significant difference in predicted pathway activity between tumors from a FVB and AKXD genetic background (FIGURE S 2.5). In Myc induced tumors with an EMT or squamous histology had distinct predicted pathway activities relative to tumors of papillary or microacinar histology (FIGURE S 2.6). In Neu-induced tumors, we observed a major difference in predicted pathway activity between Neu tumors using the MMTV promoter and a Tet-on system to drive oncogene expression (FIGURE S 2.7). Taken together, these data demonstrate that tumor type, genetic background, and promoter result in key differences in pathway activity.

To validate and illustrate the utility of pathway activation predictions for developing hypotheses about pathways that function in tumor progression, we identified models with clear pathway activity predictions. Previous genetic studies that correlate with these predictions are noted (TABLE 2.2). Demonstrating the validity of the gene signatures, we observe a large degree of agreement between pathways with predicted activity and results from previous investigations.

COMPARISONS TO HUMAN BREAST CANCER

With identification of pathways that function in tumor progression in mouse models, it is important to understand whether the given model is reflective of human breast cancer. To this end, we combined datasets for human breast cancer and the mouse mammary tumors in our database, removing both batch and platform effects (FIGURE S 2.8). To investigate the relationships between the mouse mammary tumors and human breast tumors, we used unsupervised hierarchical clustering. We identified a large number of mouse mammary tumor models that had similarities in gene expression profiles to human breast cancer (FIGURE 2.5). Importantly, Myc and Met induced tumors both recapitulate the heterogeneity observed in

human breast cancer. Using histological annotations, specific relationships between Myc tumor types and human breast cancer subtypes were observed (FIGURE S 2.9). For example, Myc tumors with EMT histology clustered together with human claudin low breast cancer. Extending this to the cluster of tumors predicted to have mesenchymal gene expression features (FIGURE 2.1C), we observed that a large majority of these tumors also clustered with claudin low breast cancer. Importantly, further investigation of these tumors matched marker expression for claudin low tumors (FIGURE S 2.10). Together these data demonstrated that there are mouse models that share human breast cancer heterogeneity with individual tumor types that are closely related to subsets of human breast cancer at the level of gene expression.

In addition to comparing mouse mammary tumors and human breast cancer with gene expression, we tested relationships using pathway activation predictions. Using a mixture modeling approach, we clustered human breast cancer into ten different groups based on pathway activation profiles (FIGURE 2.6). The pie chart above each heatmap shows the spectrum of the intrinsically annotated samples in each group. No single group was made up of one intrinsic subtype, illustrating the heterogeneity of pathway activation within and between intrinsic subtypes of breast cancer. After groups of human tumors were identified, the probability that an individual mouse mammary tumor belonged to a group of human breast cancer was calculated using the pathway activation profile of the mouse mammary tumor sample. Observing these probabilities with a heatmap, we noted that no single group of human breast cancer was modeled by a single mouse mammary tumor type at the pathway level. Instead, for each group of human breast cancer, multiple mouse models showed similar predicted pathway activation profiles. Further, these results demonstrated that mouse model relationships to human breast cancer extended beyond the initiating oncogene. For example, mouse tumors initiated by Myc

overexpression contained several different tumor types, each modeling a different group of human breast cancer including those groups that have lower predicted Myc activity. Moreover, Neu initiated tumors using an inducible promoter frequently model a single group of human breast cancer (FIGURE S 2.11), while other Neu models have diverse pathway activation profiles leading to relationships with several different groups of human breast cancer. These results considered together highlight the similarity and differences between mouse models and human breast cancers.

DISCUSSION

Here we have described the genomic analysis of a dataset composed of publicly available gene expression data for mouse models of breast cancer. These data have been analyzed through a variety of mechanisms to ask how mouse models are distinct, what properties they share and how they reflect human breast cancer. These data indicate that great care should be taken to appropriately choose the mouse model to use and that a genomic and histological characterization of tumors should be completed following experimentation.

In the examination of mouse models in the database, unsupervised hierarchical clustering revealed significant heterogeneity both between models and within models and was pronounced in tumor models with a large number of samples. Between model differences were fully expected given the unique initiating events causing tumor formation. However, prior studies with relatively few samples for each model did not demonstrate extensive within model heterogeneity [105]. In comparison, we have demonstrated extensive heterogeneity within many models. In part this is due to differences between intrinsic clustering methods [144] and unsupervised hierarchical clustering. However, given that we have noted corresponding differences in fold change, GSEA predictions and pathway signature probabilities, it is likely that this is a reflection

of the number of samples used in the analysis. As such, this provides an important caution to characterize a sufficiently large population of tumors to capture heterogeneity in the analysis.

Given that there is typically a predominant histological pattern associated with a given GEM tumor type [145], it was not surprising that there was a predominant genomic pattern. Indeed, we noted for many models that histology is predictive of the genomic subtype. Interestingly, this histological and genomic interaction is capable of spanning tumor initiating events from different mouse models. Indeed, EMT and spindle-type tumors from diverse models clustered together and were distinct from the non-EMT samples originating in the same model system. Thus, it is also critical for investigators to analyze all tumors from a given model for both histological and genomic patterns.

Mouse models were also investigated individually in comparison to the entire dataset using a variety of methods. This revealed characteristic gene expression patterns at the fold change level, specific GSEA enrichment effects, and key pathway signature differences. In many cases, these results correlated with prior studies. For instance, annotation of fold change results predicted that Neu induced tumors upregulated Krox 20 which was consistent with previous ChIP results [146]. When pathway signatures were examined, there were a large number of predictions that could be made for pathways used in specific GEM tumor models. Importantly, while these pathway signatures have previously been validated [47], the model by model pathway predictions shown in TABLE 2.2 were highly consistent with previously published tests. For instance, the pathway signatures predicted a high probability of Src activation in PyMT tumors in the FVB background and recent work has demonstrated the necessity for c-Src in PyMT induced tumors [147]. Collectively, for the pathways listed in TABLE 2.2, we note agreement between the pathway signature predictions and the reported genetic crosses.

Moreover, the pathway signature predictions are also reflective of additional mutations that accumulate in the samples. This was noted in the Myc and TAG induced tumors were the Ras signature was predicted to be elevated, consistent with the large number of Ras activating mutations in these strains [106, 148]. Given that numerous published genetic tests are in agreement with the pathway predictions, the remaining cell signaling pathway predictions offer a large number of testable hypotheses. In the future, pathway predictions in the various models should prove to be an important resource for initiating studies into investigating the importance of various signaling pathways in tumor biology.

One of the key aspects of this study was the comparison between mouse models and human breast cancer. These data demonstrated similarities and differences between the two groups and should serve as an important consideration when attempting to extend the comparison of mouse models to human cancer. Taking into account the clustering data, we readily noted that the heterogeneity between human breast cancer samples was present within individual mouse models. Despite capturing the genomic diversity of the samples, we noted several samples with no genomic similarity to human breast cancer, including tumors from strains with other samples that had clear similarity to human breast cancer. This clearly suggests that if conclusions are to be drawn from mouse models of breast cancer, that the mouse samples should be compared and clustered with a variety of human tumors.

In addition to clustering of genomic data, we compared mouse models to human breast cancer through signaling pathway activation predictions. These results showed that for any given group of human breast cancer samples, there was a mouse model with similar pathway activation profiles. Using these results, it is possible to select the mouse model that most closely represents a group of human breast cancer for the signaling pathways of interest. However, it is critical to

consider both clustering and pathway activation and to combine these methods to choose the most appropriate model to mimic human breast cancer. For example, to model HER2+ breast cancer and to study the role of HER2 in tumor development, research initially used the MMTV-Neu mice [134]. However, the gene expression data reveals that this strain does not associate with the HER2+ human samples through genomic clustering. However, mixture modeling indicated that a proportion of HER + human cancers did group with the MMTV-Neu samples at the level of pathway activation. This indicates that in some aspects the mouse model is appropriately related to human HER2+ breast cancer. Further, recent reports demonstrate that a strain of mice with conditional activation of Neu under the control of the endogenous promoter which undergo amplification [135] far more closely recapitulate human HER2+ breast cancer [149]. Taken together, these data illustrate the importance of fully characterizing and using all genomic information to select the appropriate model for examination.

Recent reports have described the development of serially transplantable human breast cancer samples that are grown in a murine host with clear genomic similarity to the primary human breast cancer samples [150] and obviously this is an optimal model for specific studies. However, there is clear utility for GEM models, especially with regard to the ability to ask defined genetic questions with regard to key signaling pathways in tumor biology. As such, the prior characterization of mouse and human breast cancer similarities was a critical development [105]. The expanded number of samples and methods of analysis in this report have clearly illustrated additional components of mouse breast cancer biology that require careful consideration. Indeed, the extent of genomic heterogeneity was only appreciated previously for select models [106, 107, 137, 140], but our work indicates that this is a general characteristic across the majority of breast cancer model systems. As such, this work underscores the

requirement to fully characterize mouse tumor biology at histological and genomic levels before a valid comparison to human breast cancer may be drawn. Thus, we have provided the complete files for all of the comparisons made in this manuscript, from fold change between models to GSEA and pathway predictions, with the intent of this being used as a resource to choose and compare mouse models in breast cancer research.

Collectively, our work demonstrates genomic heterogeneity in mouse mammary tumor models. As an additional outcome of this research, we have provided a large scale predictive resource for each of the mouse models in the database. With heterogeneity driving a variety of relationships between individual mouse mammary tumors and human breast cancer, this work highlights the necessity of fully characterizing mouse tumor biology at molecular, histological and genomic levels before a valid comparison to human breast cancer may be drawn.

METHODS

COMBINATION OF DATASETS

Datasets (GSE10450, GSE11259, GSE13221, GSE13231, GSE13259, GSE13553, GSE13916, GSE14226, GSE14457, GSE14753, GSE15119, GSE15263, GSE15632, GSE15904, GSE16110, GSE17916, GSE18996, GSE20465, GSE20614, GSE21444, GSE22150, GSE22406, GSE23938, GSE24594, GSE25488, GSE27101, GSE30805, GSE30866, GSE3165, GSE31942, GSE32152, GSE34146, GSE34479, GSE6453, GSE6581, GSE6772, GSE7595, GSE8516, GSE8828, GSE8863, GSE9343, GSE9355 GSE37954, GSE2034, GSE2603, GSE4922, GSE6532, AND GSE14020) were downloaded from Gene Expression Omnibus. E-TABM-683 and E-TABM-684 were downloaded from Array Express. For Affymetrix data, Bayesian Factor Regression Methods (BFRM) [25] was used to combine datasets and remove batch effects.(http://www.stat.duke.edu/research/software/west/bfrm/download.html). Agilent data was

merged with Affymetrix data using Chip Comparer (http://chipcomparer.genome.duke.edu/) and Filemerger (http://filemerger.genome.duke.edu/)To remove platform effects between Affymetrix and Agilent data and batch effects between individual Agilent studies we used COMBAT [24] (http://www.bu.edu/jlab/wp-assets/ComBat/Download.html). Batch effects and batch correction were visualized by principle component analysis in Matlab (for code see Additional file 1, https://www.msu.edu/~andrech1/BCR_Supplemental/BCR_Supplemental.html).

DATA ANALYSIS

Unsupervised hierarchical clustering was done using Cluster 3.0 and exported using Java Tree View. The color scheme for the heatmap and sample legends were made using Matlab . Human breast cancer sample intrinsic subtypes were classified according to protocol [58]. Prior to clustering mouse models with human breast cancer, we clustered the human breast tumor samples on their own, to identify genes that would organize the breast tumors according to their intrinsic subtype in the combined dataset. We used these genes to filter the mouse and human combined gene expression dataset for unsupervised hierarchical clustering.

Significance analysis of microarrays [28] was used for fold change analysis. Settings for each comparison can be found in the excel download for each model (Additional files 2, 3, https://www.msu.edu/~andrech1/BCR_Supplemental/BCR_Supplemental.html). Gene ontology and TRANSFAC predictions were made using GATHER (http://gather.genome.duke.edu/). GSEA was done using Genepattern (http://genepattern.broadinstitute.org/gp/pages/login.jsf). The gene-set describing mammary cell-types was derived from [151].

Pathway activation was predicted according to previous studies [37, 47] . For mouse samples, specific conditions for each pathway signature can be found in Additional file 4. For human breast tumor samples, pathway activation was predicted using Score Signatures

(https://genepattern.genome.duke.edu/gp/pages/login.jsf) and conditions can be found [47].

Mixture modeling was implemented according to[47].

CHAPTER 3:

THE E2F TRANSCRIPTION FACTORS REGULATE TUMOR DEVELOPMENT AND METASTASIS IN A MOUSE MODEL OF METASTATIC BREAST CANCER.

ABSTRACT

While the E2F transcription factors have a clearly defined role in cell cycle control, recent work has uncovered new functions. Using genomic signature methods, we predicted a role for the activator E2F transcription factors in the MMTV-PyMT mouse model of metastatic breast cancer. To genetically test the hypothesis that the E2Fs function to regulate tumor development and metastasis, we interbred MMTV-PyMT mice with the knockouts of E2F1, E2F2 and E2F3. With the ablation of individual E2Fs we noted alteration of tumor latency, histology, and vasculature. Interestingly, we noted a striking reduction in metastatic capacity and circulating tumor cells in both E2F1 and E2F2 knockout backgrounds. Investigating E2F target genes that mediate metastasis, we found that E2F loss led to decreased levels of Vegfa, Bmp4, Cyr61, Nupr1, Plod 2, P4ha1, Adamts1, Lgals3, and Angpt2. These gene expression changes indicate that the E2Fs control expression of genes critical to angiogenesis, remodeling of the extracellular matrix, tumor cell survival and tumor cell interactions with vascular endothelial cells to facilitate metastasis to the lungs. Taken together, these results reveal that the E2F transcription factors have key roles in mediating tumor development and metastasis in addition to their well characterized roles in cell cycle control.

INTRODUCTION

Breast cancer remains a leading cause of death for women, with high mortality rates attributed to distant metastasis [65]. To simplify the examination of signaling pathways requirements in metastatic breast cancer, research has turned to mouse model systems. Previous studies in mouse models of breast cancer have begun to reveal the mechanistic features of breast cancer metastasis and *in vivo* selection has demonstrated the ability to select for tumors that metastasize to a specific location [64, 84, 85, 152]. Yet, we lack a complete understanding of the

pathways that govern the molecular circuitry of metastatic breast cancer. One model that has been integral in examining metastatic progression is the MMTV-Polyoma Virus Middle T (PyMT) model. Originally described with rapid tumor onset and a high degree of pulmonary metastasis [110], this model has since been used to examine a number of facets of metastasis. For example, work using the MMTV-PyMT model led to the discovery of the prometastatic signaling exchange between tumors and macrophages [153]. In addition, the metastatic contribution of individual signaling molecules, such as TGF-beta, AKT, and adiponectin, has also been uncovered using this model [111, 154, 155]. Given that PyMT can activate multiple signaling pathways with relevance to human breast cancer [136] there is clear utility in this model for characterizing pathways contributing to breast cancer metastasis.

Identification of signaling pathways contributing to tumor progression has been enhanced by recent progress in bioinformatic methods. One such method is the development of genomic signatures for determining signaling pathway activation status [36, 37]. By generating gene expression training data for cell signaling pathways, a signature can be created and applied to subsequent gene expression datasets to predict whether the pathway in question is activated. This method has demonstrated heterogeneity in human breast cancer [47], moving beyond and refining the intrinsic classification of breast cancer [57, 59]. In addition, this method demonstrated tumor heterogeneity in mouse models of breast cancer [106, 141]. As a predictive tool, genetic signatures allow the identification of signaling pathways that may contribute to tumor development. Indeed, applying genomic signatures to Myc induced tumors revealed that E2F transcription factors were predicted to function in tumorigenesis and a genetic test of this prediction demonstrated that E2Fs were involved in tumor onset and progression [108]. Using a

similar approach, our current study predicted a role for the activator E2F transcription factors in MMTV-PyMT induced tumorigenesis.

The E2F transcription factor family is broadly classed into transcriptional activators (E2F1-3A) and repressors (E2F3B-8) and the family members have been well characterized as regulators of the cell cycle [156-158]. Prior implications in human cancer show that E2Fs are important regulators of apoptosis and proliferation [159]. However, recent work has identified roles for E2Fs beyond simple cell cycle regulation [160]. For example, a xenograft study utilizing shRNA knockdown of E2F1 in melanoma cell lines, showed that knockdown of E2F1 significantly reduced the size of pulmonary metastases [161]. In esophageal squamous cell carcinoma, it was shown that patients with tumors that immunostained positive for E2F1 had a worse overall survival rate than patients with E2F1 negative tumors [162]. Similarly, in prostate cancer patients with detectable nuclear staining for E2F3 have worse prognosis than patients where E2F3 is undetectable [163]. Furthermore, we recently demonstrated that E2Fs also play a role in human breast cancer relapse free survival time [108]. Together, these prior studies demonstrate the clinical significance of the E2Fs in human cancer. Here we used genomic signatures to predict that E2F transcription factors are involved in a mouse model of breast cancer metastasis. We then genetically demonstrated that E2F loss in MMTV-PyMT tumors alters tumor development, progression, and metastasis. Taken together, these data indicate that there is a critical role for E2F transcription factors in the regulation of metastasis.

RESULTS

To identify signaling pathways associated with the metastatic progression of breast cancer, we applied a number of genomic signaling signatures to gene expression data from mouse models of breast cancer. Using previously described training data and methods [36, 37,

47, 106], we predicted signaling pathway activity across these models, which included the highly metastatic PyMT tumors. When we examined the pathway activation predictions in the PyMT tumor model (FIGURE 3.1A), we noted activation in a number of pathways known to be critical in PyMT tumors, such as AKT [155]. We also observed a surprising degree of genomic heterogeneity in the tumor samples. Despite this heterogeneity, we found that virtually all samples had high levels of predicted activity for the E2F1 transcription factor. The two other activator E2Fs were also observed to have elevated probability of activation in a subset of tumor samples. With activation of E2F signatures in the majority of samples, this indicated that the E2F transcription factors may be involved in PyMT-mediated tumorigenesis.

PyMT induced tumors are highly metastatic and have previously been used to examine the metastatic process. Accordingly, we used the Kaplan-Meier Plotter tool (kmplot.com) to screen human breast cancer Distant Metastasis Free Survival (DMFS) clinical data for association between E2F1, E2F2, and E2F3 individual probes and upregulated signature genes with DMFS times in human breast cancer. Importantly, elevated expression levels of either E2F1 (Hazard ratio: 1.52, 95% CI: 1.22-1.88, p=0.00016), E2F2 (Hazard ratio: 1.33, 95% CI: 1.07-1.66, p=0.012), or E2F3 (Hazard ratio: 1.41, 95% CI: 1.15-1.74, p=0.00095) were individually associated with reduced time to distant metastasis in breast cancer compared to patients with tumors with low levels of expression (FIGURE 3.1B-D, respectively). Similarly, we assessed the upregulated genes in E2F1, E2F2, E2F3 pathway signatures. Elevation of E2F1 signature genes (Hazard ratio: 1.46, 95% CI: 1.19-1.79, p=0.00024) and E2F2 signature genes (Hazard ratio: 1.52, 95% CI: 1.25-1.87, p=0.000037) also correlated with reduced time to distant metastasis in breast cancer patients compared to patients with low expression of these genes (FIGURE S 3.1 A, B). In contrast, high expression of E2F3 signature genes (Hazard ratio: 0.73,

95% CI: 0.59-0.91, p=0.0052) correlated with a prolonged time to distant metastasis in breast cancer patients (FIGURE S 3.1 C). Taking into account intrinsic subtype status, we found that high levels of E2F1 and E2F1 signature genes associated with a decreased time to distant metastasis in luminal A and luminal B breast cancers (FIGURE S 3.2). High levels of E2F2 predicted a decrease time to metastasis in luminal A breast cancer while E2F2 signature genes were similarly associated in both luminal subtypes (FIGURE S 3.3). For E2F3, elevated probe levels and signature genes did not associate with a decreased time to metastasis in a particular subtype (FIGURE S 3.4). Considered with the PyMT mouse model data, this strongly suggested that activator E2F transcription factors have a functional role in breast cancer progression and metastasis.

To test the hypothesis that E2Fs regulate metastatic breast cancer, we interbred MMTV-PyMT mice into E2F1, E2F2 and E2F3 knockout backgrounds. Due to embryonic lethality, E2F3 mice were maintained in the heterozygous state. To study E2F function in mammary tumor development, we examined mammary whole mounts of 35 day old MMTV-PyMT virgin mice in the various E2F backgrounds (FIGURE 3.2A-D). In all E2F backgrounds transformation of the ductal tree was evident through whole mount analysis. In comparison to the E2F wild type control background (FIGURE 3.2A), normal ductal epithelium was consistently absent as indicated by representative $E2F1^{-/-}$ mammary glands (FIGURE 3.2B). Unlike E2F1, loss of E2F2 or E2F3 did not result in mammary glands appreciably different from the control (FIGURE 3.2C-D respectively).

Given that loss of E2F1 resulted in transformation of the entire mammary epithelium we postulated that this should result in acceleration of tumor onset. Mammary glands were regularly palpated for the presence of mammary tumors. Consistent with the mammary whole mount

results, we noted a significant ($p<0.0001$) acceleration in tumor onset in the E2F1$^{-/-}$ mice compared to the control wild type E2F background (FIGURE 3.2E). Indeed, 50% of control mice had developed tumors by 42 days while the loss of E2F1 resulted in tumors in 50% of mice by 35 days (MMTV-PyMT median latency= 42 days, MMTV-PyMT E2F1$^{-/-}$ median latency= 35 days, hazard ratio= .2507, 95% CI= .02660-.1458). In contrast, no latency differences were associated with the loss of E2F2 (median latency = 40 days, hazard ratio= .8195, 95%CI= .4109-1.461) (FIGURE 3.2F). However, E2F3 heterozygous (+/-) mice were noted to have a significant delay in tumor onset ($p=0.004$, median latency= 48 days, hazard ratio= 1.955, 95% CI= 1.297-4.124) (FIGURE 3.2G). Together these data demonstrate the differential roles of the E2Fs during the initiation of tumor development.

In order to examine the role of the E2F transcription factors in tumor proliferation, we compared the growth rate of the PyMT induced primary tumors in the various E2F backgrounds. Despite the differences in tumor onset, no significant alterations in the time from tumor palpation to end stage were observed with loss of the E2Fs (FIGURE S 3.5A). In addition, we assessed tumor burden at endpoint when the primary tumor reached 20 mm in the largest dimension. While no differences were observed for the E2F1$^{-/-}$ or E2F2$^{-/-}$ background, fewer tumors developed in the E2F3$^{+/-}$ background ($p=0.01$) (FIGURE S 3.5B). However, when total tumor volume was observed the E2F3 mutants were indistinguishable from the wild type E2F background (FIGURE S 3.5C). Moreover, KI67 (FIGURE S 3.6) and TUNEL (FIGURE S 3.7) staining in early stage (tumor diameter =6mm) and end-stage tumors indicated that E2F1 loss had no effect on tumor cell proliferation or apoptosis. Together these data indicate that despite alterations to tumor latency there were surprisingly few effects of E2F loss on growth rate, tumor burden, proliferation or apoptosis in MMTV-PyMT tumors.

Since E2F loss had no impact on these features of tumor growth, we investigated whether compensatory upregulation of other E2F family members had occurred. To do this, we assayed levels of E2F1, E2F2, E2F3A, and E2F3B by qRT-PCR across tumor genotypes (FIGURE 3.3). Compared to $E2F^{WT/WT}$ (n=4) tumors, $E2F1^{-/-}$ tumors (n=4) showed similar levels of E2F2 and E2F3B, but had significant upregulation of E2F3A (p=0.0232). In $E2F2^{-/-}$ tumors (n=4), we detected a significant decrease in E2F1 levels (p=0.0016) and significant upregulation of E2F3A (p=0.0105). In $E2F3^{+/-}$ tumors, expression levels of E2F1 and E2F2 were similar to $E2F^{WT/WT}$ tumors. Interestingly, $E2F3^{+/-}$ mice had upregulation of E2F3A bordering statistical significance (p=.0641) and significant downregulation of E2F3B (P=0.0175). $E2F3^{+/-}$ 35 day old mammary glands (n=4) show a modest decrease in E2F3 protein levels in $E2F3^{+/-}$ mice compared to $E2F^{WT/WT}$ mammary glands (n=4) (FIGURE S 3.8A). In early stage $E2F3^{+/-}$ tumors (n=3, 6mm diameter), we detected similar levels of E2F3 as $E2F^{WT/WT}$ tumors (n=3, FIGURE S 3.8B). Similar observations were made in end stage tumors (n=4 for each genotype, 20 mm diameter), though E2F3 protein levels were more variable (FIGURE S 3.8C). As a whole, these results demonstrate that E2F loss in these tumors led to compensatory upregulation of the E2F3A isoform.

In addition to effects on latency, histological patterns observed in the tumors indicated that the E2F transcription factors play a role in tumor heterogeneity. In all backgrounds the most common histological type of tumor was the microacinar subtype (FIGURE 3.4A). In addition, we frequently noted adenosquamous tumors (FIGURE 3.4B) as well as a number of other types at reduced frequency across genetic backgrounds (FIGURE 3.4C). The frequency of adenosquamous tumors was noticeably affected in the $E2F1^{-/-}$ and $E2F2^{-/-}$ mice. Indeed, loss of E2F1 significantly reduced the frequency of adenosquamous tumors from 8% to 1% (p=0.001)

(FIGURE 3.4D). Conversely, loss of E2F2 significantly increased the proportion of adenosquamous tumors to 21% of all tumors (p=0.0003). In contrast to the alterations in histology found with E2F1 and E2F2, no effects on tumor histology were noted for E2F3 mutants.

Given that we predicted a high probability of E2F activation in the highly metastatic MMTV-PyMT mouse model and that we noted high E2F levels being associated with decreased time to distant metastasis in human breast cancer, we hypothesized that the E2Fs were involved in breast cancer metastasis. To test this hypothesis we examined the lungs of MMTV-PyMT mice in the various E2F backgrounds at endpoint. Metastatic tumors were readily observed on the surface of the lungs in control MMTV-PyMT mice (FIGURE 3.5A). Interestingly, we did not observe these metastases in the $E2F1^{-/-}$ or $E2F2^{-/-}$ backgrounds (FIGURE 3.5B, C respectively) but did note metastatic tumors on the surface of the lungs in the $E2F3^{+/-}$ background (FIGURE 3.5D). Histology for matched sections of the lungs was examined, demonstrating widespread metastasis in the lungs of $E2F^{WT/WT}$ mice (FIGURE 3.5E) with the histology for the median number of metastatic lesions being shown. Consistent with the gross observations, there was a readily apparent decrease in the number of metastatic lesions in the $E2F1^{-/-}$ and $E2F2^{-/-}$ mice (FIGURE 3.5F, G) but not in the E2F3 mutant mice (FIGURE 3.5H). Metastases in the boxed areas are shown at higher magnification (FIGURE 3.5I-L). To quantitate the number of metastases, a representative section of the lung was counted for each of the tumor bearing mice (n=37 for $E2F^{WT/WT}$, n=21 for $E2F1^{-/-}$, n=21 for $E2F2^{-/-}$, n=23 for $E2F3^{+/-}$). This revealed a significant reduction in the number of metastases observed in the lung in both $E2F1^{-/-}$ ($p<0.0001$) and $E2F2^{-/-}$ (p=0.002) backgrounds (FIGURE 3.5M). Given that many of the noted metastases were smaller in the $E2F1^{-/-}$ and $E2F2^{-/-}$ backgrounds, we also examined the area

occupied by the metastases as a function of the total lung area. This demonstrated that loss of either E2F1 or E2F2 significantly reduced metastatic burden ($p<0.0001$ for both) (FIGURE 3.5N).

To determine the stage at which E2F1 or E2F2 loss blocked metastasis, we assayed the number of circulating tumor cells (CTCs) at endpoint in the tumor bearing mice. To detect CTCs, we collected blood from a cardiac puncture and cultured the CTCs in a colony forming assay. Compared to age matched wild type tumor free controls (n=6) where no colonies were detected (FIGURE 3.6A), the MMTV-PyMT strain (n=14) was found to have a number of discreet colonies (FIGURE 3.6B). The number of CTCs was visibly reduced in both the $E2F1^{-/-}$ (n=7, FIGURE 3.6C) and in the $E2F2^{-/-}$ mice (n=10, Fig 3.6D) but not in the $E2F3^{+/-}$ (n=10) mice (FIGURE 3.6E). Quantitation revealed that this was a significant decrease in CTC colonies relative to the PyMT control in the $E2F1^{-/-}$ ($p=0.02$) and in the $E2F2^{-/-}$ ($p=0.006$) mice (FIGURE 3.6F). A reduction in CTCs was confirmed by extracting RNA from blood at tumor end stage and using qRT-PCR to assay transgene expression as an indicator of circulating tumor cells. This demonstrated that $E2F1^{-/-}$ mice had a 4.8 fold reduction in transgene levels in blood compared to controls (FIGURE S 3.9), indicating a reduction in circulating tumor cells. Further, in FVB negative controls we observed no amplification of the transgene. Taken together, these results strongly suggest that loss of E2F1 or E2F2 inhibits metastasis by reducing the number of circulating tumor cells.

To test whether E2F1 or E2F2 also regulate colonization ability, we injected 5.0×10^5 cells derived from $E2F^{WT/WT}$ PyMT, $E2F1^{-/-}$ PyMT, and $E2F2^{-/-}$ PyMT tumors into the circulation of wild type control mice. In mice injected with $E2F^{WT/WT}$ PyMT cells (n=8), we observed robust lung colonization (FIGURE 3.7A). In contrast we observed a vast reduction in colonization for

mice injected with E2F1$^{-/-}$ PyMT tumor cells (n=9) and E2F2$^{-/-}$ PyMT tumor cells (n=7) (FIGURE 3.7 B and C). Mice injected with E2F$^{WT/WT}$ PyMT tumor cells contained an average of 8.5 metastases per section of lung, while a significant reduction was observed in mice injected with E2F1$^{-/-}$ (average number of metastases=1.9, p=0.01) and E2F2$^{-/-}$ tumor cells (average number of metastases=0.85, p=0.02) (FIGURE 3.7D). Additionally, we measured the area of the occupied by metastasis in these mice and observed a 18-fold reduction for mice injected with E2F1$^{-/-}$ tumor cells and a 50-fold reduction for mice injected with E2F2$^{-/-}$ tumor cells in comparison to mice receiving E2F$^{WT/WT}$ PyMT tumor cells (FIGURE 3.7E). Interestingly, the reduction in circulating tumor cells (FIGURE 3.6) or in tumor cell colonization ability (FIGURE 3.7) was not related to defects in cell migration measured through scratch assays (FIGURE S 3.10) and transwell migration assays (FIGURE S 3.11) as these experiments revealed no difference in motility between E2F$^{WT/WT}$ and E2F$^{-/-}$ tumor cells. Nonetheless, these data provide a clear demonstration that E2F1 and E2F2 are necessary for tumor cell metastasis.

To further illustrate the significance of the E2Fs for pulmonary colonization, we performed qRT-PCR to compare E2F1 and E2F2 levels in E2F$^{WT/WT}$ pulmonary metastases relative to E2F$^{WT/WT}$ primary tumors. Interestingly, E2F1 expression levels were nearly 7 times higher in lung metastases compared to primary tumors (p=0.0004, FIGURE 3.8A). E2F2 levels were similar between lung metastases and primary tumors (FIGURE 3.8B). We also applied our gene signatures for E2F1 and E2F2 activity to previously published gene expression data for MMTV-PyMT tumors and lung metastases [164]. Consistent with qRT-PCR results, predicted E2F1 activity was significantly higher (p=0.0007) in lung metastases compared to primary tumors (FIGURE 3.8C), while no differences in activity were observed for E2F2 (FIGURE 3.8D).

To test whether metastatic regulation by E2F1 or E2F2 is cell autonomous or is a result of E2F-associated tumor microenvironment defects, we assayed metastatic progression utilizing a tumor transplant study. For this experiment, viable frozen tumor samples ($E2F^{WT/WT}$, n=4; $E2F1^{-/-}$, n=4; $E2F2^{-/-}$, n=4) from transgenic mice were transplanted into MMTV-Cre $E2F^{WT/WT}$ control mice. At primary tumor endpoint, metastasis was analyzed. Histological sections of lungs of mice implanted with an $E2F^{WT/WT}$ tumor demonstrated extensive metastasis (FIGURE 3.9A). In contrast, metastatic lesions were rarely observed in lungs of mice receiving an $E2F1^{-/-}$ (FIGURE 3.9B) or PyMT $E2F2^{-/-}$ (FIGURE 3.9C) tumor. These results revealed a significant reduction in the number of metastases observed in the lungs of mice implanted with an $E2F1^{-/-}$ ($p=0.003$) or $E2F2^{-/-}$ ($p=0.01$) tumor compared to mice implanted with an $E2F^{WT/WT}$ tumor (FIGURE 3.9D). Further, these metastatic defects also resulted in a dramatic reduction in the proportion of the lungs occupied by metastasis in mice implanted with $E2F1^{-/-}$ or $E2F2^{-/-}$ tumors compared to mice receiving $E2F^{WT/WT}$ tumors (FIGURE 3.9E).

To determine if compensation by other E2F family members was occurring, E2F expression in transplanted tumors was analyzed (FIGURE S 3.12). PyMT $E2F1^{-/-}$ tumors (n=4) were observed to have a significant increase in E2F2 ($p=0.0032$) and E2F3A ($p=0.0254$) expression with a significant decrease in E2F3B ($p=0.0358$). Similar to the spontaneous tumors, $E2F2^{-/-}$ transplanted tumors had a significant decrease in E2F1 expression ($p=0.0046$). Interestingly, E2F3A upregulation in $E2F2^{-/-}$ tumors was not statistically significant. However, there was significant downregulation ($p=0.0024$) of E2F3B in these tumors.

To test for tumor microenvironment effects, we also performed F4/80 staining. Consistent with observations from the tumor transplant study, we observed no differences in macrophage infiltration across E2F mutant backgrounds (FIGURE S 3.13). Together, this data

suggested that metastatic defects associated with E2F1 or E2F2 loss were intrinsic to the tumor cells.

To begin to investigate the intrinsic mechanistic features of metastatic defects, we performed CD31 staining to assay tumor vasculature in end stage tumors. In PyMT $E2F^{WT/WT}$ tumors (n=5), we observed well defined and continuous staining for CD31, indicating well developed vasculature structure (FIGURE 3.10A). In contrast, PyMT $E2F1^{-/-}$ tumors (n=5) showed remarkably altered tumor vasculature (FIGURE 3.10B). This was accompanied by a significant reduction (p=0.0002) in gene expression of the pro-angiogenic signaling molecule Vegfa in $E2F1^{-/-}$ tumors (n=6) compared to $E2F^{WT/WT}$ (n=6) tumors (FIGURE 3.10C). Jointly, these results indicate that loss of Vegfa expression in $E2F1^{-/-}$ tumors has altered blood vessel development.

To further investigate potential mechanisms of metastasis, we utilized an informatics approach to identify potential E2F targets mediating metastatic potential. As outlined in FIGURE 3.11A, we utilized E2F signature gene expression data with published ChIP-Seq and ChIP-Chip data to identify direct E2F target genes. Next, we filtered our potential targets using gene sets for metastasis available on MSigDB. Testing potential target genes via qRT-PCR we found that E2F1 (n=6) and E2F2 (n=6) tumors have significantly lower levels of; Bmp4 ($p=0.0002$, $p<0.0001$, respectively), Cyr61 ($p=0.0009$, $p=0.0006$, respectively), and Nupr1 ($p<0.0001$, $p<0.0001$, respectively), Plod 2 isoform 1($p<0.0001$, $E2F2^{-/-}$ only), Plod 2 isoform 2($p=0.0122$, $p=0.0015$, respectively), P4ha1($p=0.0006$, $p<0.0001$, respectively), Adamts1 ($p<0.0001$, $p<0.0001$, respectively), Lgals3($p<0.0001$, $p<0.0001$, respectively), and Angpt2 ($p=0.0065$, E2F1-/- only) (FIGURE 3.11 B-J). These results show which E2F target genes associated with metastatic function are downregulated in $E2F1^{-/-}$ and $E2F2^{-/-}$ tumors. In addition, these results

provide important mechanistic insight into the molecular basis of E2F regulation of tumor metastasis.

DISCUSSION

In this study we applied genomic signaling pathway signatures to mouse tumor model microarray data and predicted a role for the E2F transcription factors in PyMT induced tumors. By genetically testing this prediction, we found that E2F1 loss enhanced ductal transformation and accelerated tumor onset. In contrast to this, we noted delayed tumor onset in E2F3$^{+/-}$ mice. Histologically, we observed that loss of E2F1 resulted in a significant decrease in the incidence of adenosquamous tumors, while E2F2 loss led to an increase in the frequency of this tumor type. In some of the most striking findings, we identified a role for the E2F transcription factors in tumor metastasis. Indeed, loss of E2F1 or E2F2 dramatically reduced the metastatic capacity of MMTV-PyMT tumors. These metastatic defects were associated with a reduction in circulating tumor cells and were cell autonomous. Together, these data demonstrate a significant role for individual E2Fs in tumor development and progression.

E2F transcription factors have previously been associated with defined roles in cell cycle control, proliferation and apoptosis. While exploring the role of E2F1 in tumor development, we found that loss of E2F1 enhanced ductal transformation and accelerated tumor onset. In a previous study of Myc-mediated tumorigenesis, we noted a similar acceleration of tumor onset with loss of E2F1 [108]. Consistent with E2F1's role in Myc induced apoptosis [165], we noted that loss of E2F1 reduced apoptosis and caused Myc tumors to grow more quickly. Unlike the Myc model, we did not observe any effects on apoptosis with E2F1 loss (FIGURE S 3.7). This suggests that E2Fs can respond to different pathway stimulus uniquely to differentially regulate specific genes and separate cellular processes.

With a delay in tumor onset in E2F3$^{+/-}$ mice and previously defined roles for the E2Fs in cell cycle progression and apoptosis, alterations in tumor growth were expected but not observed in E2F mutant mice. However, levels of the various E2F alleles indicated that there was significant compensation in the E2F3$^{+/-}$ strain. Given the previously noted potential for E2F compensation [166], the lack of defects associated with tumor growth in the E2F mutant mice may not be surprising. Consistent with our data, it has been shown that with E2F1 loss E2F3A is upregulated [166]. Further work has also demonstrated that E2F3A can compensate for the loss of E2F1 to sustain cell proliferation [167], while E2F3A is necessary cellular proliferation [167, 168]. With upregulation of E2F3A in each of the E2F mutant tumors, this previous data supports our tumor growth and proliferation observations.

In addition to tumor development, we found that loss of E2F transcription factors altered tumor heterogeneity. As apparent in FIGURE 3.4C, PyMT induced tumors with wide histological heterogeneity. Interestingly, E2F1 and E2F2 had opposite effects on the incidence of adenosquamous tumors. Together the latency, histological differences and metastatic capacity clearly illustrate that the E2Fs have unique and individual roles. Given that the E2F DNA binding site is conserved in all E2Fs [169] and ChIP-seq studies have demonstrated that various E2Fs bind the same targets [157], this reinforces the idea that cooperating transcription factors such as YY1 are required for specificity of function [170].

Perhaps most notable of the experimental findings was the identification of E2Fs as regulators of breast cancer metastasis. Importantly, high levels of E2F1, E2F2, and E2F3 were predictive of accelerated onset of distant metastasis in human breast cancer. While our data for E2F1$^{-/-}$ and E2F2$^{-/-}$ mice corroborate these predictions, we did not observe metastatic impairment in the E2F3$^{+/-}$ mice. However, E2F1, E2F2, and E2F3 levels were maintained in the E2F3$^{+/-}$ mice

(FIGURE 3.3), potentially allowing other E2Fs to compensate, resulting in normal regulation of metastatic progression. Importantly, our tumor transplant study provides evidence that the E2Fs regulate tumor metastasis in a cell autonomous manner. Indeed, E2F loss had no effects on tumor growth rate, tumor burden, proliferation, apoptosis, and macrophage staining while reducing metastatic potential. This indicated that the metastatic defects are not a result of altered tumor development but are due to inherent properties of the tumor.

While E2F1$^{-/-}$ and E2F2$^{-/-}$ mice each present a reduction in tumor metastasis, our data suggest that E2F1 loss is responsible for the metastatic deficiency. Indeed, E2F1 levels and activity were elevated in lung metastases compared to primary tumors. In addition, PyMT E2F1$^{-/-}$ tumors transplanted into wild type recipients had significantly higher levels of E2F2, yet still had significantly lower number of metastases. Meanwhile, both spontaneous and transplanted PyMT E2F2$^{-/-}$ tumors had significantly lower levels of E2F1 while failing to metastasize. Interestingly, E2F1$^{-/-}$ and E2F2$^{-/-}$ tumors shared expression patterns for most metastatic target genes we surveyed. Together, this may indicate the mechanism behind metastatic defects noted in both E2F1$^{-/-}$ and E2F2$^{-/-}$ tumors may be mediated primarily by E2F1 regulated genes.

Noting both the reduction in CTCs and tumor cell colonization ability, our data suggests E2Fs regulate metastasis in early and late stages of metastatic progression. Through qRT-PCR of E2F target genes that have noted metastatic properties, we have identified molecular features of metastasis regulated by E2Fs. In the early steps of metastasis, we found that E2F1$^{-/-}$ tumors had a reduction in Vegfa and as a result, the tumor vasculature was altered. Indeed, previous reports have shown that E2F1 controls angiogenesis through the VEGF signaling axis [171]. In addition, we detected other altered E2F target genes noted to function in tumor angiogenesis. Previous studies have demonstrated Angpt2 is able to promote angiogenesis [172, 173] and

tumor cell invasion [174]. In addition, E2F1 tumors have significantly low levels of Cyr61 (cysteine-rich angiogenic inducer 61) and blocking Cyr61 function decreased metastasis in a xenograft of the MDA-MB231 human breast cancer cell line [175]. Further, given the finding that Cyr61 can regulate tumor angiogenesis independent of Vegfa expression [176], it seems that loss of Cyr61 together with loss of Vegfa and Angpt2 provides a molecular context for the pronounced tumor vasculature effects associated with E2F loss. With blood vessel recruitment being a key rate limiting step for metastasis [177], these data reveal the molecular alterations contributing to the angiogenesis defects associated with E2F loss and indicate one of the mechanisms by which E2Fs are involved in mediating the early steps of metastasis.

In addition to tumor angiogenesis, we detected gene expression changes that indicate that loss of E2F1 or E2F2 may impact tumor cell remodeling of the extracellular matrix (ECM). Specifically, we noted $E2F1^{-/-}$ and $E2F2^{-/-}$ tumors had a 2-fold reduction in the extracellular metalloprotease, Adamts1. Adamts1 has been ablated in the MMTV-PyMT mouse model, revealing the requirement for Adamts1 during tumor metastasis [178]. Importantly, this work demonstrated that Adamts1 remodels the ECM to facilitate the transition from ductal carcinoma in-situ to invasive and metastatic disease. Also critical to remodeling of the extracellular matrix to facilitate a metastatic niche is collagen deposition [179]. The PyMT tumors in the $E2F1^{-/-}$ and $E2F2^{-/-}$ background had significantly lower expression of P4ha1 and Plod2, whose products both function as collagen hydroxylases. Recent work has demonstrated the necessity for these genes in tumor cell collagen deposition [180]. Further, knock down of these genes in metastatic human breast cancer cell lines reduced the number CTCs in the blood, as well as pulmonary and lymphatic metastasis in a xenograft study [181]. Taken together, these data indicate the reduction

in metastatic potential in E2F1$^{-/-}$ and E2F2$^{-/-}$ may be due to an inability to form extracellular fibrillar collagen resulting from loss of expression of E2F target collagen hydroxylase genes.

In addition to molecular signals that recruit blood vessels and remodel the extra-cellular matrix, we found that major cell-signaling pathways were impacted by E2F loss. Our qRT-PCR analysis suggests that pathways related to the TGF-β super family and Smad activation have been impacted by loss of E2F1 and E2F2. BMP4 expression was reduced more than 3 fold in the E2F1$^{-/-}$ and E2F2$^{-/-}$ tumors. BMP4, a member of the TGF-β super family, leads to the activation of Smad1, Smad5, and Smad8 [182, 183] and studies have illustrated a role for BMP4 breast cancer cell invasion [184, 185]. These data may indicate loss of BMP4 signaling has reduced the invasive potential of the MMTV-PyMT tumor cells and thus contributing to the observed reduction in CTCs. In addition, we also detected over a 3 fold reduction in Nupr1**.** Nupr1 expression has been shown to increase in response to TGFβ1 and helps facilitate Smad transactivation [186]. The pro-metastatic functions for Nupr1 were originally identified in work that showed that Nupr1 supports the growth of human breast cancer cells after seeding a distant organ [187]. In light of these findings, loss of NUPR1 expression may lend itself to the observed defects in tumor cell colonization of the lungs in tumors lacking E2F1 or E2F2.

We also detected significantly lower expression levels of Lgals3 expression in E2F1$^{-/-}$ and E2F2$^{-/-}$ tumors. A wide array of pro-metastatic functions have been described for LGALS3 with relevance to early and late steps of the metastatic cascade [188]. For instance, it has been shown that the LGALS3 protein, galectin-3, mediates tumor cell adhesion to the ECM [189] and promotes the dissemination of tumor cells from the primary tumor [190]. Galectin-3 has also been shown to be critical for recognition and interaction of human breast cancer cells with endothelial cells of the vasculature [191, 192] and those tumor cell-endothelial cell interactions

are necessary for tumor cell invasion and metastasis [193]. In the later steps of metastasis, galectin-3 induces apoptosis in T-cells and monocytes [188]. This suggests that LGALS3 deficiency in $E2F1^{-/-}$ and $E2F2^{-/-}$ tumor cells may have allowed immune cells to reduce the number CTCs and eliminate cells before they could colonize the lungs in our study using retro orbital injection of tumor cells into the bloodstream.

In conclusion, these data demonstrate that E2F1 and E2F2 play a critical role in MMTV-PyMT mediated tumorigenesis with E2F loss leading to alterations in tumor latency, histology, and metastasis. Importantly, E2F1 and E2F2 were shown to be critical regulators of intrinsic cell signaling that allows tumor cells to progress through both early and late steps of metastasis. Importantly, we identified the potential E2F target genes that associate with these changes in metastatic ability. These gene expression changes suggests that the E2F transcription factors mediate expression of genes critical to tumor angiogenesis, tumor cell remodeling of the extracellular matrix, tumor cell survival and tumor cell interactions with vascular endothelial cells to facilitate metastasis to the lungs. Taken together, these findings indicate E2Fs regulate key functions involved in metastasis in both mouse models and human breast cancer and as such, add to the paradigm of E2F function.

METHODS

BIOINFORMATICS

Gene expression data for MMTV-PyMT tumor samples and cell lines was obtained from the Gene Expression Omnibus (GEO) under accession numbers GSE13553, GSE13221, and GSE14457. This data was combined with other mouse mammary tumor models with accession numbers GSE11259, GSE13230, GSE15904, GSE24594, GSE22406, GSE6246, GSE6453, GSE6581, GSE7595, GSE8516, GSE8828, GSE8863, GSE9343, GSE9355, GSE10450,

GSE13259, GSE13916, GSE14226, GSE14457, GSE14753, GSE15119, GSE15263, GSE15632, GSE16110, and GSE17916 . Batch effects were removed between Affymetrix derived datasets using BFRM [25]. COMBAT [24] was used to remove batch effects between Agilent and Affymetrix datasets. Pathway predictions were conducted as previously described using a gene signature approach [37, 47, 133]. Briefly, pathway predictions were made using gene expression training data for individual oncogenic pathways. By comparing oncogene overexpression to control samples, the training data provided the transcriptional response of oncogenic pathway activation. For each training dataset, metagene scores were calculated and signature genes were identified. Bayesian fitting of probit binary regression models were used to map MMTV-PyMT tumors to the metagene signature and calculate probability of activation of individual oncogenic pathways utilizing BINREG software in Matlab. Metastasis gene sets were downloaded from MSigDB (www.broadinstitute.org). ChIP-seq data and ChIP-chip data for E2F1, E2F2, and E2F3 were obtained from their respective publications [194-197]. Unsupervised hierarchical clustering was performed with Cluster 3.0 and results were visualized with JavaTreeView. Figures were converted to a full-spectrum color scale using Matlab. To examine human breast cancer, we used kmplot.com [198] to examine distant metastasis free survival.

ANIMAL STUDIES

All animal work has been conducted according to national and institutional guidelines. All mice were in the FVB background. MMTV-PyMT634 mice were obtained from Jackson Labs. PyMT transgenics were interbred with $E2F1^{-/-}$ [199], $E2F2^{-/-}$ [200], or $E2F3^{+/-}$ [168] mice. A very small number of mice with runted growth from each genotype were excluded from the experiment due to failure to thrive. Tumors were detected through weekly palpation and tumor growth was measured twice per week. Kaplan–Meier curves for tumor latency were generated

using GraphPad Prism. Tumor growth rate was measured by the amount of time for the primary tumor to reach a volume of 2,000 mm^3 after palpation.

Mammary whole mounts were conducted as previously described [133]. Once the primary tumor reached the approved endpoint the number of lung metastases and the percent area of the lungs occupied by metastasis were quantified across both lobes of a single section of H&E stained lungs. GraphPad Prism was used to conduct the Mann-Whitney statistical test.

Immunohistochemistry was performed on sections of mammary glands and representative end-stage tumors using the following antibodies; F4/80 rat monoclonal antibody from Serotec (Q61549), Ki67 from Abcam (Ab15580) and CD31 from HistoBioTec (Dianova DIA-310). TUNEL staining was done using the In Situ Cell Death Detection Kit from Roche (11684817910) and a DAB substrate kit from Vector Labs (SK-4100). For western blots assessing E2F3 protein levels, the primary antibody was purchased from Santa Cruz (sc-878) and was imaged using a Licor goat anti-rabbit secondary antibody (Licor 326-32211).

A colony forming assay was conducted based upon previously published methods [201]. For this experiment, mice with end stage tumors are used. Blood is collected by cardiac puncture from the right atrium using a 25 G 5/8 syringe tip. Blood was put in heparin coated tubes (BD Microtainer tubes with lithium heparin REF 365965) and then 200 uL of blood was added into 800 uL of DMEM , 3.5 g/L D-glucose , 3.7 g/L NaHCO3, Ab/Am 15% FBS media in sterile Eppendorf tube to be centrifuged at 1000rpm for 5min. The supernatant is discarded, and the remaining is cultured in the media described above with regular maintenance for 10 days. On the 11^{th} day, equilibrate plate with 1X PBS for 1 minute, then fix with 10% formalin for 5 minutes. Wash away formalin using 1X PBS, and stain with hematoxylin. To wash away hematoxylin, use

1% acetic acid. Colonies were quantified using Image J software. GraphPad Prism was used to conduct the non-parametric t-test using Mann-Whitney methods.

RNA extraction from mammary glands, tumors, and blood cells was performed using the QIAamp RNA blood mini kit (Qiagen 52304). Quantitative RT-PCR was performed using the QuantiTect SYBR Green RT-PCR kit (Qiagen 204243). Primer sequences for the SV40poly-A transgene marker are SV40-F: GGAACCTTACTTCTGTGGTGT; SV40-R: GGAAAGTCCTTGGGGTCTTCT. Primer sequences are Actin-F: CATCATGCGTCTGGACCTG; Actin-R: CTCACGTTCAGCTGTGGTCA; Adamts1-F: GATAATGGACACGGGGAATG; Adamts1-R: GATAATGGACACGGGGAATG; Angpt2-F: GCCCAAGTACTAAACCAGACG; Angpt2-R: CACTGGTCTGATCCAAAATCTG ;Bmp4-F: CAATGGAGCCATTCCGTAGT; Bmp4-R: CATGATTCTTGGGAGCCAAT; Gapdh-F: TCATGACCACAGTGGATGCC; Gapdh-R: GGAGTTGCTGTTGAAGTCGC; Cyr61-F: ACGAGGACTGCAGCAAAACT; Cyr61-R: TGAGCTCTGCAGATCCCTTT; Lgals3-F: CAACGCAAACAGGATTGTTC; and Lgals3-R: CGTGTTACACACAATGACTCTCC Nupr1-F: CAATACCAACCGCCCTAGC; Nupr1-R: CCTTATCTCCAGCTCCGTCTC; P4ha1-F: ACCTGTGAAGTTCCCCAAGA; P4ha1-R: CAGTCATCTGACCAATTGACGTA; Plod2 Isoform 1-F: CCCCAAAGGGTGTGTTTATG; Plod2 Isoform 1-R: TTCAAAAATCTGCCAGAAGTCA; Plod2 Isoform 2-F: TCAAGGAAAGACACTCCGATCT; Plod2 Isoform 2-R: AACACACCCATATCTCTAGCATTG; and Vegfa-F: CAGGCTGCTGTAACGATGAA; Vegfa-R: GCATTCACATCTGCTGTGCT.

Tumor cells were obtained from viably frozen tumor tissue [202] and cultured for two passages prior to colonization assays. Adherent proliferating tumor cells were counted and 5.0 x 10^5 cells were injected retro-orbitally into MMTV-CRE wild-type control mice. Mice were

examined 35 days post injection and lungs were paraffin embedded prior to routine H&E staining.

Tumor transplant studies were conducted using viable frozen tumor samples generated during the study of tumor development and metastasis using the transgenic mice. Tumors (E2F WT, $E2F1^{-/-}$, or $E2F2^{-/-}$) were implanted into the mammary fat pad of MMTV-CRE wild-type control mice. Once the primary tumor reached endpoint, lung metastases and the percent area of the lungs occupied by metastasis was examined with routine histology and was quantified. GraphPad Prism was used to conduct the non-parametric t-test using Mann-Whitney methods.

IN VITRO ASSAYS

Tumor cells were obtained from viable frozen tumor tissue and cultured for use in a wound healing assay in the presence of 2ug/mL Mitomycin C using standard methods [203]. Photomicrographs were taken at 0 hour, 12 hour, 24 hour, 36hour, and 48 hours.

Tumor cells were obtained from viable frozen tumor tissue, cultured, and serum starved for 24 hours for use in a transwell invasion assay according to standard protocols [203]. Serum starved cells were re-suspended in serum free media with 2ug/mL Mitomycin C and seeded at a density of 3.0×10^5 cells on the insert. DMEM with 10% FBS and 2ug/mL Mitomycin C was used as a chemo attractant. Cells were allowed to migrate for 6 hours prior to 3% paraformaldehyde fixation and crystal violet staining.

CHAPTER 4:

IDENTIFYING THE MECHANISTIC FEATURES BY WHICH THE E2F1 TRANSCRIPTION FACTOR REGULATES BREAST CANCER METASTASIS.

ABSTRACT

In human breast cancer, the major cause of lethality is the metastasis of tumor cells to distant organs. Because of this, current research has revealed many of the molecular features that promote distant metastasis, yet we lack a complete understanding of the circuitry that regulates progression to metastasis. Previously, we used genomic signatures to predict that E2F transcription factors are involved in a mouse model of breast cancer metastasis. Testing these predictions, we genetically demonstrated that loss of E2F1 in MMTV PyMT tumors inhibits metastatic potential at multiple steps of metastatic advancement including a reduction in tumor angiogenesis, circulating tumor cells, as well as the ability of tumor cells in the bloodstream to colonize the lungs. In addition, using a tumor transplant approach we found that the metastatic defects associated with E2F1 loss occurred in mechanisms intrinsic to the tumor cells. Collectively, these data illustrate E2F1 as a key regulator of the gene expression changes required to progress tumor cells to distant organ metastasis. In more recent work, we identify the genes by which E2F1 coordinates tumor cell metastasis by using Affymetrix gene expression arrays to compare metastatic E2F WT MMTV-PyMT tumors with non-metastatic E2F1 KO tumors. These results revealed differential expression of known regulators of tumor metastasis. In addition, several potential E2F target genes that have not been tested for driving metastasis to the lungs were identified. Importantly, these genes had multiple E2F binding sites within their promoter region and high expression of these genes correspond to early human breast cancer metastasis events in a large clinically annotated dataset. As a result, we hypothesized these genes are also key effectors of E2F1's regulation of metastasis. To test this hypothesis, we utilized CRISPR to knockout expression of these genes in a MMTV-PyMT cell line. Using cell migration assays, we show that elimination of these E2F1 target genes stunt cell migration. More

importantly, we tested tumor cells in vivo via tail vein injection and find that knockout of these genes block tumor cell colonization of both the lungs and liver of recipient mice. The significance of this work is twofold: we reveal the genomic features by which E2F1 regulates metastasis and identify new functions for several E2F1 target genes by highlighting their contribution to the metastatic ability of tumor cells.

INTRODUCTION

Stage 4 breast cancer is exemplified by distant organ metastasis and corresponds to poor overall prognosis, with only 22% of patients surviving for at least five years [67]. Comparatively patients with less advanced stages carry a considerably more favorable prognosis. This illustrates that metastasis to distant organs as a major cause of breast cancer fatality. As a result, there is great deal of interest in understanding the process of metastasis so that strategies to limit metastatic potential and treat tumors that have formed in distant organs can be realized.

In previous work, we uncovered a new function for the E2F transcription factors [204] . Using genomic signature methods, we predicted a role for the activator E2F transcription factors in the MMTV-PyMT mouse model of metastatic breast cancer. Genetically testing the hypothesis that the E2Fs function to regulate metastasis, we interbred MMTV-PyMT mice with the knockouts of E2F1, E2F2 and E2F3 and noted that the ablation of individual E2Fs were associated with alteration of tumor latency, histology, and vasculature. More importantly, we found that tumors from the E2F1 and E2F2 knockout backgrounds had severely diminished metastatic capacity. For example, we detected that tumor bearing E2F1 and E2F knockout mice had a significant reduction metastasis to the lungs and this was accompanied by a significant decrease in circulating tumor cells. Bypassing the invasion defects, we tested the ability E2F knockout tumor cells to colonize the lungs by injecting tumor cells into the bloodstream. This

experiment showed that E2F1 and E2F2 were critical for tumor cells to colonize the lungs. This also suggested that the E2F transcription factors regulation of metastasis was associated with mechanisms intrinsic to the tumor cells and not the tumor microenvironment. We confirmed this notion using a tumor transplant strategy; finding that E2F1 and E2F2 knockout tumors failed to metastasize in wild type recipient mice. This led to the hypothesis that the E2F transcription factors regulate gene expression programs within tumor cells that are critical throughout metastatic progression. Using qRT-PCR to investigate E2F target genes that have been demonstrated to mediate metastasis, we found that the E2Fs control expression of genes critical to angiogenesis, remodeling of the extracellular matrix, tumor cell survival and tumor cell interactions with vascular endothelial cells to facilitate metastasis to the lungs.

While have uncovered some of the molecular aspects associated with the E2Fs regulation of metastasis, our previous study did not incorporate analysis of global analysis of transcriptional alterations associated with E2F loss. Given that these transcription factors have been demonstrated to bind thousands of individual target genes [205], we sought to characterize the gene expression profiles of MMTV-PyMT tumors using microarray technology. We hypothesized that this approach would allow us to identify the genes that are controlled by E2F1 that contribute to metastatic ability. In testing this hypothesis, we had two primary goals. One was to identify genes that are differentially regulated with E2F loss that correspond to human breast cancer metastasis events and have been demonstrated to regulate metastasis. The second was to test altered target genes that have not yet been demonstrated to regulate breast cancer metastasis. In support of our hypothesis, we reveal the E2F1 target genes that are associated with E2F1's metastasic function and identify new functions for several E2F1 target genes by showing their contribution to the metastatic ability of tumor cells. This chapter reflects work currently in

progress and data obtained at the time of writing the thesis. Additional controls and experiments will be done to ensure the accuracy of the conclusions presented here. For example, as a control for non-specific effects associated with Crispr, lentiviral expression contructs have been assembled to reintroduce Fgf 13 and Adm into their respective knockout cells. Characterization of these cells for rescue phenotypes will confirm if this genes truly block metastasis.

RESULTS

GENOMIC COMPARISON OF E2F $^{WT/WT}$ TUMORS AND E2F $^{-/-}$ TUMORS

To determine the global gene expression response to E2F loss, we analyzed MMTV-PyMT from E2F $^{WT/WT}$, E2F1 $^{-/-}$, E2F2 $^{-/-}$, and E2F3 $^{+/-}$ backgrounds on Affymetrix microarrays. Using unsupervised hierarchical clustering, we investigated the gene expression relationships amongst the various tumors (FIGURE 4.1A). Interestingly, tumors with squamous histology showed a distinct gene expression profile, separating into their own cluster. Co-clustering amongst the E2F2 $^{-/-}$, E2F3 $^{+/-}$, and a subset of the E2F $^{WT/WT}$ tumors was observed. However, tumors with a E2F1 $^{-/-}$ genotype tended to cluster with one another separating from a majority of the E2F $^{WT/WT}$ tumors; indicating prominent gene expression differences between E2F1 $^{-/-}$ and E2F $^{WT/WT}$ tumors.

To test if these differences corresponded to activation major cell signaling pathways, we utilized gene signatures to predict pathway activation across the mouse mammary tumors. Unsupervised hierarchical clustering again separated out squamous tumors and led to co-clustering amongst E2F2 $^{-/-}$, E2F3 $^{+/-}$, and a subset of the E2F $^{WT/WT}$ tumors (FIGURE 4.1B). In this analysis more E2F $^{WT/WT}$ tumors clustered with E2F1 $^{-/-}$ tumors. Focusing in on the cluster where the majority of E2F1 $^{-/-}$ tumors were found indicates that E2F1 $^{-/-}$ tend to have high activity of E2F4, E2F5, and p53 pathways and low activity of Src, p110, EGFR, Ras, RhoA, beta-

catenin, and Tgfb signaling pathways.

To test for the genes that were significantly differentially regulated between E2F1 $^{-/-}$ and E2F $^{WT/WT}$ tumors we used significance analysis of microarrays or SAM [28]. Using SAM, we identified the statistically significant gene expression changes between E2F1 $^{-/-}$ and E2F $^{WT/WT}$ tumors (FIGURE 4.1C). Since we hypothesized that E2F1's role in regulating metastasis was by transcriptional activation of target genes, we were particularly interested in the 226 genes that were significantly downregulated in the E2F1 $^{-/-}$ tumors . To begin characterizing these genes for metastatic potential, we used Kaplan Meier analysis of clinically and intrinsically annotated human breast cancer gene expression data [86]. To identify E2F1 target genes, we used ChIPBase [206]. This database contains results from a variety of ChIP-Seq experiments including those for E2F1 [205]. Out of the 226 differentially regulated genes, 98 were E2F1 targets.

To characterize E2F1's role in regulating metastasis by transcriptional activation of target genes, we focused on genes where high expression in tumors correlated with a decreased time to distant metastasis across all breast cancers as well as within with individual intrinsic subtypes of tumors. We found that 94/226 were associated with earlier metastasis in at least one subtype of breast cancer, although many genes were not uniform when comparing predictions from one subtype of breast cancer to the next. Eliminating any conflicting predictions left 55 genes with pro-metastatic predictions (TABLE 4.1); 34 of which had demonstrated E2F1 binding E2F1 [205]. By using a fisher's exact test, we saw that the distribution of direct E2F1 targets was significantly higher in the genes concordant human breast cancer predictions ($p=0.001$) than genes either discordant or not predictive of human breast cancer metastasis. This illustrates that the E2F1 target genes altered in these tumors are associated with human breast cancer metastatic

potential.

Investigating these 55 genes for possible function, we tested for overlap with gene sets on the molecular signatures database from the broad institute. The top scoring overlap corresponded to gene sets for hypoxia with 16 direct target genes and 22 total genes being upregulated in response to hypoxia (p- value =6.01 e $^{-25}$, FDR q-value = 2.84 e $^{-21}$, marked in TABLE 4.1). In addition, we found a significant association with glycolysis (p-value=9.03 e $^{-12}$, FDR q-value = 8.54 e $^{-15}$). Importantly, all eight of the glycolysis genes were also associated with hypoxia response. We also tested whether hypoxia response was significantly associated the 55 genes with pro-metastatic associations as opposed to the genes that have either no or discordant metastatic predictions. This revealed hypoxia response was significantly ($p<0.0001$) associated with the gene expression changes in E2F1 $^{-/-}$ tumors that are concordantly predictive of human breast cancer metastasis. This data illustrates that E2F1 $^{-/-}$ tumors have a defect in their ability to activate genes of the hypoxia response gene expression programs.

We also detected significant overlap with other gene sets as well. Importantly, amongst the top scoring gene sets were those for cell signaling pathways that we had predicted low activity for in the E2F1 $^{-/-}$ tumors. For example, 13 genes (9 direct targets) corresponded to gene that upregulated in response to Tgfb1 stimulation (TABLE 4.1, p-value = 6.18 e $^{-16}$, FDR q-value= 7.3 e $^{-13}$). We also identified 13 genes that correspond to the RhoA signaling pathway(TABLE 4.1, p-value = 1.23 e $^{-12}$, FDR q-value = 5.83 e $^{-10}$). In agreement with lower EGFR activity, E2F1 $^{-/-}$ tumors have reduced expression of 13 genes normally upregulated by Egfr signaling. Importantly, among these genes were Egfr ligands amphiregulin (AREG) and heparin-binding EGF-like growth factor (HBEGF) [207]. To further investigate how our fold-change data corresponds with the pathways that showed low activity in E2F1 $^{-/-}$ tumors , we

mapped the altered genes and pathways to an interaction network including Rb1 and E2F1 using string-db[208]. Out of the 55 genes that regulate metastasis, 34 genes mapped to the interaction network with Rb1, E2F1, Src, p110, EGFR, Ras, RhoA, and Tgfb pathways (FIGURE 4.1D. Meanwhile, many of the genes associated with hypoxia and glycolysis appeared in a node mostly associated with an Rb1, Tgfb, Egfr, p110, beta-catenin and Vegfa node. Importantly, EGFR, CTNNB1, TGFB1 also appear in hypoxia gene sets on MSigDB. Collectively, this data shows which pathways the pro-metastatic genes altered by E2F1 loss are coordinated with.

To further detail the metastatic defects in E2F1 $^{-/-}$ tumors, we cross referenced the 55 genes associated with human breast cancer metastasis with the literature to identify which of the molecular changes have already been demonstrated to regulate breast cancer metastasis *in vivo*. As depicted in TABLE 4.2, Vegfa [209], Hbegf [210], Hspb1[211], Flt1[212], L1cam [213], and Plaur [214] have all previously been shown to regulate breast cancer metastasis *in vivo.* As summarized in TABLE 4.2, we see that a number of the cell signaling pathways with low activity in E2F1 $^{-/-}$ tumors, the Tgfb [215], Src [216], beta-catenin [142], RhoA [217] and Egfr [218] pathways have been shown to function in breast cancer metastasis *in vivo.* Additionally, there were genes that had been shown to have pro-metastatic features *in vitro* such as Areg [219], Tead1 [220], Coro1C [221], Lama5 [222], Tgm2 [223], and Fgf 7 [224] Taken together, this shows that E2F1$^{-/-}$ tumors have low expression of genes and low activation of pathways demonstrated to promote breast cancer metastasis.

TESTING ADDITIONAL GENES FOR METASTATIC FUNCTION

To identify candidate genes for further testing and possibly reveal new metastatic regulators, we focused in on genes that had not been demonstrated to regulate breast cancer metastasis in vivo, that were E2F target genes, and properly correlated with a decreased time to

distant metastasis across several human breast cancer human breast cancer subtypes. With this approach, we noted that high expression of adrenomedullin (Adm) was associated with earlier onset of distant metastasis across all (FIGURE 4.1E, HR=1.88 (1.53-2.3), Log rank P= 5.6 e -10), basal, luminal A and Luminal B subtypes of breast cancer. Importantly, this E2F1 target gene was part of the hypoxia response network and expressed 2.2 times lower in E2F1 -/- tumors (q-value =.007). In addition to being shown to be bound E2F1 in a chip-seq experiment, analysis of the 500 bp sequence upstream of the transcriptional start site (TSS), we found six E2F binding motifs in human and one in mouse for this gene. A second E2F1 target gene with a compelling association with human breast cancer metastasis was fibroblast growth factor 13 (Fgf 13). This gene was associated with earlier onset of distant metastasis across all (FIGURE 4.1F, HR=1.59 (1.28-1.97), Log rank P= 2 e -05), basal, Luminal B, and Her-2 positive subtypes of breast cancer. Also demonstrated to be bound by E2F1 in a chip-seq experiment, analysis of the 500 bp sequence upstream of the transcriptional start site (TSS), we found seven E2F binding motifs in human and ten in the mouse sequence for this gene.

To test these genes for metastatic behavior, we utilized a PyMT-derived cell lines (PyMT 419 cells) [225] and a CRISPR (clustered regularly interspaced short palindromic repeats) approach to create Adm and Fgf 13 knockout cells. FIGURE 4.2A shows an example of sequence trace for Adm $^{WT/WT}$ cells and for Adm $^{-/-}$ cells . In total we generated four Adm $^{-/-}$ clones and the sequence alignment is shown in FIGURE 4.2B. FIGURE 4.2C shows an example of sequence trace for FGF13 $^{WT/WT}$ cells and for FGF13 $^{-/-}$ cells and FIGURE 4.2D shows the sequence alignment for the two FGF13 $^{-/-}$ clones that were identified.

To characterize the impact of Adm and Fgf13 loss, we began with *in vitro* assays. We started by seeding 100,000 cells and performed cell counts over 3 days to study the doubling

times of the Adm $^{-/-}$ clones (FIGURE 4.3A). As a trend, cell counts were variable amongst the clones, with two clones closely resembling the control and two having fewer cells at day 4. Using a similar approach, we studied the Fgf 13 $^{-/-}$ clones and found that Fgf13 loss led to reduced cell counts over the 3 day period (FIGURE 4.3B).

An important feature to metastasis is cell migration, as such we measured the migratory ability of Adm $^{-/-}$ and Fgf 13 $^{-/-}$ clones. Monitoring the control cells, the wound was fully closed by 18 hours following the scratch. In Adm $^{-/-}$ clones, cell migration was variable amongst the clones, two clones were able to close or mostly close the wounds by 18 hours and two clones showed a prominent decrease in wound closure (FIGURE 4.3C,D). For Fgf 13-/- clones, we observed a significant defect in cell migration, with both clones failing close the wound by 18 hours (FIGURE 4.3 E,F). This data suggests that Adm loss does not impact migration mechanisms. Meanwhile, this data supports the possibility that Fgf13 loss may impact mechanisms associated with cell migration.

To test the clones for metastatic capability in vivo, we utilized a tail vein injection strategy. For this experiment, we injected 50,000 cells monitored metastasis after 21 days following injection. In mice receiving control cells, metastasis was observed predominately at the lungs (FIGURE 4.4A) and liver (FIGURE 4.4B). In addition, metastatic tumors were found sporadically throughout the mouse at other sites (data not shown). Results for the Adm KO clones revealed that loss of Adm severely limited the ability of the tumor cells to form metastases at the lung as depicted by representative photos in FIGURE 4.4A, with only small micro metastasis (FIGURE 4.4A blue arrowheads, inset) observed in two mice from two separate clones. No metastases were observed in the liver (FIGURE 4.4B) or other sites (not shown) for the mice receiving Adm KO cells. Similarly for mice receiving Fgf 13 KO clones, loss of Fgf 13

dramatically limited the ability of the tumor cells to form metastases at the lung as illustrated by representative sections in FIGURE 4.4A, again with only small micro metastasis (FIGURE 4.4A blue arrowheads, inset) observed in three mice receiving the 3H1 clone. No metastases were observed in the liver as shown by representative examples in FIGURE 4.4B or other sites (not shown) for the mice receiving Fgf 13 KO cells. Quantifying metastatic outcomes for Adm KO and Fgf 13 KO clones showed that significantly fewer mice presented metastasis at the lungs (FIGURE 4.4C) or liver (FIGURE 4.4D). Similarly, looking at the number of metastases, mice receiving Adm KO and Fgf 13 KO clones presented a dramatic reduction in the number of lesions present in the lungs (FIGURE 4.4E) and liver (FIGURE 4.4F). Together, these results show that Adm and Fgf 13 are key participants in E2F1's regulation of metastasis and reveal new roles for these genes.

INVESTIGATING ADM FUNCTION

To begin investigating which parts of the metastatic circuitry these genes are part of, we used a weighted correlation network analysis also known as WGCNA [226] for Adm and Fgf13 in MMTV-PyMT E2F $^{WT/WT}$ tumors. The rationale for this approach is that genes with expression patterns that go up or down together across a large collection of samples are likely coordinated in expression due to participating in similar molecular response networks to control key cellular functions. As a result, WGCNA will allow us to identify the molecular response networks Adm and Fgf 13 are part of and relate these networks to the cellular functions they are associated with. WGCNA for Adm identified 114 probes with gene significance score higher than 0 .6 (p-value < 0.008). After removing duplicate gene results, we had 81 genes that are part of the Adm covariance network.

Testing for significant association with functional gene sets on the broad institute's MSigDB, we detected significant overlap with a number of gene sets for hypoxia, including the hallmark hypoxia gene set (p-value =2.34 e $^{-50}$, FDR q-value 1.12 e $^{-46}$). In addition, we detected a significant association with glycolysis (p-value = 4.04 e $^{-29}$, FDR q-value 1.17 e $^{-48}$). Association with functional gene sets is summarized in FIGURE 4.5A and illustrates the Adm network is strongly associated with hypoxia and hypoxia related glycolysis gene expression programs.

To identify if the Adm network was associated with any of the pathways altered in E2F1$^{-/-}$ tumors, we first used string-db to create an interaction network for the altered pathways and the Adm network (FIGURE 4.5B). Importantly, 50 of the Adm network genes mapped to these pathways. Most of the hypoxia response genes sat in a node shared by PI3K, Egfr, Src, Ras, Tgfb, and beta-catenin. While the glycolysis genes were in this same network, they were more removed from these nodes. To further investigate the relationship between the Adm network and these pathways, we tested whether any of these genes were upregulated by these pathways by combining gene sets for each pathway using data from MSigDB. In doing this we found that the Adm network was not closely related to gene expression responses from the Src (only one gene overlap), beta-catenin (only one gene overlap), or PI3K pathways (only two gene overlap). Instead, the Adm network maybe more tightly associated with Egfr (16 gene overlap),Ras (5 gene overlap), RhoA (16 gene overlap), and Tgfb (12 gene overlap)signaling pathways. As illustrated by FIGURE 4.5C, there was a large degree overlap between which of these pathways the genes corresponded to. In addition, there was a large degree of overlap between ADM network genes that corresponded to RhoA, Egfr, and TGFB pathways and the ADM network genes that corresponded to hypoxia and glycolysis (FIGURE 4.5 D). Testing these genes as a

signature for association with human breast cancer metastasis events, we found that high expression of these genes were significantly associated with a shorter time to distant metastasis across all breast cancers (FIGURE 4.5E, HR=1.49 (1.22-1.83) Log rank P= 9.4 e -05) as well as in basal, luminal A, and luminal B breast cancer. Altogether this data suggests that Adm is part of the pro-metastatic hypoxia response and is possibly associated with Tgfb, RhoA, EGFR, and Ras activation upstream of E2F1.

INVESTIGATING FGF13 FUNCTION

As a result of investigating Fgf13's network with WGCNA, we identified 38 probes corresponding to 35 genes that tightly correlate with Fgf 13 expression (gene significance threshold =0 .6, p-value <0.0005). Testing for significant association with functional gene sets on the broad institute's MSigDB, we didn't detect significant overlap with hypoxia pathways for the Fgf 13 network. Further, manually testing for overlap combined gene sets for hypoxia, angiogenesis, and glycolysis showed only two genes overlapping with these gene sets, indicating Fgf13 operates in a separate mechanism. Instead and in agreement with our scratch assay results, testing for overlap on MSigDB revealed a significant association with the Rac1 cell motility pathway (p-value = 5.38 e^{-7}, FDR q-value = 1.27 e^{-3}). Key to this association, was strong covariance with the Rac1 g-protein, the GTPase activating protein chimerin 1 (Chn1), and Wasf1 (which acts downstream of Rac1 to regulate the cytoskeleton). In support of this association, FGF13 was part of the Kegg pathway for regulation of the actin cytoskeleton. Testing these alterations for the presence in an interaction network with Rb-E2F1 and the pathways altered in E2F1 $^{-/-}$ tumors, illustrated a relationship between these pathways (FIGURE 4.6A). Importantly, this also included the expected interaction with RhoA and Rac1. Importantly, testing these genes as a signature for association with human breast cancer metastasis events, we found that high

expression of these genes were significantly associated with a shorter time to distant metastasis across all breast cancers (FIGURE 4.6B) as well as in basal, luminal A, luminal B, and Her-2+ breast cancer. Together, this data suggests that Fgf 13 functions in metastasis mechanisms associated with cell movement and cytoskeleton control.

DISCUSSION

By using a gene expression microarrays, we expanded on previous work where we identified that the E2Fs regulate breast cancer metastasis [204]. Here we used bioinformatic analysis to compare global gene expression differences between E2F $^{WT/WT}$ MMTV-PyMT tumors and E2F1 $^{-/-}$ MMTV PyMT tumors. Leveraging clinical annotations and gene expression data from tumors from breast cancer patients, we identify which of the genes downregulated with loss of E2F1 are properly correlated with a shorter time until distant metastasis. By focusing in on genes with no opposing associations across intrinsic subtype, we identified 55 genes correlating with a faster progression to metastatic disease. Importantly, we also identified that loss of E2F1 led to decreased activity in several key signaling pathways previously demonstrated to regulate metastasis. Perhaps most important of our findings, was that the genes correlating with a faster progression to metastatic disease were significantly associated with the gene expression response to hypoxia, proving a context for why E2F1 -/- tumors fail to metastasize. Also of note, we demonstrate two E2F1 target genes that were downregulated in E2F1 $^{-/-}$ tumors as new molecular participants in breast cancer metastasis.

In comparing global gene expression differences between E2F $^{WT/WT}$ MMTV-PyMT tumors and E2F1 $^{-/-}$ MMTV PyMT tumors, we identified which genes were downregulated in response to E2F1 loss and separated genes into categories: those that correspond to human breast cancer metastasis events with no opposing predictions and those that do not. Here we found that

a majority of the genes that were direct targets were distributed in the category of genes that correctly predict human breast cancer metastasis. This illustrates the gene expression changes associated with metastasis are due to E2F1's direct regulation of target genes as opposed to indirect effects.

Focusing in on the altered pro-metastatic genes in the literature, we identified a number of genes that been previously tested *in vitro* and *in vivo* that can explain the metastatic defects we observed in E2F1 $^{-/-}$ tumors. The *in vitro* studies showed that Areg [219] , Tead1 [220] , Coro1C [221], Lama5 [222], Tgm2 [223] and Fgf 7 [224] are involved in cell migration and invasion features of tumor cells. The reduced expression of these genes involved in invasion phenotypes may provide a mechanistic relationship to our previous finding that E2F1 $^{-/-}$ tumors had possible invasion/intravasation problems indicated by a reduction in circulating tumor cells [204].

The gene expression results also further explain the stunted and reduced tumor vasculature we observed in E2F1 $^{-/-}$ tumors. Indeed, in our previous report we observed a 5-fold reduction in Vegfa levels using qRT-PCR which provided a preliminary look into the molecular basis for the angiogenesis defects. Indeed, microarray analysis also showed reduced Vegfa expression in E2F1 -/- tumors. However, we also now see these defects were compounded by reduced expression of other E2F1 target genes associated with angiogenesis. This includes adrenomedullin and Flt1. Flt1 is a vascular endothelial cell growth factor receptor that binds Vegfa and placental growth factor and mediates endothelial cell migration [227]. While our report focused on adrenomedullin's regulation of breast cancer metastasis to the lungs and liver, previous studies have described adrenomedullin's angiogenic roles [228-230]. As a result our gene expression analysis identified additional genes associated with the angiogenesis defects we observed in E2F1 -/- tumors. Furthermore, the pro-metastatic gene expression changes depict a

much larger picture illustrating a defect in the hypoxia response pathway with E2F1 loss.

By testing the 55 genes that correlate with human breast cancer metastasis for function, we detected a significant association with hypoxia response gene sets. Importantly, the majority of the genes associated with hypoxia were also direct E2F1 target genes. Hypoxia has been described as a master regulator of metastasis due to the result of gene expression changes brought about by hypoxia response[231]. These gene expression changes enable tumor cells to progress through a variety of rate limiting steps in metastasis. This includes promoting the tumor angiogenesis, epithelial to mesenchymal transition, tumor cell invasion, remodeling of the extra cellular matrix, and increasing cell migration [231-234]. In addition, hypoxia also activates genes associated with facilitating tumor cell intravasation, survival in the blood stream, extravasation and colonization at distant organs [231]. Consistent with processes associated with hypoxia response, we observed angiogenesis defects, a decrease in circulating tumor cells suggesting intravasation defects, and inability of E2F1 -/- tumor cells to colonize the lungs even when injected into the bloodstream. Altogether, this lends to the possibility that these metastatic defects could be attributed to the lack of expression of key hypoxia response genes that facilitate metastatic progression.

In support of this, all but one (hspb1) of the genes that had all already been shown to regulate metastasis *in vivo* corresponded to MSigDB gene sets for the hypoxia gene expression response. Looking at these genes in more detail paints a picture as to how the inability to properly respond to hypoxia has led to an inability of the E2F1 -/- tumors to metastasize. E2F1 $^{-/-}$ tumors showed a significant reduction in Vegfa. While this clearly played a role in tumor angiogenesis, Vegfa has other pro-metastatic functions such as increasing vascular permeability to enable intravasation and extravasation that also factor in to metastatic ability[235]. Another

hypoxia response gene altered was Flt1. Importantly, treating mice with a Flt-1 inhibiting peptide decreased metastasis of MDA-MB-231 cells in a xenograft study [212]. In addition to blocking metastasis, inhibition of Flt 1 also resulted in decreased cell migration and invasion through matrigel. Importantly, hypoxia can promote invasion through the extracellular matrix by upregulating proteases [231]. E2F1-/- had reduced expression of the plasminogen activator, urokinase receptor gene which is also known as UPAR or Plaur. Plaur converts the zymogen plasminogen to plasmin [236]. Plasmin is a wide-reaching protease and promotes remodeling of the extracellular matrix. Importantly, a study of overexpressing this Plaur in breast cancer cells from rat displayed increased metastasis to the lungs following orthotopic injection and increased invasion through matrigel [214]. .Another hypoxic response gene with reduced expression was L1Cam. Importantly this gene was shown to have increased expression during hypoxia and function in metastasis to the lungs using a tail vein injection [213]. Further, this work showed that L1Cam mediates metastasis to the lungs in part by facilitating adherence to endothelial cells, possibly providing further context for why E2F1 -/- cells fail to colonize the lungs when injected into bloodstream. Taken together, these studies highlight how the hypoxia responsive E2F1 target genes contribute to metastasis and may explain the invasion and colonization deficits of $E2F1^{-/-}$ tumors.

Importantly, we did not detect any fold change in for the hypoxia inducible factors with E2f1 loss. And yet, we detect low expression of hypoxia related genes with loss of E2F1, while this can be explained by the fact that they are E2F1 target genes. However, what is still unclear is what upstream pathways couple E2F1 to the hypoxia response. Importantly, we observed reduced activity of cell signaling pathways associated with breast cancer metastasis as summarized in TABLE 2.2. By mapping the metastatic gene expression changes to these

pathways using gene sets from MSigDB and interaction networks, we were able to begin piecing together the gene expression changes associated with hypoxia and metastasis with what pathways they are possibly associated with. While few of the metastasis associated genes were mapped to the Ras, Src, PI3K, and beta-catenin pathways. We did find a strong overlap between the hypoxia response/metastasis genes with gene sets for genes upregulated by the Tgfb, RhoA, and Egfr pathways. While we did not detect alteration of known activators for the Tgfb and RhoA pathways, it may be possible that E2F1 acts downstream of these pathways carrying out the transcriptional consequences of pathway activation during hypoxia. This possibility is supported by the fact that both pathways are activated by hypoxia[231, 237, 238] and a number of the hypoxia related genes that mapped to these pathways were also E2F1 target genes.

Another potential model for directing E2F1 to hypoxia response involves the Egfr pathway, a pathway critical to invasion and metastasis[218]. As shown in FIGURE 1B, predicted that E2F1 $^{-/-}$ tumor have low Egfr activity. Illustrating Egfr signaling as a key portion of hypoxia response, it has been shown that Egfr is upregulated in response to hypoxia [239] Related to the low Egfr activity in E2F1 $^{-/-}$ tumors, was the reduced expression of Hbegf and Areg, both Egfr ligands [207]. Hbegf mapped to hypoxia response using MSigDB, and while Areg did not, literature supports this gene as part of the hypoxia gene expression response [240]. Further, Hbegf was an E2F1 target gene determined by a chip-seq experiment [205] and Transfac analysis of Areg shows E2F1 binding sites in both mouse and human. This could suggest that E2F1 loss contributes to low Egfr activity from an inability to initiate high transcription of these Egfr ligands during hypoxia. In addition, the number of genes that mapped to gene sets (shown in TABLE 4.1) for by induced by Egfr signaling like Adm and Plaur could also therefore be impacted by this; with E2F1 target genes being impacted in a feedback loop. Furthermore, this

mechanism may be acting upstream of RhoA, as Egfr has been shown to regulate RhoA [241]. Thus providing an explanation for the large degree of overlap for genes mapping to Egfr and Rho A pathways as shown in TABLE 4.1. Taken together, these results suggest one possible molecular mechanism for reduced metastatic capacity with loss of E2F1.

In addition to this genomic characterization, we tested two E2F1 target genes that at the time of the study had not yet been demonstrated to function in metastasis. One gene we tested was Adm, a secreted protein that was part of the hypoxic response with published roles in tumor angiogenesis [228-230, 242]. Importantly, we found that that knockout of Adm significantly reduced metastasis to the lungs and liver in a tail vein injection; illustrating its role in tumor metastasis. In support of this, a recent publication has detailed that Adm also regulates breast cancer metastasis to the bone [243] As a result, our study and the aforementioned one complement each other to depict Adm as a key factor for enabling metastasis to multiple metastatic sites. In addition, we show that Adm is relevant to multiple subtypes of human breast cancer, with an association to basal, luminal A, and luminal B breast cancer.

By investigating Adm using WGCNA, we observed a number of genes related to hypoxia gene expression response with strong overlaps with the Egfr, RhoA, and Tgfb pathways. However, adrenomedullin on its own mapped to the Egfr and RhoA pathways, possibly implicating the potential Egfr model described above. While it is clear that Adm is part of the hypoxia response, what is not completely clear is how Adm functions in metastasis to the lungs and liver. One possibility is that, similar to hypoxia response genes, that it acts on these distant organs to facilitate changes to promote a suitable microenvironment. Never the less, these results add another dimension to E2F1's regulation of metastasis as part of the hypoxia response by demonstrating that Adm is required colonization of the lung and liver.

We also demonstrated new metastatic roles for Fgf 13. Importantly, knockout of Fgf13 reduced cell migration and metastasis to the lungs and most prominently the liver. FGF13 is a nonsecretory protein of the FGF family [244]. Predicting Fgf13 molecular function using WGCNA depicted its participation in the chimerin, Wasf1, and Rac1 pathway. Importantly, this pathway is described for its role in cell motility and shows an interaction with another cell migration participant RhoA [245]. In agreement with Fgf13 being a key participant in the cell migration process, both Fgf13 knockout clones exhibited impaired cell migration in a scratch assay. Furthermore, deletion of FGF13 in mice resulted in neuronal migration defects, through stabilization of microtubules [244]. In agreement with this function, we found Mtap1B (microtubule-associated protein 1B) and Tubb6 (tubulin, beta 6) as part of the Fgf 13 network. Importantly, Map1B has also been shown to increase microtubule stability [246]. In light of the fact that microtubule stabilization activates Rac1 [247], one possible mechanism of how Fgf13 can control cell migration becomes apparent: Fgf13's stabilization of microtubules leads to activation of Rac1, Rac1 activation leads to the formation of lamellipodia at the leading edge of migrating cells to drive cell movement [248]. Importantly, using the that genes tightly correlate with Fgf13 expression as a signature depicted the Fgf13 network a general pro-metastasis mechanism as significant metastatic associations were detected in each intrinsic subtype; which is fitting for a network that regulates control of cell motility. Together, our data illustrates new functions Fgf13 in breast cancer metastasis to the lungs and liver, possibly through regulation of cell migration mechanisms.

As a whole this study shows that the metastatic defects associated with E2F1 loss are largely associated with an inability to properly initiate a gene expression response to hypoxia. Importantly, many of the genes downregulated in E2F1 $^{-/-}$ tumors associated with hypoxia

response have been shown as regulators breast cancer metastasis. In addition, we illustrate new roles for the E2F1 target gene Adm in regulating metastasis to the lungs and liver; furthering our understanding of the hypoxia response to breast cancer metastasis. In addition, we identified that the E2F1 target gene Fgf 13 controls metastasis to the lungs and liver, potentially through a cell migration mechanism; possibly providing a means for E2F1 to regulate cell migration. Collectively, this study furthers our initial characterization of E2F1s regulation of metastasis by identifying the metastasis associated gene expression response to E2F1 loss.

METHODS

RNA AND MICROARRAY

Preparation of RNA samples from flash frozen tumors was done using the Qiagen RNeasy kit after roto-stator homogenization. RNA from 17 Myc induced tumors was submitted to the Michigan State University Genomics Core facility for gene expression analysis using Mouse 430A 2.0 Affymetrix arrays.

GENE EXPRESSION ANALYSIS

Raw intensity .CEL files were processed and RMA normalized using Affymetrix Expression Console. Unsupervised hierarchical clustering was done using Cluster 3.0 and exported using Java Tree View. Pathway activation was predicted according to previous studies [37, 47, 249]. Significance analysis of microarrays [28] was used to compare E2F $^{WT/WT}$ and E2F1 $^{-/-}$ tumors in a fold change analysis. Direct E2F1 target genes were identified using ChIP-base [206] and data from a previous ChIP-seq experiment [205]. Kaplan-Meier plots were generated using Correlation with human breast cancer. Significant overlaps with previously established gene sets were detected using the molecular signatures database (MSigDB) http://www.broadinstitute.org/gsea/msigdb/annotate.jsp. Gene annotations for Hypoxia response,

angiogenesis, and cytoskeleton as well as being responsive to Egfr, Tgfb, Src, Ras, RhoA, and PI3K pathways were generated by combing gene sets from MSigDB for each term, using only upregulated genes. For example, by combining all of the genes from gene sets that identify the genes that are upregulated during hypoxia. Interaction networks were assembled using www.string-db.org [208]. Weighted correlation network analysis was implemented according to published protocols [226] and using a gene significance score threshold of 0.6 to select genes for further analysis.

CELL CULTURE

A MMTV-PyMT derived cell line, PyMT 419 cells, were used for in vitro and in vivo experiments and have been previously characterized [225]. All cells were cultured in Dulbecco's Modified Eagle's Medium, 3.7 g/L of NaHCO3, 3.5 g/L d-glucose, 5ug/mL insulin, 1ug/mL hydrocortisone, 5ng/mL Egf, 35ug/mL BPE, 50ug/mL gentamicin, 1X Antibiotic/Antimycotic, and 10% fetal bovine serum. Media was set to a pH of 7.4.

CRISPR

Sequence for Fgf13 and Adm was obtained from the UCSC genome browser [250]. Guide sequences for each gene were done by submitting exon (using only those that were common across isoforms for each gene) sequence using the CRISPR design tool at: http://crispr.mit.edu/. Oligos for guide sequence assembly were designed by adding a ‘G’ followed by ‘CACC’ at the 5’ end of the guide sequence. For the complementary DNA to the guide, add ‘CAAA’ to the 5’ end. Oligonucleotide sequences are as follows:

ADM 5’: CACCGGATAAGTGGGCGCTAAGTCGT

ADM3’: AAACACGACTTAGCGCCCACTTATCC

Fgf 13 5’: CACCGTCAGCAGCAATCCGGCCGA

Fgf 13 3': AAACTCGGCCGGATTGCTGCTGACC

Oligonucleotides for guide sequence assembly were ordered from integrated DNA technologies https://www.idtdna.com/site. Oligos were diluted to a concentration of 100uM. To anneal the oligonucleotides 5 uL of the forward and 5uL of the reverse oligo are incubated in 10uL of 2X annealing buffer (10 mM Tris, pH 7.5–8.0, 50 mM NaCl, 1 mM EDTA) at 95 degrees Celsius for 4 minutes. The annealed oligonucleotides were inserted into the PX458 vector from Addgene http://www.addgene.org/48138/.This vector is ampicillin resistant and contains a selectable marker for EGFP.

To digest the vector, 2.5 ug of the vector was incubated with 10X BBSI reaction buffer (NEBuffer 2.1), 5ul of the BBS1 restriction enzyme, and adding enough ddH2O for a total of a 25uL volume and incubated for 1 hour at 37 degrees Celsius. After one hour the digested vector was treated with 2μL of calf intestinal alkaline phosphatase (CIP). After 30 additional minutes, another 1 uL of CIP was added and allowed to incubate for an additional 30 minutes.

To phosphorylate the guide sequence DNA, 4 uL of the annealed oligonucleotides were added to 2 uL of PNK (polynucleotide kinase) buffer, 2uL of T4 PNK ((polynucleotide kinase), and 12 uL of ddH20 for a 30 minute 37 degree Celsius incubation. After 30 minutes, 80 uL of ddH20 was added to the phosphorylated nucleotides.

The digested vector was gel purified by gel electrophoresis on 0.7% agarose gel to separate the cut vector (the single band that appears at approximately 10kB). The digested vector was purified using the Qiagen QIAquick Gel Extraction Kit .

To ligate the vector and guide sequence DNA, we incubated 1uL of the gel purified vector with 2uL of the phosphorylated guide sequence, 5uL of T7 ligase buffer, 1uL T7 ligase, and 5uL of ddH20. For negative control and control for self-annealing, we set up a similar

incubation without the 2uL of the phosphorylated guide sequence and increased the ddH20 to 7uL. Incubation took place at room temperature for at least 1hour.

To clone the vector, we used heat shock delivery of the vector into competent e. coli. To each ligation tube (as described in the paragraph above), 200 uL of competent cells were added and incubated on ice for 30 minutes. To heat shock, tubes were transferred to 42 degrees Celsius for exactly 90 seconds. Following 90 seconds, tubes were brought back to ice for 2 minutes. After 2 minutes 500 uL of pre-warmed to 37 degrees Celsius SOC media (2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, and 20 mM glucose) was added to each tube and allowed to recover on a shaker at 37 degrees Celsius for 2 hours. After recovery, each culture was plated onto bacterial culture plates (LB agar with ampicillin) using aseptic technique and allowed to grow overnight at 37 degrees Celsius. Any colonies on the plate that received the negative control (as described in the paragraph above) are self-ligations and are false positives. Colonies were selected from the positive control plate and inoculated in to 2mL of LB with ampicillin. Tubes were placed on a shaker at 37 degrees Celsius overnight. Miniprep was then done using the Qiagen QIAprep Miniprep kit and protocol.

Confirmation of the vector was done using a double digest. Double digest was set up using a master mix of 1uL of the BBS1 enzyme, 1uL of the EcoR1-HF enzyme, 5uL of 10X NEB buffer 2.1and 39 uL of ddH20 for each sample. 23uL of master mix was combined with 2uL of purified vector and incubated at 37degrees for 1 hour. Digest product is separated by gel electrophoresis on a 1% agarose gel. Clones with the guide sequence inserted will display bands at 9KB and 1KB and clones without the guide sequence inserted will display bands at 6.5KB and 3.5KB.

Next, for samples where the double digest suggested our guide sequence was inserted,

confirmation of the guide sequence was done using Sanger sequencing. For sequencing, we used 1ug of the vector containing the guide sequence, 3uL of primers for the U6 promoter, and add ddH20 to bring the total volume up to 12uL. Confirmed plasmid was expanded by expanding corresponding clones and using either multiple mini preps or a maxi prep.

PyMT 419 cells were transfected using the Life Technologies Lipofectamine 3000 protocol https://tools.lifetechnologies.com/content/sfs/manuals/lipofectamine3000_protocol.pdf . To select clones, GFP positive cells were sorted into 96 well plates using fluorescence activated cell sorting. Knockouts were identified by isolating DNA from individual clones, using polymerase chain reaction of the region of a 300bp region DNA containing the area targeted by the sequence centrally located, and gel purified using gel electrophoresis on 3% agarose gel and Qiagen QIAquick Gel Extraction Kit. Primers for amplification are as follows:

Adm 5': CTGAGAGATGGTCTGGAGGTG

Adm 3': CAATCCCCAGGGTCAGAGTA

Fgf 13 5': TGTTCTAACTTCCAGAAAGGCATA

Fgf 13 3': CAGTGGTTTGGGCAGAAAAT

For sequencing, nested primers were used with sequences as follows:

Adm 5': GGCTGGGACATCACTTGAAC

Fgf 13 5': CACACCCATATAAGTATTGACTTTCA.

IN VITRO ASSAYS

For cell counts over 3 days, 100, 000 cells were seeded into 6- well plates. On each day cells were trypsinized and counted on a Nexcelom Cellometer T4 automatic cell counter. Each clone was counted in triplicate for each day.To measure cell migration, we did wound healing

assays in the presence of 2ug/mL Mitomycin C using standard methods [203]. Photomicrographs were taken at 0 hour and 18 hours.

IN VIVO ASSAYS

All animal work has been conducted according to national and institutional guidelines. All mice were in the FVB background. For tail vein injection, MMTV-Cre control mice were used to avoid immune response to the middle T antigen [136, 251, 252]. For control cells and each knockout clone, 50,000 cells were injected into the bloodstream via the tail vein. After 21 days, mice were euthanized. Lungs and liver were resected for routine H&E staining to detect metastases.

CHAPTER 5:

GENE EXPRESSION SIGNATURES PREDICT TUMOR HISTOLOGY AND HIGHLIGHT SIMILARITIES AND DIFFERENCES BETWEEN MOUSE MAMMARY TUMORS AND HUMAN BREAST CANCER.

ABSTRACT

The heterogeneity present in breast cancer establishes a complex array of distinct subtypes of tumors. With this, modeling the disease in vivo requires numerous preclinical models that effectively mimics the multiple factors inherent to human breast cancer progression and parallel the molecular profiles of human breast cancer subtypes. Using a gene expression database of mouse models and human breast cancer, we identified mouse models that parallel gene expression profiles of specific subtypes of human breast cancer. However, there are different tumor histologies observed in mouse models from those observed in human breast cancer. As such there is a need to identify how differences in tumor histology impact comparisons between mouse mammary tumors and human breast cancer. Further, much of the publicly available gene expression data for mouse mammary tumors comes without histological classification. Together, this illustrates the need to identify the genes that are intrinsic to mouse mammary tumor histology. We hypothesize that gene expression signatures can be generated to accurately predict tumor histology of mouse mammary tumors across different tumor initiating events. Using gene expression data from histologically annotated mouse mammary tumors initiated by different oncogenic events, we have developed gene expression signatures that define tumors with squamous or adenosquamous tumor histology and a signature that defines tumors with epithelial to mesenchymal transition (EMT)-like tumor histology. Testing these signatures in human breast cancer we found that human breast cancers do not have squamous gene expression features; however this signature was able to identify histologically annotated squamous tumors in other human cancer types such as lung cancer. Interestingly the EMT-like signature was conserved in a subset of human claudin low breast cancer, splitting tumors into mesenchymal and non-mesenchymal tumor types. Together, this data demonstrates robust

signatures that can used to characterize tumor histology and further our understanding of human breast cancer heterogeneity.

INTRODUCTION

One of the hallmarks of breast cancer is tumor heterogeneity. Breast cancer heterogeneity is evident at both the histological and genomic level. The histological type of the tumor refers to the morphological and cytological patterns evident within a tumor. There are a large number of distinct tumor histologies recognized for breast cancer [253, 254]. This includes ductal carcinoma in situ (DCIS), invasive ductal carcinoma, lobular carcinoma in situ (LCIS), invasive lobular carcinoma, tubular carcinoma, cribriform carcinoma, invasive lobular carcinoma, mucinous carcinoma, neuroendocrine carcinoma, papillary carcinoma, adenoid cystic carcinoma, secretory carcinoma, acinic-cell carcinoma, apocrine carcinoma, medullary carcinoma, metaplastic carcinoma with squamous metaplasia, metaplastic spindle cell carcinoma, and metaplastic matrix-producing carcinoma. The most frequently observed tumor histology is the invasive ductal carcinoma [255].

Similarly, there is a large degree of genomic heterogeneity in human breast cancer, which has been classified using gene expression analysis. Classification of breast tumors into their molecular subtypes based on unique gene expression profiles has led to tumors being described according to their "intrinsic subtype": basal, luminal A, luminal B, her-2 positive, claudin low and normal-like breast group [57-59]. Importantly, these intrinsic subtypes of breast cancer now serve as the fundamental basis by which researchers classify tumor heterogeneity. However, since the development and identification of the intrinsic subtypes of breast cancer, researchers have expanded on this work to further define tumor heterogeneity. Among these, an important study detailing the pathway activation profiles within the intrinsic subtypes, demonstrated

molecular complexity beyond the six subtypes of breast cancer [47]. Specifically, this work identified subgroups within the intrinsic subtypes, totaling up to 17 subtypes of breast cancer on the basis of predicted pathway activation profiles.

Importantly, recent work has connected the dots between intrinsic subtypes of human breast cancer and specific histological types of breast cancer [254].Chief amongst their findings was that within intrinsic subtypes of cancer were multiple histological types of cancer. For example, two histological tumor types were categorized as claudin low: medullary and metaplastic breast cancer. Further, individual tumors of the same tumor histological types corresponded to different intrinsic subtypes of breast cancer. For example, some medullary tumors were classified as basal and others were categorized as claudin low. These findings, might suggest that gene expression methods may do better job of organizing tumors into similar disease entities. Collectively, these studies demonstrate that histological and genomic heterogeneity present in breast cancer establishes a complex array of distinct subtypes of tumors.

With this, modeling the disease in vivo requires numerous preclinical models that effectively mimics the multiple factors inherent to human breast cancer progression and parallel the molecular profiles of human breast cancer subtypes. While the use of human cell lines and patient derived xenografts offer the opportunity to study human breast cancer in-vivo, they rely on immunocompromised hosts. The use of genetically engineered mouse models of cancer offer the advantage and the opportunity to study tumor progression in an immuno-competent system. As a result, a major focus has been to establish which genetically engineered mouse models have parallels in human breast cancer. For example, work from the Perou lab identified similarities between human and mouse mammary tumors using immunohistochemistry for key biomarkers [105]. Expanding upon these findings with additional tumor models and samples, numerous

reports have documented mouse and human counterparts at the level of gene expression [129, 140, 141, 249, 256]. Despite gene expression similarities one aspect that needs to be addressed is how the tumor histology of mouse mammary tumors factor into these relationships.

As seen in human breast cancer, a large number of tumor histologies are observed for mouse mammary tumors [257] . This includes glandular, acinar, cribriform, papillary, solid, squamous, fibroadenoma, adenomyoepithelioma, adenosquamous, microacinar, adenocarcinoma, comedoadenocarcinoma, and medullary [141, 204, 225, 257] . The prevalence of tumor histology is dependent on the particular mouse model, for example tumors initiated by activated Neu tend to form comedoadenocarcinas [106, 141] . In comparison of mouse and human histologies there are noticeable differences, for example, squamous tumors are not observed in human breast cancer [253, 255]. As such, it is important to begin to understand how mouse and human tumor histology impact the genomic relationships between mouse models and human breast cancer.

With this in mind our goal was to identify the genes that define specific tumor histologies in the mouse. In previous work, we observed that unsupervised hierarchical clustering of Myc initiated tumors organized tumors into subclasses that correlated with their histology [106]. Further, even in the presence of loss of the activator E2F transcription factors, clustering arranged tumors according to histology, rather than genotype[108]. This suggests that there are unique gene expression components inherent to individual tumor histologies. We hypothesize that gene expression signatures can be generated to accurately predict tumor histology of mouse mammary tumors across regardless of differences in tumor initiating oncogenic events. Using gene expression data from histologically annotated mouse mammary tumors initiated by different oncogenic events, we have developed gene expression signatures that define tumors

with squamous or adenosquamous tumor histology and a signature that defines tumors with epithelial to mesenchymal transition (EMT)-like tumor histology. Testing these signatures in human breast cancer we found that human breast cancers do not have squamous gene expression features; however this signature was able to identify histologically annotated squamous tumors in other human cancer types such as lung cancer. Interestingly the EMT-like signature was conserved in a subset of human claudin low breast cancer, splitting tumors into mesenchymal and non-mesenchymal tumor types. Further, applying these signatures to our published database [249] of mouse mammary tumors we identified additional tumors that have are of either a squamous or a EMT-like histology. Together, this data demonstrates robust signatures that can used to characterize tumor histology and further our understanding of human breast cancer heterogeneity.

RESULTS

ASSEMBLY OF THE SQUAMOUS HISTOLOGY SIGNATURE

To build a gene expression signature that could identify squamous tumors, we utilized histologically annotated tumors from a MMTV-PyMT mouse model that had several different genotypes. In this dataset there were tumors that were either $E2F^{WT/WT}$ or $E2F2^{-/-}$. Using a significance analysis of microarrays (SAM) we executed the following comparisons: $E2F^{WT/WT}$ squamous tumors compared to $E2F^{WT/WT}$ non-squamous tumors, $E2F2^{-/-}$ squamous tumors compared to $E2F2^{-/-}$ non-squamous tumors, $E2F2^{-/-}$ squamous tumors compared to $E2F^{WT/WT}$ non-squamous tumors, and $E2F^{WT/WT}$ squamous tumors compared to $E2F2^{-/-}$ non-squamous tumors. In this way any of the genotypic differences can be filtered out and the genes that were consistently differentially regulated in this analysis would be the genes intrinsic to squamous identity. FIGURE 5.1 illustrates the identification of consistently differentially regulated in

squamous tumors. Focusing only on the genes detected in all four comparisons, we identified 179 genes were upregulated in squamous tumors. We did not detect any genes that were that were consistently downregulated in this comparison. As a result, we found 179 genes that potentially define squamous tumor histology.

ASSEMBLY OF THE EMT-LIKE HISTOLOGY SIGNATURE

We used a similar approach to generate a signature that defines tumors with EMT-like tumor histology. Using gene expression data from histologically annotated MMTV-Myc tumors, we used a SAM analysis to identify the genes intrinsic to tumors with an EMT-like histology making the following comparisons: EMT-like tumors compared to all non-EMT-like tumors, EMT-like tumors compared to squamous tumors, EMT-like tumors compared to papillary tumors, and EMT-like tumors compared to microacinar tumors. Signature genes focused on genes that were differentially regulated in each comparison. FIGURE 5.2A shows the overlapping genes that were upregulated in each of the comparisons. FIGURE 5.2B shows the overlapping genes that were downregulated in each of the comparisons. In total, the analysis revealed 185 genes consistently upregulated in EMT-like tumors and 175 genes consistently downregulated in EMT-like tumors.

VALIDATING THE SQUAMOUS HISTOLOGY SIGNATURE

To validate the 179 genes that define squamous histology, we tested our signature in independent datasets where squamous tumors developed in mouse mammary tumor models initiated by other oncogenes. The first test was to determine if the squamous signature could separate squamous tumors from other tumors within the MMTV-Myc mouse model using unsupervised hierarchical clustering limited to the squamous signature genes. As depicted in FIGURE 5.3A, clustering on our signature genes accurately splits out squamous tumors from the

other tumor histologies. To measure the statistical significance of this signature, we compared squamous tumors to all non-squamous tumors using gene set enrichment analysis (GSEA). As evident in FIGURE 5.3B, the squamous signature genes derived from MMTV-PyMT tumors were significantly enriched in MMTV-Myc squamous tumors (NES 1.48, nominal p-value=0.0, FDR q-value =0.029, FWER p-value=0.016). Together, this data shows the validity of the squamous signature by its ability to predict the tumor histology in tumors initiated by different oncogenes.

VALIDATING THE EMT-LIKE HISTOLOGY SIGNATURE

To validate the 185 genes upregulated in EMT-like tumors and 175 genes downregulated in EMT-like tumors, we tested these genes on gene expression data annotated MMTV-Met tumors. Using a similar validation approach, we tested if the EMT-like signature could separate EMT-like (also referred to as splindoid) tumors from other tumors within the MMTV-Met mouse model using unsupervised hierarchical clustering limited to the signature genes. As depicted in FIGURE 5.4A, the EMT-like signature genes separated out the EMT-like tumors from other Met-induced tumors. Testing for enrichment of the signature genes with GSEA illustrated a significant enrichment for upregulation of the upregulated EMT-like signature genes in Met-induced EMT-like tumors compared to non EMT-like tumors (FIGURE 5.4B, NES=1.76, nominal p-value=0.0, FDR q-value= 0.009, FWER p-value = 0.011). Likewise, GSEA found significant enrichment for downregulation of the downregulated EMT-like signature genes in Met-induced EMT-like tumors compared to non EMT-like tumors (FIGURE 5.4C, NES=-1.66, nominal p-value=0.006, FDR q-value= 0.009, FWER p-value = 0.018). As a whole, this data shows the validity of the EMT-like signature.

CLASSIFYING MOUSE MAMMARY TUMORS

With validation of both the squamous and EMT-like signatures, we wanted to use this data to classify tumor samples in our previously published mouse mammary tumor model gene expression database [249]. This database contained over 1,0000 mouse mammary tumor samples across 26 major mouse models of breast cancer. In assembling this database, only subset of samples contained histological annotations and a majority lacked histological annotations. As a result, development of these signatures presented the opportunity to predict whether tumor samples in this database had EMT-like or squamous tumor histology.

To predict which tumor samples were squamous or EMT-like we used unsupervised hierarchical clustering using our signatures for EMT-like tumors and squamous tumors (FIGURE 5.5). In this way, we could visualize expression patterns for signature genes as well use existing annotations to both predict tumor histologies for other tumor samples and to monitor the performance of the signatures. Importantly, each set of signature genes clustered tightly together. Observing expression patterns for the squamous signature genes, we identified a cluster of samples that had high expression of these genes (FIGURE 5.6A); suggesting squamous tumor histology. Indicating the validity of this prediction, Myc tumors noted to have squamous histology were found in this cluster. Amongst the other tumor found in this cluster were a subset of tumors from the MMTV-PyMT, MMTV-Wnt, large T-antigen, Cre-ETV6-NTRK3, DMBA treated, IGFIR, BRCA and p53 mutant, and p53 mutant mouse models. To test the statistical significance of this prediction, we compared tumors from this cluster (our predicted squamous cluster) to all other tumors in the database using GSEA (FIGURE 5.6B). Supporting the prediction that these tumors are squamous, GSEA detected a significant enrichment for

upregulation of the squamous genes for tumors predicted from the squamous cluster (NES=1.99, nominal p-value = 0.0, FDR q-value=0.0029, FWER p-value = 0.002).

Observing expression patterns for the EMT-like signature genes, we identified a cluster consisting of the majority of the tumor samples with an annotation for an EMT-like histology spanning several oncogenic models (FIGURE 5.7A). In agreement, tumors in this cluster had high expression of the genes upregulated in EMT-like tumors and low expression of genes downregulated in EMT-like tumors. This approach identified additional tumors with an EMT-like histology from the Myc model that had not previously been annotated. For example, we identify a minority of tumors initiated by the Neu oncogene that EMT-like gene expression features. In addition, a subset of p53 mutant, BRCA mutant, BRCA and p53 mutant,Wnt, TNP8,Stat5, and LPA induced tumors were found in this cluster. To test the statistical significance of this prediction, we compared tumors from this cluster (our predicted EMT-like cluster) to all other tumors in the database using GSEA (FIGURE 5.7B,C). Supporting the prediction that these tumors are EMT-like, GSEA detected a significant enrichment for high expression of the genes upregulated in EMT-like tumors (FIGURE 5.7B, NES=1.88, nominal p-value = 0.0, FDR q-value=0.0071, FWER p-value = 0.016) and low expression of the genes downregulated in EMT-like tumors (FIGURE 5.7C, NES=-1.89, nominal p-value = 0.0, FDR q-value=0.01, FWER p-value = 0.016).

Importantly, there were a large number of tumors that did not match expression patterns for squamous or EMT-like tumors. As shown in FIGURE 5.5, the majority of tumors in the database had high expression of genes that are down-regulated in EMT-like tumors and low expression of genes that are upregulated in squamous and EMT-like tumors. Among the tumor histologies found here, were the microacinar and papillary tumors from the MMTV-PyMT

mouse model. In addition, there were a number of tumors from other mouse models without histological annotations in the portion of the clustering including tumors from the MMTV-Neu, MMTV-PyMT, MMTV-Myc, and p53 mutant mouse models. Collectively, these results show demonstrate the squamous EMT-like signatures to pull out and predict tumors with corresponding histology without forcing non squamous or EMT-like tumors into the one or the other categories.

TESTING HISTOLOGICAL SIGNATURES IN HUMAN CANCER

With the signatures ability to detect corresponding tumor histologies in a large database of mouse mammary tumors, wanted to answer whether these gene expression signatures are conserved in human cancer. To test the squamous signature in the context of human breast cancer, we brought our training dataset containing the MMTV-PyMT squamous tumors together with a database that we assembled of over 1,000 human breast cancer samples that were annotated for intrinsic subtype while mediating the batch effects between them [249] . Clustering on the basis of squamous genes in this setting accurately identified all of squamous tumors from the PyMT mouse model as expected. However, none of the human breast cancers showed activation of the squamous genes (FIGURE 5.8). This finding was also expected given the fact that squamous histology is rarely observed in human breast cancer [253, 255].

However, squamous tumors are found in other cancer types. This includes oral, lung, and esophageal cancers. As a result, we assembled a database consisting of human cancers that include breast, lung, colorectal, thyroid, oral, cervical, endometrial, melanoma, head and neck, ovarian, and merkell cell carcinomas using appropriate batch correction methods. This database consisted of over 3,000 human tumors. Clustering these tumors on the basis of the squamous gene signature organized squamous tumors from oral, lung, and esophageal cancer into the same

cluster that had markedly higher expression of the squamous genes than tumors from other clusters (FIGURE 5.9). Interestingly, subsets of samples from melanoma, cervical cancer and merkell cell carcinoma also ordered into this cluster suggesting a squamous component in these cancers (FIGURE 5.10A). Using GSEA to test for statistical enrichment of the squamous signatures in these samples indicted these samples demonstrated that this enrichment is significant (FIGURE 5.10B, NES= 1.93, nominal p-value 0.0, FDR q-value 0.003, FWER p-value =.002). Also important to note is that while a majority of the squamous genes were highly expressed in squamous tumors, there was a cluster of genes that did not vary across tumor types. These genes that did not change may be unique to squamous tumors in mouse. The remainder may indicate which of these genes are absolutely intrinsic to squamous tumors since their activation not only spans different oncogenic events, but are conserved in squamous cancers from multiple tissues, and extend over mouse and human cancers.

Similarly, we wanted to test whether aspects of the EMT-like signature was conserved in human breast cancer. To examine this we combined our MMTV-Myc tumors that were used to establish the EMT-like signature with our database of human breast cancer and used unsupervised hierarchical clustering limited to the EMT-like signature genes (FIGURE 5.11). This revealed a subset of human claudin low breast cancer tumors that showed high expression of the EMT-like signature genes and clustered with EMT-like tumors from the MMTV-Myc mouse model(FIGURE 5.12A). Testing the significance of these gene expression patterns, GSEA revealed that the claudin low tumors that clustered with mouse EMT-like tumors were significantly enriched for high expression of genes upregulated in EMT-like tumors (figure 5.12B, NES=1.87, nominal p-value =0.0m FDR q-value=.002, FWER p-value = 0.004) and enriched for low expression of genes downregulated in EMT-like tumors (FIGURE 5.12C,

NES=-2.11, nominal p-value=0.0, FDR q-value =0.0, FWER p-value 0.0). As a result, the EMT-like signature split claudin low tumors into two subclasses: tumors have EMT-like gene expression features and those that do not.

DISCUSSION

We hypothesized that gene expression signatures can be generated to predict tumor histology. To this end, we have developed gene expression signatures that define tumors with a squamous or adenosquamous tumor histology and a signature that defines tumors with epithelial to mesenchymal transition (EMT)-like tumor histology across different oncogenic mouse models and human cancers. As a result, application of these signatures to gene expression data from mouse mammary tumors, we can annotate tumors that are of either a squamous or an EMT-like histology and those that are not. Testing these signatures in human breast cancer we found that human breast cancers do not have squamous gene expression features; however this signature was able to squamous tumors in other human cancer types. Interestingly the EMT-like signature was conserved in a subset of human claudin low breast cancer, splitting tumors into mesenchymal and non-mesenchymal tumor types. Together, this data demonstrates robust signatures that can used to characterize tumor histology and further our understanding of human breast cancer heterogeneity.

Using the squamous signature to predict the histology of non-annotated mouse mammary tumors, we separated out tumors with a significant enrichment for squamous gene expression features. This included tumors from the MMTV-Myc model previously annotated as squamous [141]. However, we also predicted squamous tumors were present in the cluster highlighted in FIGURE 5.6 from the MMTV-PyMT, MMTV-Wnt, large T-antigen, Cre-ETV6-NTRK3, DMBA treated, IGFIR, BRCA and p53 mutant, and p53 mutant mouse models. In support of this

prediction, tumors with squamous components have been reported in tumors initiated by PyMT [204, 258]. Similarly, overexpression of the Wnt pathway in the mammary gland has been demonstrated to induce squamous tumors [259]. Also, indicating the predictions for a subset of tumors in the Cre-ETV6-Ntrk3 model are correct, tumors with squamous metaplasia were amongst the variety of histologies observed [260]. DMBA treated mammary tumors [261]and p53 mutant tumors [262] were also reported to exhibit squamous histology . In regards to large T-antigen tumors, inactivation of the tumor suppressor protein retinoblastoma resulted in tumors with a squamous component [263]. Together, these previous studies support the prediction that squamous signature is accurately predicting squamous tumor histology in a large database of mouse mammary tumor samples supports labeling tumors in this cluster as squamous.

Like the squamous signature, the signature for EMT-like tumors showed a high degree of accuracy. Correctly, it identified tumors with a previous EMT annotation. In addition, this approach identified additional tumors with an EMT-like histology that had not previously been annotated. For example, we identify a minority of tumors initiated by the Neu oncogene that EMT-like gene expression features. Additionally, a subset of p53 mutant, BRCA mutant, BRCA and p53 mutant, Wnt, TNP8,Stat5, and LPA induced tumors we predicted to have EMT-like histology. While Neu induced tumors are generally comedoadenocarcinomas, in unpublished work from in our lab we observed a minor subset of Neu induced tumors with EMT histology in tumors with displaying intra-tumor heterogeneity. Thus, it is possible that the tumors we detected in the EMT-like cluster resulted from array analysis of similarly EMT like portions of the Neu induced tumors. In agreement with the predictions, tumors characterized by elongated spindle shaped cells were observed amongst the mammary tumors in initiated by Wnt[259], as well tumors from the p53 mutant mouse model [262] . Taken together, this illustrates our EMT-like

signature is able to identify tumors with an EMT-like tumor histology across the expansive database of mouse mammary tumor samples and supports annotating tumors in this cluster as EMT-like.

Perhaps most important about this work is that we show signatures for tumor histology derived from mouse mammary tumors have conserved features in human cancer. Specifically, we identified a subset of human claudin low tumors that clustered with MMTV-Myc EMT-like tumors and we enriched for the EMT-like signature. In agreement, previous studies have found that there are claudin low tumors that resemble mouse mammary tumors with EMT-like histology at the level of gene expression [129, 140, 141] . This suggests these claudin low tumors indeed have mesenchymal identity. However, our signature split claudin low tumors into subgroups. In previous work, we observed a split in claudin low tumors, those that clustered with mouse EMT-like tumors and those that do no not while detecting key differences in Myc activity [141] .Importantly, this work expands upon those findings, subdividing claudin low tumors into those with mesenchymal gene expression features and those that are not mesenchymal. In agreement, previous work has found that claudin-low tumors correlate with two different tumor histologies: medullary and metaplastic [254] . Histologically, metaplastic tumors resemble the elongated spindle shaped cells that are typical of tumors with EMT-like tumor histology. In light of this, it is possible that the tumors enriched for the EMT-like signature are metaplastic and those that are not are medullary. As gene expression data for medullary tumors become available similar gene signature approaches would be able to confirm this speculation.

In addition to the EMT-like signature, we found that the squamous signature also translates to human cancer. The squamous signature was not enriched in any human breast cancer samples, however, this finding was expected and further shows the accuracy of the

signature as this histology is not common in human breast cancer [253, 255]However, the squamous signature did organize squamous tumors from oral, lung, and esophageal cancer into the same cluster. While demonstrating the signature can predict human cancer tumor histology, this finding lends itself to a far more important conclusion. The most critical lesson of this work is that it demonstrates the presence of unifying gene expression features for tumors of the same histology. This conclusion is upheld by the fact that our signatures were consistent in mouse mammary tumors of the same histology despite differences in the initiating event that leads to carcinogenesis in each of the mouse models. Further, they span the gap from mouse to human as observed for the enrichment of the EMT-like signature in claudin low tumors and the squamous signature in variety human squamous tumors. Finally, they spanned cancers from several different tissues, as observed for oral, lung, and esophageal cancer. This demonstrates that there are gene expression programs that are intrinsic to tumor histological types. The significance of this finding may have important clinical significance as it might suggest common pathway activation, metabolic, and extrinsic dependencies for these human tumors. As a whole, our findings demonstrate the utility of gene expression signatures to characterize tumor histology and further our understanding of human cancer heterogeneity.

METHODS

MICROARRAY DATA

Details for assembling the mouse mammary tumor model and human breast cancer database can found [249] . Gene expression data from squamous and non-squamous MMTV-PyMT tumors were prepared by isolation of RNA samples from flash frozen tumors using the Qiagen RNeasy kit after roto-stator homogenization. RNA was submitted to the Michigan State University Genomics Core facility for gene expression analysis using Mouse 430A 2.0 Affymetrix arrays.

Gene expression data from MMTV-Myc EMT-like and non EMT-like tumors is described in previous work [141]. Gene expression data from human squamous and non-squamous tumors was accessed on the Gene Expression Omnibus under the following accession numbers: GSE10245, GSE10300, GSE14020, GSE17025, GSE18520, GSE2034, GSE20347, GSE21422, GSE21653, GSE2280, GSE2603, GSE27155, GSE27678, GSE29044, GSE30219, GSE30784, GSE3292, GSE33630, GSE3524, GSE35896, GSE37745, GSE39491, GSE39612, GSE43580, GSE45670, GSE4922, GSE50081, GSE51010, GSE6532, and GSE7553. These datasets were normalized using Affymetrix Expression Console. Bayesian Factor Regression Methods (BFRM) [25] was used to combine datasets and remove batch effects.(http://www.stat.duke.edu/research/software/west/bfrm/download.html).

DATA ANALYSIS

Gene expression signatures were derived using significance analysis of microarrays [28] to detect the genes that were differentially regulated for each tumor histology. For the squamous signature we executed the following comparisons: E2F $^{WT/WT}$ squamous tumors compared to E2F $^{WT/WT}$ non-squamous tumors, E2F2 $^{-/-}$ squamous tumors compared to E2F2 $^{-/-}$ non-squamous tumors, E2F2 $^{-/-}$ squamous tumors compared to E2F $^{WT/WT}$ non-squamous tumors, and E2F $^{WT/WT}$ squamous tumors compared to E2F2 $^{-/-}$ non-squamous tumors. For the EMT-like signature, we made the following comparisons: EMT-like tumors compared to all non-EMT-like tumors, EMT-like tumors compared to squamous tumors, EMT-like tumors compared to papillary tumors, and EMT-like tumors compared to microacinar tumors. Unsupervised hierarchical clustering was done using Cluster 3.0 and Java Tree View. The color scheme for the heatmap and sample legends were made using Matlab. Gene set enrichment analysis [34] was done by converting our gene expression data and gene lists to the specified formats using Gene Pattern.

CHAPTER 6:

POSSIBLE FUTURE DIRECTIONS

While much of the work done during the course of this dissertation has been published, there are a number of possible avenues for extending the research presented in each chapter. As such, this section will highlight key directions and results that can be followed up on to lead to new discovery.

In chapter 1, A mouse model with T58A mutations in Myc reduces the dependence on KRas mutations and has similarities to claudin-low human breast cancer, there is more to learn regarding the significance of the molecular alterations in the EMT-like mouse mammary tumors. For example, as shown in FIGURE S 1.16, both the MMTV-Myc EMT-like tumors and human claudin low breast cancer show high probability of Ras activation. However, despite the high probability of Ras activation, both of these tumors predict that they are not dependent upon this pathway for tumorigenesis. Clearly, answering whether or not silencing/inhibiting this pathway has any consequence on tumor progression or viability is warranted due to this prediction. Further, this presents an opportunity to explore the functional significance of Ras activation in both the mouse EMT-like tumors and human claudin low breast cancer. Another potential avenue for investigation in this area deals with the effectors of Ras signaling in the EMT-like tumors where Myc shows low protein levels and activity (FIGURE 1.4). We know a majority of these tumors have activating mutations in KRas (FIGURE 1.3). As demonstrated by Rosie Sears group, a transcriptional effector of Ras signaling is Myc [122, 123]. This leads to the question: what transcription factors are now activated in KRas mutant tumors where Myc activity is low? Another interesting result to follow up on is the differential activation of Myc target genes across histological subtypes (FIGURE 1.5). This lends to a host of questions such as: Why do the different lines of Myc feature different utilization of Myc target genes? Does the mutation of Myc at threonine 58 cause it? Does the half-life of Myc impact which targets Myc is able to

bind? Does Myc have access to specific promoters based on tumor histology? Do the different utilization of Myc target genes cause the changes in histology or is it a result of histological change? Answering these questions and further investigation of these results would add to our understanding of how Ras and Myc function in tumor progression.

Extending the work presented in the chapter 2, A genomic analysis of mouse models of breast cancer reveals molecular features of mouse models and relationships to human breast cancer, has a large number of possibilities. For example, the strategy used to assemble this database and the analytic approach could be mimicked to assess other mouse models outside of breast cancer. This could answer, for example, how well genetically engineered mouse models of prostate cancer reflect human prostate cancer.

In addition, the mouse model gene expression database itself could be subject to further analysis. For example, chromosomal alterations like amplifications and deletions could be predicted using existing software called ACE (analysis of copy number abnormalities by expression data)[264]. This approach arranges altered genes according to chromosomal location and scores them based on fold change and the degree to which they are nearby on the chromosome to predict amplification or deletion. This approach could identify "hot spots" for alterations in the mouse genome, identify common DNA events from mouse to human, and provide additional areas to test in the mouse model. Similarly, it would be desirable to complement the mouse gene expression data with sequencing data to identify mutations. Together these data would be well served to be integrated with the expression data analysis and made accessible with an online tool much in the image of cBIOPORTAL.

In addition to these big picture follow up projects, there are more specific areas for follow up as well. To begin, there are a number of individual bioinformatic predictions that can be

tested. For example, as shown in FIGURE S 2.5 beta-catenin has high activity in FVB background MMTV-PyMT mice. A genetic test, could reveal the role of beta-catenin in this mouse model. Similarly, p53 has high activity in nearly all MMTV-PyMT tumors. How do we reconcile this prediction, given that these tumors have rapid onset, low levels of apoptosis, and relatively fast growth rate? Again, a genetic test of p53 may provide some interesting insight into p53 functions in aggressive tumor types.

In chapter 3, The E2F transcription factors regulate tumor development and metastasis in a mouse model of metastatic breast cancer, there were some results that could lead to follow up projects. In particular, we see that E2F1 KO tumors develop more quickly (FIGURE 3.2). Why? What is causing enhanced hyperplasia/transformation of the mammary gland? These results may support sequencing early stage transgenic mammary glands and tumors to determine if specific mutations are occurring in E2F1 KO tumors that are causing a fast tumor onset.

Another interesting result was the opposing impact of E2F1 and E2F2 on adenosquamous tumor formation. As depicted in FIGURE 3.4, E2F1 loss led to a reduction in formation of this tumor type and E2F2 loss increased adenosquamous tumor frequency. Why is E2F1 needed for adenosquamous tumor formation, and why does E2F2 seemingly inhibit these tumor types? One possible area for exploration is the role of the E2Fs in luminal progenitor cells. Preliminary bioinformatic analysis comparing E2F2 -/- adenosquamous tumors to non-adenosquamous tumors predicts that the adenosquamous tumors have enrichment for gene expression features of luminal progenitor cells (not shown). Perhaps, the E2F1 is needed for expansion of the luminal progenitor cell population that is needed for adenosquamous tumor development (likewise, maybe E2F2 represses luminal progenitor expansion). Investigating this possibility would provide insight into the non-overlapping functions of E2F1 and E2F2 and how they function in a

key mammary cell type.

Looking at other data from this project, I observed compensatory upregulation of E2F3A in each of the E2F -/- backgrounds (FIGURE 3.3). An interesting investigation may be to determine what knockout of E2F3A specifically in the mammary gland does to tumor progression. One possibility given that this isoform is important for cell cycle progression [167, 168]is that it may slow tumor growth.

Moving on to the work presented in Chapter 4, identifying the mechanistic features by which the E2F1 transcription factor regulates breast cancer metastasis, there is still plenty of data that can be utilized for additional projects. The first is that the characterization focused on E2F1 $^{-/-}$ tumors. There still is fold change data for the E2F2 $^{-/-}$ tumors that could feed new projects and further our understanding as to how E2F2 functions in tumor metastasis. In addition, there were many E2F1 target genes that I did not get to test for metastatic function, but had designed guide sequences and constructs for Crispr knockout that could be readily tested; this includes Trim24, Tgm2, Fzd5, Bmp4, MAFF, and Ttyh1. Like the investigation of Fgf13 and Adm testing these previously mentioned genes could identify new regulators of metastasis and further our understanding as to how E2F1 governs metastasis.

To extend the work from Chapter 5, gene expression signatures predict tumor histology and highlight similarities and differences between mouse mammary tumors and human breast cancer, it would be ideal to turn these signatures into a resource. One thing I am working on is developing additional signatures for other tumor histology. Once complete, it would be useful to develop a website where tumor gene expression data can be uploaded and screened using the signatures to predict and annotate tumor histology. Such a resource could provide annotations for previously unclassified data and also confirm H&E and based classifications of tumor types.

CHAPTER 7:

CONCLUSION

As a whole, the work that I have completed during my dissertation identifies specific mouse models for studying human breast cancer. Furthermore, I demonstrate an integrative approach and training that can be used to solve the complex problems inherent to tumor progression and metastasis. Importantly, this allowed me to demonstrate the usefulness of mouse models of breast cancer and reveal E2F1 and E2F2 as critical regulators of breast cancer metastasis.

APPENDIX

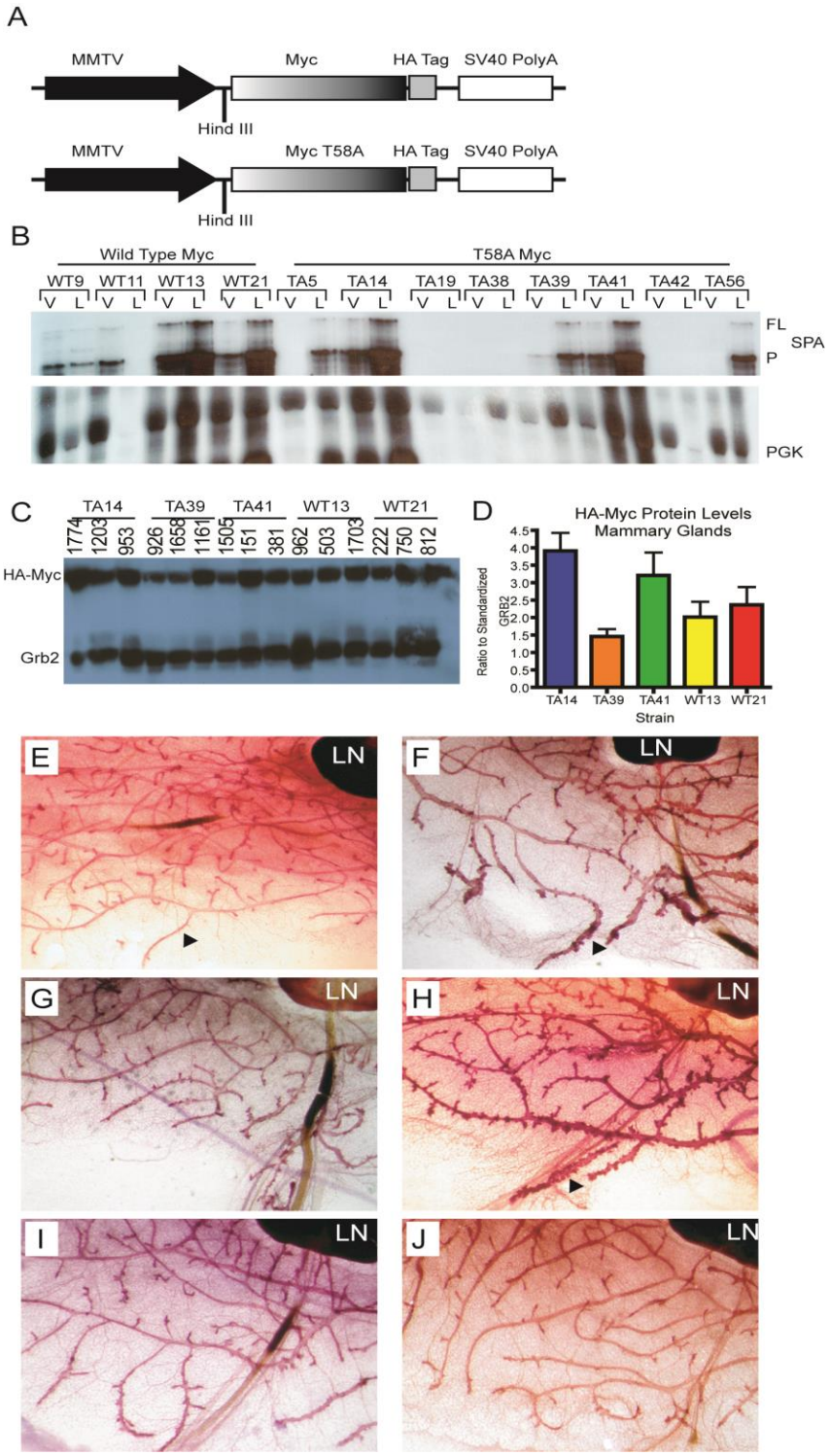


FIGURE 1.1: GENERATION OF MYC AND MYC T58A TRANSGENIC MOUSE MODELS

The construct to generate transgenic mice consist of Myc cDNA or Myc T58A placed under the

FIGURE 1.1 (cont'd)

transcriptional control of the MMTV promoter enhancer and is followed by a HA tag and a SV40 polyadenylation sequence (A). RNase protection assay for transgene expression in both 8 week virgin (v) and lactating (l) mammary glands is shown for the SV40 polyA signal with a PGK internal control (B). Western blot analysis for Myc and Grb2 from mammary glands from the various strains is shown (C). Standardization of Myc protein levels to the Grb2 control show Myc protein levels was completed (D). Mammary gland whole mounts in comparison to the wild type control (E) are shown for WT13, WT21, TA14, TA41 and TA39 lines at 8 weeks of age (F-J respectively). Differences in sidebud formation relative to wild type control (E) in the WT13 (F) and TA14 (H) strains are highlighted with arrowheads.

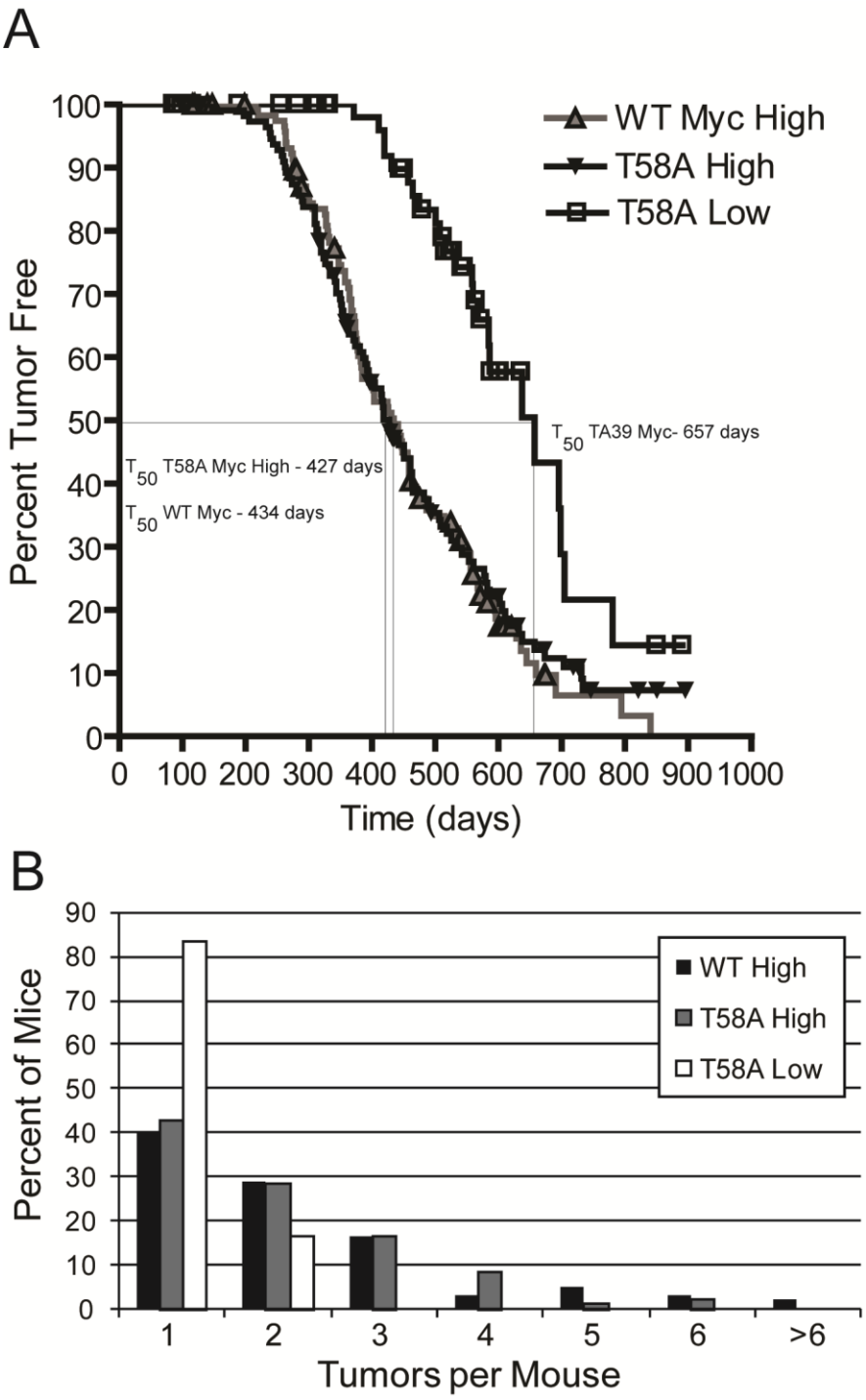


FIGURE 1.2: REDUCED TUMOR LATENCY IN LOW LEVEL T58A MYC TRANSGENIC MICE

Tumor latency was monitored over time and is presented in a Kaplan Meier plot for the two Myc strains (WT13 and WT21), the two high T58A Myc strains (TA14 and TA41) and the low

FIGURE 1.2 (cont'd)

expression T58A strain (TA39) (A). The number of tumors per mouse was assessed in the WT Myc, T58A high and T58A low transgenic strains (B).

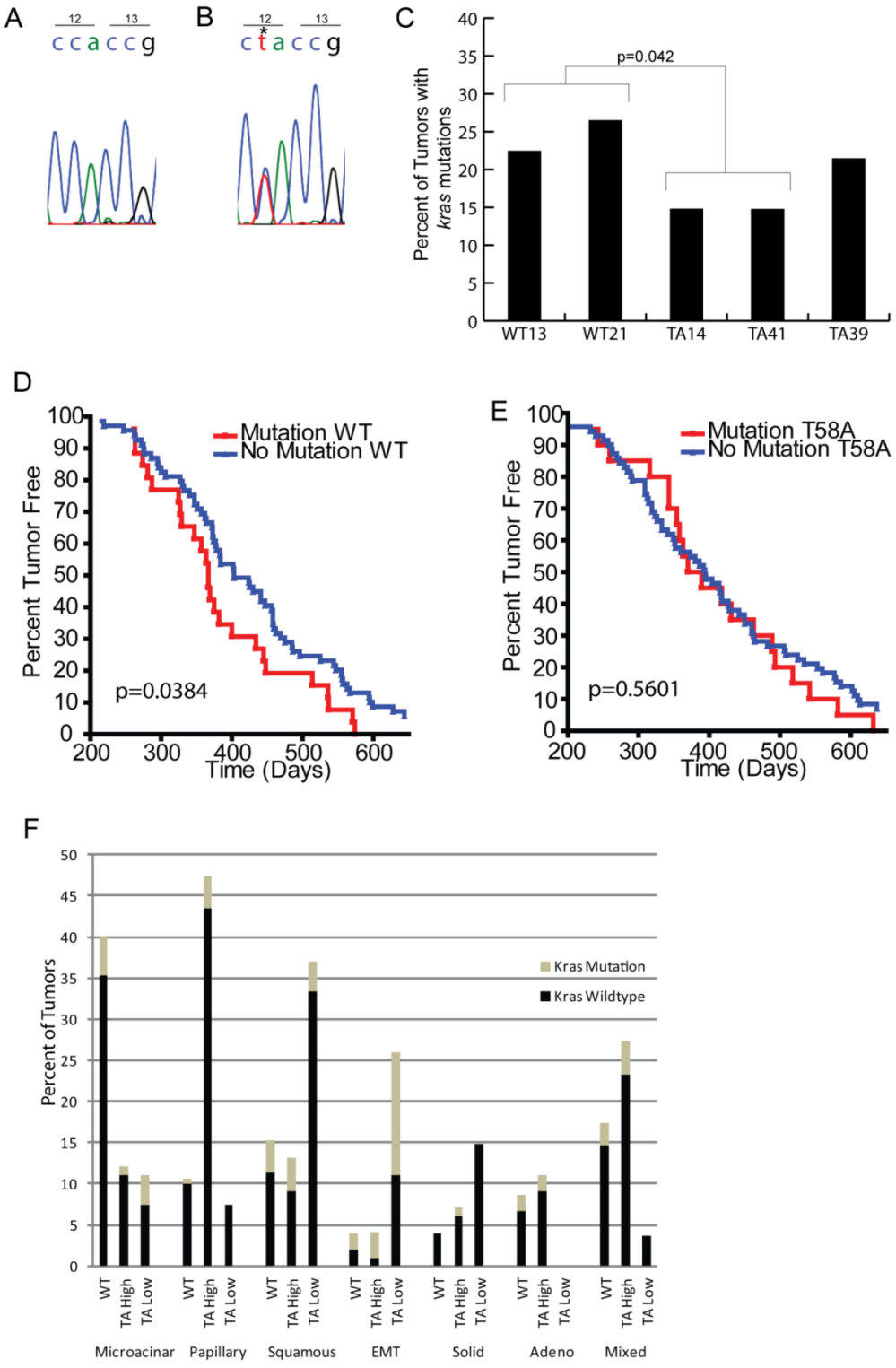


FIGURE 1.3: REDUCED DEPENDENCE UPON ACTIVATING MUTATIONS IN KRAS IN T58A MYC TRANSGENIC MICE

FIGURE 1.3 (cont'd)

To establish the frequency of activating mutations in KRas, we sequenced the RT-PCR KRas product. The sequence trace for codons 12 and 13 from a wild type (A) and tumor with a mutation is shown (B). The percentage of tumors with KRas mutations for WT13, Wt21, TA14, TA41, and TA39 strains was determined (C). To examine tumor latency effects from activating mutations in KRas, Kaplan Meier plots comparing tumor onset between tumors with(n=26) and without KRas mutations (n=69) in the WT Myc transgenic strains (WT13 and WT21) were generated (D). Similarly, Kaplan Meier plots comparing tumor onset between mice with (n=20) and without KRas mutations (n=71) in the T58A high level expression transgenic strains (TA14 and TA41) was examined (E). The histological subtype and percent of each tumor subtype with KRas mutations across the various tumor histologies and transgenic lines. This was grouped into WT MMTV-Myc (WT), T58A high (TA14 and TA41) level (TA High) and low (TA39) level (TA Low) strains (F).

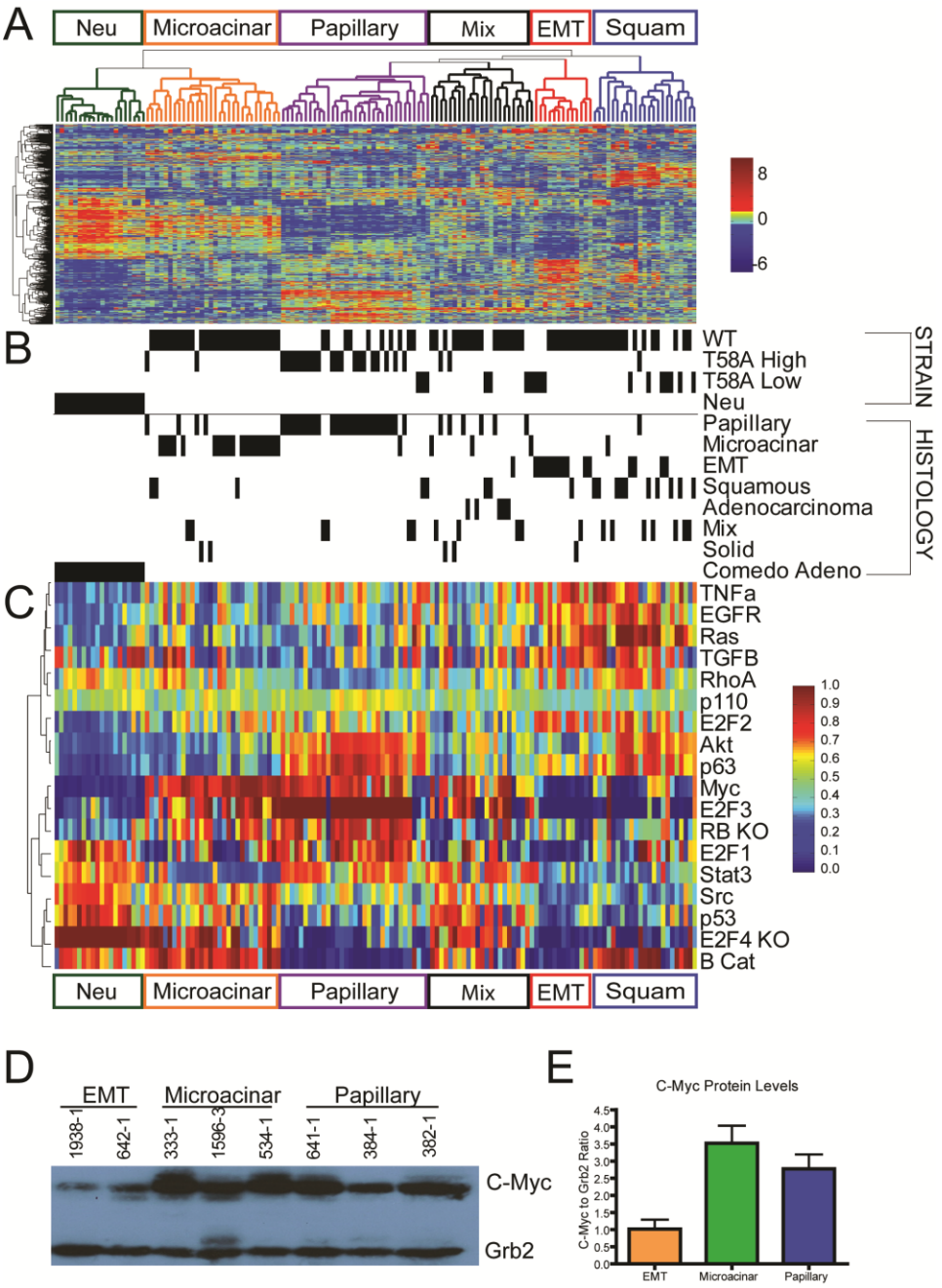


FIGURE 1.4: MOLECULAR HETEROGENEITY OF MYC INDUCED TUMORS

Unsupervised hierarchical clustering of RMA normalized gene expression levels from various MMTV-Myc transgenic lines, as well as MMTV-Neu controls, is shown (A). The identity of the strain and histological type in the clustering analysis is represented with a vertical black bar (B). Maintaining the same order, we present the corresponding pathway activation probabilities for

FIGURE 1.4 (cont'd)

tumor samples in a heatmap with high probabilities in red and low probabilities in blue (C). An example of variable Myc levels between tumor subtypes is shown through a Western blot of EMT, microacinar, and papillary tumor lysates from the TA41 strain (D). Standardization of Myc protein levels to the Grb2 control for each of the subtypes demonstrates relative Myc protein levels (E).

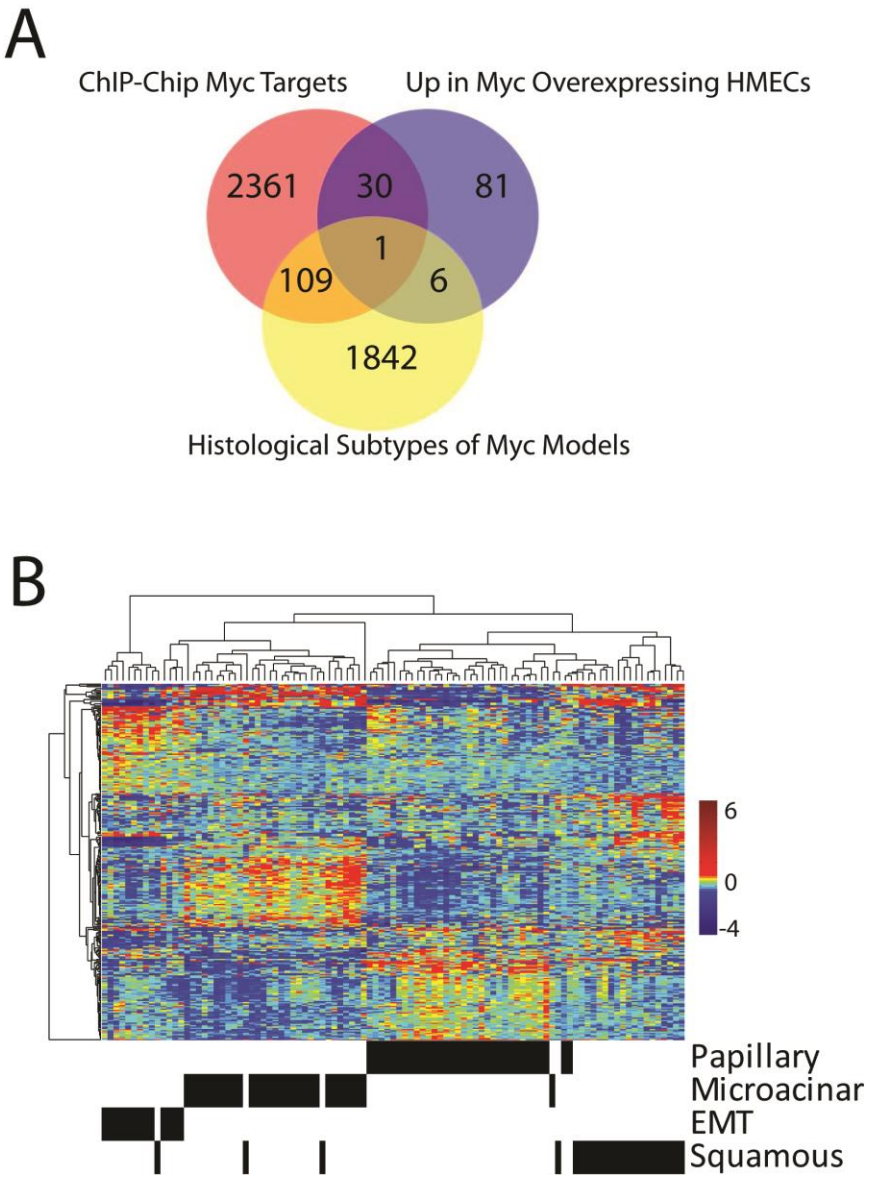


FIGURE 1.5: MYC TARGET UTILIZATION VARIES BETWEEN HISTOLOGICAL SUBTYPES OF MYC INDUCED TUMORS

The Venn Diagram illustrates overlap between Myc target genes defined through a ChIP-Chip experiment, genes upregulated as a result of Myc overexpression in HMECs and genes that are differentially regulated between the major histological types of Myc induced tumors (A). Unsupervised hierarchical clustering of the top 70 differentially regulated potential Myc target genes from each of the papillary, microacinar, EMT, and squamous histological subtypes,

FIGURE 1.5 (cont'd)

representing a total of 280 genes is depicted(B). For the comparison of specific tumor-types versus other Myc tumors the FDR was minimized to 0. This is with the exception for the squamous tumors where the FDR was minimized to 7.05. The fold change of genes upregulated in the EMT-type of tumors ranged from a minimum of 1.254 to a maximum of 5.106; while genes upregulated in microacinar tumors ranged from 1.368 to 6.054. The fold change of the top 70 upregulated genes in papillary tumors ranged from 1.272 to 7.333. Finally, the fold change of genes upregulated in squamous tumors ranged from 1.299 to 4.483.

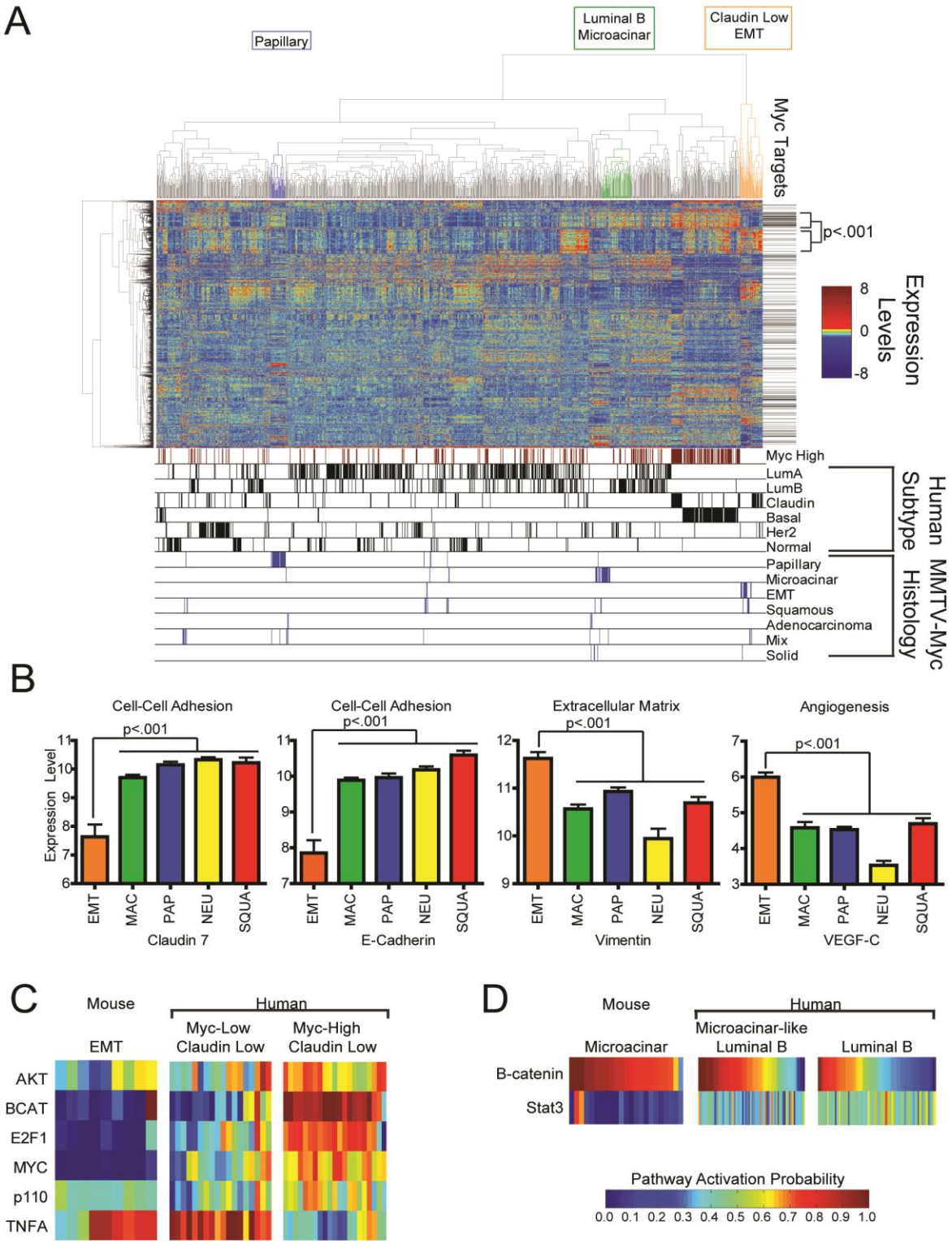


FIGURE 1.6: MYC INDUCED MOUSE TUMOR MODELS GENE EXPRESSION SIMILARITIES TO HUMAN BREAST CANCER

FIGURE 1.6 (cont'd)

Unsupervised hierarchical clustering of Myc induced tumors and human breast cancer reveals relationships between mouse models and human breast cancer(A). The papillary tumors (blue in dendrogram) display unique gene expression patterns while a cluster of human luminal B tumors and microacinar mouse tumors (green in dendrogram) illustrate similar gene expression profiles. A subtype of human claudin low tumors and Myc induced EMT tumors cluster together (orange in dendrogram). Below the heatmap, tumors with a high probability (>60%) of Myc signaling are marked by red bars. Various subtypes of human breast cancer are labeled by black bars. The Myc-induced mouse mammary tumors are depicted with blue bars. Putative Myc target genes are shown with the presence of a horizontal bar at the right of the heatmap (A). RMA normalized expression levels for markers of claudin low tumors were compared between the various histological types of Myc-induced mouse mammary tumors(B). A comparison of signal pathway activation between EMT-type of mouse mammary tumors, Myc-Low human claudin low tumors, and Myc-High human claudin low tumors reveals key differences in human tumors and similarities to the mouse model (C). Likewise, a comparison of the microacinar Myc induced tumors, the human microacinar-like luminal B tumors, and other luminal B tumors for patterns of B-catenin and Stat3 pathway activation reveals similarities (D).

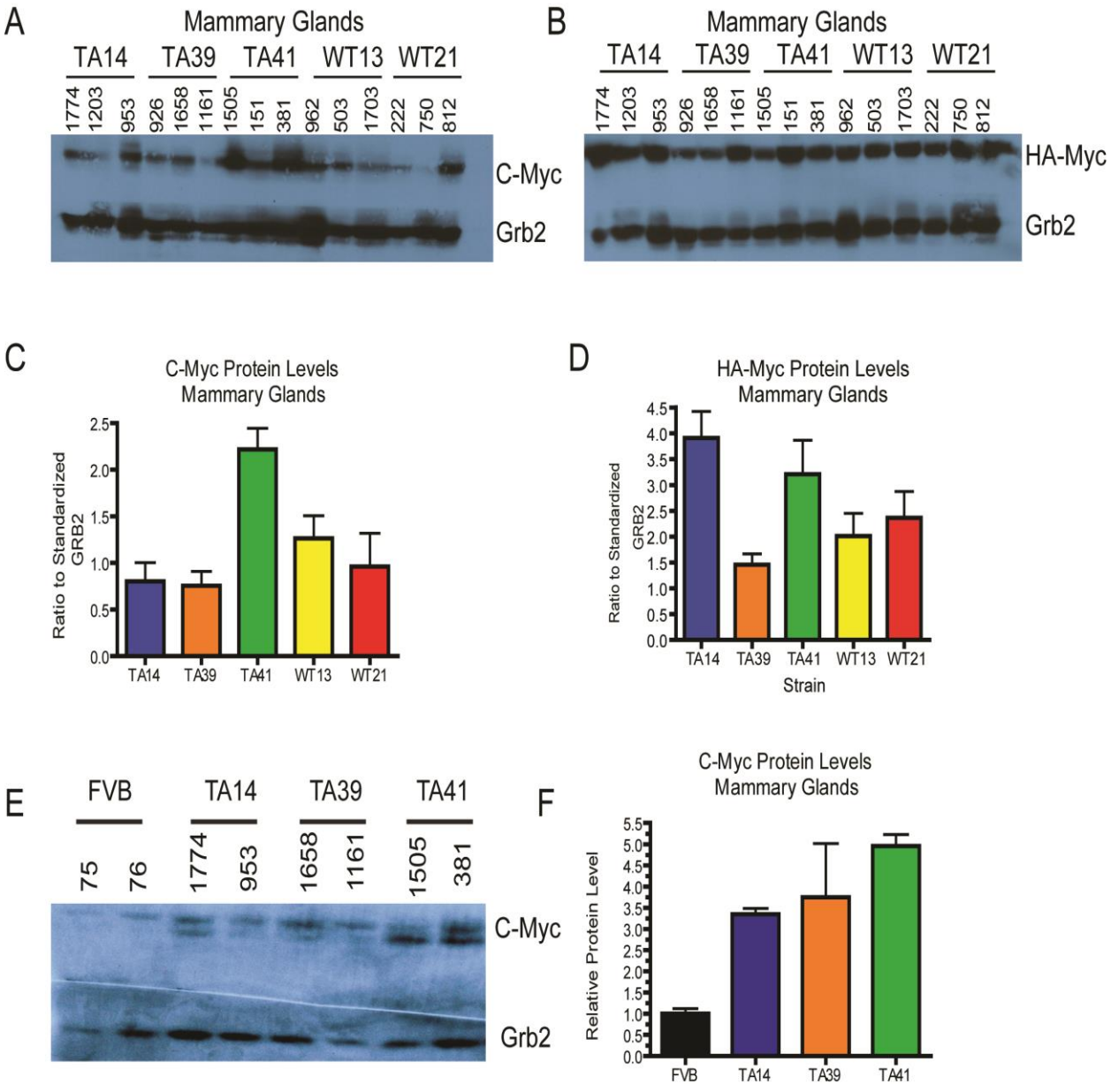


FIGURE S 1.1: MYC LEVELS IN FVB AND TRANSGENIC MAMMARY GLANDS

Western blot analysis of total Myc protein (A) and exogenous HA-tagged Myc (B) for the mammary glands from the various strains of MMTV-Myc mice demonstrates Myc levels. Quantification of total Myc (C) and exogenous HA-tagged Myc (D) protein levels in relation to the Grb2 standard is shown. Western blot analysis of lysates from non-transgenic FVB mammary glands and mammary glands from the transgenic T58A lines demonstrates the

FIGURE S 1.1 (cont'd)

increase in expression relative to wild type controls (E). Quantification of Myc protein levels in the T58A transgenic lines is also shown(F). FVB Myc levels were standardized to 1.0 in order to reveal the degree of Myc overexpression in the transgenic lines.

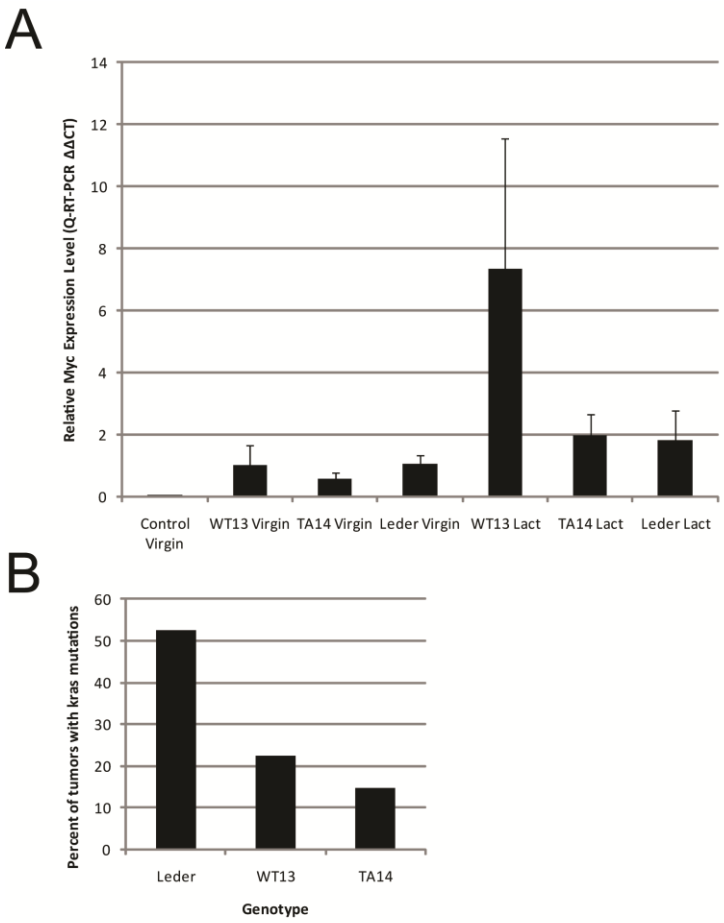


FIGURE S 1.2: THE RELATIONSHIP BETWEEN MYC EXPRESSION AND KRAS MUTATIONS

Quantitative RT-PCR reveals relative levels of Myc expression in virgin and lactating mammary glands of WT13, Leder Myc, and TA14 strains. Leder Myc refers to the original line of MMTV-Myc mice. Expression levels were standardized to the virgin WT13 strain and error bars represent the standard deviation between the three experimental replicates (A). The percentage of tumors with KRas mutations in Leder Myc, WT13, and TA14 strains (B).

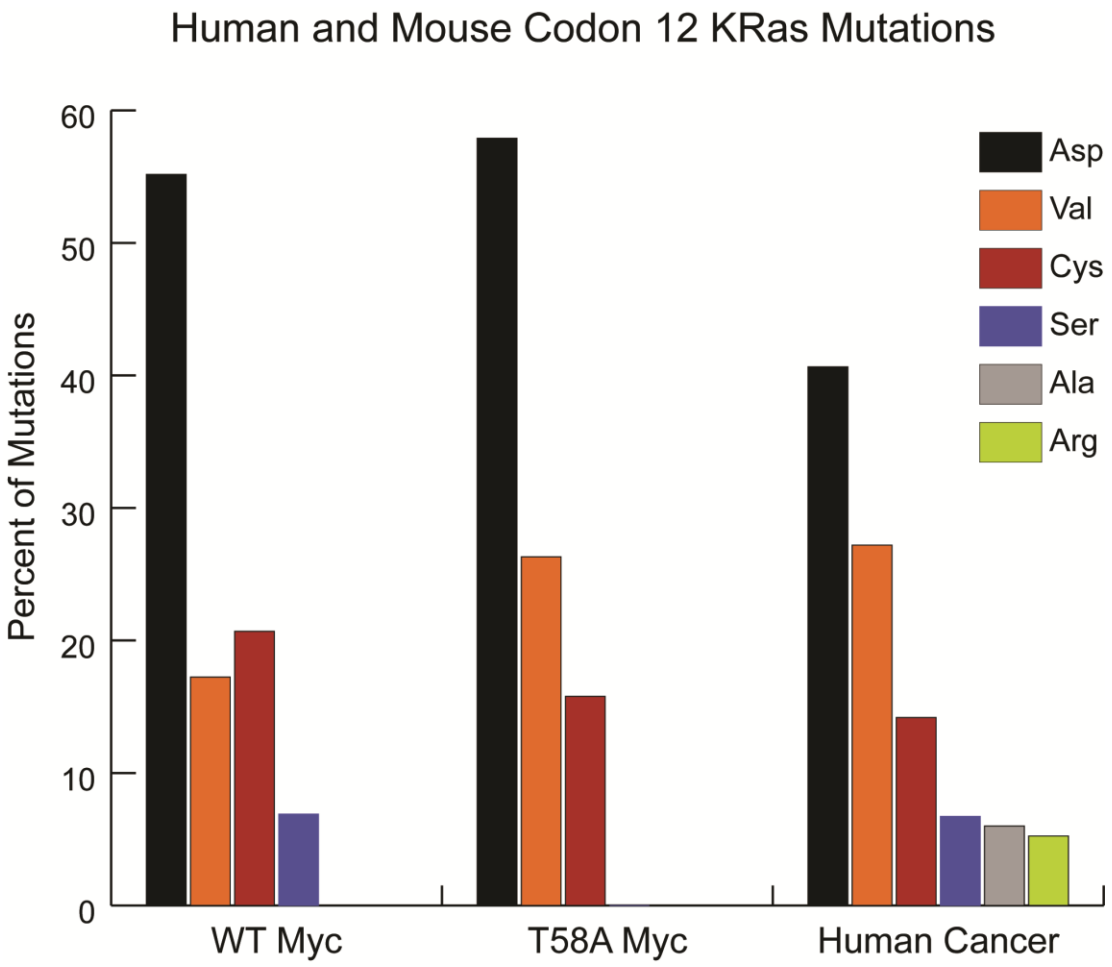


FIGURE S 1.3: TYPES OF ACTIVATING MUTATIONS IN KRAS IN MYC INDUCED TUMORS AND HUMAN BREAST CANCER

The percentage and type of KRas mutations in tumors as reported by the COSMIC database (9056 total tumors) for human cancers as compared with wild type MMTV-Myc (33 tumors) and MMTV-Myc T58A (24 tumors).

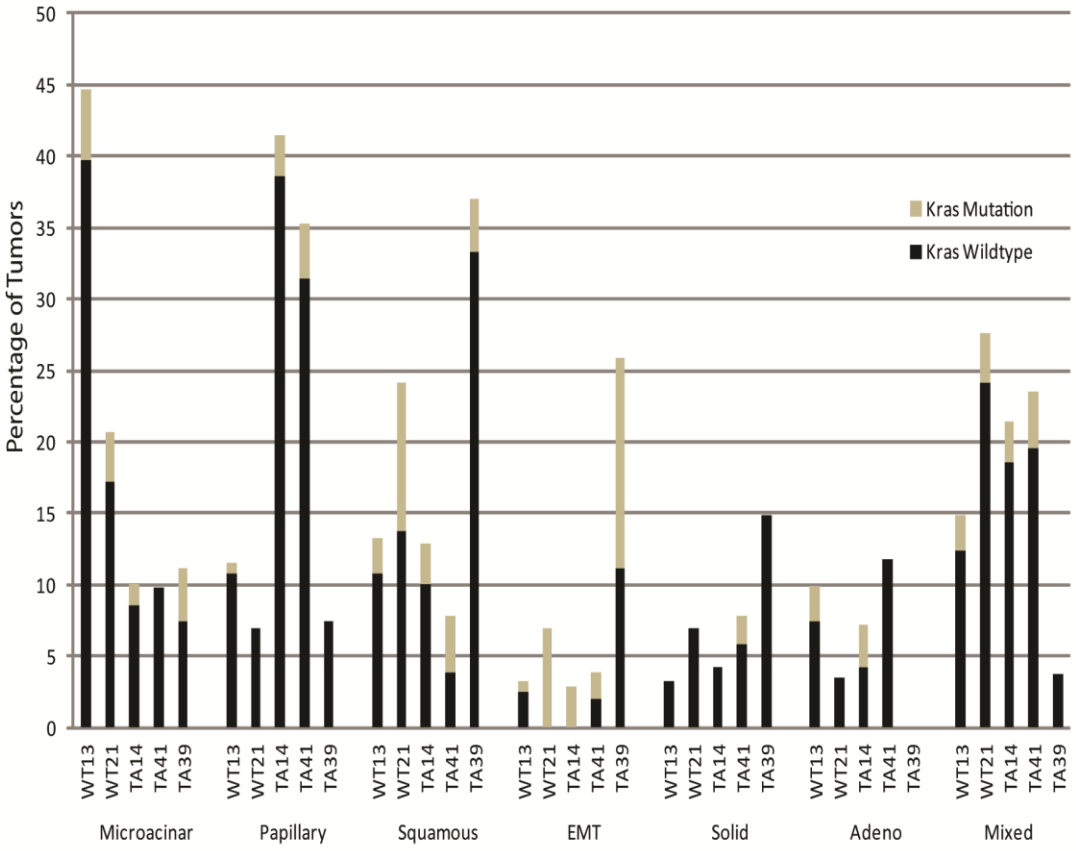


FIGURE S 1.4: KRAS MUTATIONS OCCUR MOST FREQUENTLY IN THE TA39 STRAIN OF EMT TUMORS

The percentage of tumors featuring KRas mutations for each histological type is shown for the individual strains of MMTV-Myc tumor prone mice.

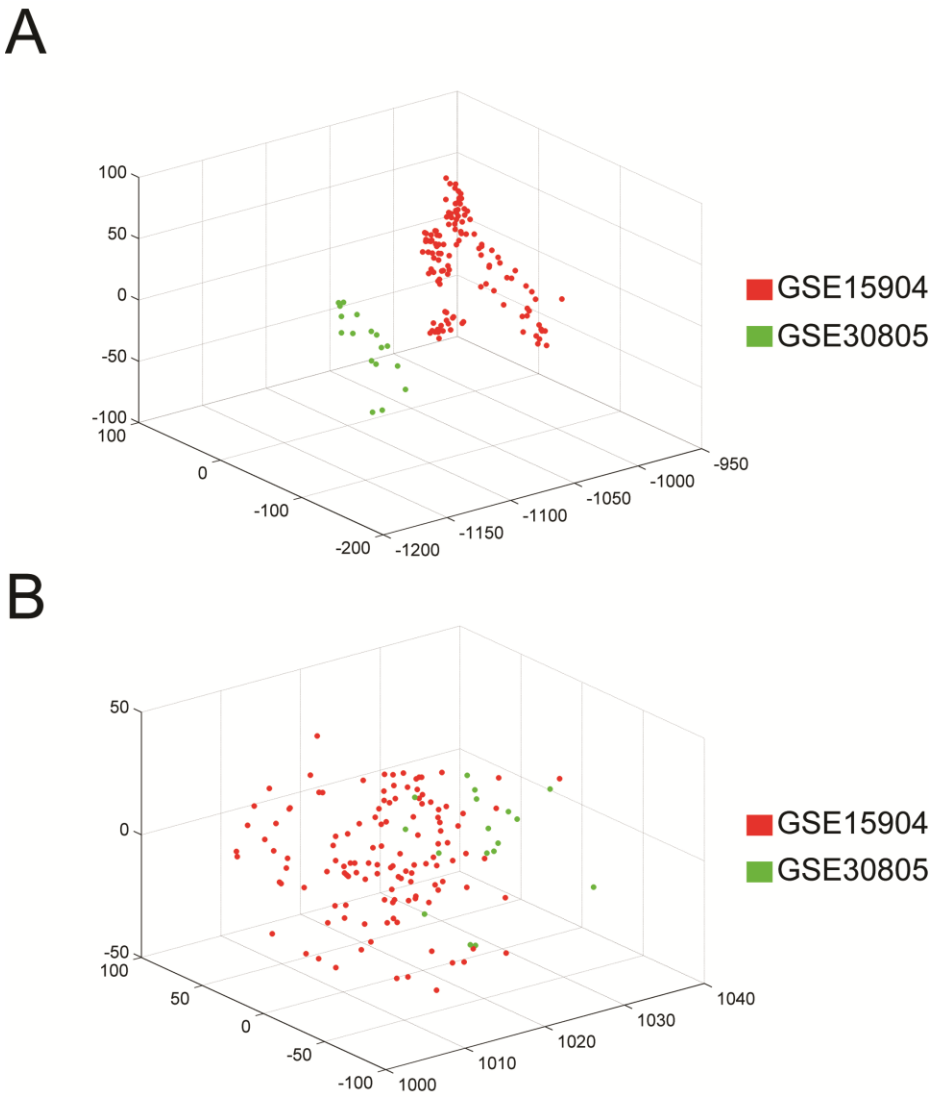


FIGURE S 1.5: PRINCIPLE COMPONENTS ANALYSIS OF GENE EXPRESSION DATA FROM MYC INDUCED TUMORS

Batch effects between separate array experiments are illustrated with a principle components analysis(A) . Using Bayesian factor regression methods, batch effects were removed as illustrated in the principle components analysis (B).

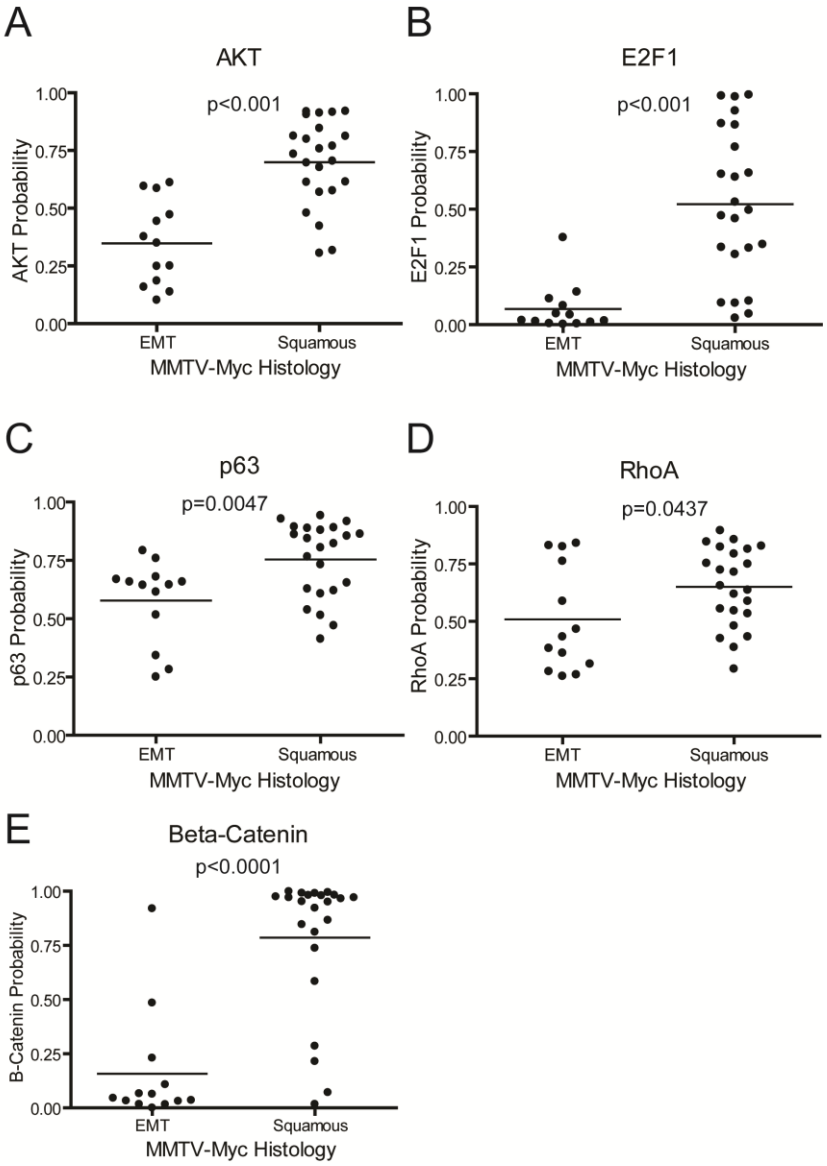


FIGURE S 1.6: STATISTICAL ANALYSIS OF PATHWAY PROBABILITIES IN EMT AND SQUAMOUS TUMORS

A statistical comparison between EMT and squamous tumors for AKT ($p<0.001$) (A), E2F1 ($p<0.001$)(B), p63($p=0.0047$)(C), and RhoA ($p=0.0437$)(D) signal pathway activation probability (A-D respectively).

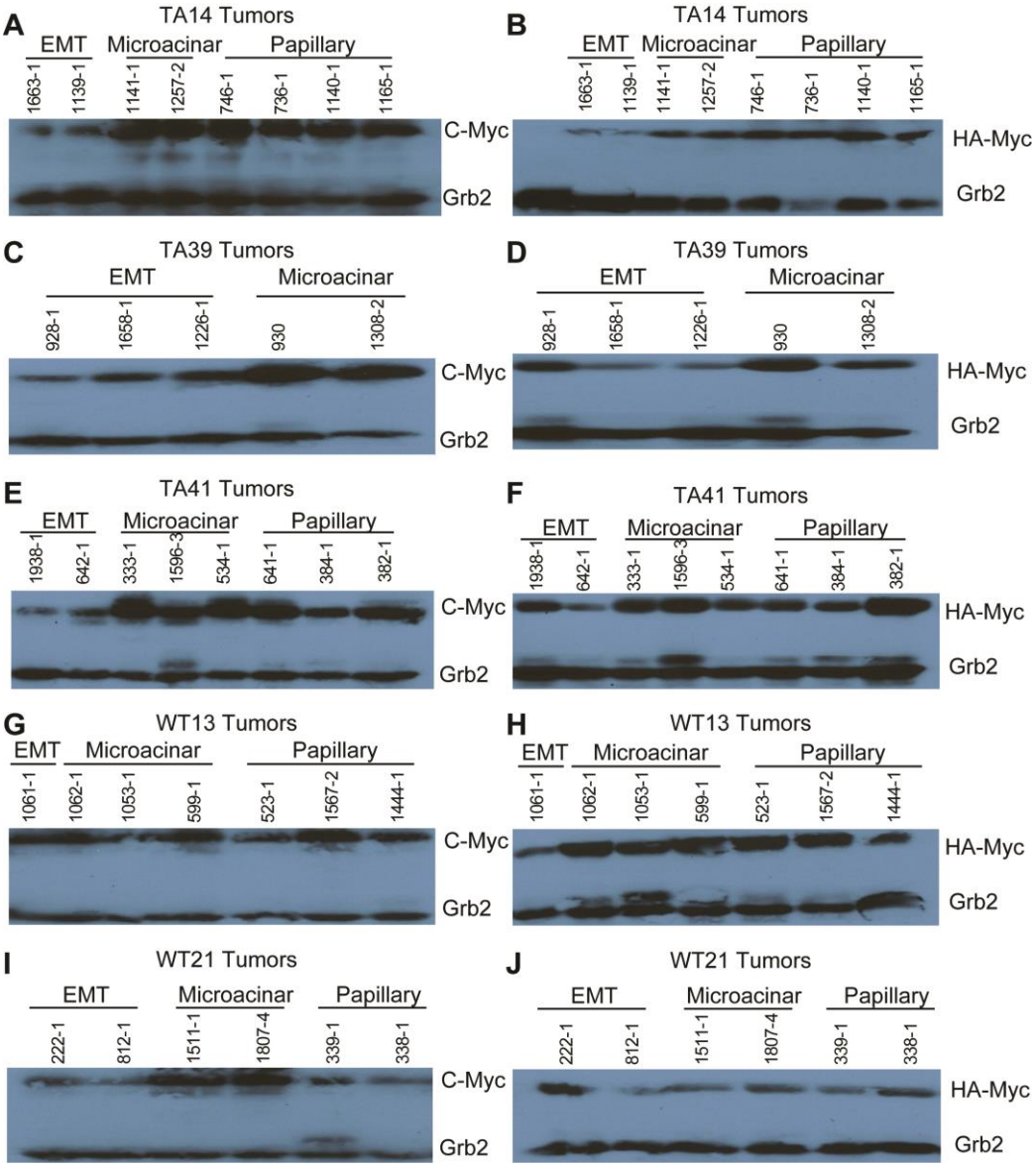


FIGURE S 1.7: THE STABILIZATION OF MYC IN T58A MAMMARY GLANDS AND THE DECREASE OF MYC PROTEIN LEVELS IN EMT-TYPE TUMORS

A Western blot analysis of total Myc protein and exogenous HA-tagged Myc for tumors from the various strains of MMTV-Myc mice is shown (A-J). Total Myc is shown on the left and HA-Myc on the right for EMT, microacinar and papillary tumors (when available) for each strain.

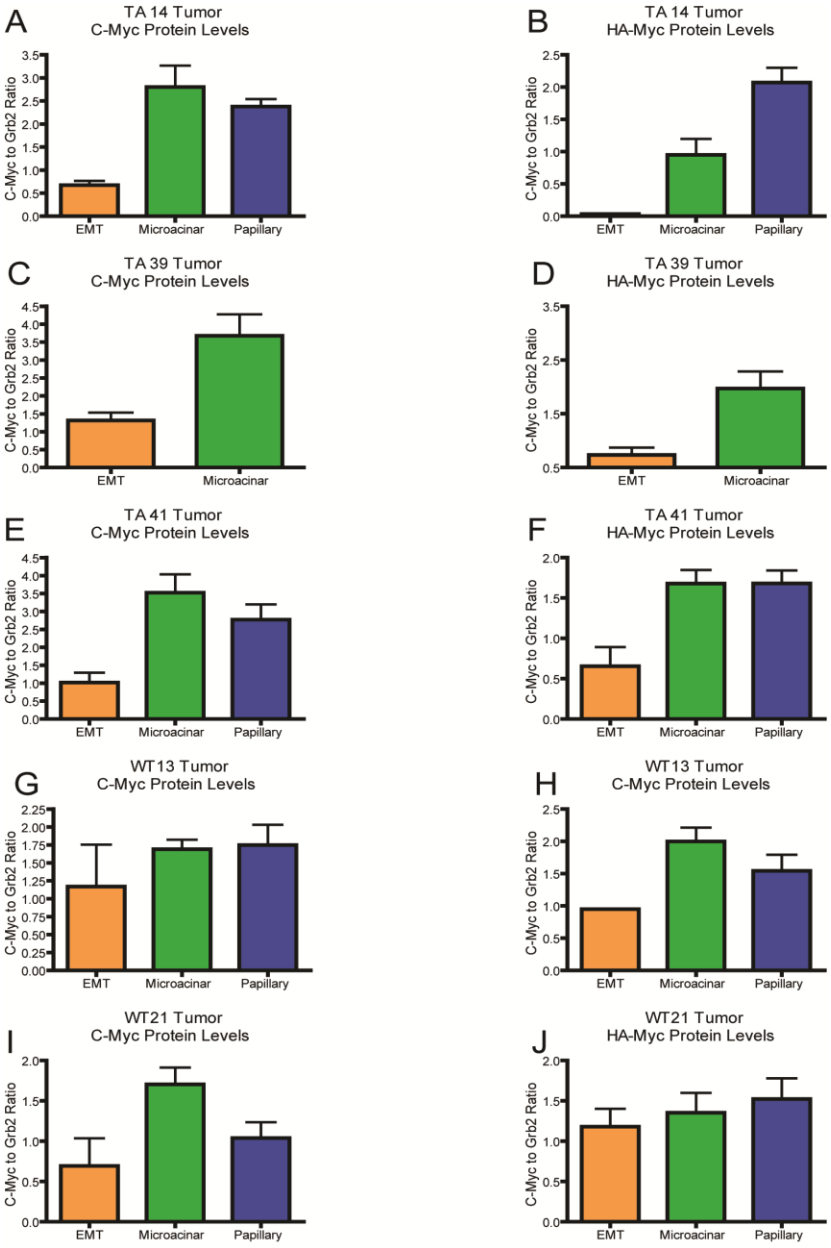


FIGURE S 1.8: QUANTIFICATION OF WESTERN BLOT ANALYSIS REVEALS STABILIZATION OF MYC IN T58A MAMMARY GLANDS AND LOSS OF MYC IN EMT-TYPE TUMORS

Quantification of protein levels from FIGURE S 1.7 for total Myc and HA-Myc are shown for tumors from the various strains of MMTV-Myc mice, in the same order as the previous figure (A-J).

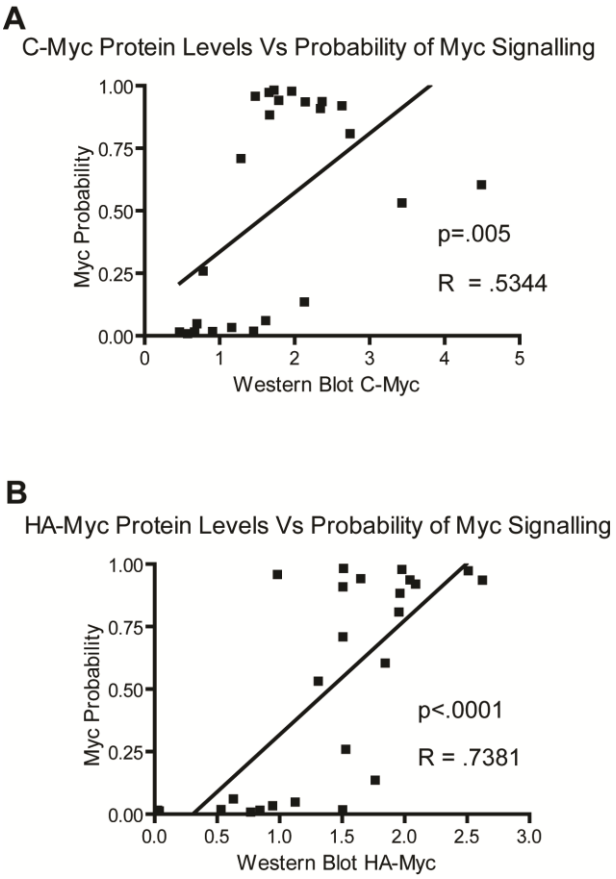
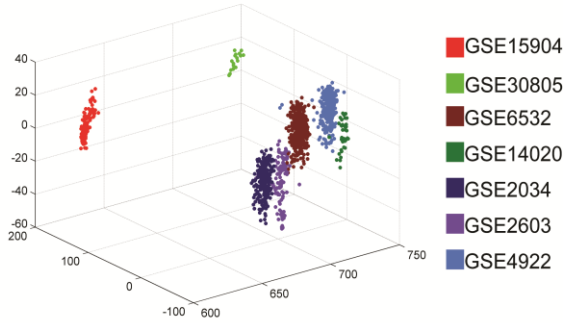


FIGURE S 1.9: PROTEIN LEVELS OF TOTAL AND EXOGENOUS MYC CORRELATE WITH GENE SIGNATURES OF MYC PATHWAY ACTIVATION

Protein levels of total Myc significantly correlate (p=.005, R=.5344) with predicted Myc signaling pathway activation by application of gene signatures (A). Similarly, protein levels of HA-Myc also significantly correlate (p<.0001, R=.7381) with predicted Myc signal pathway activation by application of gene signatures (B).

A



B

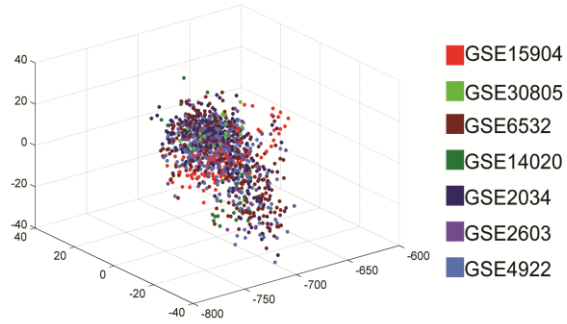


FIGURE S 1.10: PRINCIPLE COMPONENTS ANALYSIS OF GENE EXPRESSION DATA FROM MYC INDUCED TUMORS AND HUMAN BREAST CANCER

Batch and platform effects between mouse tumor and human breast cancer gene expression data is illustrated with a principle components analysis (A). Using Bayesian factor regression methods, batch and platform effects were removed as illustrated in the principle components analysis (B).

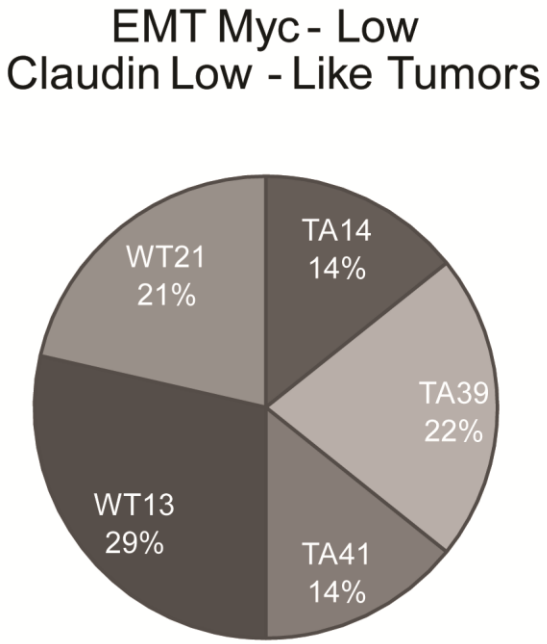


FIGURE S 1.11: VARIOUS STRAINS OF MMTV-MYC MICE DEVELOP EMT-TYPE TUMORS THAT ARE SIMILAR TO MYC-LOW HUMAN CLAUDIN LOW BREAST CANCER

MMTV-Myc tumors that are EMT-type and are similar to the Myc-low claudin low human breast cancer subtype contain a demographic of 21% of tumors from the WT21 strain, 29% of tumors from the WT13 strain, 14% of tumors from the TA41 strain, 22% of tumors from the TA39 strain, and 14% of tumors from the TA14 strain.

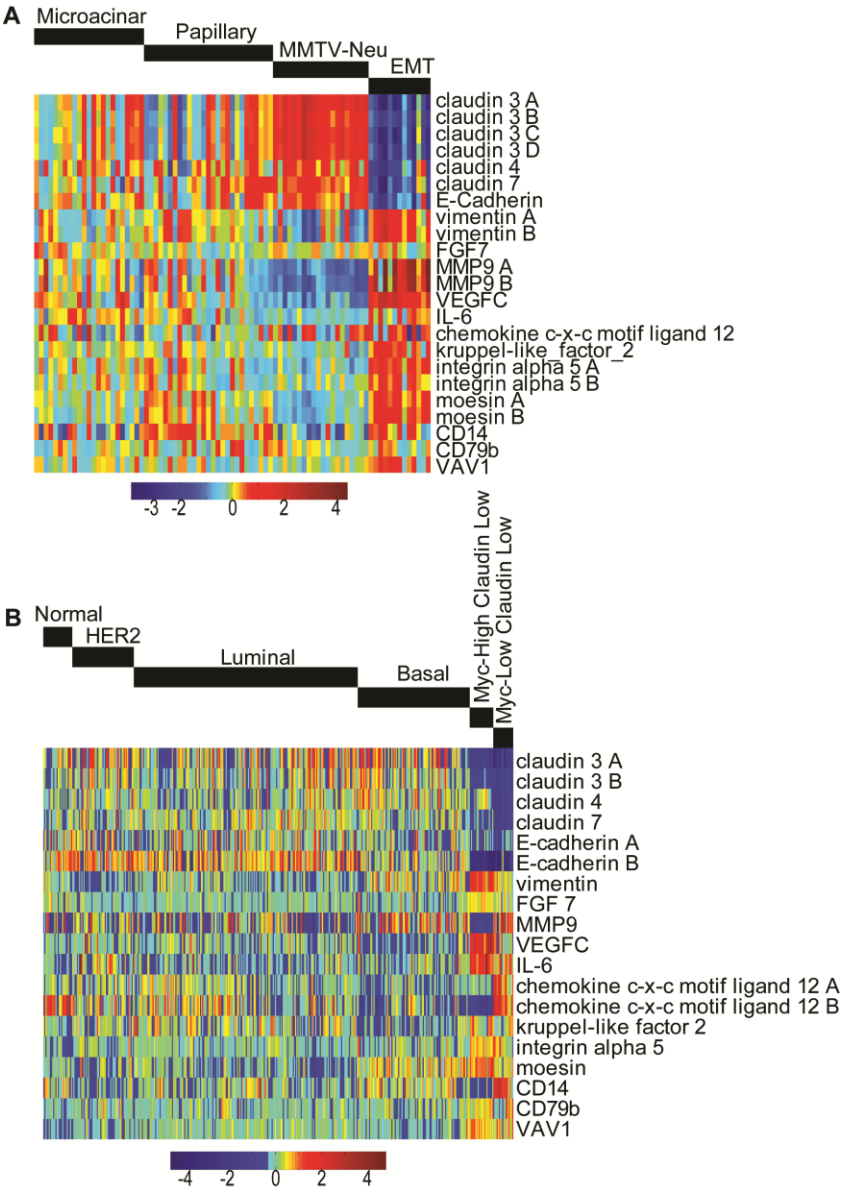


FIGURE S 1.12: EXPRESSION OF HUMAN CLAUDIN LOW TUMOR MARKERS IN MYC INDUCED EMT-TYPE TUMORS AND CLAUDIN LOW TUMORS

A comparison of various tumor types of mouse mammary tumors for RMA normalized gene expression levels of markers of human claudin low tumors (A). Similarly a comparison between intrinsic subtypes of human breast tumors for expression levels of markers for human claudin low tumors (B).

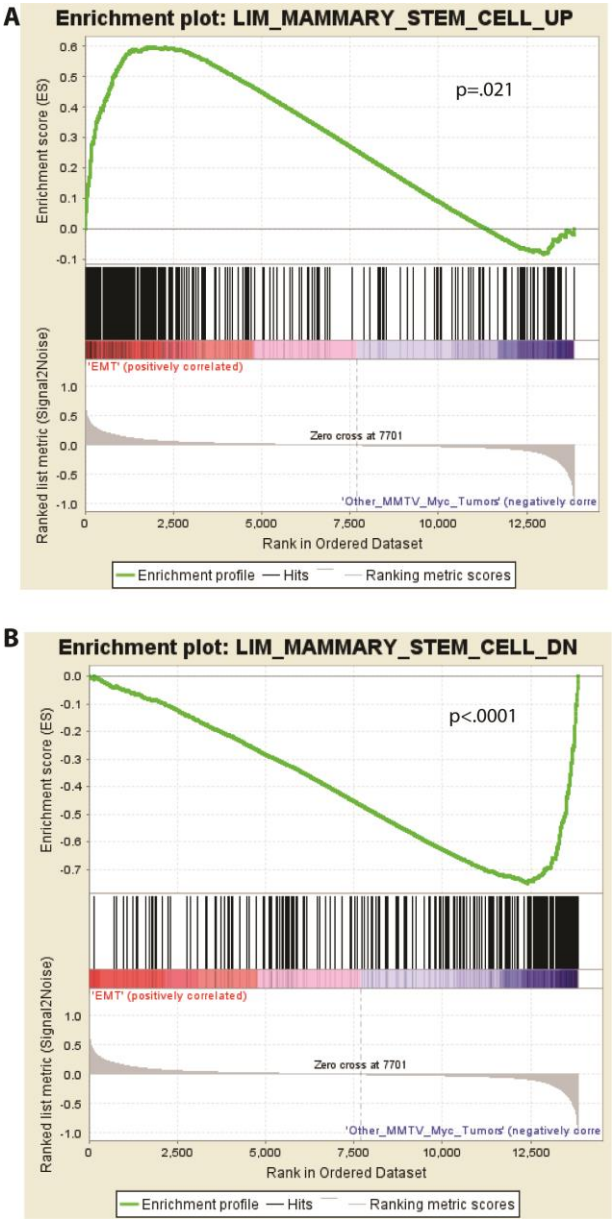


FIGURE S 1.13: MMTV-MYC TUMORS OF THE EMT-TYPE FEATURE STEM CELL-LIKE PROPERTIES

Gene set enrichment analysis comparing EMT-types of MMTV-Myc tumors with other histological types of Myc induced using a signature for genes that are upregulated in mammary stem cells (A). Likewise, MMTV-Myc tumors of the EMT-type compared to other histological

FIGURE S 1.13 (cont'd)

types of Myc induced tumors using a signature for genes that are downregulated in mammary stem cells(B).

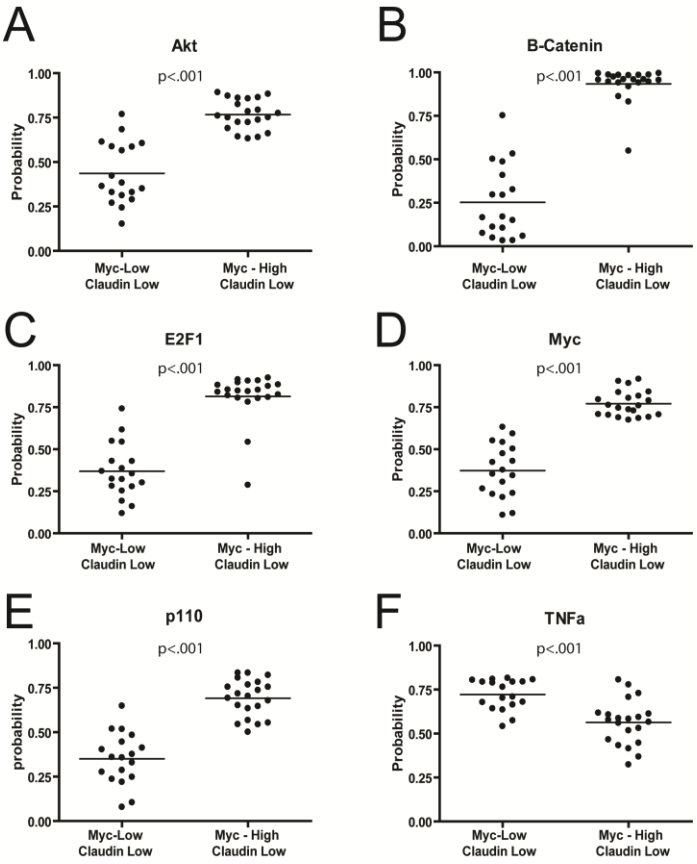


FIGURE S 1.14: STATISTICAL ANALYSIS OF PATHWAY PROBABILITIES IN MYC-LOW CLAUDIN LOW TUMORS AND MYC-HIGH CLAUDIN LOW TUMORS

A comparison between the Myc-low and Myc-High human claudin low tumors statistical differences in predicted pathway activation for Akt, β-catenin, E2F1, Myc, p110, and TNFα (A-F respectively) probabilities.

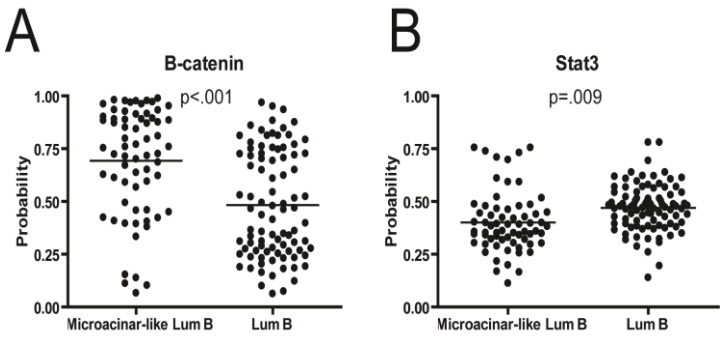


FIGURE S 1.15: STATISTICAL ANALYSIS OF PATHWAY PROBABILITIES IN MICROACINAR-LIKE LUMINAL B TUMORS AND OTHER LUMINAL B TUMORS

A comparison of β-catenin ($p<0.001$) and Stat3 ($p=0.009$) pathway activation between a cluster of predominantly human Luminal B breast cancers that are similar to Myc induced microacinar tumors and another cluster of mainly human luminal B breast cancer (A,B respectively).

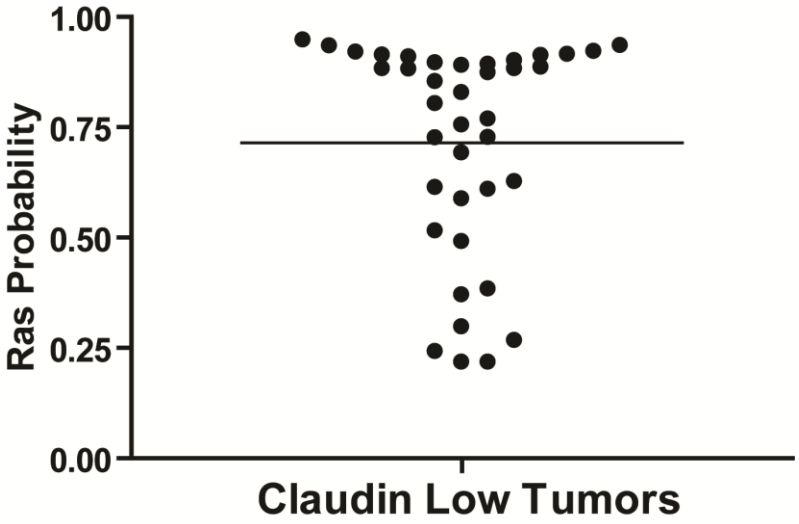


FIGURE S 1.16: HUMAN CLAUDIN LOW BREAST CANCERS HAVE A HIGH PROBABILITY OF RAS SIGNALING PATHWAY ACTIVATION

Ras pathway activation probabilities are shown for human claudin low breast cancer and reveal that nearly 80% of human claudin low tumors have greater than a 0.50 probability of Ras signaling pathway activation.

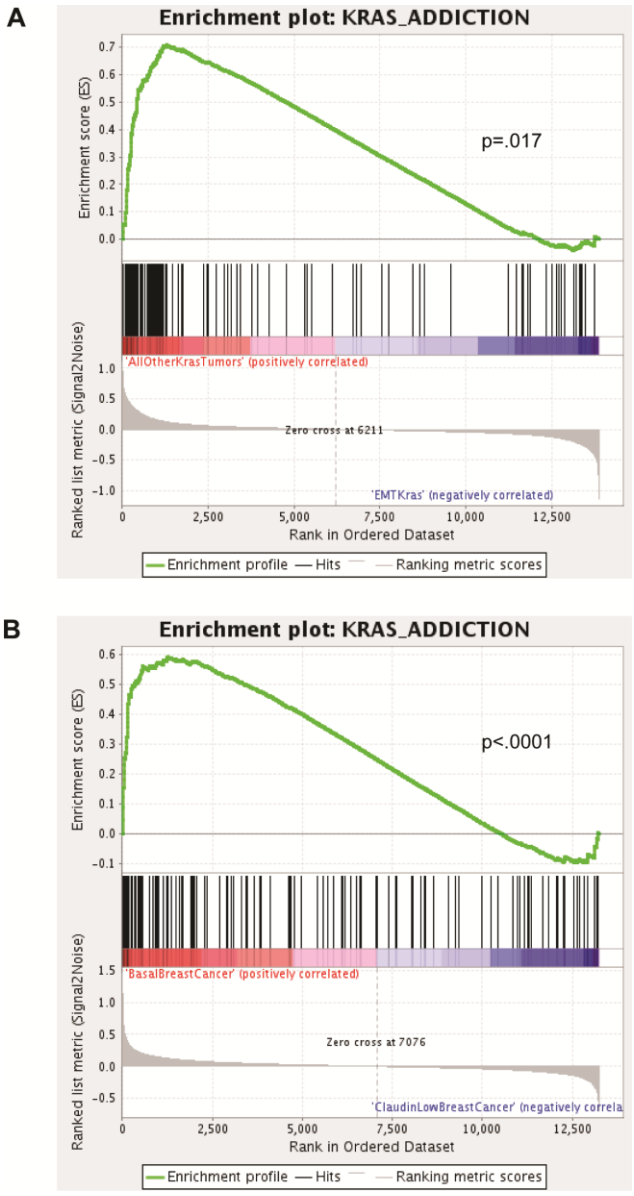


FIGURE S 1.17: HUMAN CLAUDIN LOW BREAST CANCER AND KRAS MUTANT MMTV-MYC TUMORS OF THE EMT-TYPE ARE NOT KRAS ADDICTED

Gene set enrichment analysis with a KRas addiction signature compares MMTV-Myc KRas mutant tumors and MMTV-Myc EMT-type KRas mutant tumors by GSEA using a KRas addiction signature(A). A comparison between human claudin low tumors and basal breast cancer (B).

TABLE 1.1: ACTIVATING MUTATIONS IN KRAS

Myc Type	Mouse	Codon 12 / 13	Protein	Myc Type	Mouse	Codon 12 / 13	Protein
Control	Control	GGTGGC	Gly / Gly				
WT	2811	GATGGC	Asp / Gly	T58A Low	15681	GATGGC	Asp / Gly
WT	14451	GATGGC	Asp / Gly	T58A Low	947	GATGGC	Asp / Gly
WT	14453	GATGGC	Asp / Gly	T58A Low	928	GGTGAC	Gly / His
WT	15682	GATGGC	Asp / Gly	T58A Low	1308	GATGGC	Asp / Gly
WT	18073	GATGGC	Asp / Gly	T58A Low	13082	GATGGC	Asp / Gly
WT	10873	GATGGC	Asp / Gly	T58A Low	9261	TGTGGC	Cys / Gly
WT	5991	GATGGC	Asp / Gly	T58A High	11562	GGTGAC	Gly / His
WT	817	GATGGC	Asp / Gly	T58A High	4931	GGTTGC	Gly / Cys
WT	5052	GATGGC	Asp / Gly	T58A High	12562	GATGGC	Asp / Gly
WT	13562	GATGGC	Asp / Gly	T58A High	21801	GATGGC	Asp / Gly
WT	593	GATGGC	Asp / Gly	T58A High	1139	GATGGC	Asp / Gly
WT	4213	GATGGC	Asp / Gly	T58A High	643	GATGGC	Asp / Gly
WT	222	GATGGC	Asp / Gly	T58A High	13142	GATGGC	Asp / Gly
WT	8402	GATGGC	Asp / Gly	T58A High	956	GATGGC	Asp / Gly
WT	1062	GATGGC	Asp / Gly	T58A High	10972	GATGGC	Asp / Gly
WT	10863	GATGGC	Asp / Gly	T58A High	7302	GATGGC	Asp / Gly
WT	854	TGTGGC	Cys / Gly	T58A High	1938	GATGGC	Asp / Gly
WT	812	TGTGGC	Cys / Gly	T58A High	1798	GATGGC	Asp / Gly
WT	1016	TGTGGC	Cys / Gly	T58A High	957	GATGGC	Asp / Gly
WT	4214	TGTGGC	Cys / Gly	T58A High	11653	GTTGGC	Val / Gly
WT	1855	TGTGGC	Cys / Gly	T58A High	4932	GTTGGC	Val / Gly
WT	1085	TGTGCC	Cys / Gly	T58A High	737	GTTGGC	Val / Gly
WT	5182	GTTGGC	Val / Gly	T58A High	589	GTTGGC	Val / Gly
WT	5183	GTTGGC	Val / Gly	T58A High	384	GTTGGC	Val / Gly
WT	8404	GTTGGC	Val / Gly	T58A High	16631	TGTGGC	Cys / Gly
WT	10525	GTTGGC	Val / Gly	T58A High	642	TGTGGC	Cys / Gly
WT	1003	GTTGGC	Val / Gly	T58A High	384	TGTGGC	Cys / Gly
WT	4212	AGTGGC	Ser / Gly				
WT	10852	AGTGGC	Ser / Gly				
WT	8192	GGTGAC	Gly / His				
WT	1066	GGTCGC	Gly / Arg				
WT	10532	GGTCGC	Gly / Arg				

Myc Type	Mouse	Codon 61	Protein
Control	Control	CAA	Gln
WT	10882	CGA	Arg
T58A High	3332	CAT	His
T58A High	3333	CAC	His
T58A High	1315	CGA	Arg

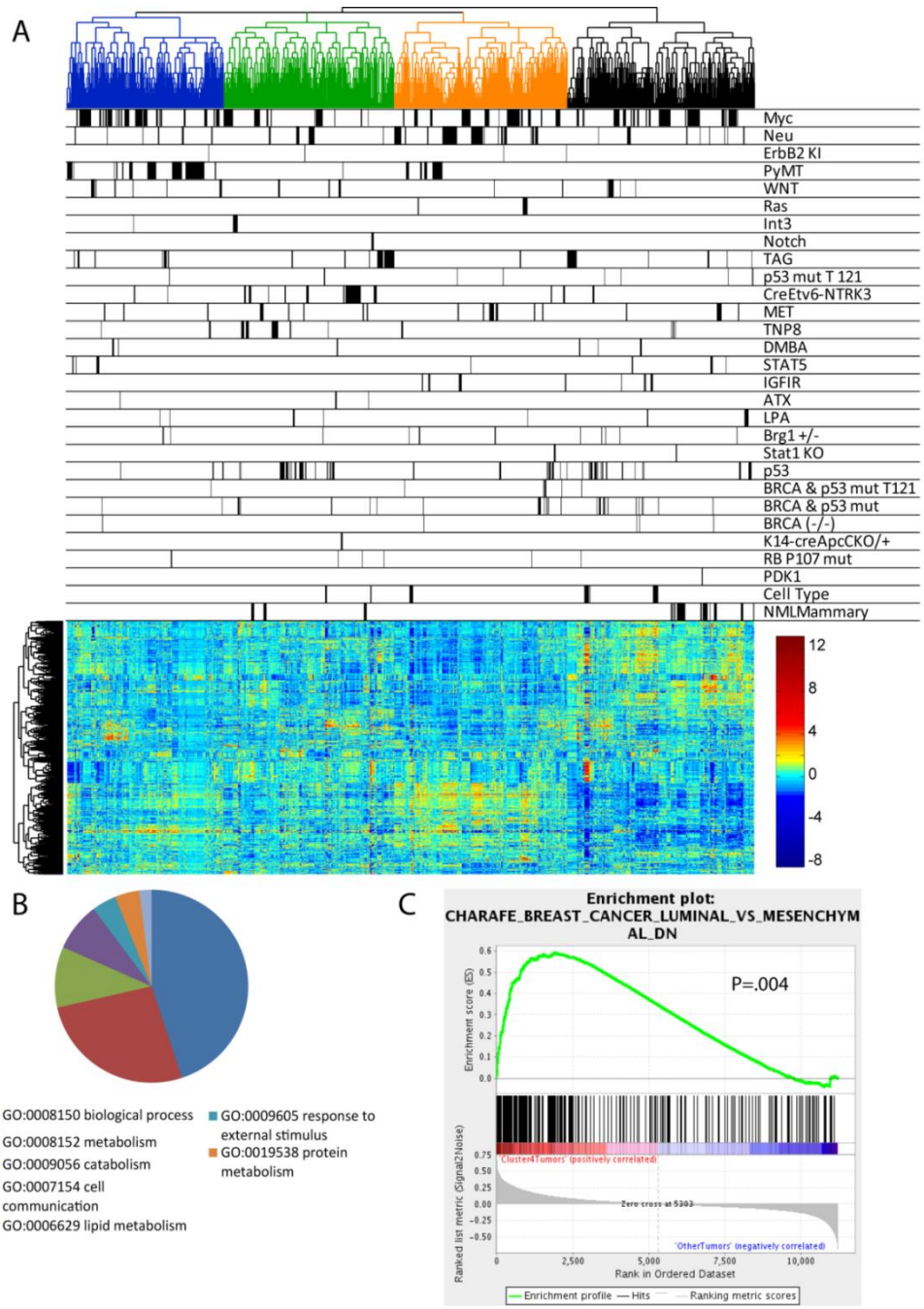


FIGURE 2.1: ANALYSIS OF RELATIONSHIPS BETWEEN MOUSE MAMMARY TUMOR MODELS

(A)The unsupervised hierarchical clustering analysis of gene expression data for mouse mammary tumors, cell types, and normal mammary gland is shown. The dendrogram across the top illustrates relationships between samples and is color-coded to itemize the four main clusters.

FIGURE 2.1 (cont'd)

Below the dendrogram, black bars label samples from each corresponding model on the same line. Gene expression values are illustrated with the heatmap, according to the scale shown. The vertical dendrogram beside the heatmap illustrates genes with similar patterns of expression across the samples in the dataset. (B) The pie chart illustrates the gene ontologies of the genes that are significantly (q=0, fdr=0) overexpressed as identified by SAM in the blue cluster of tumors compared to tumors in other clusters. (C) The gene set enrichment plot comparing tumors from cluster 4 (black) to tumors in the other clusters shows significant enrichment for high expression of a gene set that defines mesenchymal breast cancer (p=.004).

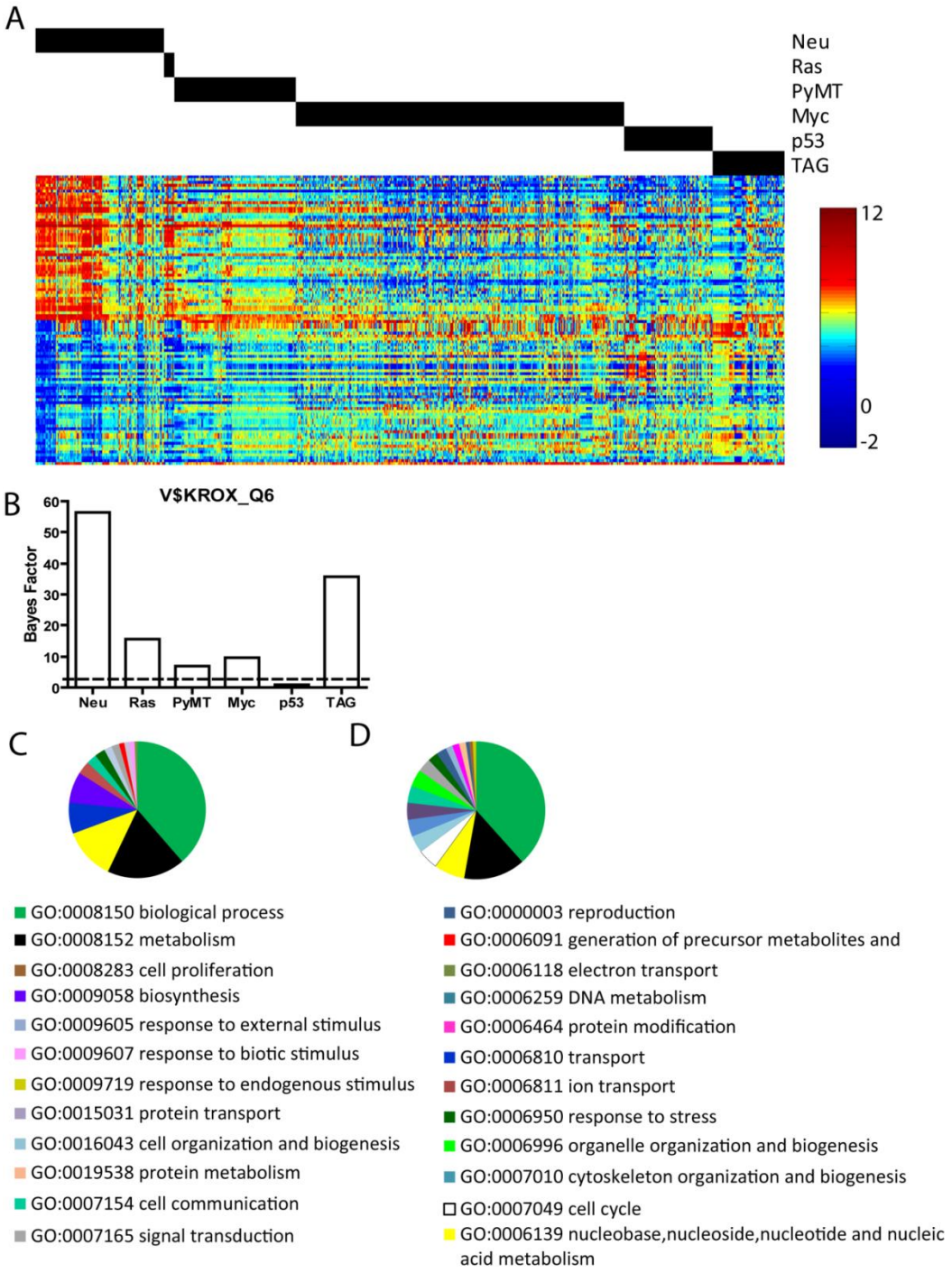


FIGURE 2.2: FOLD CHANGE ANALYSIS OF NEU INDUCED TUMORS COMPARED TO OTHER TUMOR MODELS

(A)The expression pattern for the top 50 significantly (q=0, fdr=0) upregulated and down regulated genes for Neu-induced tumors as identified by SAM are illustrated with the heatmap.

FIGURE 2.2 (cont'd)

Above the heatmap, black bars denote the model each sample corresponds to. Expression levels are depicted according to the color bar beside the heatmap. (B) The bar graph shows the Bayes factor measuring the enrichment of predicted binding sites for the Krox family of transcription factors within upregulated genes from each model. The dotted line indicates a Bayes factor of 2.0. (C). Gene ontologies for upregulated genes in Neu induced tumors are depicted in the pie chart according to the color-coded categories. (D) Gene ontologies for upregulated genes in TAG induced tumors are depicted in the pie chart according to the listed color-coded categories.

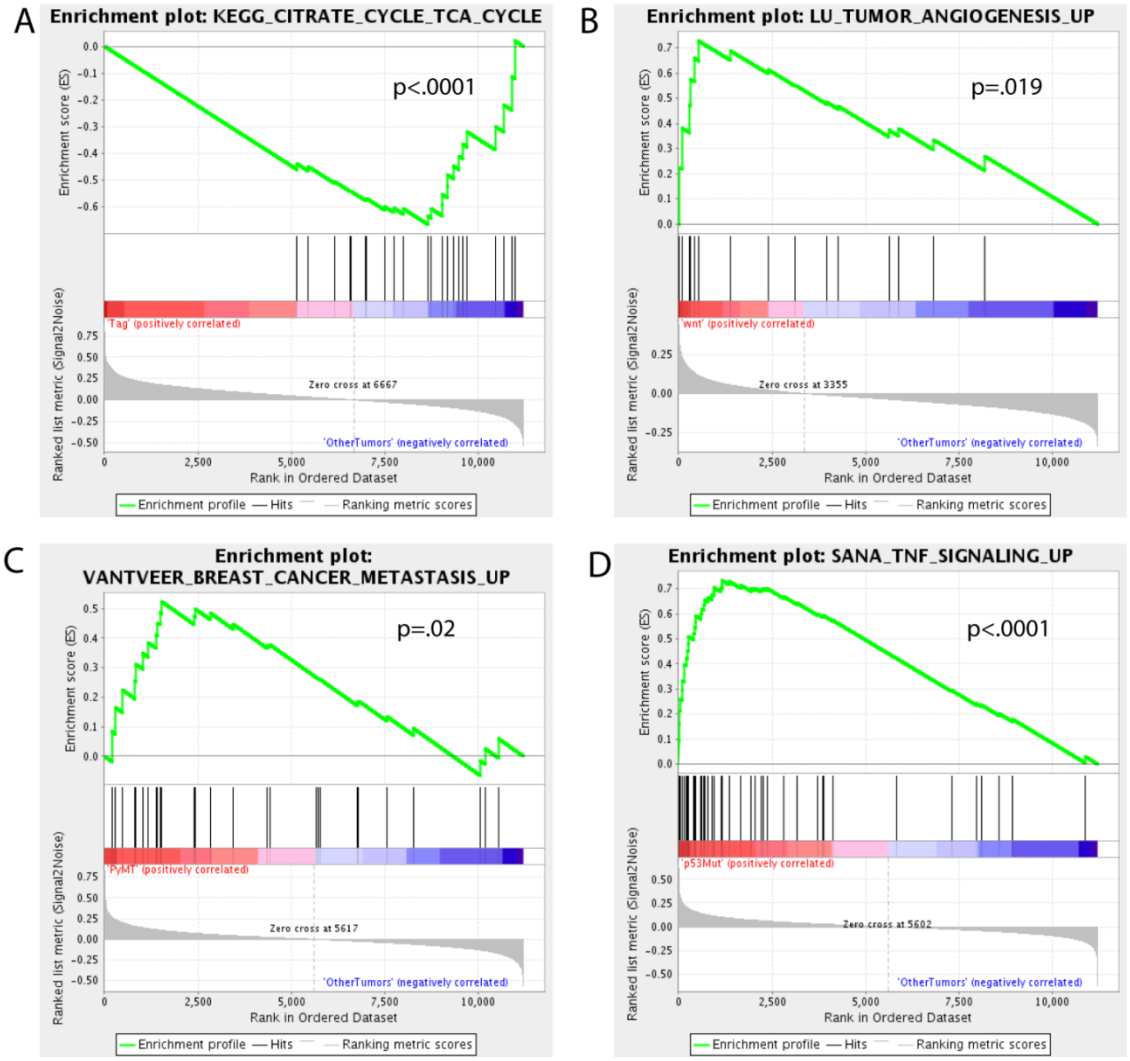


FIGURE 2.3: GENE SET ENRICHMENT ANALYSIS OF MOUSE MAMMARY TUMOR MODELS

(A)Gene set for genes involved in the TCA cycle are significantly enriched (P<.0001) for low expression in TAG tumors. (B) A gene set for genes upregulated during tumor angiogenesis are significantly enriched (p=.019) for high expression in Wnt induced tumors. (C)A gene set for genes upregulated in breast cancer metastasis is significantly enriched (p=.02) for high expression in PyMT induced tumors. (D) A gene set for genes that upregulated as a result of TNF signaling is significantly enriched (p<.0001) for high expression in p53 mutant tumors.

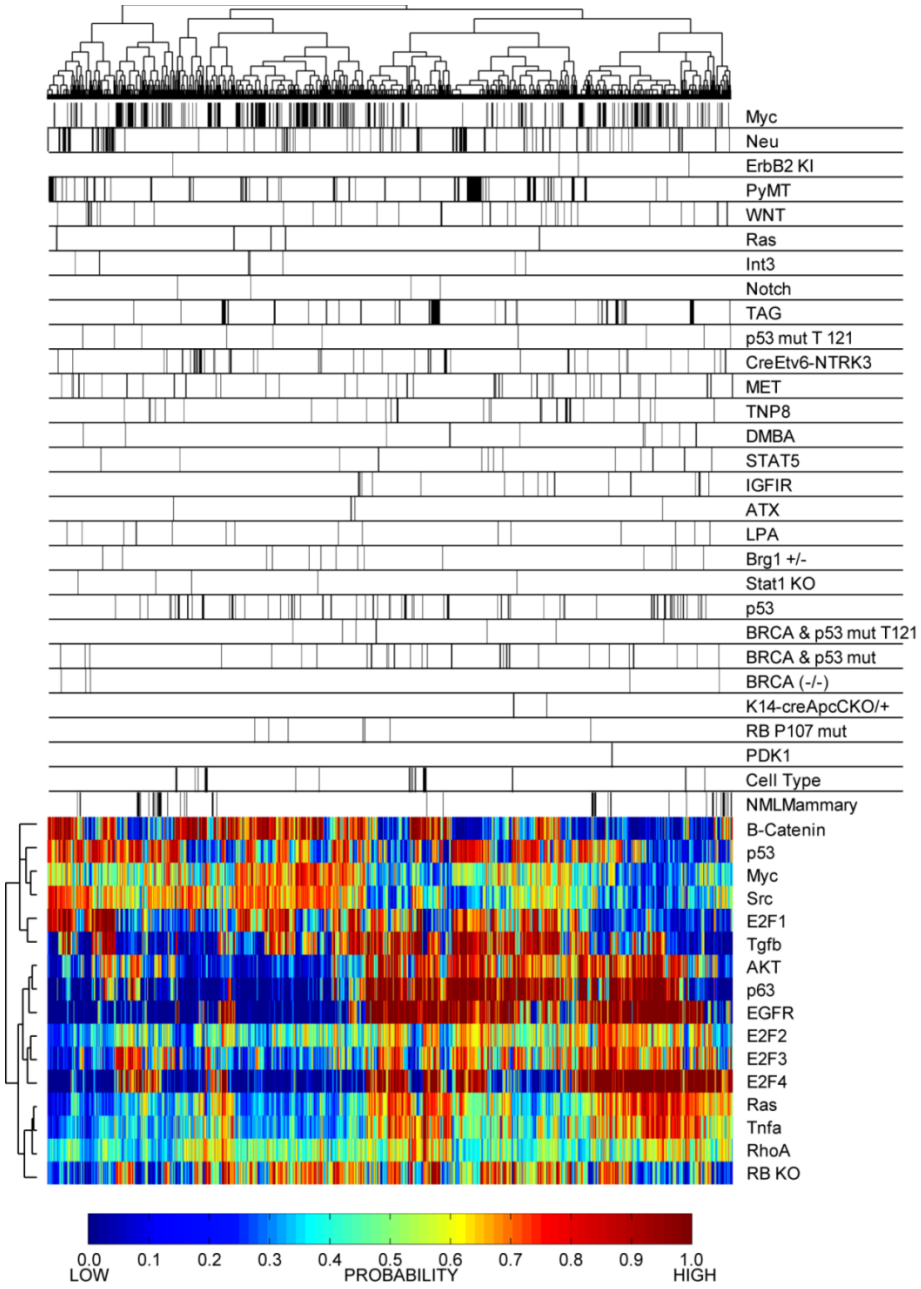


FIGURE 2.4: UNSUPERVISED HIERARCHICAL CLUSTERING OF PATHWAY ACTIVATION PREDICTIONS IN MOUSE MAMMARY TUMORS

The dendrogram across the top illustrates the relationship between samples based on predicted pathway activation profiles. Below the dendrogram, the black bars mark tumor samples corresponding to the model listed on the same line. The heatmap illustrates the probability of

FIGURE 2.4 (cont'd)

pathway activation according to the color bar provided below the heatmap. The vertical dendrogram beside the heatmap illustrates pathways with similar predicted activity across the samples in the dataset.

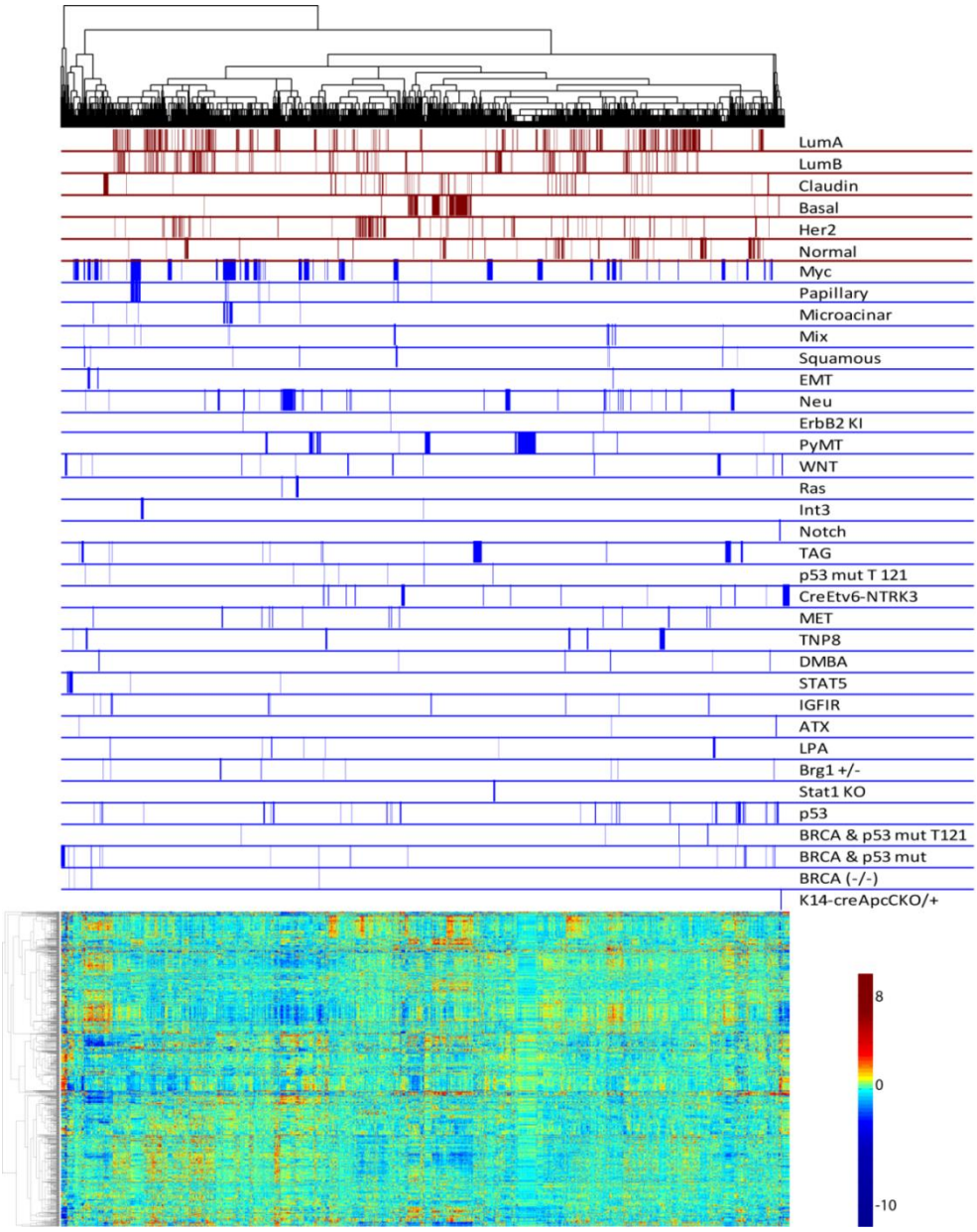


FIGURE 2.5: UNSUPERVISED HIERARCHICAL CLUSTERING OF MOUSE MAMMARY TUMOR AND HUMAN BREAST CANCER GENE EXPRESSION DATA

Across the top, the dendrogram illustrates the relationship between human and mouse tumor samples on the basis of gene expression profiles. The red bars mark the intrinsic subtype of each human tumor sample according the annotation on the same line. The blue bars correspond to the mouse mammary tumor type. Below this, a heatmap shows the gene expression patterns for each

FIGURE 2.5 (cont'd)

sample, with expression values illustrated according to the color bar on the right. The dendrogram beside the heatmap shows the correlation between genes based on expression patterns across the samples in the dataset.

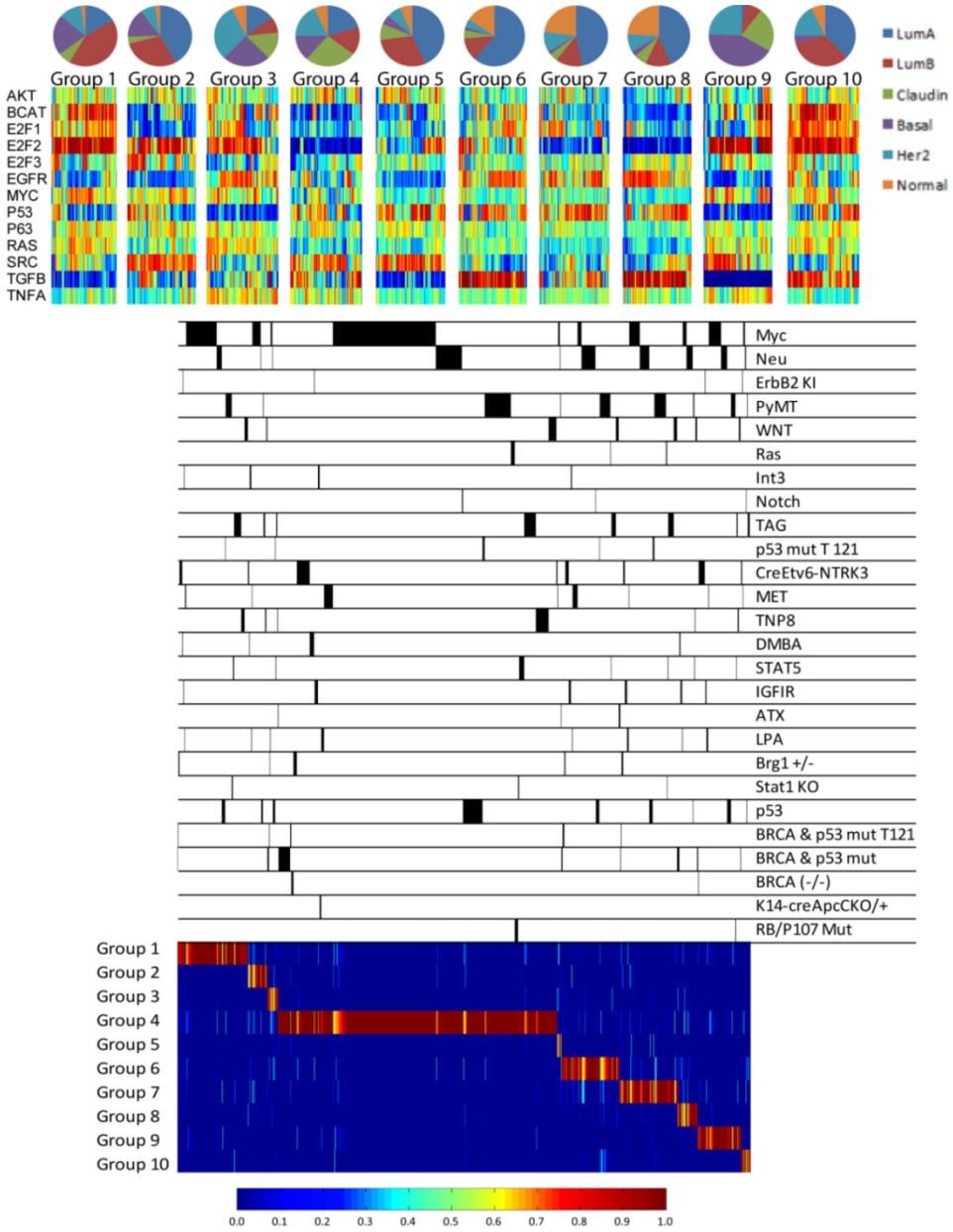


FIGURE 2.6: MIXTURE MODELING ANALYSIS OF HUMAN BREAST CANCER PATHWAY HETEROGENEITY AND RELATIONSHIPS TO MOUSE MODELS OF BREAST CANCER

Pie charts above each heatmap illustrate the distribution of the intrinsic subtype of samples in each group, according to the color-coded legend. The heatmap for groups 1-10 show predicted pathway activity with probabilities corresponding to the color bar at the bottom of the figure.

FIGURE 2.6 (cont'd)

Below this black bars mark the samples corresponding to annotations on the same line. Following the samples down to the heatmap below the black bars, the probability that a mouse model has similar pathway activation profiles is shown for each group. Probabilities for this heatmap are shown according to the color bar at the bottom of the figure.

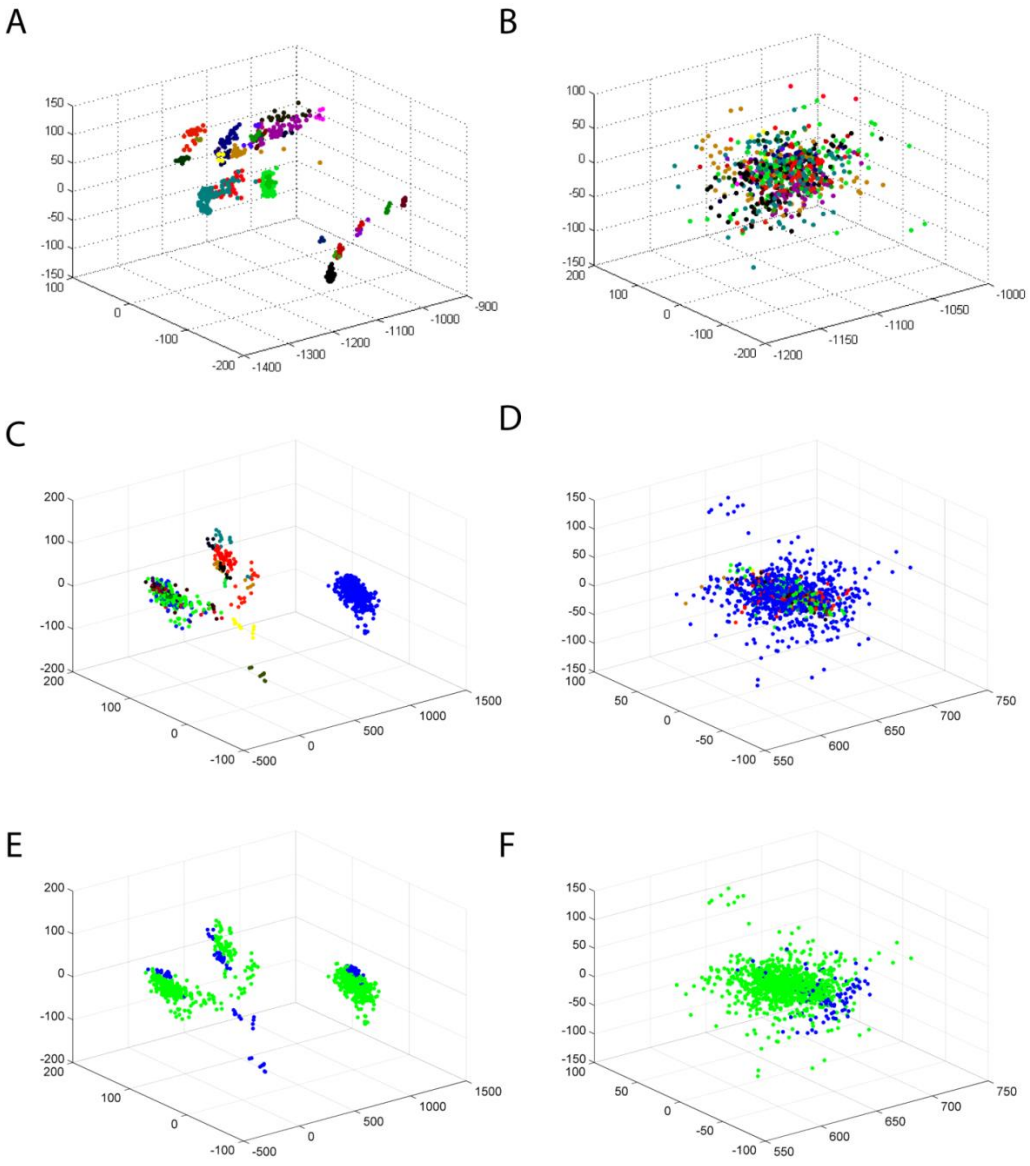


FIGURE S 2.1: REMOVAL OF BATCH EFFECTS FROM AFFYMETRIX DATASETS

(A)Affymetrix datasets color coded according to the study of origin in a principle components analysis plot prior to BFRM batch effect correction. (B) Affymetrix datasets color coded according to the study of origin in a principle components analysis plot after BFRM batch effect correction. (C) Affymetrix datasets are color coded together in blue after BFRM batch effect correction, the various Agilent gene expression datasets are color-coded and plotted along with

FIGURE S 2.1 (cont'd)

Affymetrix data on the three principle components to illustrate platform and batch variance. (D) Agilent and Affymetrix color-coded data plotted after COMBAT removed batch and platform technical variance.(E) Neu-induced tumors are color coded in blue and all other tumors are in green, illustrating variance between similar tumor types on the basis of platform and batch artifacts. (F) Neu-induced tumors are color coded in blue and all other tumors are in green illustrating mediation of batch and platform effects.

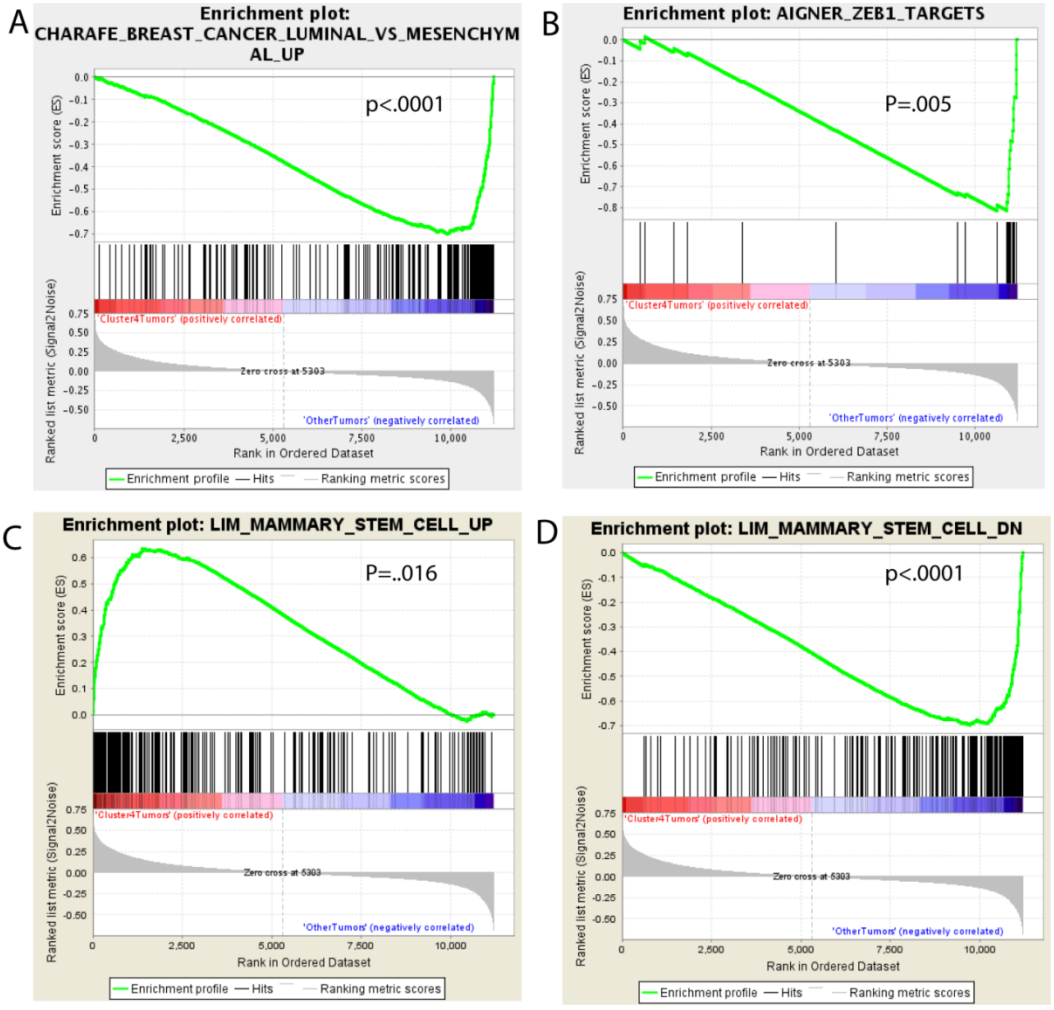


FIGURE S 2. 2: GENE SET ENRICHMENT ANALYSIS FOR MOUSE MAMMARY TUMORS IN THE BLACK COLOR-CODED CLUSTER

(A)A gene set for downregulated genes in mesenchymal breast cancer is significantly enriched (p<.0001) and downregulated in the black cluster (cluster4) of tumors. (B) A gene set for Zeb1 target genes is significantly enriched (p=.005) for low expression for the tumors in the black cluster. (C) A gene set for genes highly expressed in mammary stem cells is significantly enriched (p=.016) and upregulated in tumors from cluster 4 (black). (D) A gene set for genes that are downregulated in mammary stem cells is significantly enriched (p<.0001) and also downregulated in the cluster 4(black) tumors.

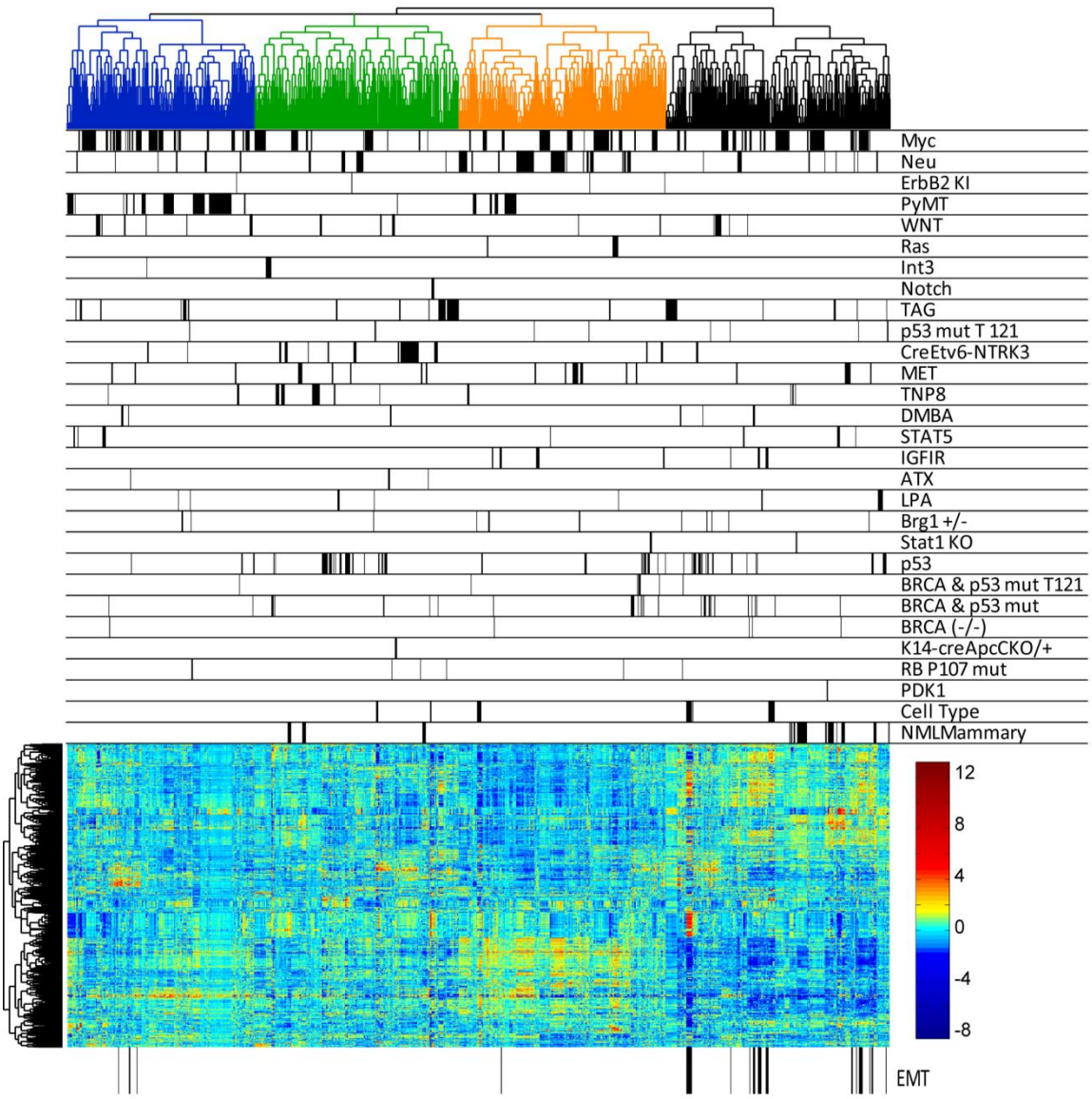


FIGURE S 2.3: TUMORS THAT WERE CLASSIFIED FOR MESENCHYMAL HISTOLOGY CLUSTER INTO THE BLACK CLUSTER

Highlighting prior histological annotations for mesenchymal or EMT-like tumors across the Myc, IGF-IR, DMBA, and p53 mutant models show that a large majority of these tumors cluster together in the black cluster.

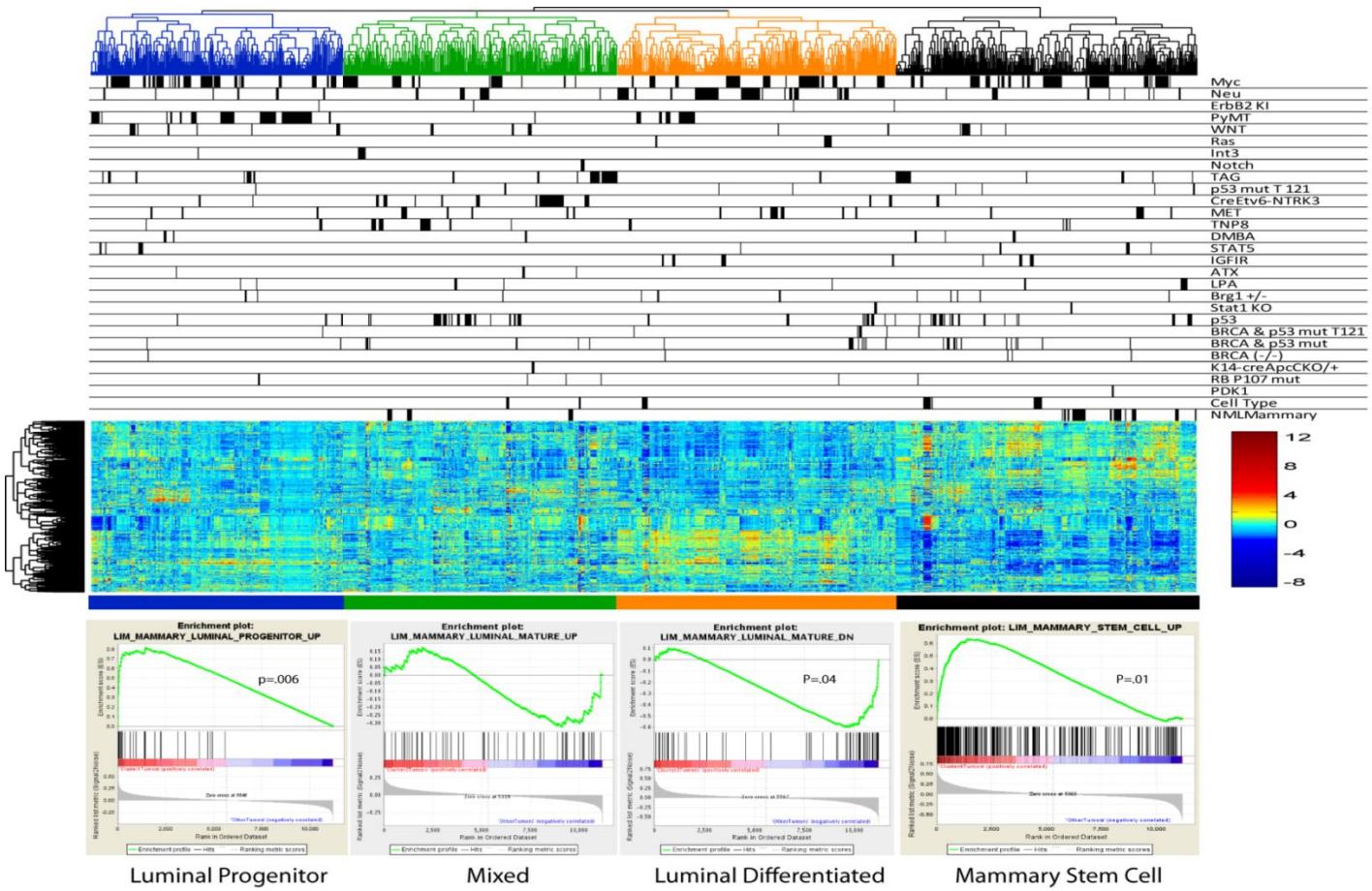


FIGURE S 2.4: GENE SET ENRICHMENT ANALYSIS FOR MAMMARY CELL TYPES ACROSS MAJOR CLUSTERS OF MOUSE MAMMARY TUMORS

GSEA for tumors in blue cluster compared to all other clusters show significant enrichment for a mammary luminal progenitor cell gene expression signature (p=.006). Similarly, tumors from the green cluster associate with a mixture of luminal cell gene expression

FIGURE S 2.4 (cont'd)

features, while tumors in the orange cluster are significantly enriched for gene expression features of mature luminal cells (p=.04). Lastly, tumors in the black cluster are significantly enriched for gene expression features of mammary stem cells (p=.01).

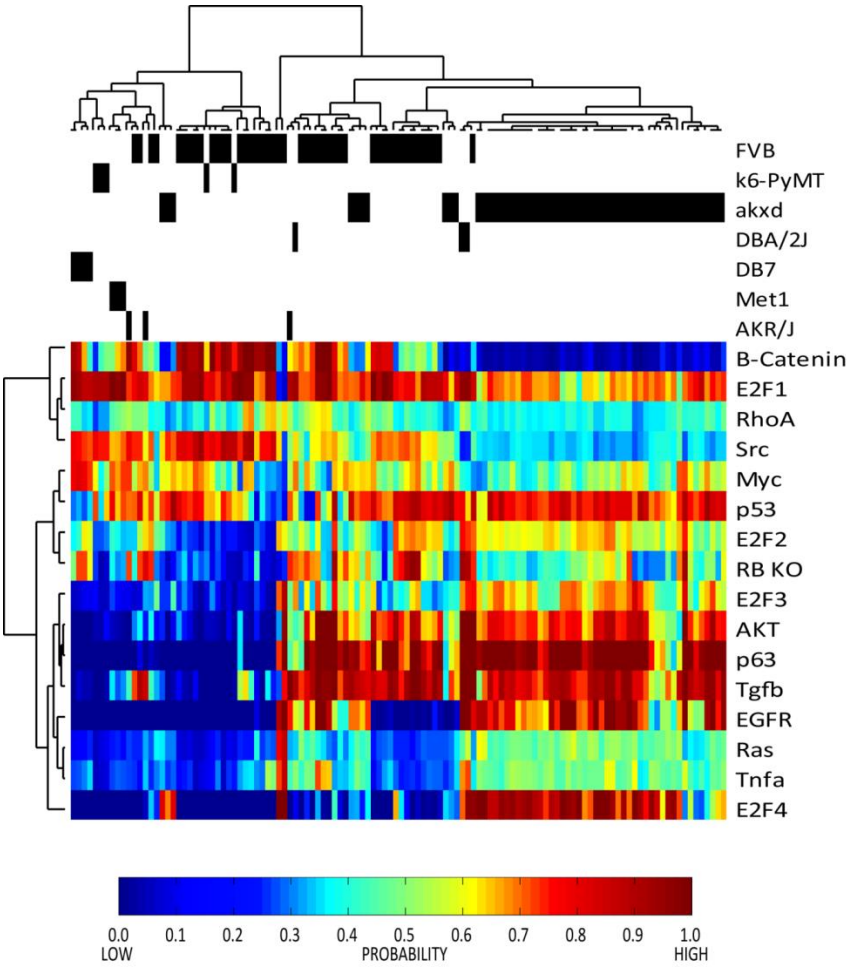


FIGURE S 2.5: UNSUPERVISED HIERARCHICAL CLUSTERING OF PATHWAY PROBABILITIES FOR PYMT INDUCED TUMORS

The dendrogram across the top illustrates the relationship between PyMT tumor types on the basis of pathway activation profiles. Below the dendrogram black bars correspond to sample details on the same line, annotating the genetic background and sample type for each sample. The heatmap shows the predicted pathway activity according to the probabilities listed on the color bar below the heatmap. Directly beside the heatmap, a vertical dendrogram illustrates the degree of correlation between pathways across the samples.

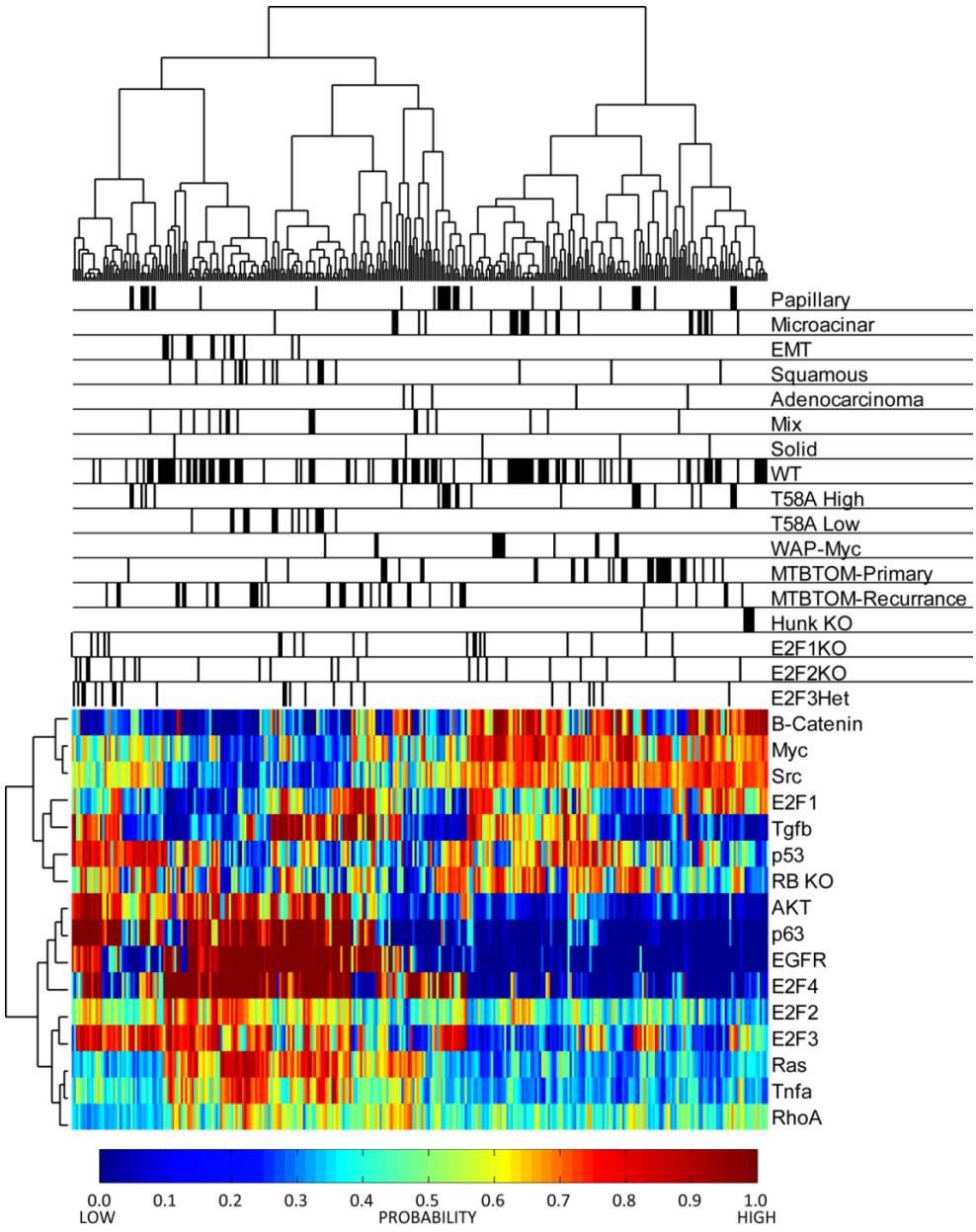


FIGURE S 2.6: UNSUPERVISED HIERARCHICAL CLUSTERING OF PATHWAY PROBABILITIES FOR MYC INDUCED TUMORS

The dendrogram across the top illustrates the relationship between Myc tumor types on the basis of pathway activation profiles. Below the dendrogram black bars correspond to sample details on the same line, annotating the tumor histology (if known), specific form of Myc expression, recurrence status, and additional modifications. The heatmap shows the predicted pathway

FIGURE S 2.6 (cont'd)

activity according to the probabilities listed on the color bar below the heatmap. Directly beside the heatmap, a vertical dendrogram illustrates the degree of correlation between pathways across the samples.

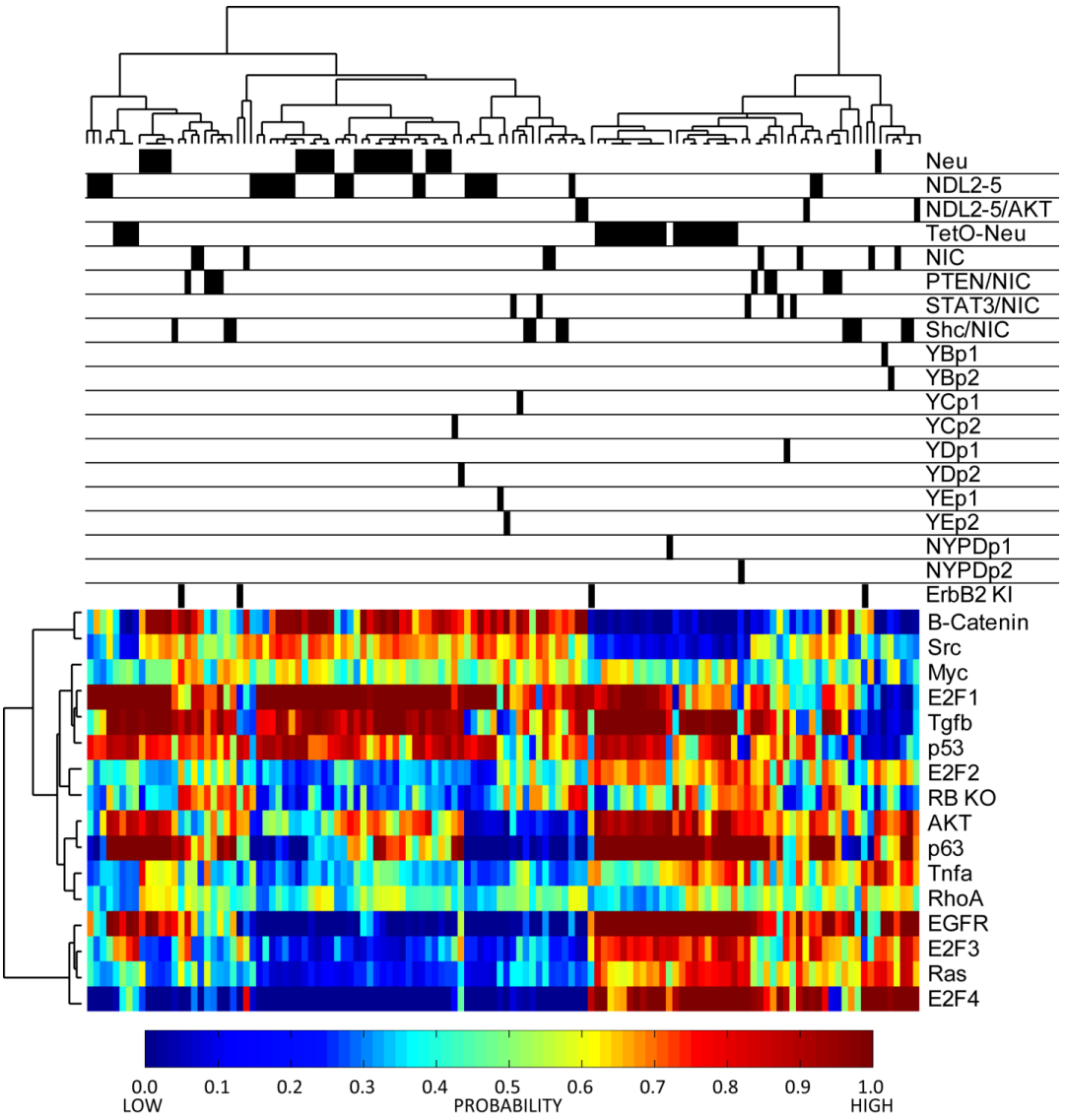


FIGURE S 2.7: UNSUPERVISED HIERARCHICAL CLUSTERING OF PATHWAY PROBABILITIES FOR NEU INDUCED TUMORS

The dendrogram across the top illustrates the relationship between Neu tumor types on the basis of pathway activation profiles. Below the dendrogram black bars correspond to sample details on the same line, annotating the specific form of Neu, and additional modifications. The heatmap shows the predicted pathway activity according to the probabilities listed on the color bar below the heatmap. Directly beside the heatmap, a vertical dendrogram illustrates the degree of correlation between pathways across the samples.

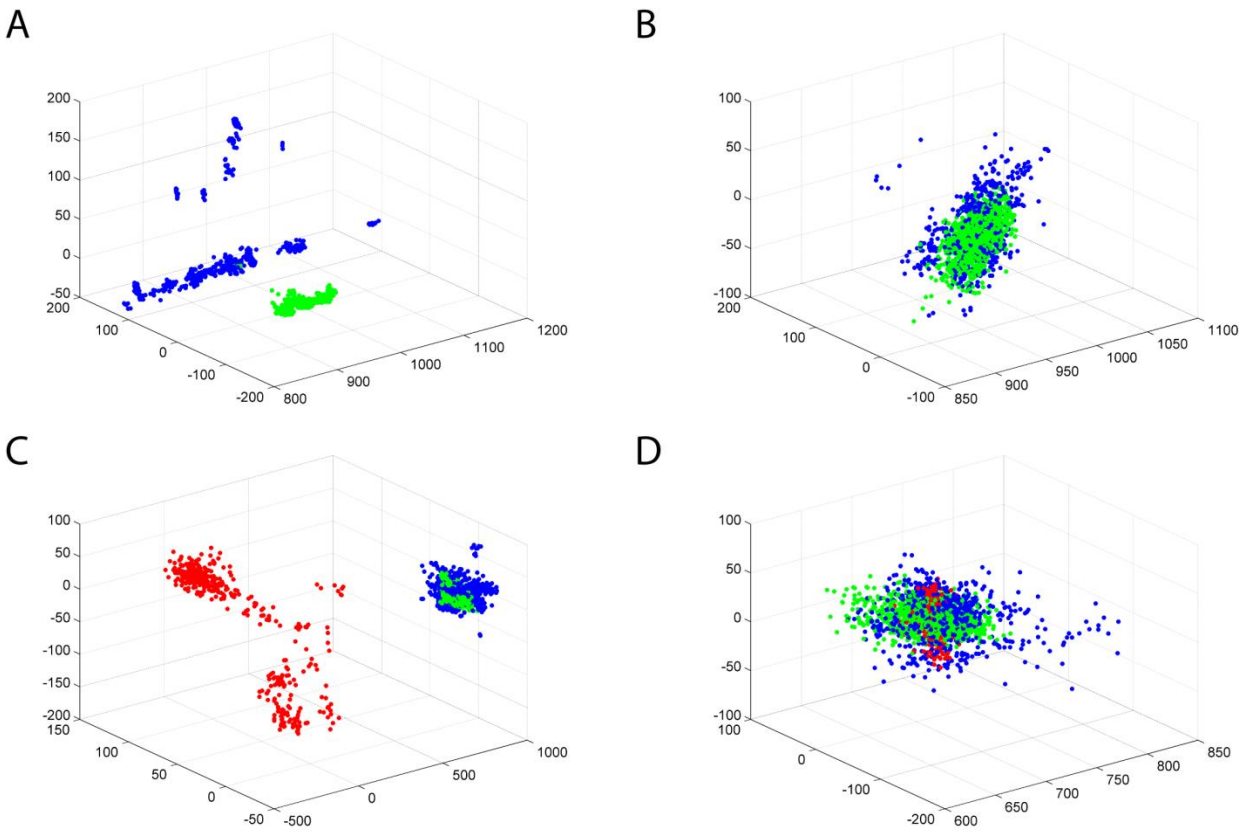


FIGURE S 2.8: REMOVAL OF BATCH EFFECTS BETWEEN MOUSE AND HUMAN BREAST CANCER DATASETS

(A)Mouse (blue) and human (green) Affymetrix data gene expression variance plotted onto three principle components prior to BFRM. (B) Mouse (blue) and human (green) Affymetrix data gene expression variance plotted onto three principle components after BFRM. (C) Human (green) and mouse (blue) Affymetrix data after BFRM correction put together with mouse Agilent data (red) prior to COMBAT. (D) Human (green) and mouse (blue) Affymetrix data after BFRM correction put together with mouse Agilent data (red) after COMBAT artifact correction.

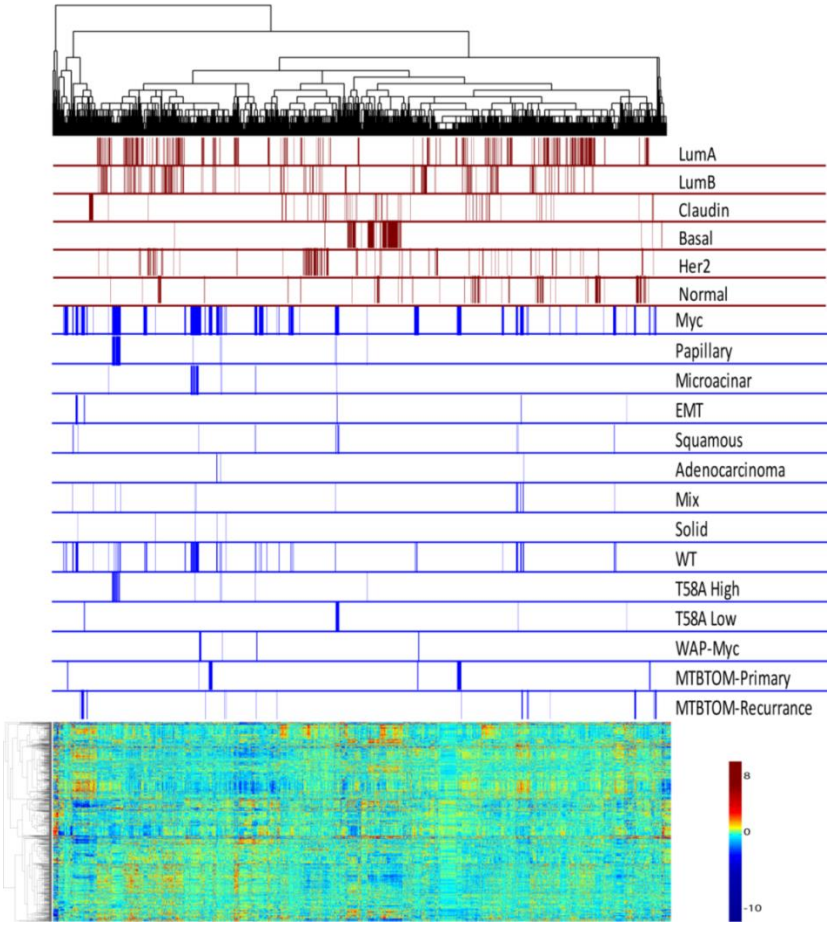


FIGURE S 2.9: UNSUPERVISED HIERARCHICAL CLUSTERING OF MYC MOUSE MAMMARY TUMORS AND HUMAN BREAST CANCER GENE EXPRESSION DATA

Across the top, the dendrogram illustrates the relationship between human and mouse tumor samples on the basis of gene expression profiles. The red bars mark the intrinsic subtype of each human tumor sample according the annotation on the same line. The blue bars correspond to the Myc mouse mammary tumor type. Below this, a heatmap shows the gene expression patterns for each sample, with expression values illustrated according to the color bar on the right. The dendrogram beside the heatmap shows the correlation between genes based on expression patterns across the samples in the dataset.

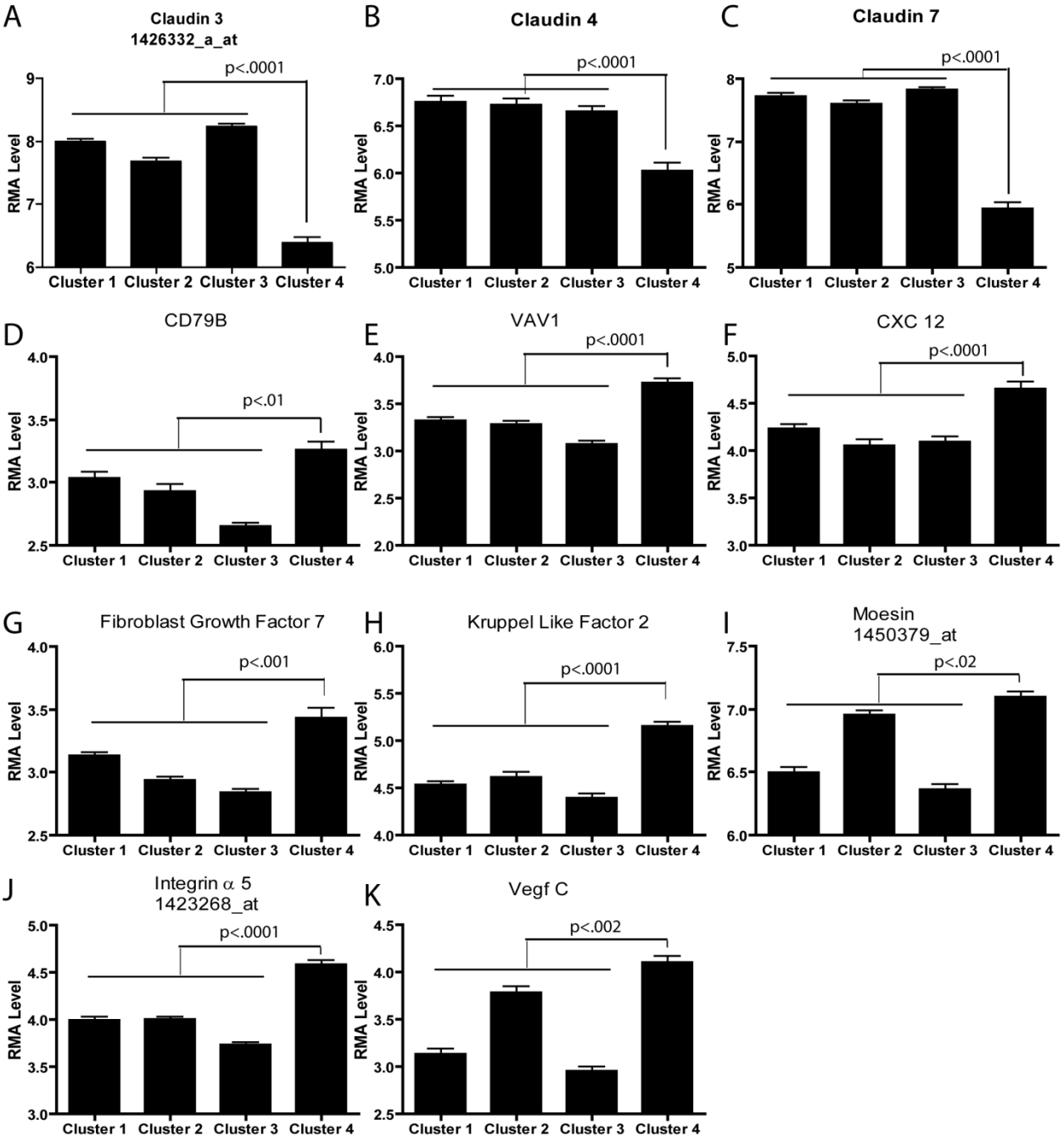


FIGURE S 2.10: CLAUDIN LOW MARKER EXPRESSION IN THE BLACK CLUSTER MOUSE MAMMARY TUMORS

Claudin low marker expression comparisons for cluster 4 (black) tumors compared to tumors in all other clusters as defined by FIGURE 2.1 A. (A-C) Cell adhesion markers that have low expression in claudin low human tumors are also downregulated in cluster 4(black tumors),

FIGURE S 2.10 (cont'd)

p<.0001. (D-E) Genes that are involved with the immune system that are found to be highly expressed in claudin low human tumors are highly expressed in mouse cluster 4 tumors (black), p<.01 for CD79B and p<.0001 for VAV1. (F) chemokine [C-X-C motif] ligand 12, involved in cell communication and previously shown to be highly expressed in claudin low tumors, is upregulated in cluster 4(black) mouse mammary tumors, p<.0001. (G) Fibroblast growth factor 7, an extracellular matrix related factor and previously shown to be highly expressed in claudin low tumors, is upregulated in cluster 4(black) mouse mammary tumors, p<.0001. (H-I) Cell migration markers previously shown to be highly expressed in human claudin low tumors are upregulated in mouse cluster 4 (black) tumors, p<.02 for moesin and p<0001 for integrin α5. (K) Angiogenesis marker, VEGFC, was previously shown to be upregulated in human claudin low tumors and is highly expressed in mouse cluster 4(black) tumors.

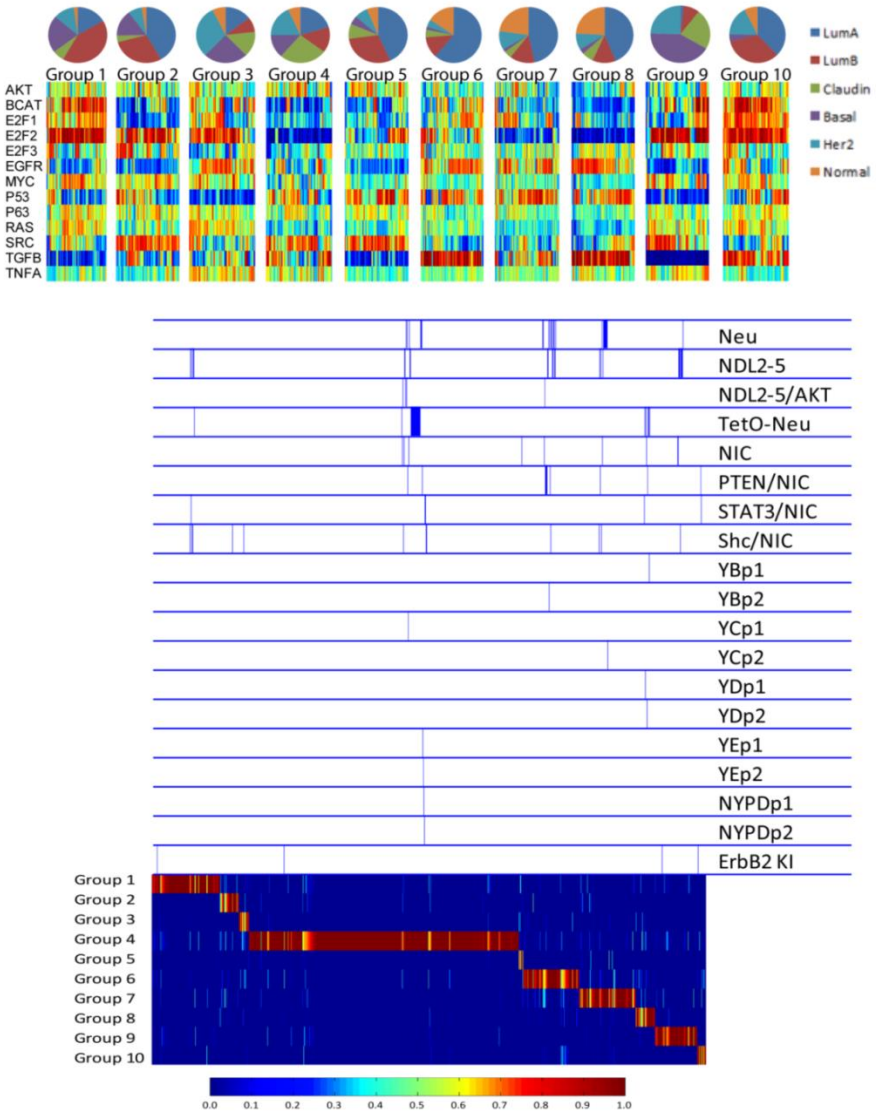


FIGURE S 2.11: MIXTURE MODELING HIGHLIGHTING PATHWAY RELATIONSHIPS BETWEEN HUMAN BREAST CANCER AND SPECIFIC MODELS OF NEU MEDIATED TUMORIGENESIS

Pie charts above each heatmap illustrate the distribution of the intrinsic subtype of samples in each group, according to the color-coded legend. The heatmap for groups 1-10 show predicted pathway activity with probabilities corresponding to the color bar at the bottom of the figure. Below this blue bars mark the samples corresponding to annotations on the same line. Following

FIGURE S 2.11 (cont'd)

the samples down to the heatmap below the blue bars, the probability that a specific type of Neu model has similar pathway activation profiles is shown for each group. Probabilities for this heatmap are shown according to the color bar at the bottom of the figure.

TABLE 2.1 : LIST OF MOUSE MODELS IN THE DATASET

Model	Arrays	Promoter	Description	References
Myc	319	MMTV WAP / Dox	Myc mammary tumors of various histological types, expression levels and stability with variable Kras mutations.	[102, 106, 108, 113, 141, 265, 266]
Neu	124	MMTV	Induction of adenocarcinomas with pulmonary metastasis.	[106, 115, 129, 266-271]
PyMT	119	MMTV K6/RCAS MMTV/RCAS	Rapid induction of luminal-type mammary tumors with pulmonary metastasis.	[110, 266, 272-274]
SV40 Large T Antigen	107	C3 WAP	Induction of mammary tumors with similarities to human basal type breast cancer.	[129, 266, 275-277]
p53	92	Null	Tumors with similarities to human basal type breast cancer.	[129, 266, 278, 279]
CreEtv6 / NTRK3	63	WAP	Fusion oncoprotein transforms through activation of AP1.	[260]
MET	52	MMTV	Diverse histologies with similarities to human basal breast cancer.	[107]
BRCA / p53	46	WAP MMTV BLG	CKO of BRCA1 in a p53 null background. Tumors similar to human basal breast cancer.	[94, 266]
Wnt	35	MMTV	Induction of mammary tumors with diverse gene expression patterns.	[266, 280-282]
IGF-IR	26	MTB	Basal-like mammary tumors. Recurrent tumors resemble human claudin-low.	[283]
LPA	16	MMTV	ER positive, metastatic tumors.	[284]
Stat5	16	BLG	Induction of mammary tumors	NA
Brg1 (+/-)	14	Mutant	Heterogeneous breast cancers.	[285]
DMBA	12	Chemical	Mammary carcinomas with three phenotypes: adenocarcinoma, squamous cell carcinoma, and myoepithelial cell carcinoma.	[286]
Ras	10	MMTV	Induction of mammary tumors with rapid tumor onset.	[266]
Int3	9	WAP	Metastatic tumors.	[287]
RB/ p107	7	CKO	Adeno and adenosquamous carcinomas similar to luminal B or basal.	[263]
APC CKO	6	K14-Cre	CKO results in adenocarcinomas with histological and molecular heterogeneity.	[288]
Autotaxin (ATX)	5	MMTV	ER+ metastatic mammary tumors.	[284]
BRCA	5	CKO	Tumors similar to human basal type breast cancer.	NA
STAT1	5	Knockout	ERa+ PR+, hormone dependent like human ERa+ luminal.	[289]
Notch	4	Dox	Induction of mammary adenocarcinomas.	NA
PDK1	2	MMTV	Induction of mammary tumors	[290]
Normal Mammary Gland	47	Not Applicable	Normal mammary gland samples from FVB, BalbC, and CD1 genetic backgrounds.	[107, 271, 284]

*Dox = Doxacycline inducible MMTV-Rtta system. CKO - conditional knockout

TABLE 2. 2 : VALIDATION OF PATHWAY PREDICTIONS

Model	Pathway	Effect	References
APC cKO	B-Catenin	Demonstrated high activation of β-catenin signaling in these tumors.	[288]
APC cKO	Myc	High levels of Myc demonstrated by IHC in these mammary tumors.	[288]
BRCA & P53 mut	EGFR	Using IHC, EGFR was shown to be overexpressed in this mouse model.	[291]
DMBA	Ras	Observation of H-Ras mutations in mammary hyperplastic outgrowths after treatment with DMBA	[147]
DMBA	EGFR	Using western blot and IHC, EGFR signaling was shown to be active in DMBA induced mammary tumors.	[148]
ETV6-Ntrk3	Src	ETV6-Ntrk3 binds to and activates c-Src, and inhibition of c-Src activation blocks EN transforming activity using mouse engineered mouse embryonic fibroblasts.	[292]
Myc	Ras	Activating mutations in K-Ras found in a subset MMTV-Myc induced tumors with an predicted elevation of Ras signaling.	[106]
Myc	B-Catenin	IHC analysis demonstrates higher expression of B-Catenin in the microacinar histology of Myc driven tumors.	[106]
Myc	E2F1 E2F2 E2F3	E2F loss altered tumor latency and Myc proliferative effects on the mammary gland.	[108]
Neu	Akt	Akt loss effects tumor development in the MMTV-Neu mouse model.	[293]
Neu	B-Catenin	Using a beta-gal reporter, ß-catenin/TCF-dependent transcription was shown to be elevated in MMTV-Neu mouse mammary glands.	[144]
Notch	B-Catenin	Knocking down Notch in a human breast cancer cell line also impacted levels of beta-catenin.	[145]
PyMT	Tgfb	Blockade of TGF-beta inhibits mammary tumor metastasis.	[146]
PyMT	Src	Loss of c-Src greatly reduced the occurrence of mammary tumors in the MMTV-PyMT mouse model.	[150]
Tag	Ras	K-ras amplifications observed in large t-antigen mediated tumorigenesis.	[294]
Tag	E2F2 E2F3 RB KO	Large T Antigen simulates loss of Rb by leading to deregulated activation of the E2F transcription factors.	[295]
Wnt	p53	MMTV-Wnt1 mammary tumors with mutant p53 exhibited a superior clinical response compared to tumors with wild-type p53.	[296]

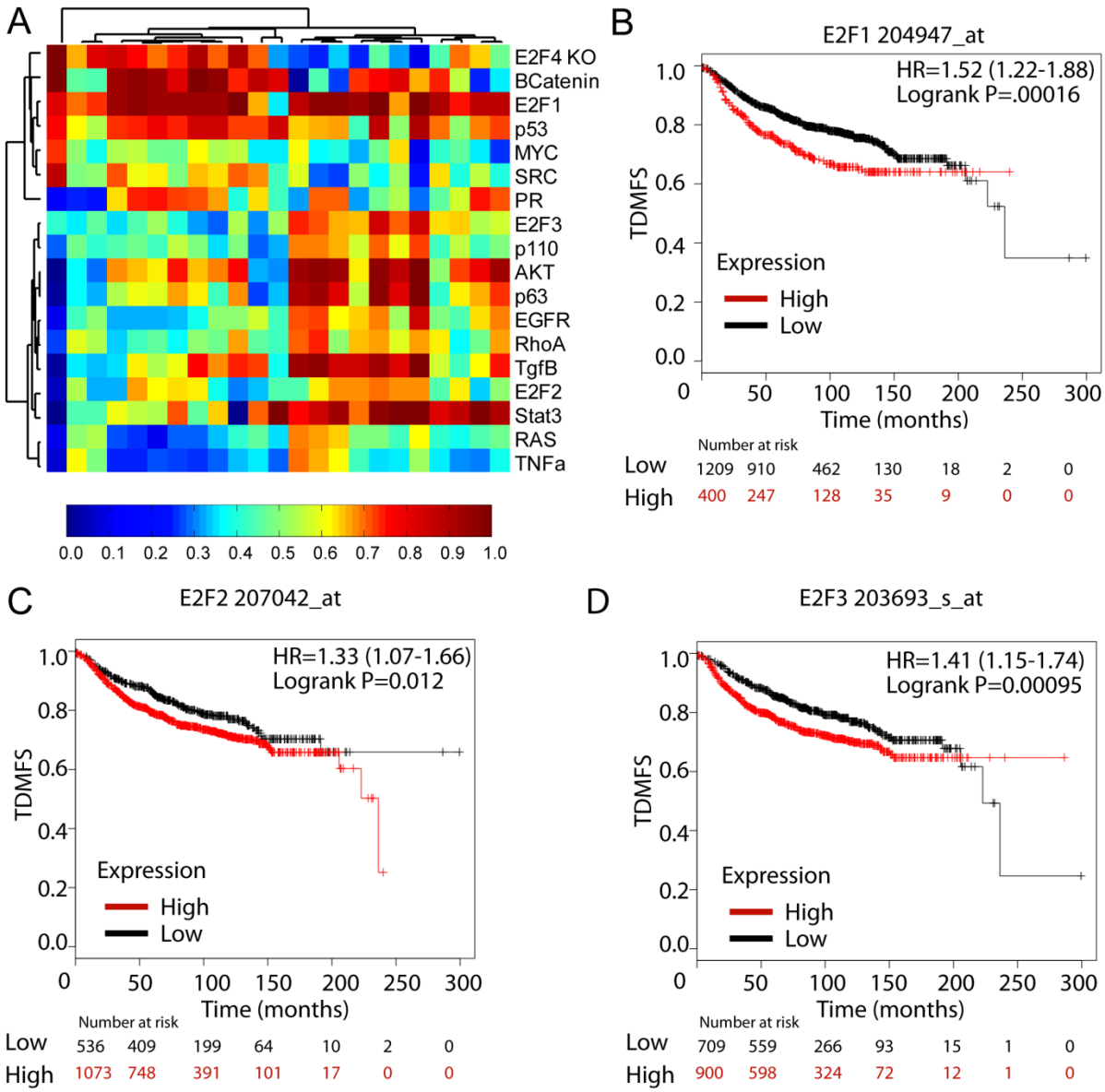


FIGURE 3.1: PATHWAY SIGNATURES PREDICT E2F ACTIVATION IN METASTATIC BREAST TUMORS

(A) Probability of pathway activation in transgenic MMTV-PyMT tumors are shown for the given signaling pathways listed on the right axis of the heatmap. Below the heatmap, a scale bar depicts the range of probabilities from 0 to 1. The probabilities were used in unsupervised hierarchical clustering of both pathways with clusters identified on the basis of pathway activation in the transgenic tumor samples. (B) The Kaplan-Meier plot shows metastasis free

FIGURE 3.1 (cont'd)

survival for breast cancer patients (n=1610) stratified on expression of E2F1, Affymetrix probe 204947_at, p=0.00016 . (C) Patients are stratified based on E2F2, Affymetrix probe 207042_at, expression levels, p=0.012. (D) Patients are stratified on the basis of E2F3, Affymetrix probe 203693_s_at, p=0.00095. Patient stratification was conducted using the www.kmplot.com auto selection option.

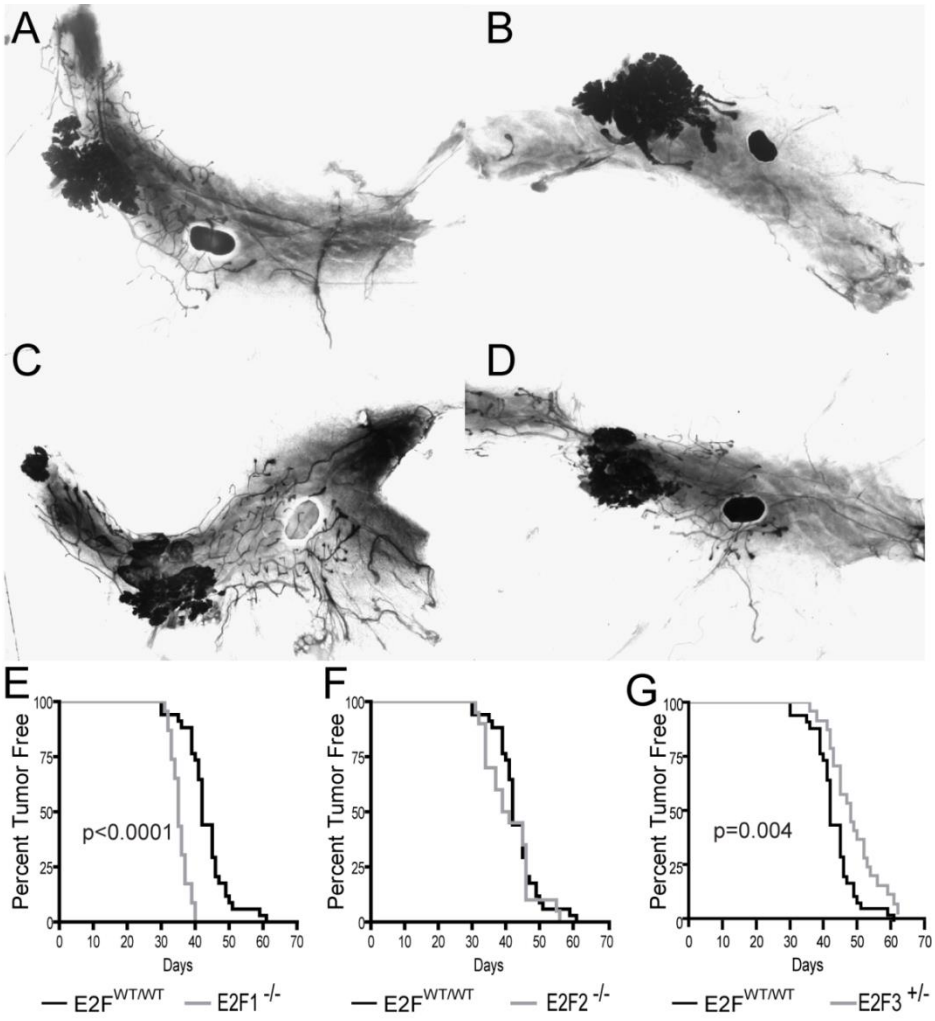


FIGURE 3.2: LOSS OF E2FS ALTER TUMOR ONSET

Representative mammary whole mounts from 35 day old virgin MMTV-PyMT $E2F^{WT/WT}$ (A), $E2F1^{-/-}$ (B), $E2F2^{-/-}$ (C), and $E2F3^{+/-}$ (D) mice are shown. (E) Tumor Latency was compared in a Kaplan-Meier plot for PyMT induced tumors in a wild type ($^{WT/WT}$) (n=34) and E2F1 -/- (n=22) background revealing a significant acceleration with loss of E2F1 ($p<0.0001$). (F) Kaplan-Meier plot for tumor onset between $E2F^{WT/WT}$ and $E2F2^{-/-}$ backgrounds (n=34 and 20 respectively). (G) Kaplan-Meier plot for tumor onset between $E2F^{WT/WT}$ and $E2F3^{+/-}$ backgrounds (n=34 and 24 respectively) with a significant delay ($p=0.004$).

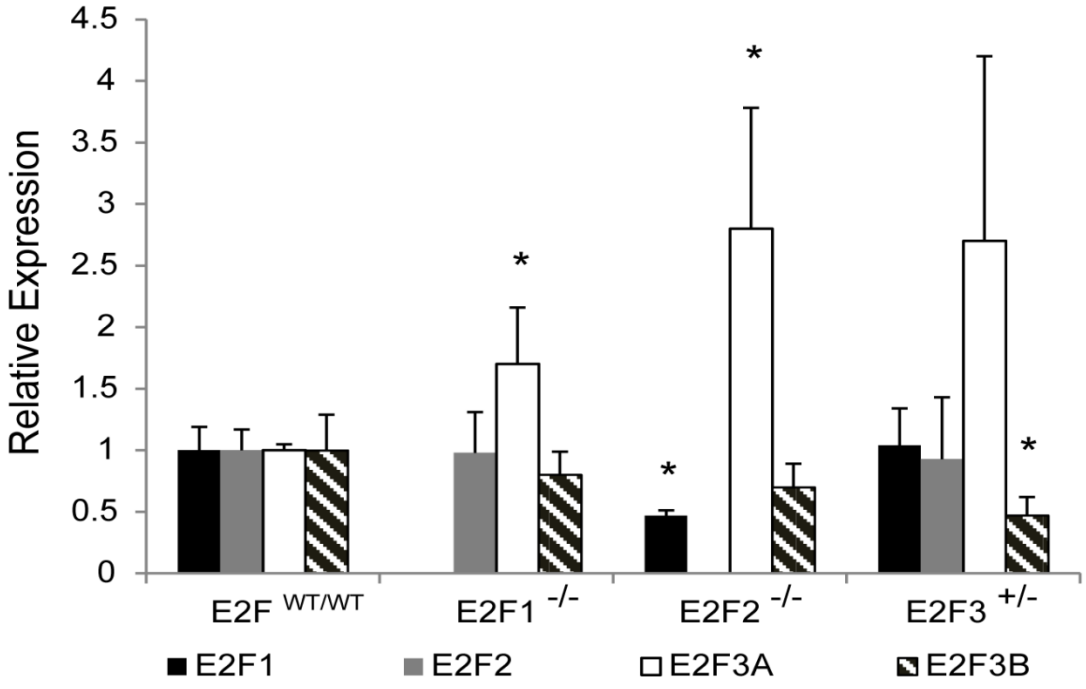


FIGURE 3.3: E2F LOSS RESULTS IN GENE EXPRESSION CHANGES OF OTHER E2F FAMILY MEMBERS

Quantitative PCR results are shown depicting relative expression of E2F1, E2F2, E2F3A and E2F3B in $E2F^{WT/WT}$ (n=4) tumors, $E2F1^{-/-}$ tumors (n=4), $E2F2^{-/-}$ tumors (n=4),and $E2F3^{+/-}$ tumors (n=4). $E2F1^{-/-}$ tumors had upregulation of E2F3A(p=0.0232). In $E2F2^{-/-}$ tumors, a decrease in E2F1 levels (p=0.0016) and significant upregulation of E2F3A (p=0.0105) is shown. In $E2F3^{+/-}$ tumors, significant downregulation of E2F3B (P=0.0175) is shown.

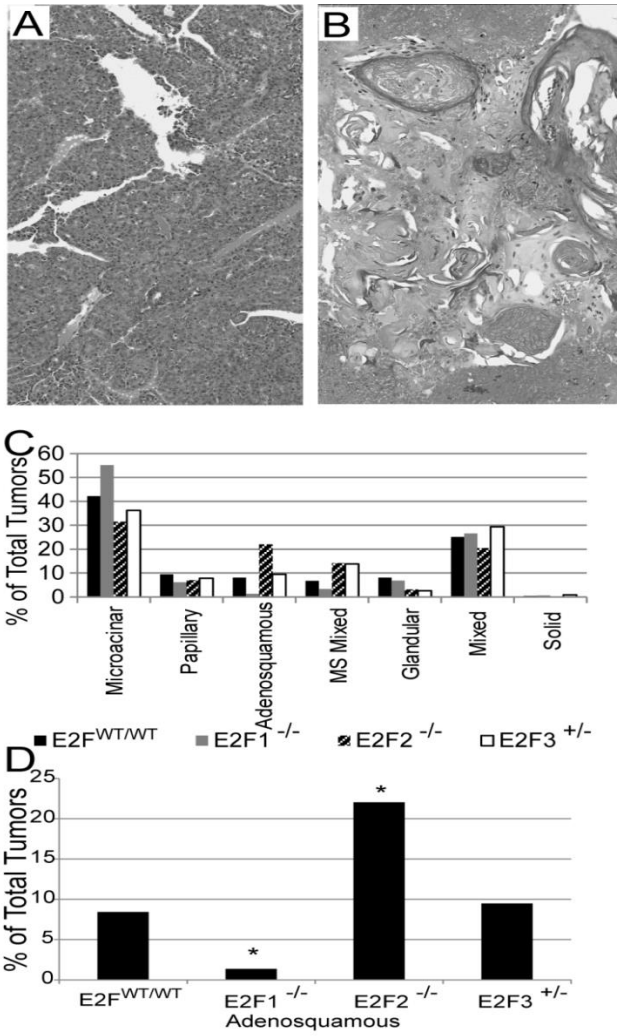


FIGURE 3.4: LOSS OF E2FS ALTERS TUMOR HISTOLOGY

H&E staining revealed microacinar tumors (A) and adenosquamous tumors (B) among other tumor types. (C) The proportion of histological types of tumors in the corresponding genotypes of MMTV-PyMT mice are shown (MS Mixed = Microacinar and adenosquamous mixture). (D) The percentage of adenosquamous tumors within the population of the E2F wild type and mutant mice. Fisher's exact test was used to compare $E2F^{WT/WT}$ with $E2F1^{-/-}$ populations (p=0.001), as well as with $E2F2^{-/-}$ populations (p=0.0003).

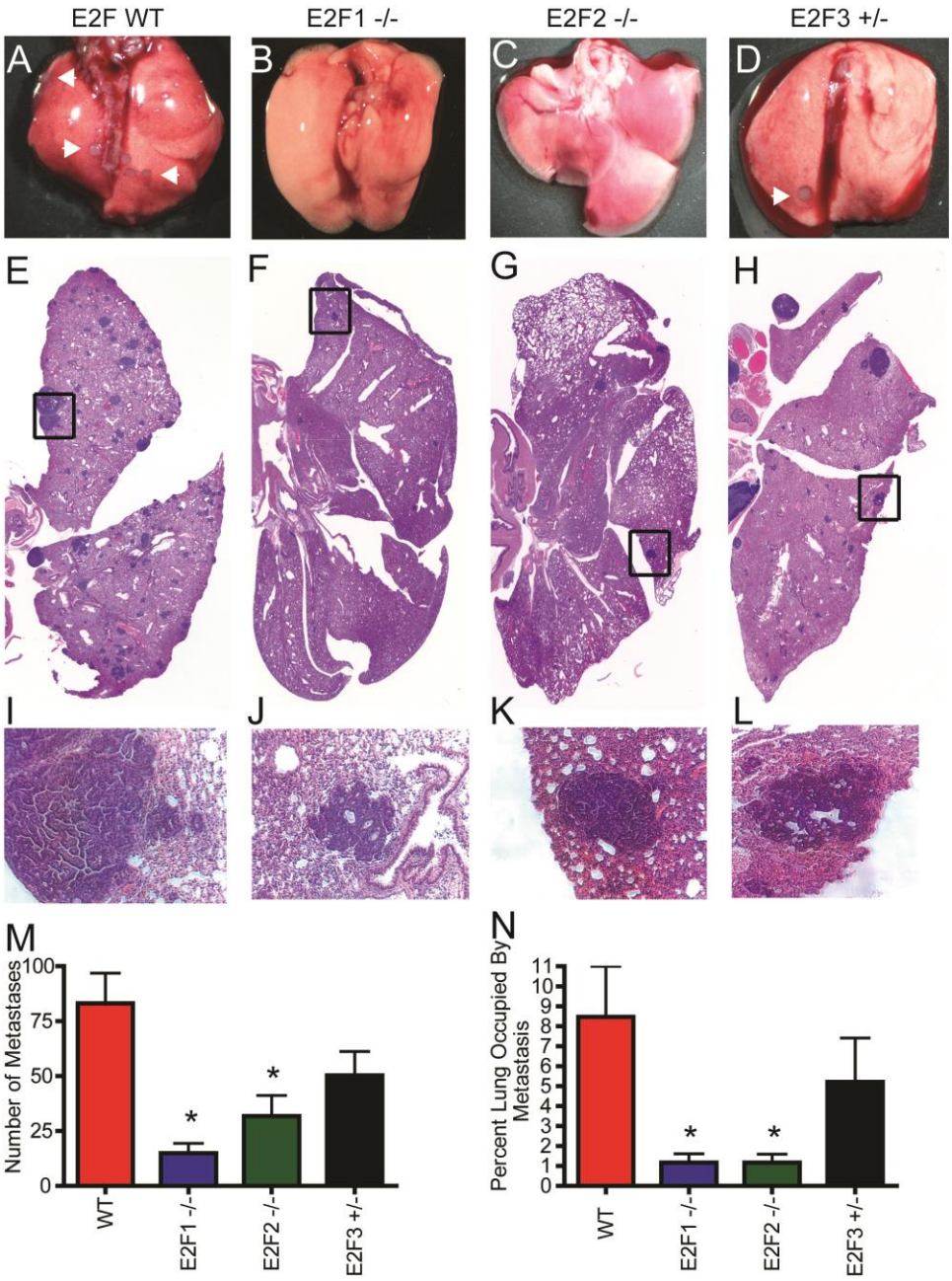


FIGURE 3.5: LOSS OF E2FS DECREASE PULMONARY METASTASIS IN MMTV-PYMT MICE

Representative wet mount images for $E2F^{WT/WT}$, $E2F1^{-/-}$, $E2F2^{-/-}$, and $E2F3^{+/-}$ mice (A-D, respectively). White arrows indicate surface metastases. (E-H)Representative images are shown for H&E stained sections of lungs. (I-L) 10X images of the regions within black boxes in E-H

FIGURE 3.5 (cont'd)

show histology of tumor metastases. (M) A comparison of the average number of lung metastases and SEM in pulmonary sections from $E2F^{WT/WT}$ mice compared to E2F mutant backgrounds revealed significant differences in $E2F1^{-/-}$ mice ($p<0.0001$) and $E2F2^{-/-}$ mice ($p=0.002$). (N) A comparison of the average percentage and SEM of the lung occupied by metastasis in pulmonary sections from $E2F^{WT/WT}$ mice compared to $E2F1^{-/-}$ mice ($p<0.0001$), $E2F2^{-/-}$ mice ($p<0.0001$), and $E2F3^{+/-}$ mice is shown.

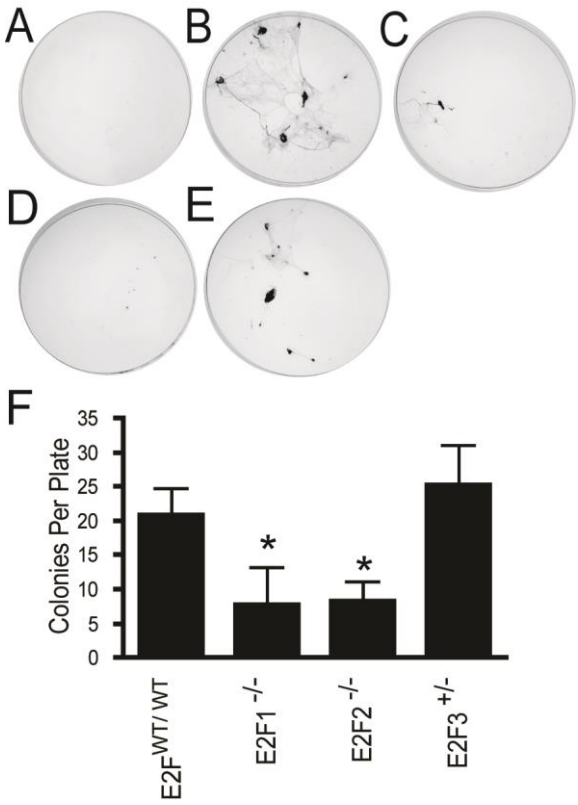


FIGURE 3.6: LOSS OF E2FS DECREASE CIRCULATING TUMOR CELLS IN MMTV-PYMT MICE

(A) Representative colony forming assay plate for negative control, non-transgenic FVB mice(n=6). Representative plates are shown for MMTV-PyMT E2F$^{WT/WT}$ (n=14) (B), MMTV-PyMT E2F1$^{-/-}$(n=7) (C), MMTV-PyMT E2F2$^{-/-}$(n=10) (D), and MMTV-PyMT E2F3$^{+/-}$(n=10)(E) mice. (F) A comparison of the average number of circulating tumor cells and SEM detected in MMTV-PyMT E2F$^{WT/WT}$ mice, MMTV-PyMT E2F1$^{-/-}$ mice (p=0.02), MMTV-PyMT E2F2$^{-/-}$ mice (p=0.006), and MMTV-PyMT E2F3$^{+/-}$ mice are shown.

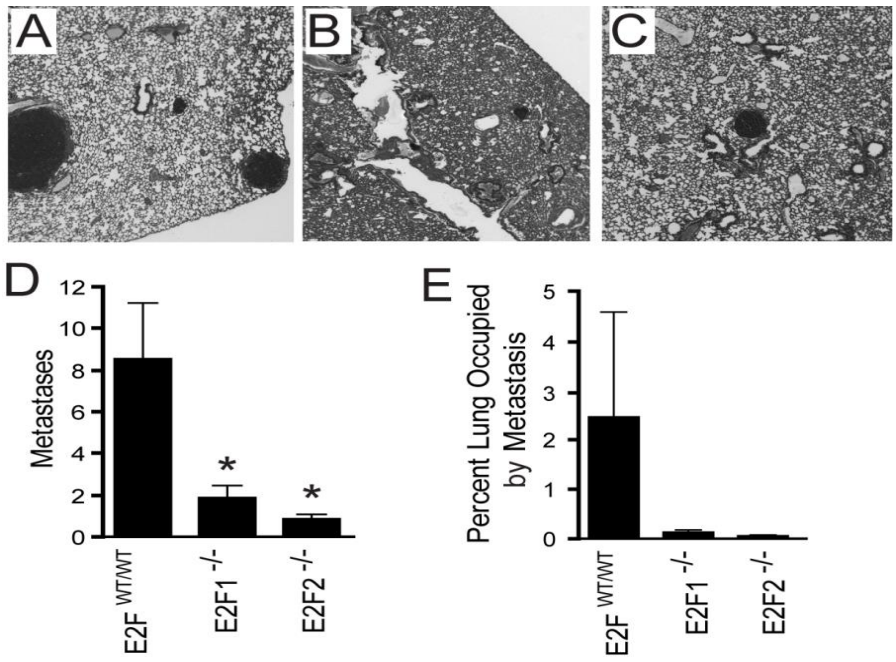


FIGURE 3.7: LOSS OF E2FS DECREASE TUMOR CELL PULMONARY COLONIZATION

Representative sections of lungs are shown for mice injected with $E2F^{WT/WT}$ tumor cells, n=8 (A), $E2F1^{-/-}$ tumor cells, n=9 (B), and $E2F2^{-/-}$ tumor cells, n=7 (C). (D) A comparison of the average number and SEM of metastases detected in sections of lungs of mice injected with $E2F^{WT/WT}$ tumor cells, $E2F1^{-/-}$ tumor cells (p=0.01), and $E2F2^{-/-}$ tumor cells (p=0.02). (E)A comparison of the average percentage and SEM of the lung occupied by metastasis in pulmonary sections from mice injected with $E2F^{WT/WT}$ tumor cells compared to mice injected with $E2F1^{-/-}$ and $E2F2^{-/-}$ tumor cells is shown.

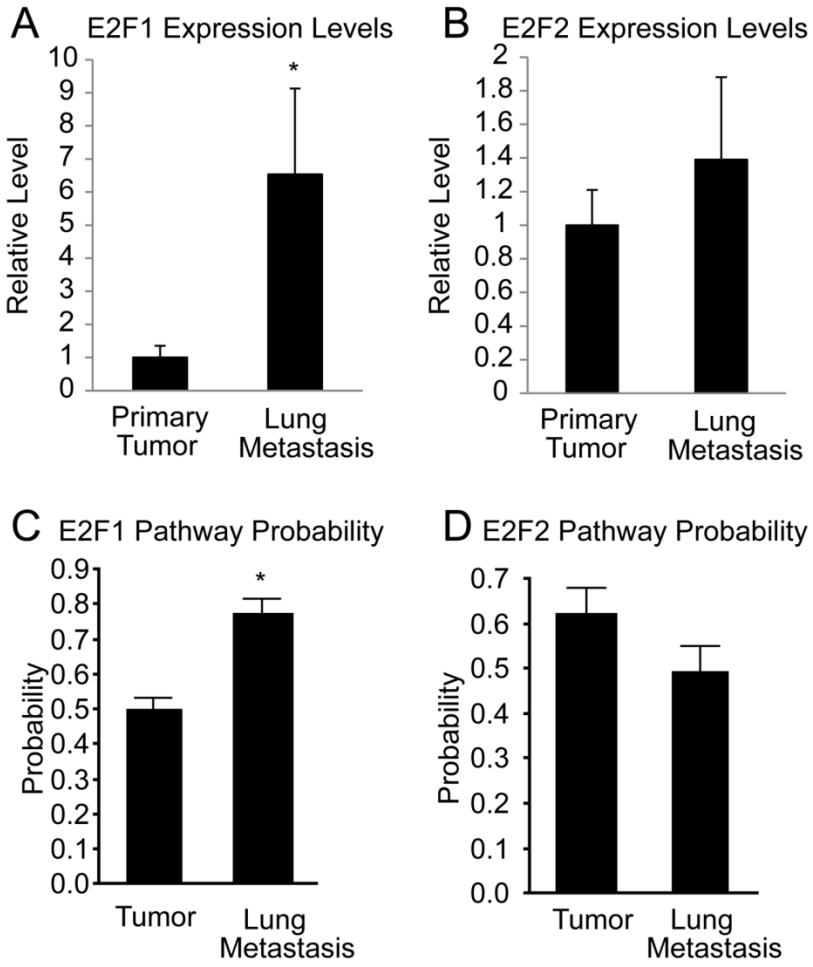


FIGURE 3.8: E2F1 EXPRESSION LEVELS AND PATHWAY ACTIVITY ARE ELEVATED IN LUNG METASTASES

(A) Quantitative RT-PCR results showing the relative expression of E2F1 in primary tumors (n=6) compared to lung metastases (n=6, p=0.0004). (B) Quantitative RT-PCR results showing the relative expression of E2F2 in primary tumors compared to lung metastases. (C) Pathway signature for E2F1 shows predicted E2F1 activation in primary tumors (n=6) and lung metastases (n=4) within GEO dataset GSE43566, p= 0.0007. (D) Pathway signature for E2F2 shows predicted E2F2 activation in primary tumors and lung metastases within GEO dataset GSE43566.

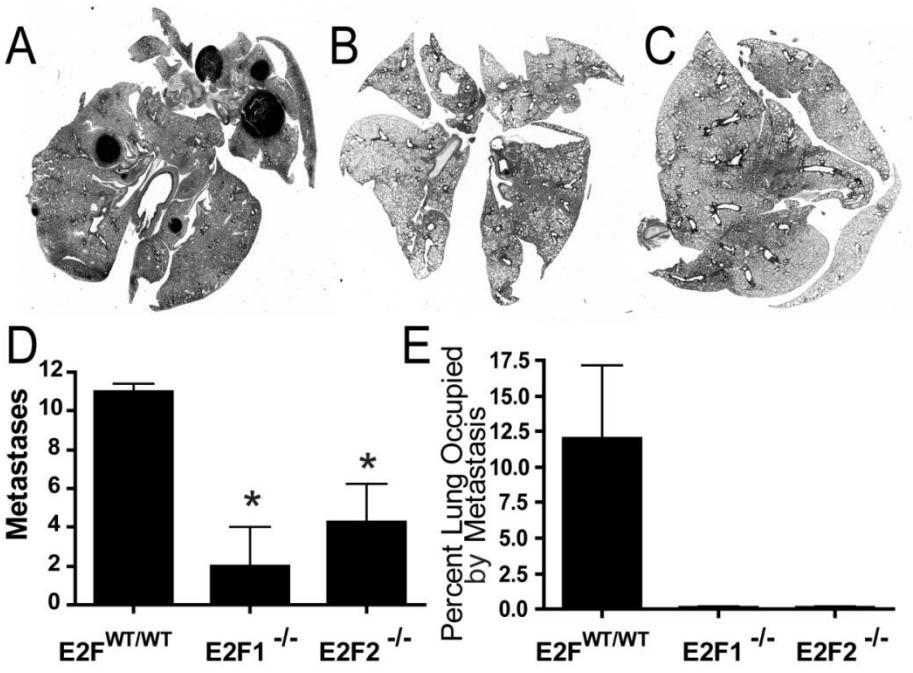


FIGURE 3.9: TRANSPLANT OF TUMORS INTO E2F WILD TYPE MICE SHOWS E2F REGULATION OF METASTASIS IS CELL AUTONOMOUS

Viable frozen tumor samples ($E2F^{WT/WT}$, n=4; $E2F1^{-/-}$, n=4; $E2F2^{-/-}$, n=4) were used for transplant into E2F wild type MMTV-Cre control mice. (A) Representative histological sections for lungs of mice implanted with an $E2F^{WT/WT}$ tumor, $E2F1^{-/-}$ tumor (B) or $E2F2^{-/-}$ tumor (C) tumor. (D)Quantification revealed a significant reduction in the number of metastases observed in the lungs of mice implanted with an $E2F1^{-/-}$ ($p=0.003$) or $E2F2^{-/-}$ ($p=0.01$) tumor compared to mice implanted with an $E2F^{WT/WT}$ tumor. (E) Quantification of the percentage of lungs occupied by metastasis shows reduced metastatic burden in mice receiving $E2F1^{-/-}$ and $E2F2^{-/-}$ tumors.

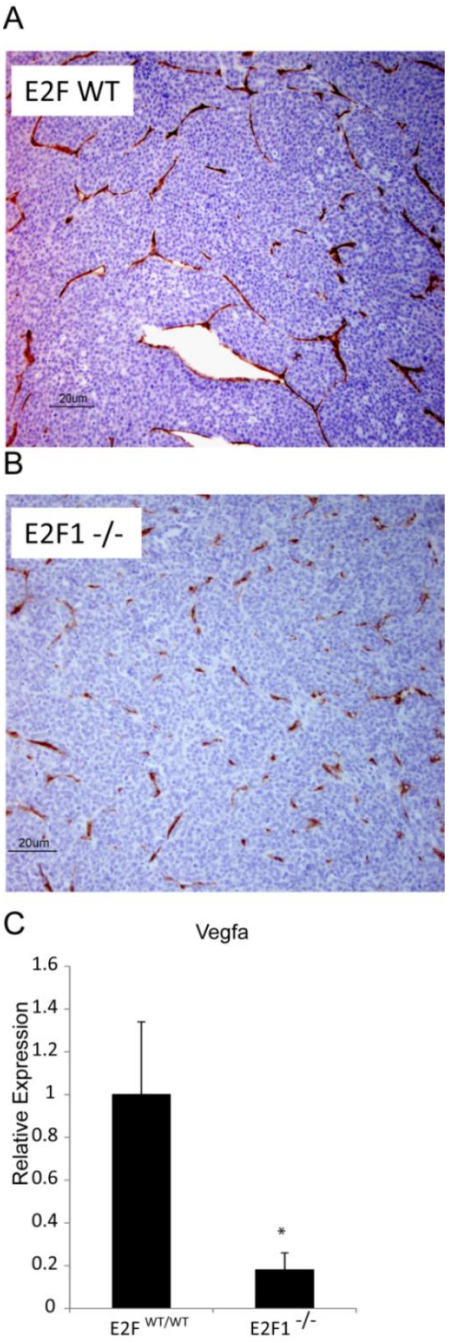


FIGURE 3.10: E2F1 LOSS ALTERS CD31 STAINING AND REDUCES VEGFA EXPRESSION IN MMTV-PYMT TUMORS

(A) Representative section of $E2F^{WT/WT}$ tumor stained for CD31 to reveal vascular structure. (B) Representative section of $E2F1^{-/-}$ tumor stained for CD31 to reveal vascular structure. (C)

FIGURE 3.10 (cont'd)

Quantitative RT-PCR results depicting the relative expression levels of VEGFA in $E2F^{WT/WT}$ (n=6) and $E2F1^{-/-}$ (n=6) tumors (p=0.0002).

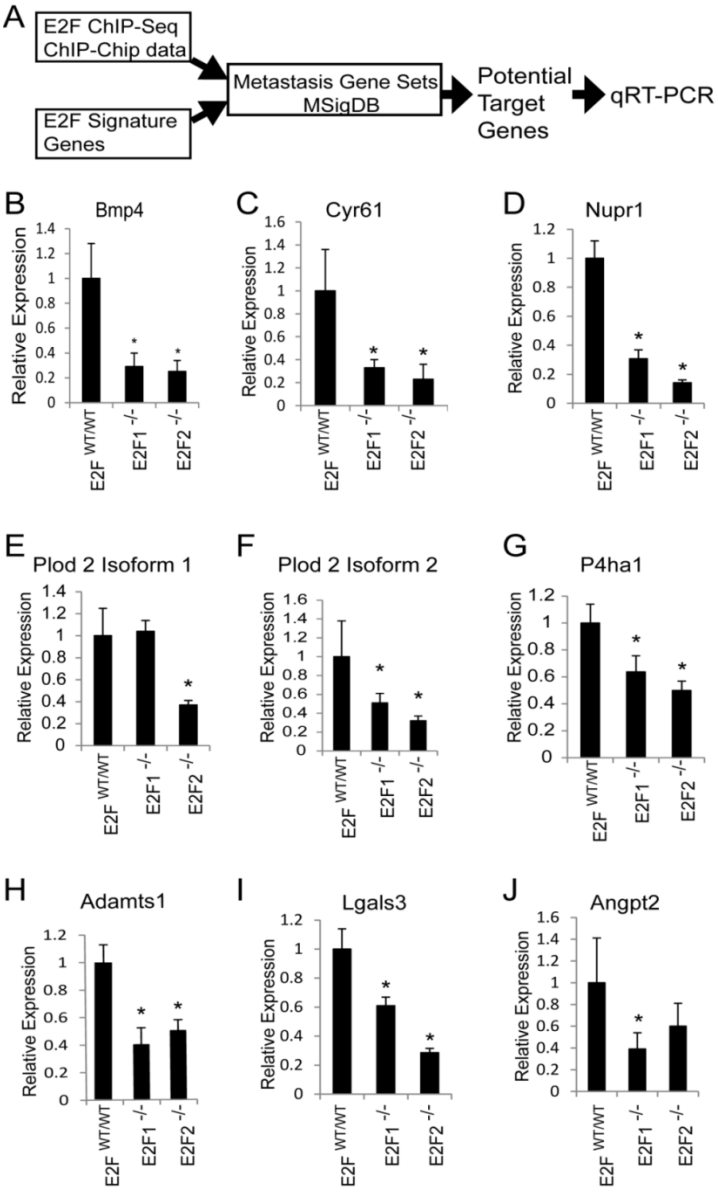


FIGURE 3.11: ANALYSIS OF E2F TARGET GENES REVEALS EXPRESSION CHANGES IN PRO-METASTATIC GENES WITH E2F LOSS

(A). Informatics pipeline for filtering candidate genes for qRT-PCR testing. Testing potential target genes via qRT-PCR we found that E2F1$^{-/-}$ (n=6) and E2F2$^{-/-}$ (n=6) tumors have significantly lower levels Bmp4 ($p=0.0002$, $p<0.0001$, respectively), Cyr61 ($p=0.0009$, $p=0.0006$, respectively), and Nupr1 ($p<0.0001$, $p<0.0001$, respectively), Plod 2 isoform 1($p<0.0001$, E2F2$^{-/-}$ only), Plod 2 isoform 2($p=0.0122$, $p=0.0015$, respectively),

FIGURE 3.11 (cont'd)

P4ha1($p=0.0006$, $p<0.0001$, respectively), Adamts1 ($p<0.0001$, $p<0.0001$, respectively), Lgals3($p<0.0001$, $p<0.0001$, respectively), and Angpt2 ($p=0.0065$, $E2F1^{-/-}$ only)(Fig 11 B-J).

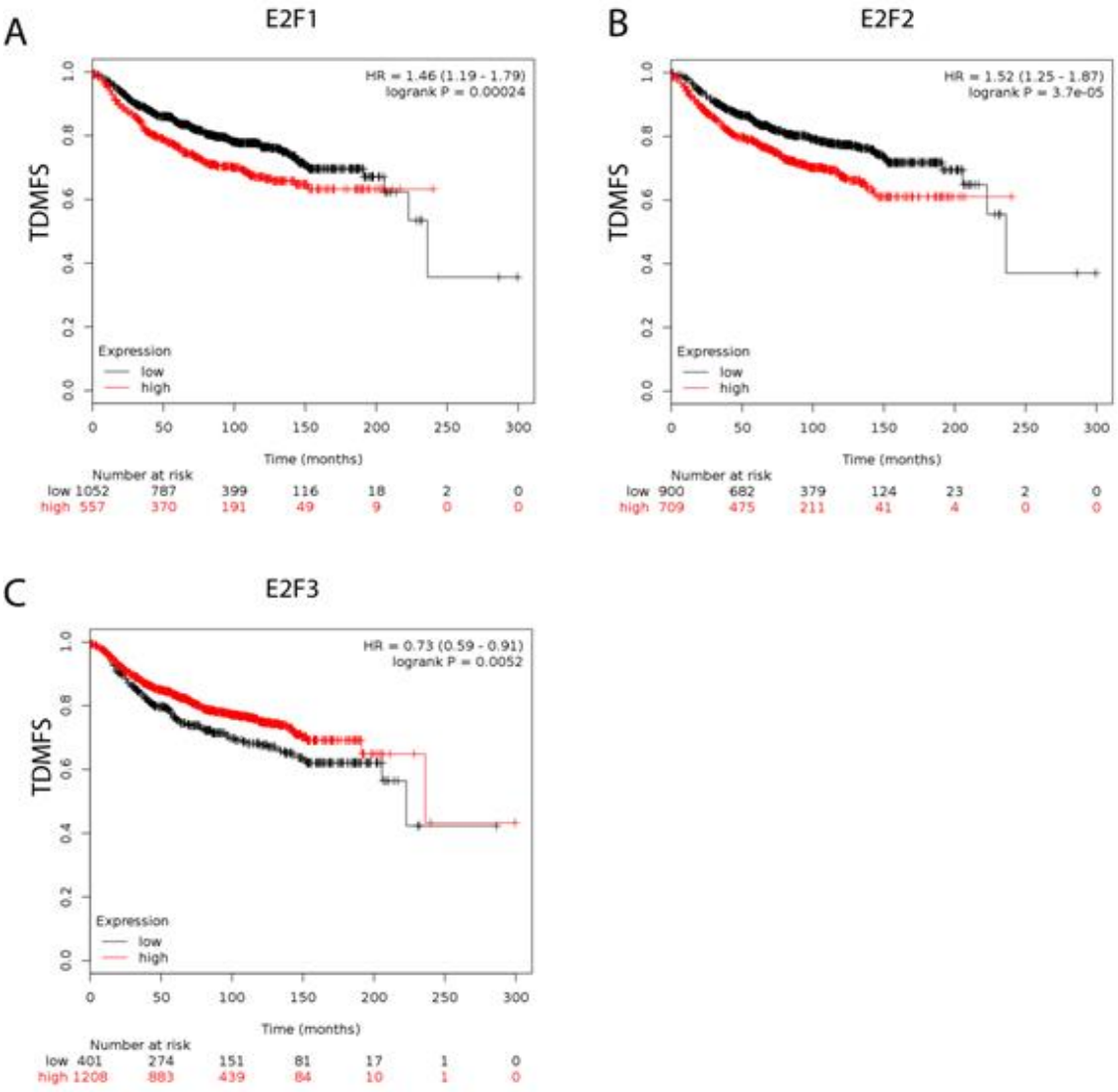


FIGURE S 3.1: ASSOCIATION E2F SIGNATURE GENES WITH HUMAN BREAST CANCER DISTANT METASTASIS FREE SURVIVAL TIMES

(A)Using the average expression of the upregulated genes in the E2F1 signature, a correlation (p=0.00024) with a shorter time to distant metastasis in human breast cancer patients (n=1610) was detected. (B) Using the average expression of the upregulated genes in the E2F2 signature, a correlation (p=0.000037) with a shorter time to distant metastasis in human breast cancer patients (n=1610) was detected. (C)Using the average expression of the upregulated genes in the E2F3 signature, a correlation (p=0.0052) with a longer time to distant metastasis in human breast

FIGURE S 3.1 (cont'd)

cancer patients (n=1610) was detected. Patient stratification was conducted using the www.kmplot.com auto selection tool.

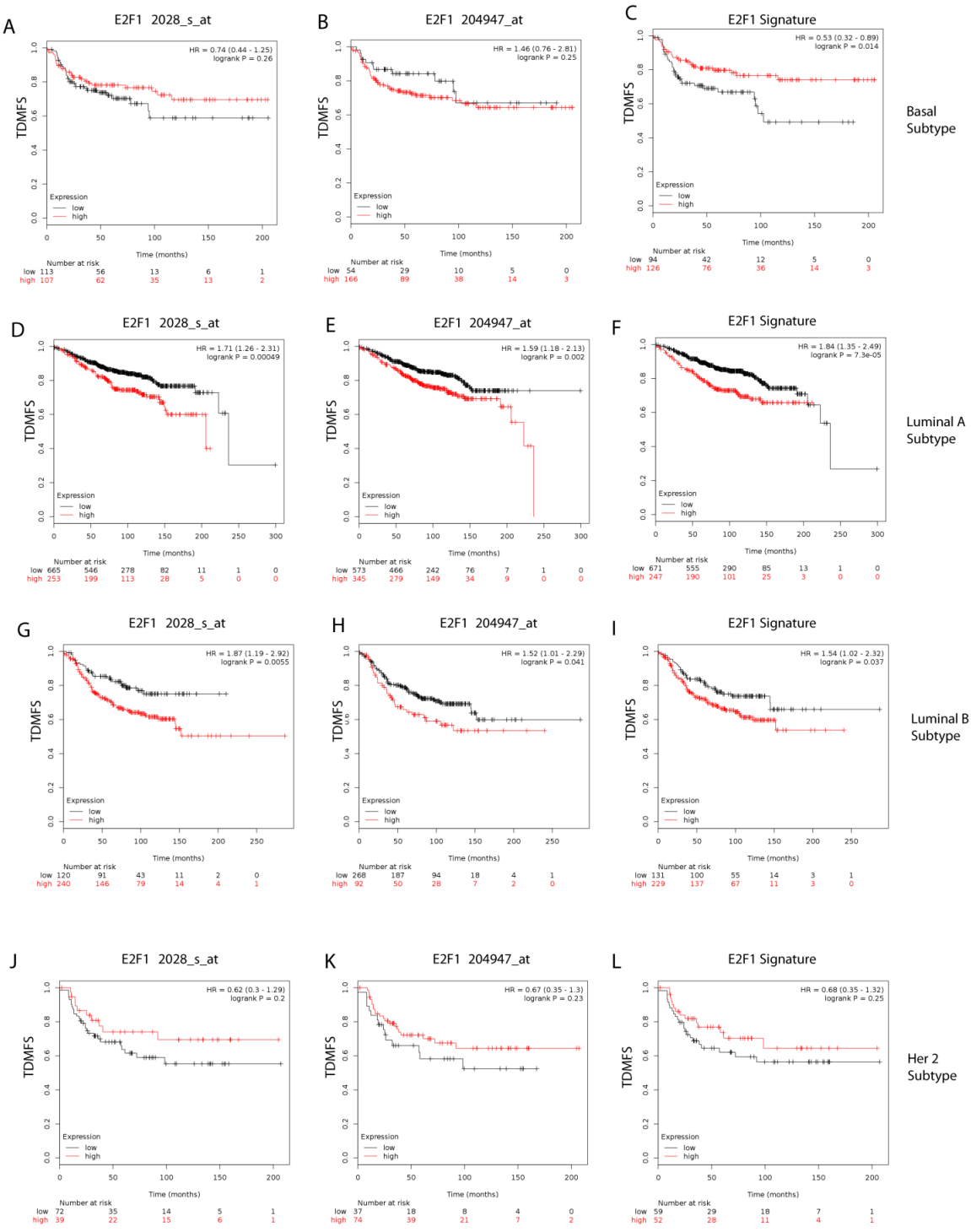


FIGURE S 3.2: ASSOCIATION OF E2F1 LEVELS AND SIGNATURE GENES DISTANT METASTASIS FREE SURVIVAL TIMES WITHIN SPECIFIC INTRINSIC SUBTYPES OF HUMAN BREAST CANCER

FIGURE S 3.2 (cont'd)

For each analysis using signature genes the average expression of the upregulated genes in the E2F1 signature were used during stratification. (A-C) The association of E2F1 levels and signature genes within the basal subtype of breast cancer are shown. (D-F) The association of E2F1 levels and signature genes with metastasis events in the luminal A subtype of breast cancer are shown(p=0.0049 for 2028_s_at, p=0.002 for 204947_at, and p=0.000073 for the analysis using the signature genes). (G-I) The association of E2F1 levels and signature genes with metastasis events in the luminal B subtype of breast cancer are shown(p=0.0055 for 2028_s_at, p=0.041 for 204947_at, and p=0.037 for the analysis using the signature genes). (J-L) The association of E2F1 levels and signature genes within the Her2 subtype of breast cancer are shown.

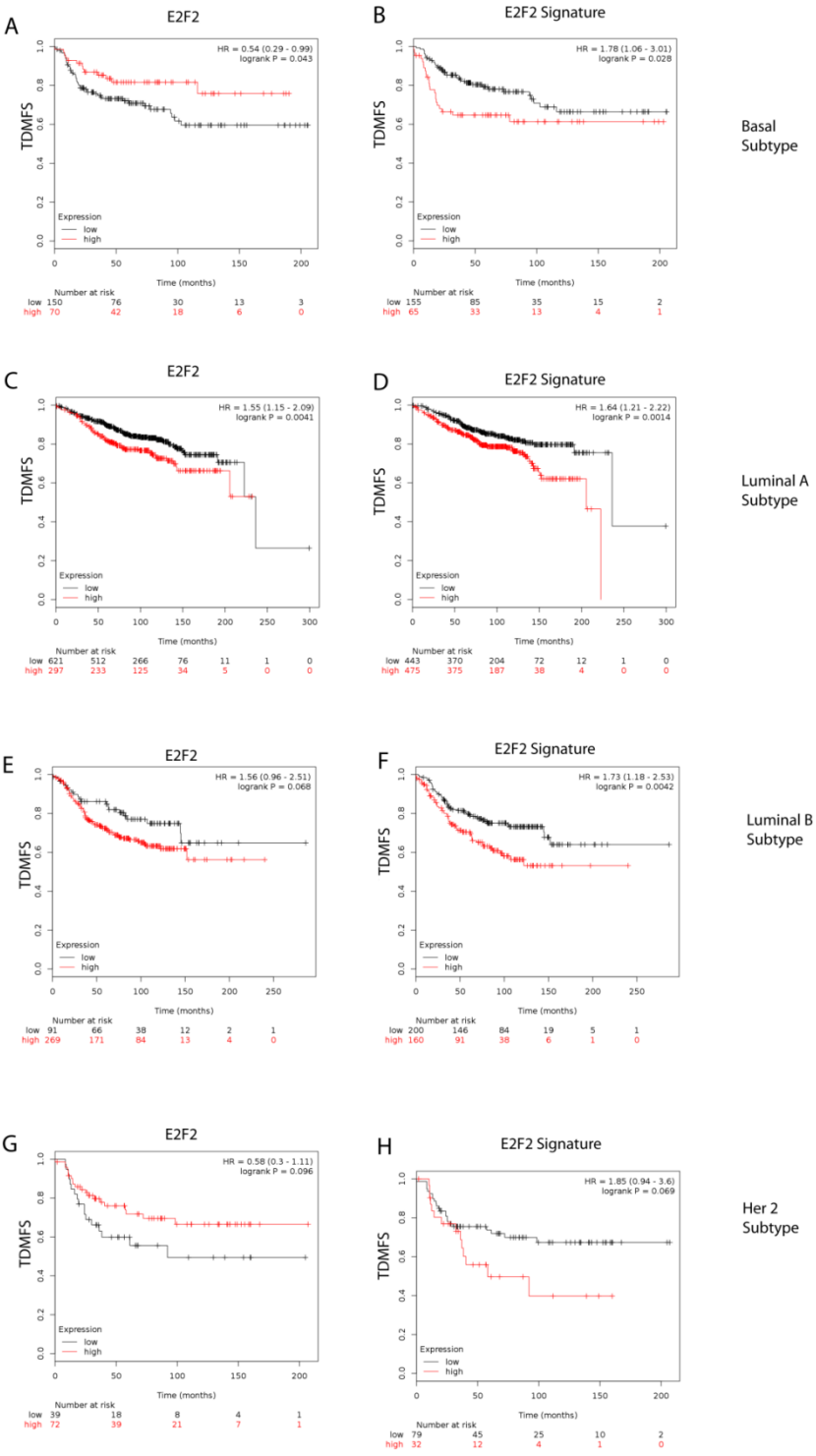


FIGURE S 3.3: ASSOCIATION OF E2F2 LEVELS AND SIGNATURE GENES DISTANT METASTASIS FREE SURVIVAL TIMES WITHIN SPECIFIC INTRINSIC SUBTYPES OF HUMAN BREAST CANCER

FIGURE S 3.3 (cont'd)

For each analysis using signature genes the average expression of the upregulated genes in the E2F2 signature were used during stratification. (A-B) The association of E2F2 levels and signature genes within the basal subtype of breast cancer are shown. While high levels of E2F2 are protective (p=0.043, high expression E2F2 signature genes predict faster onset of distant metastasis(p=0.028) (C-D) The association of E2F2 levels and signature genes with metastasis events in the luminal A subtype of breast cancer are shown(p=0.0041 for E2F2 levels, p=0.0014 for the analysis using the signature genes). (E-F) The association of E2F2 levels and signature genes with metastasis events in the luminal B subtype of breast cancer are shown(p=0.0042 for the analysis using the signature genes). (G-H) The association of E2F2 levels and signature genes within the Her2 subtype of breast cancer are shown.

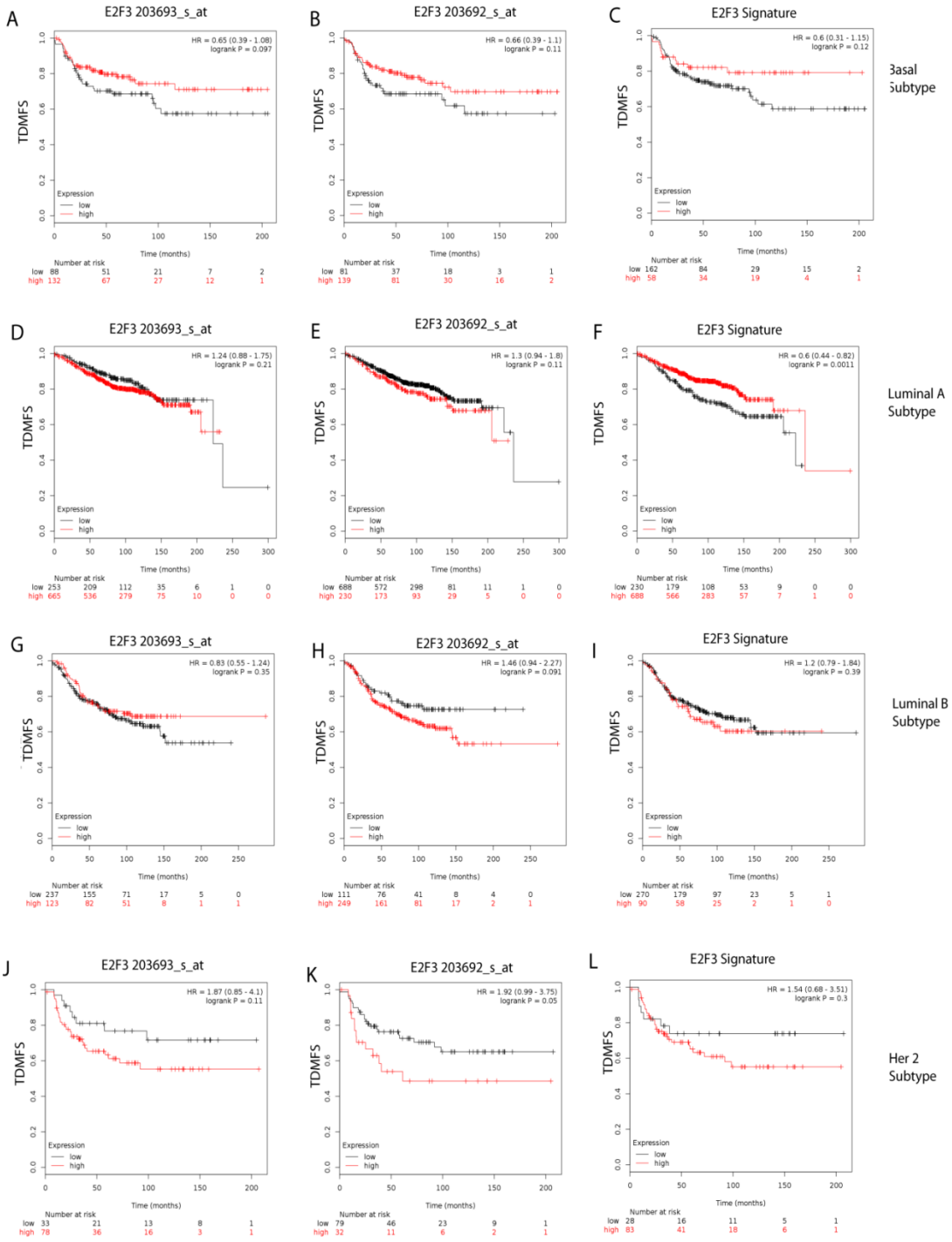


FIGURE S 3.4: ASSOCIATION OF E2F3 LEVELS AND SIGNATURE GENES DISTANT METASTASIS FREE SURVIVAL TIMES WITHIN SPECIFIC INTRINSIC SUBTYPES OF HUMAN BREAST CANCER

FIGURE S 3.4 (cont'd)

For each analysis using signature genes the average expression of the upregulated genes in the E2F3 signature were used during stratification. (A-C) The association of E2F3 levels and signature genes within the basal subtype of breast cancer are shown. (D-F) The association of E2F3 levels and signature genes with metastasis events in the luminal A subtype of breast cancer are shown. (G-I) The association of E2F3 levels and signature genes with metastasis events in the luminal B subtype of breast cancer are shown. (J-L) The association of E2F3 levels and signature genes within the Her2 subtype of breast cancer are shown.

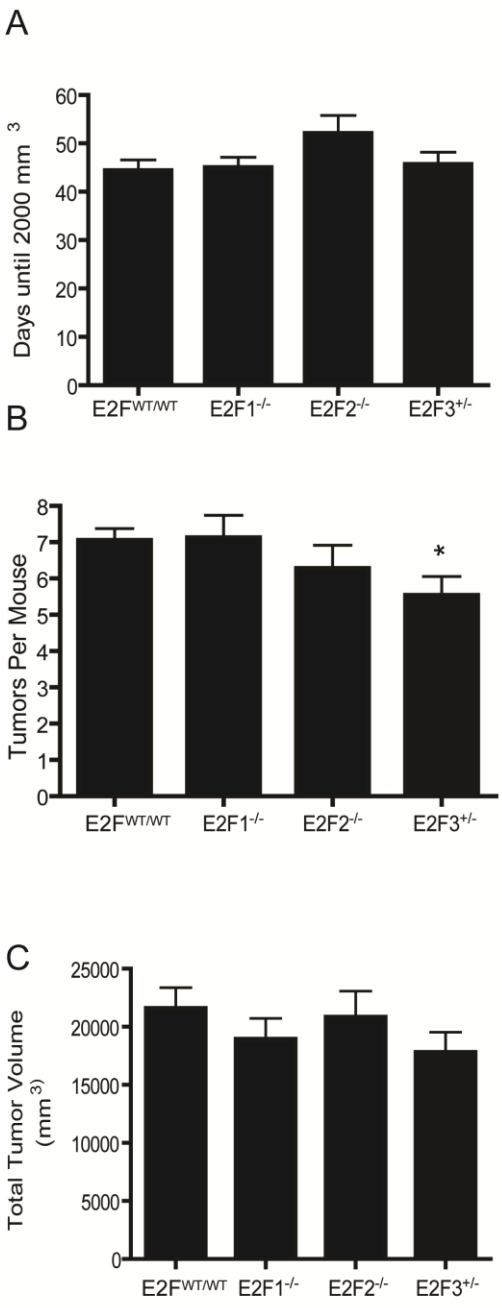


FIGURE S 3.5: LOSS OF E2FS DOES NOT AFFECT TUMOR GROWTH RATE OR TUMOR BURDEN

(A)The average tumor growth rate is depicted by the days until the primary tumor reaches 2,000mm^3 from initial palpation. At tumor endpoint, the number of tumors present on each mouse was counted. (B) The average number of tumors per mouse for each population of

FIGURE S 3.5 (cont'd)

$E2F^{WT/WT}$ or mutant mice. (C) The average volume (mm^3) from the sum of each tumor present within a mouse. All error bars represent SEM. All comparisons were made between E2F wild type controls and E2F knockout mice using a T-Test.

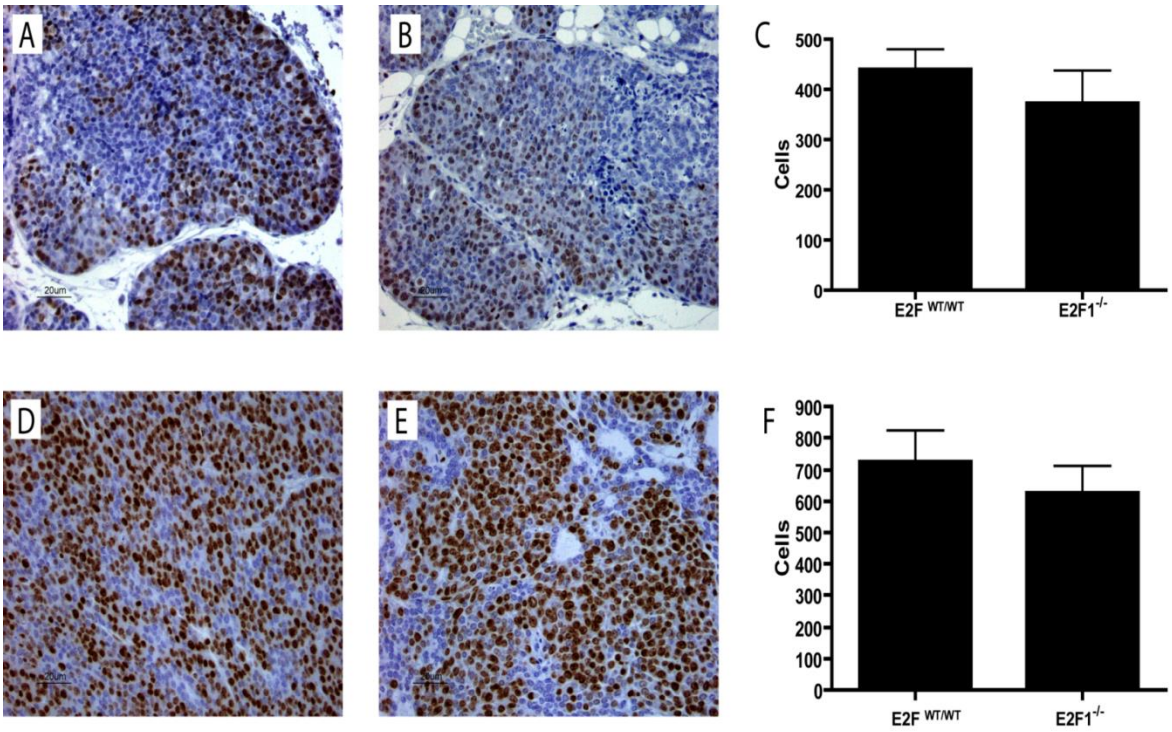


FIGURE S 3.6: E2F1 LOSS HAS NO EFFECT ON KI67 STAINING IN EARLY OR LATE STAGED TUMORS

(A) Representative IHC staining for KI67 in early stage (diameter=6mm) $E2F^{WT/WT}$ tumors (n=5). (B) Representative IHC staining for KI67 in early stage (diameter=6mm) $E2F1^{-/-}$tumors (n=5).(C) Quantification of IHC results showing the average number of KI67 positive cells per field in $E2F^{WT/WT}$ and $E2F1^{-/-}$ early stage tumors. (D) Representative IHC staining for KI67 in end stage (diameter=20mm) $E2F^{WT/WT}$ tumors (n=5). (E) Representative IHC staining for KI67 in end stage (diameter=20mm) $E2F1^{-/-}$tumors (n=5). (F) Quantification of IHC results showing the average number of KI67 positive cells per field in $E2F^{WT/WT}$ and $E2F1^{-/-}$ end stage tumors.

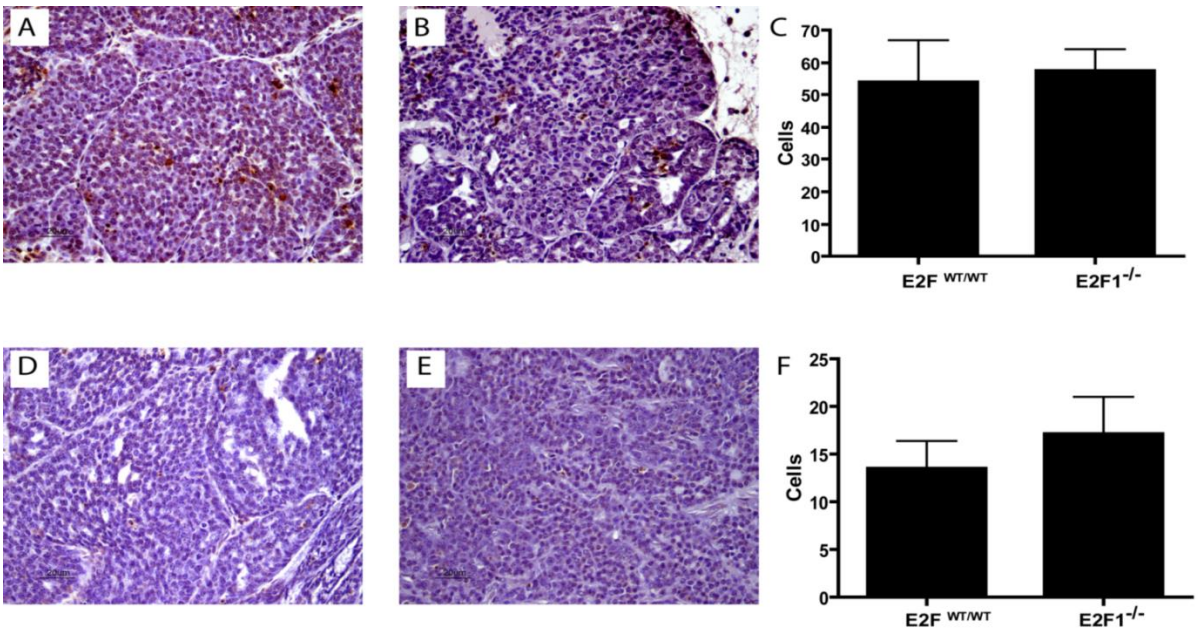


FIGURE S 3.7: E2F1 LOSS HAS NO EFFECT ON TUNEL STAINING IN EARLY OR LATE STAGED TUMORS

(A) Representative IHC staining for TUNEL in early stage (diameter=6mm) $E2F^{WT/WT}$ tumors (n=5). (B) Representative IHC staining for TUNEL in early stage (diameter=6mm) $E2F1^{-/-}$ tumors (n=5).(C) Quantification of IHC results showing the average number of TUNEL positive cells per field in $E2F^{WT/WT}$ and $E2F1^{-/-}$ early stage tumors. (D) Representative IHC staining for TUNEL in end stage (diameter=20mm) $E2F^{WT/WT}$ tumors (n=5). (E) Representative IHC staining for TUNEL in end stage (diameter=20mm) $E2F1^{-/-}$tumors (n=5). (F) Quantification of IHC results showing the average number of TUNEL positive cells per field in $E2F^{WT/WT}$ and $E2F1^{-/-}$ end stage tumors.

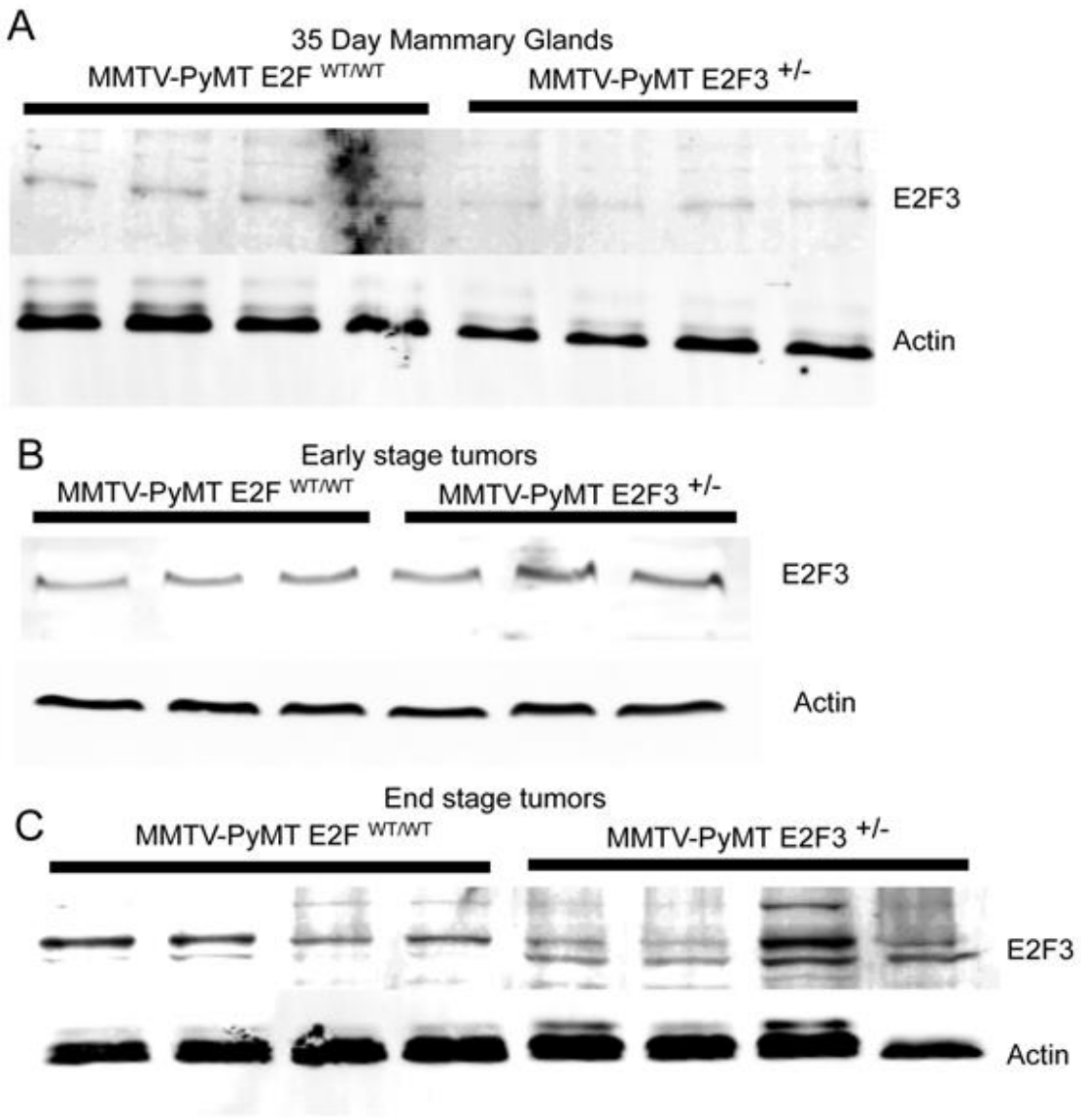


FIGURE S 3.8: WESTERN BLOT ANALYSIS SHOWS E2F3 PROTEIN LEVELS AT VARIOUS STAGES OF MMTV- PYMT TUMOR DEVELOPMENT

(A) A comparison of E2F3 protein levels in 35 day old mammary glands in E2F WT (n=4) and $E2F3^{+/-}$ (n=4) glands. (B) A comparison of E2F3 protein levels in early stage (diameter= 6mm) tumors from $E2F^{WT/WT}$ (n=3) and $E2F3^{+/-}$ (n=3) mice. (C) A comparison of E2F3 protein levels in end stage (diameter= 20mm) tumors from $E2F^{WT/WT}$ (n=4) and $E2F3^{+/-}$ (n=4) mice.

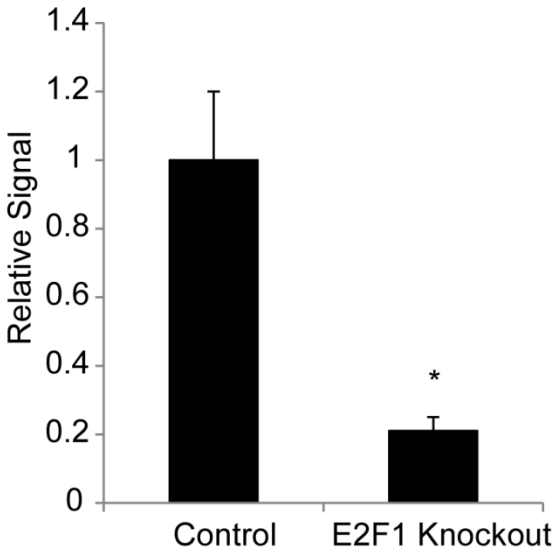


FIGURE S 3.9: LOSS OF E2FS REDUCES TRANSGENIC SIGNAL FOR CIRCULATING TUMOR CELLS

A comparison of qRT-PCR signal for expression of the transgene circulating tumor cell marker in transgenic control mice and E2F1 knockout mice reveals that loss of E2F1 significantly ($p<.0001$) reduces circulating tumor cells.

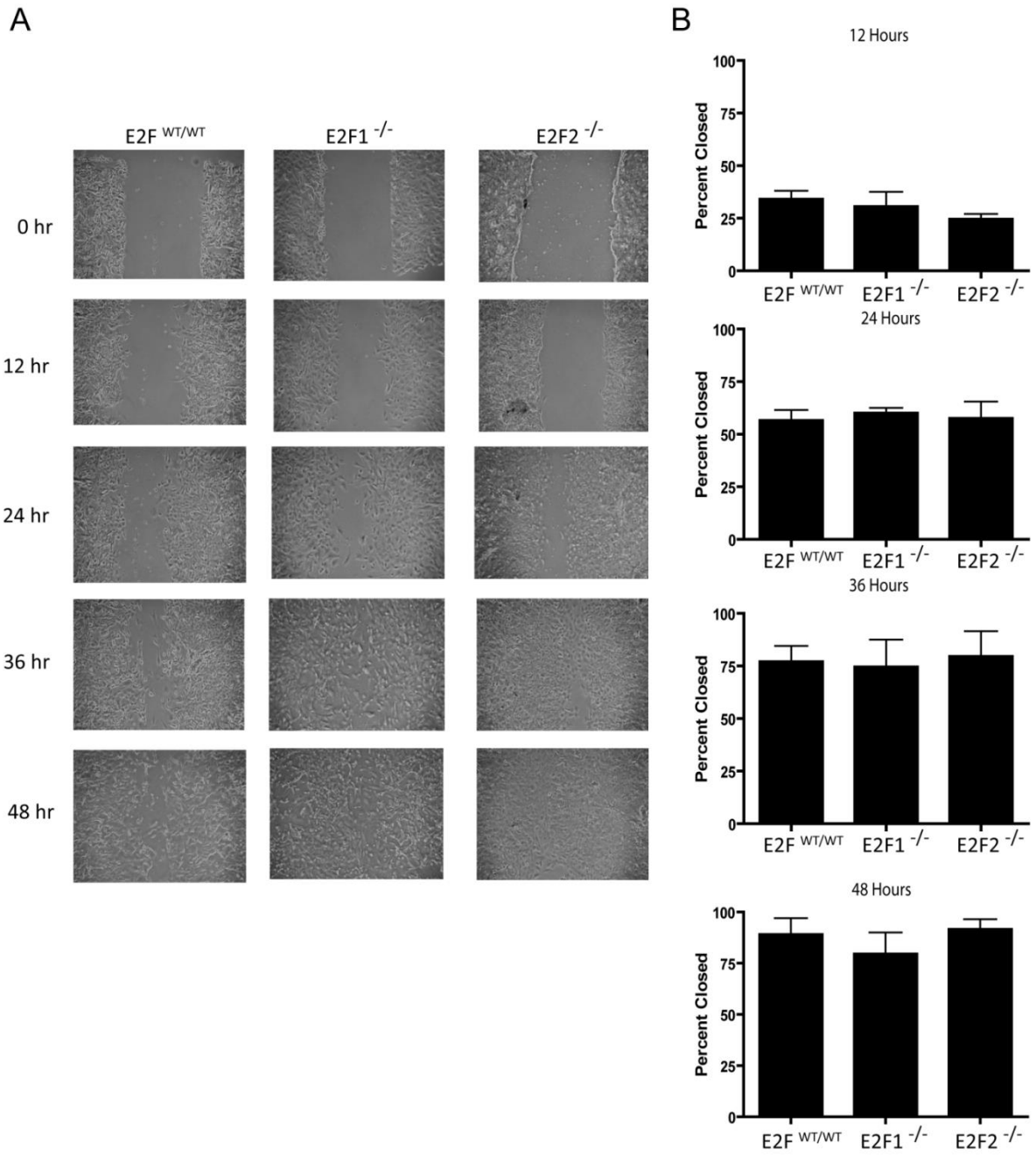


FIGURE S 3.10- WOUND HEALING ASSAY SHOWS MIGRATORY ABILITY OF TUMOR DERIVED CELLS FROM E2F$^{WT/WT}$, E2F1$^{-/-}$, AND E2F2$^{-/-}$ MICE

(A) Representative pictures of wound closure at indicated time points for E2F$^{WT/WT}$, E2F1$^{-/-}$, and E2F2$^{-/-}$ tumor cells. (B) Quantification of wound closure at indicated time points for E2F$^{WT/WT}$, E2F1$^{-/-}$, and E2F2$^{-/-}$ tumor cells.

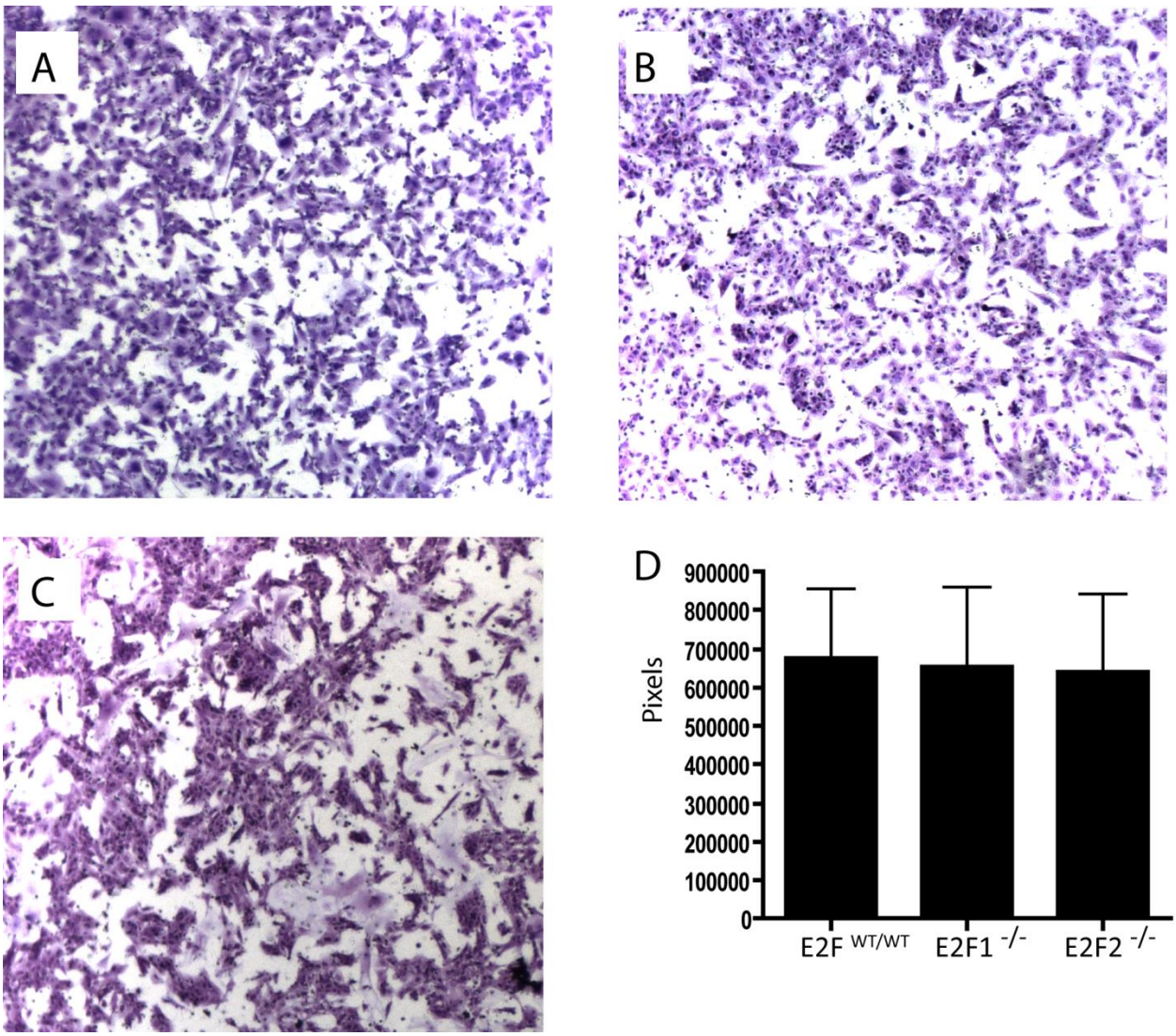


FIGURE S 3.11: TRANSWELL INVASION ASSAY SHOWS MIGRATORY ABILITY OF TUMOR DERIVED CELLS FROM E2F$^{WT/WT}$, E2F1$^{-/-}$, AND E2F2$^{-/-}$ MICE

(A) Representative picture for E2F$^{WT/WT}$ tumor cell migration in the transwell invasion assay. (B) Representative picture for E2F1$^{-/-}$ tumor cell migration in the transwell invasion assay. (C) Representative picture for E2F2$^{-/-}$ tumor cell migration in the transwell invasion assay. (D) Quantification of transwell invasion assay for E2F$^{WT/WT}$, E2F1$^{-/-}$ and E2F2$^{-/-}$ tumor cells.

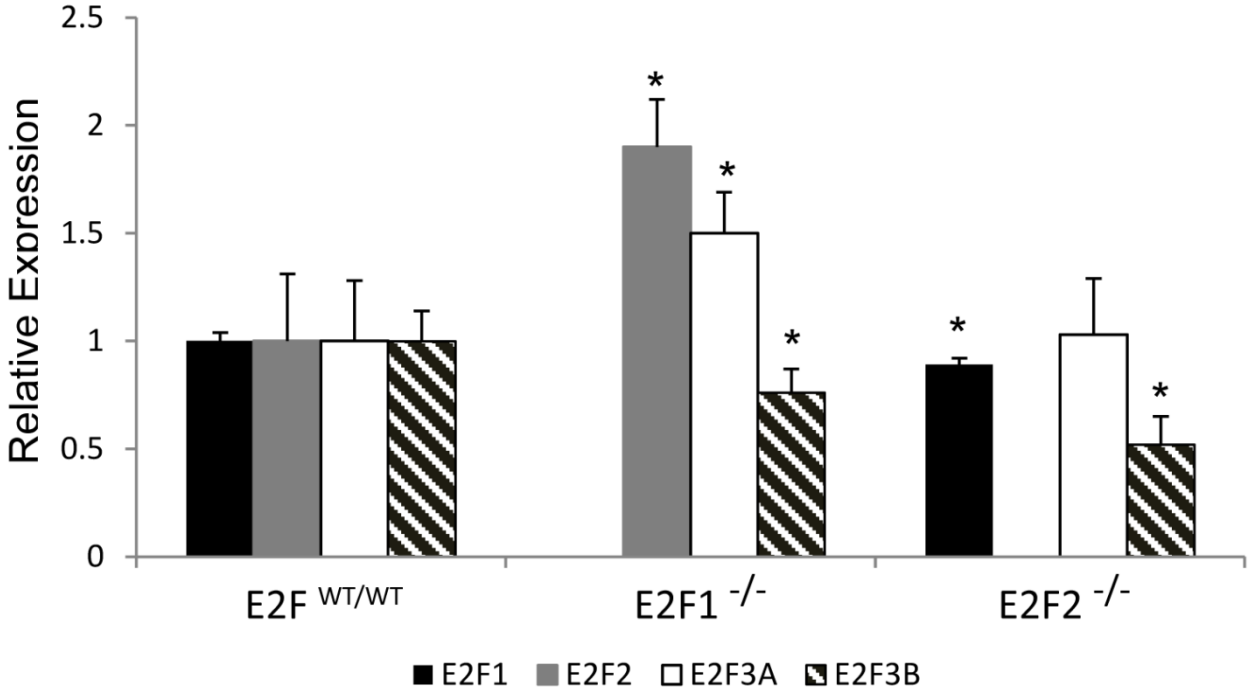


FIGURE S 3.12: RELATIVE EXPRESSION OF E2F1, E2F2, E2F3A, AND E2F3B IN MMTV-PYMT TRANSPLANTED TUMORS.

$E2F1^{-/-}$ tumors (n=4) had a significant increase in E2F2 (p=0.0032) and E2F3A (p=0.0254) expression with a significant decrease in E2F3B (p=0.0358). Similar to the spontaneous tumors, $E2F2^{-/-}$ transplanted tumors had a significant decrease in E2F1 expression (p=0.0046) and a significant downregulation (p=0.0024) of E2F3B.

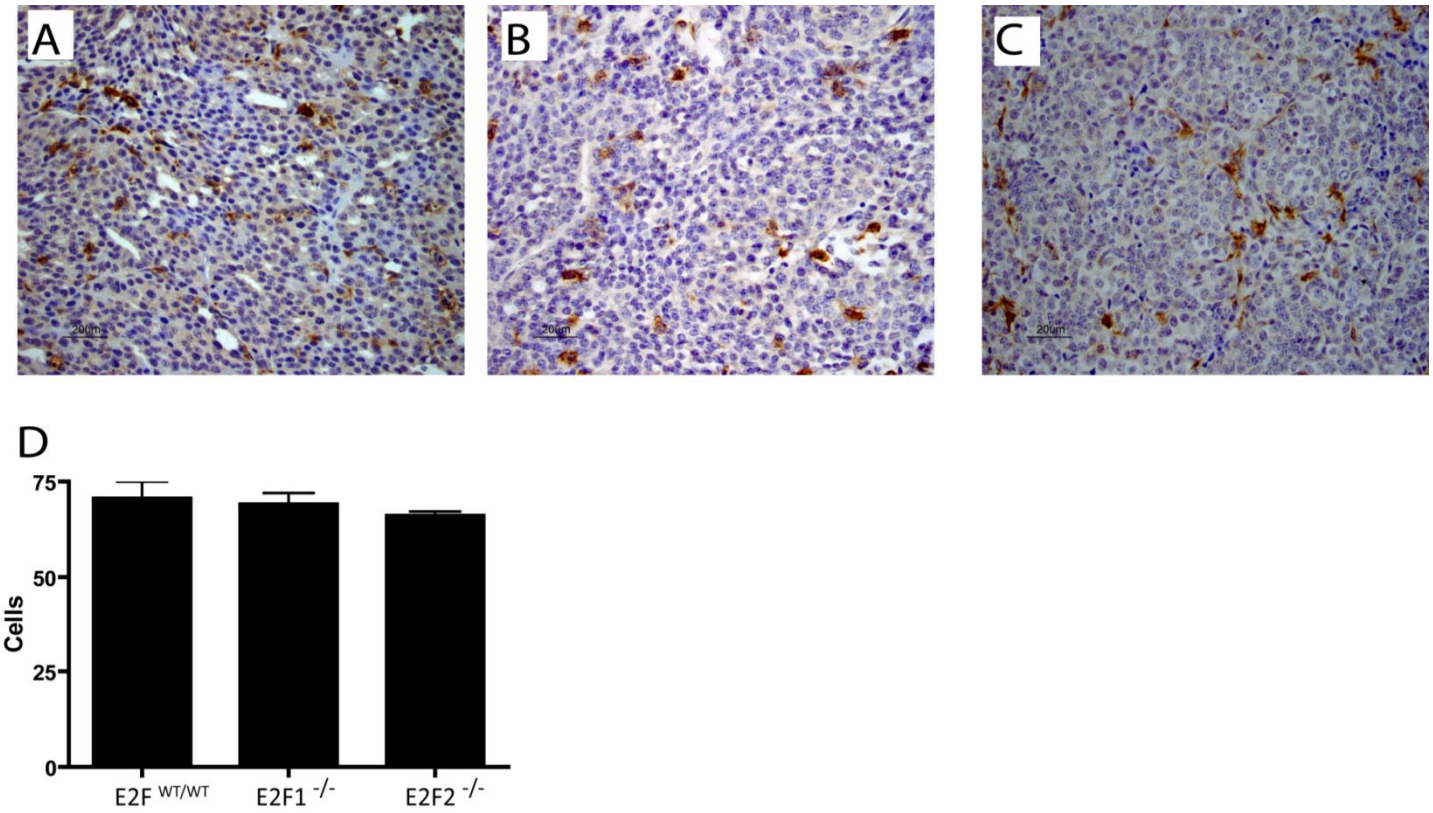


FIGURE S 3.13: E2F LOSS HAS NO EFFECT ON F4/80 STAINING IN END STAGE TUMORS

(A) Representative IHC staining for F4/80 in end stage (diameter=20mm) $E2F^{WT/WT}$ tumors (n=5). (B) Representative IHC staining for F4/80 in end stage (diameter=20mm) $E2F1^{-/-}$ tumors (n=5).(C) Representative IHC staining for F4/80 in end stage (diameter=20mm) $E2F2^{-/-}$ tumors (n=5) (D) Quantification of IHC results showing the average number of F4/80 positive cells per field across the various indicated genotypes.

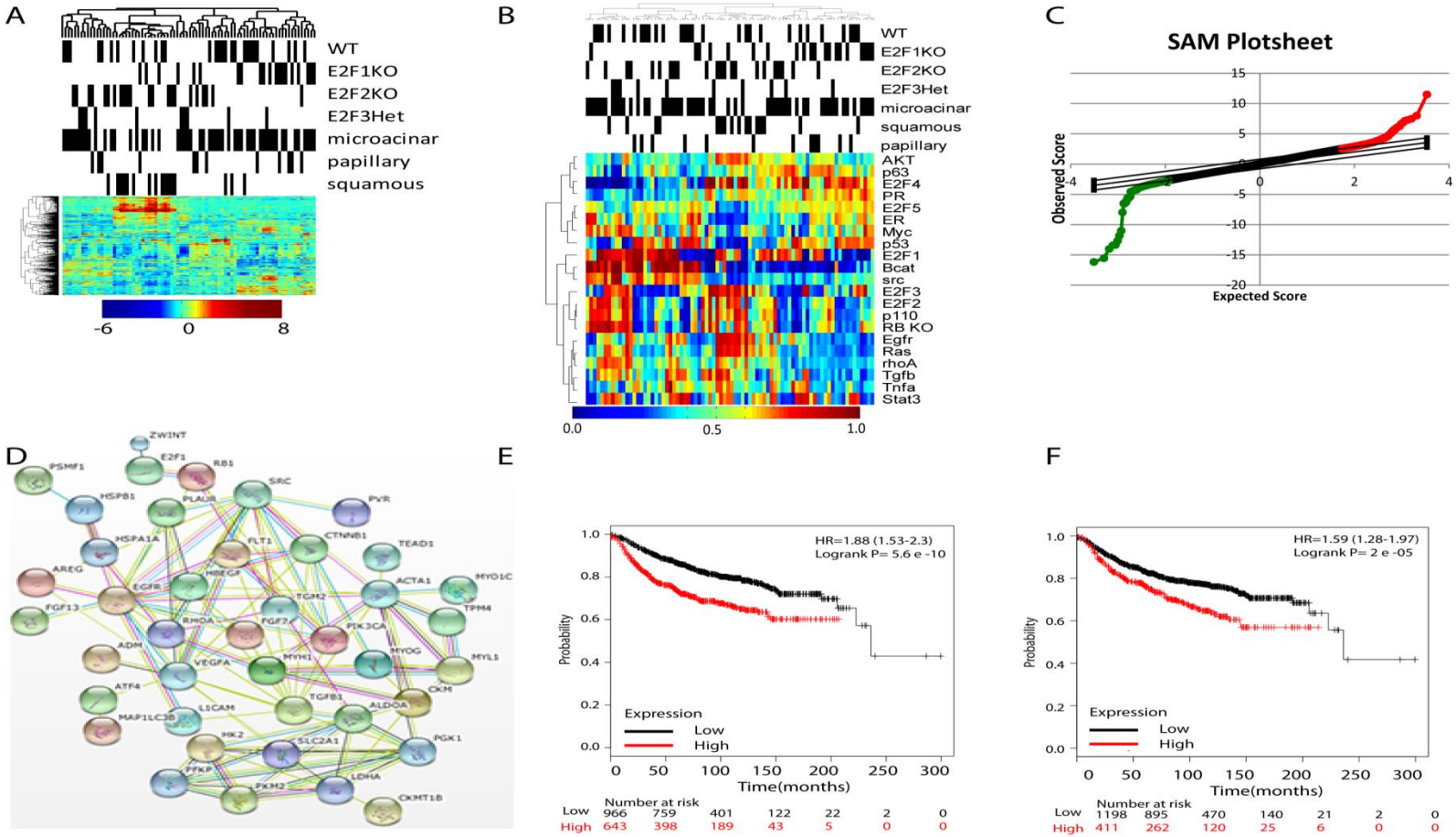


FIGURE 4.1: GENE EXPRESSION ANALYSIS OF MMTV-PYMT TUMORS REVEALS GENOMIC RESPONSE TO E2F1 LOSS AND POTENTIAL METASTATIC REGULATORS

FIGURE 4.1 (cont'd)

(A) Unsupervised hierarchical clustering of MMTV-PyMT tumor gene expression data. Black bars indicate the position of individual tumor samples and corresponding genotype and histology. The color bar depicts the range of expression values for the heatmap above. (B) Unsupervised hierarchical clustering of MMTV-PyMT tumor pathway activation predictions. Black bars indicate the position of individual tumor samples and corresponding genotype and histology. The color bar depicts the range of probability values for the heatmap above. (C) The significance analysis of microarrays plot illustrates the significant gene expression changes with loss of E2F1 in MMTV-PyMT tumors. The red bar indicates the genes that are upregulated with E2F1 loss and the green bar illustrates the genes that are downregulated with E2F1 loss. (D) Mapping the 55 genes that had concordant metastasis predictions to the Rb1, E2F1, Src, p110, EGFR, Ras, RhoA, and Tgfb pathways using the String-DB interaction network tool. (E) Kaplan Meier analysis of Adm using annotations for time until human breast cancer patients developed a distant metastasis. This analysis did not select for tumors of a specific intrinsic subtype. (F) Kaplan Meier analysis of Fgf13 using annotations for time until human breast cancer patients developed a distant metastasis. This analysis did not select for tumors of a specific intrinsic subtype.

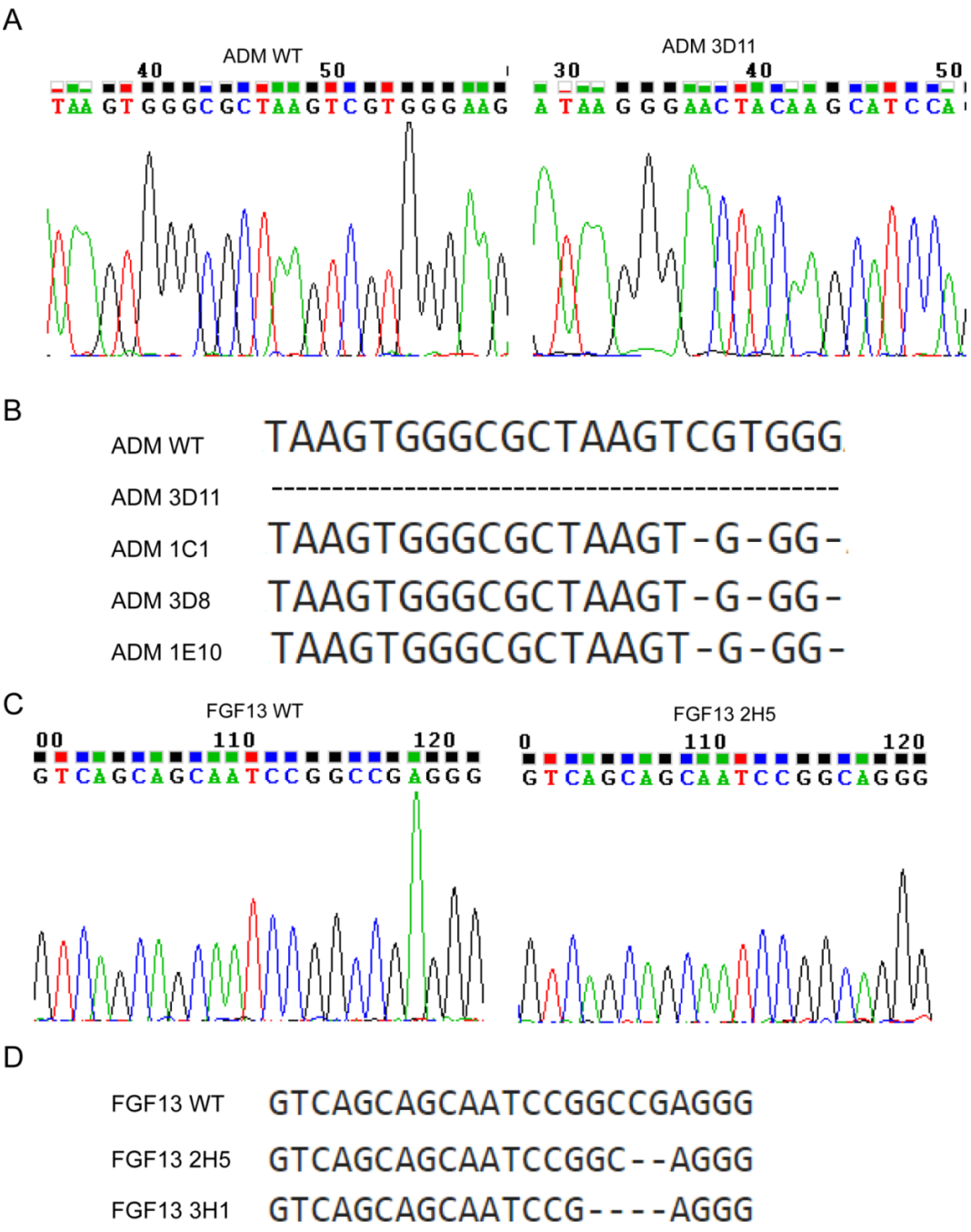


FIGURE 4.2: SEQUENCE TRACE AND ALIGNMENT FOR CRISPR-MEDIATED ADM AND FGF13 KNOCKOUT

(A) Sequence trace for Adm WT cells and Adm knockout clone 3D11. (B) Sequence alignment for the four Adm knockout clones. (C) Sequence trace for Fgf13 WT cells and Fgf 13 knockout clone 2H5. (D) Sequence alignment for the two Fgf 13 knockout clones.

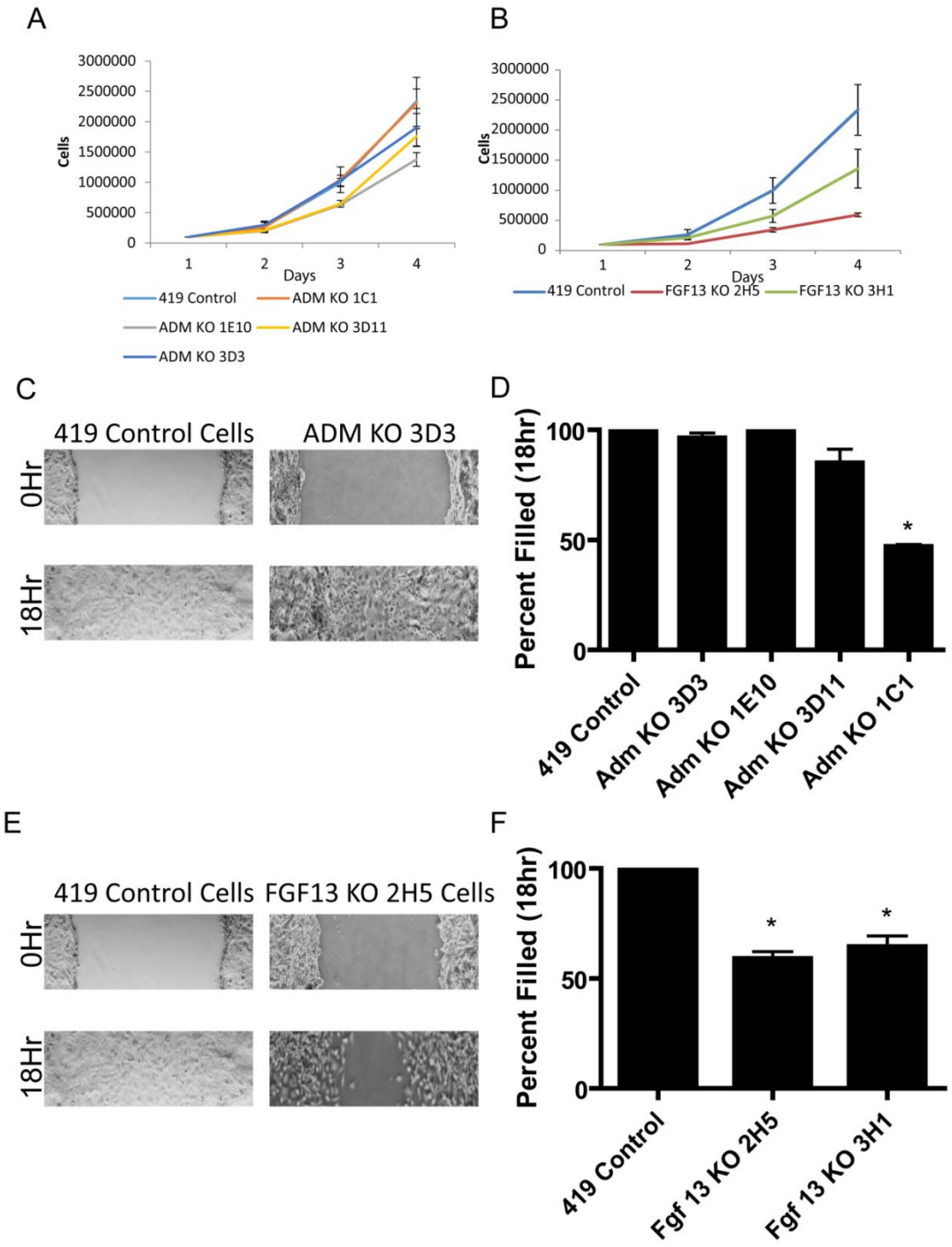


FIGURE 4.3: IN VITRO CHARACTERIZATION OF ADM AND FGF 13 KNOCKOUT CLONES

(A) Cell counts for 419 control cells and Adm KO clones over 4 days. For this experiment, 100,000 cells were seeded on day one. Cells were counted on each of the three following days.

FIGURE 4.3 (cont'd)

This experiment was done in triplicate. (B) Cell counts for 419 control cells and Fgf 13 KO clones over 4 days. For this experiment, 100,000 cells were seeded on day one. Cells were counted on each of the three following days. This experiment was done in triplicate. (C) Example wound healing assay results at 0 and 18 hours for 419 control cells and the Adm KO clone 3D3. (D) Quantification for wound healing at 18 hours for 419 control cells and Adm KO cells. This experiment was done in triplicate. (*= $p>0.0001$) (E) Example wound healing assay results at 0 and 18 hours for 419 control cells and the Fgf 13 KO clone 2H5. (F) Quantification for wound healing at 18 hours for 419 control cells and Fgf 13 KO cells. This experiment was done in triplicate. (*= $p>0.0001$)

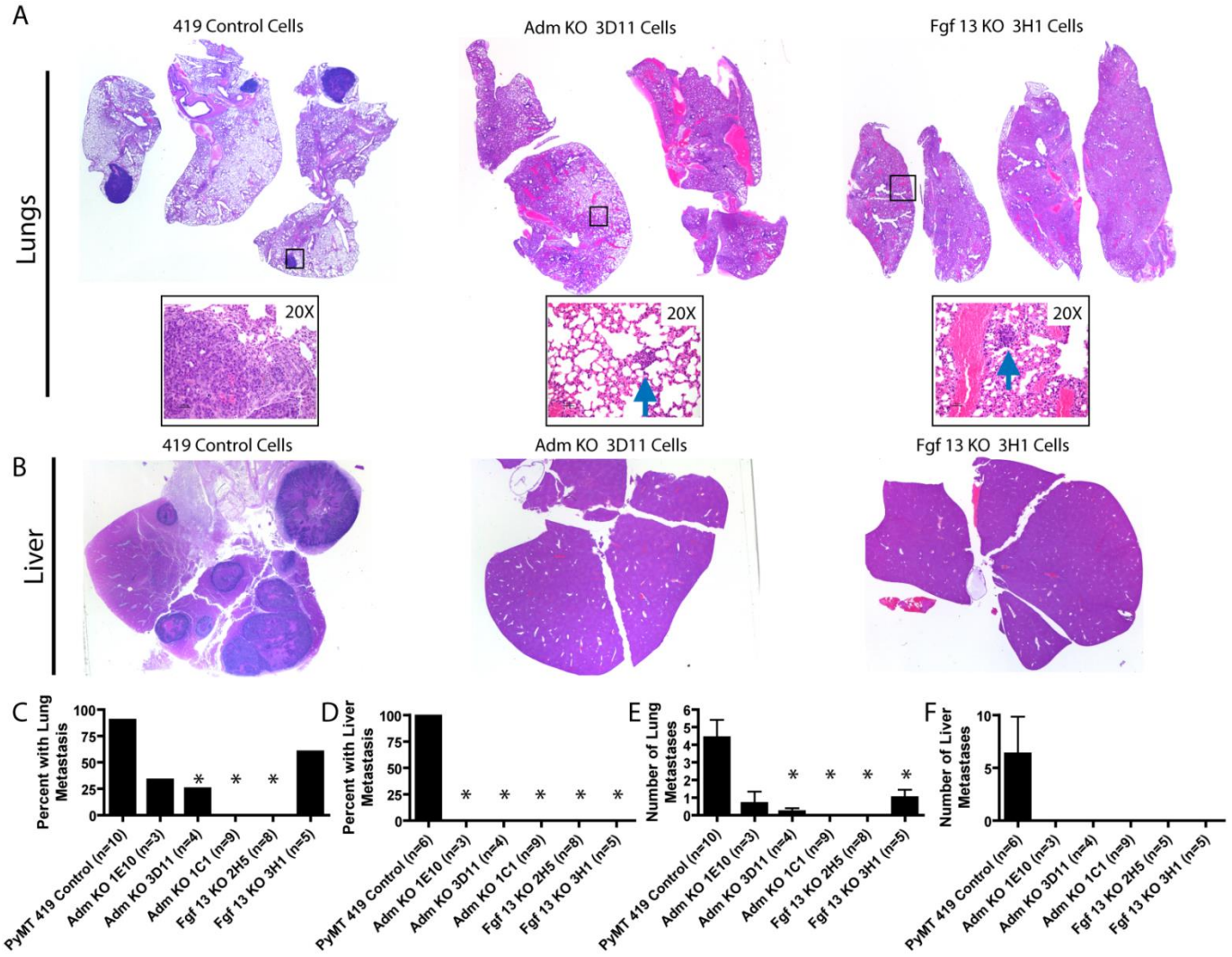


FIGURE 4.4: KNOCKOUT OF ADM OR FGF 13 INHIBITS METASTASIS TO THE LUNGS AND LIVER

FIGURE 4.4 (cont'd)

(A) Representative histological sections for the lungs of mice receiving MMTV-PyMT 419 control cells, Adm KO cells, and Fgf 13 KO cells by tail vein injection of 50,000 cells. Mice were euthanized 21 days following injection. Black boxes highlight the location of the inset. Inset photos were taken at 20X and blue arrow heads highlight the presence of a micro metastasis. (B) Representative histological sections for the liver of mice receiving MMTV-PyMT 419 control cells, Adm KO cells, and Fgf 13 KO cells. (C) Quantification for the percentage of mice with lung metastasis detected in a single section of lungs or as detected during necropsy. (D) Quantification for the percentage of mice with lung metastasis detected in a single section of liver. (E) The number of lung metastases detected in a single section of lungs from mice receiving individual control, Adm KO, and Fgf 13 KO clones. (F) The number of liver metastases detected in a single section of liver from mice receiving individual control, Adm KO, and Fgf 13 KO clones.

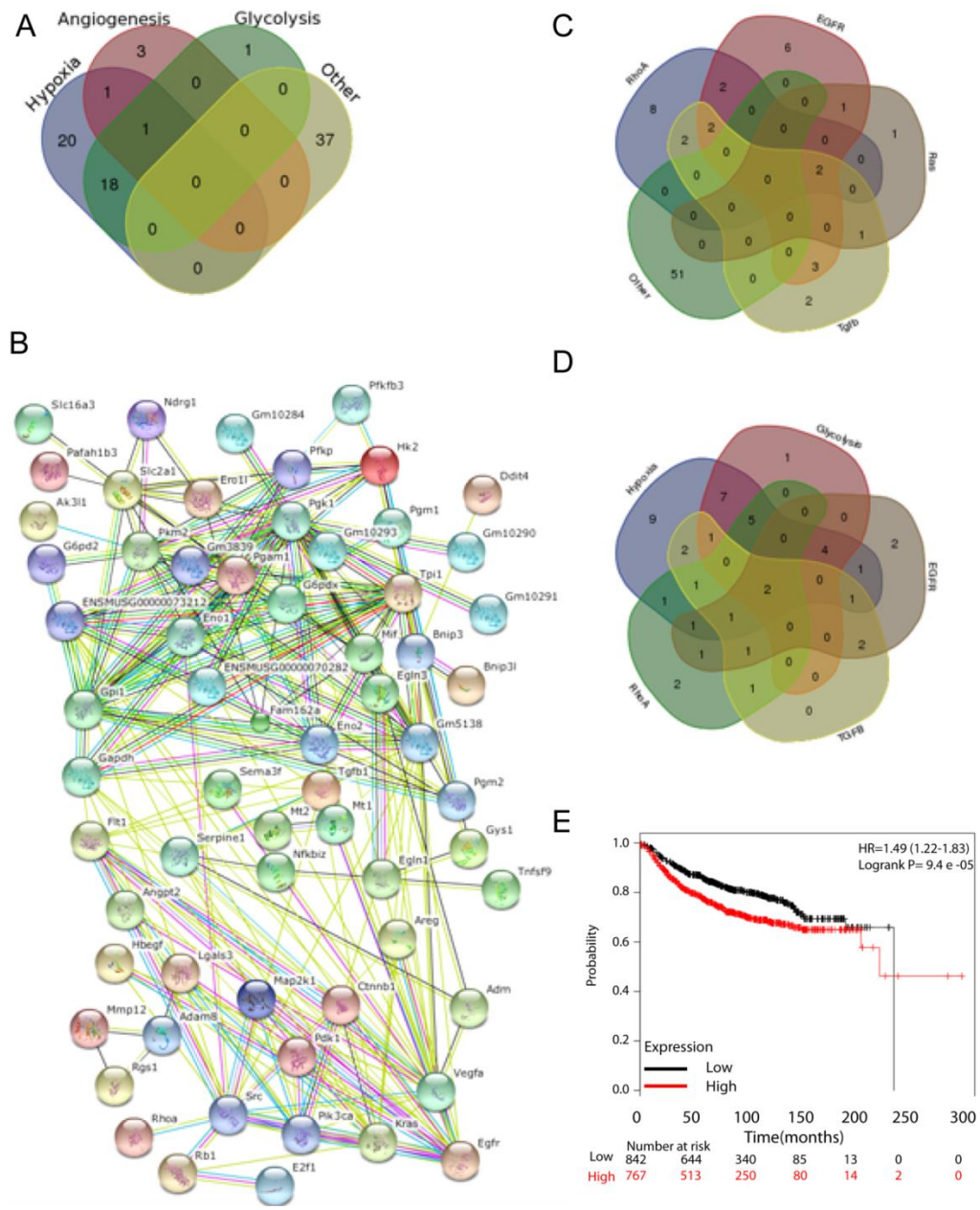


FIGURE 4.5: ANALYSIS OF THE ADM COVARIANCE NETWORK REVEALS AN ASSOCIATION WITH HYPOXIA RESPONSE, MAJOR CELL SIGNALING PATHWAYS, AND ACCELERATION OF TIME UNTIL DISTANT METASTASIS IN HUMAN BREAST CANCER

(A) Venn diagram analysis depicting the Adm covariance network association with hypoxia response, angiogenesis, glycolysis or other categories. (B) Interaction network analysis of the

FIGURE 4.5 (cont'd)

Adm covariance network genes with the Egfr, Beta-catenin, Ras, Src, Rb, E2F1, PI3K, RhoA, and Tgfb pathways. (C) Venn diagram illustrating the number of Adm covariance network genes that are upregulated in response to RhoA, Egfr, Ras, or Tgfb pathways. (D) Venn diagram showing the relationship of the Adm covariance network genes that are upregulated in response to RhoA, Egfr, or Tgfb pathways that are also glycolysis or hypoxia response genes. (E) Kaplan Meier analysis of Adm covariance network genes as a signature to test for correlation with annotations for time until human breast cancer patients developed a distant metastasis. This analysis did not select for tumors of a specific intrinsic subtype.

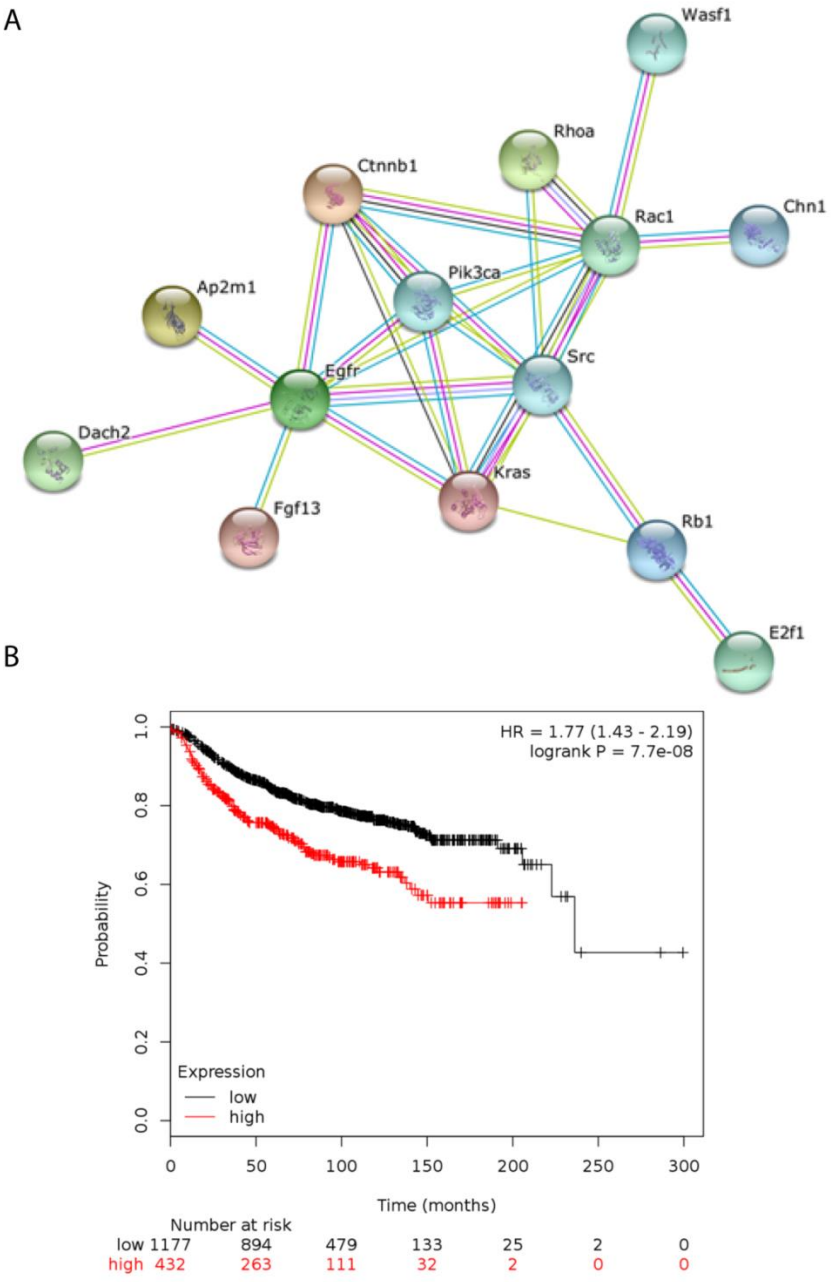


FIGURE 4.6: THE FGF 13 COVARIANCE NETWORK ASSOCIATES WITH MAJOR CELL SIGNALING PATHWAYS AND EARLIER HUMAN BREAST CANCER METASTASIS EVENTS

(A) String interaction network for Fgf 13 covariance network genes and cell signaling pathways with low activity in E2F1 $^{-/-}$ tumors reveals an association for a subset of Fgf13 covariance

FIGURE 4.6 (cont'd)

network genes and the Ras, Egfr, RhoA, Src, Rb, E2F1, and beta-catenin pathways. (B) Kaplan Meier analysis for the Fgf 13 covariance network genes as a signature shows these genes are significantly associated with earlier human breast cancer metastasis. This analysis did not select for tumors of a specific intrinsic subtype.

TABLE 4.1: PRO-METASTATIC GENES SIGNIFICANTLY DOWNREGULATED IN E2F1 -/- TUMORS COMPARED TO E2F WT/WT TUMORS

Gene Symbol	q-value(%)	Fold Change	Human Breast Cancer Metastasis Prediction Upregulation Accelerates Metastasis Onset	E2F Target	Hypoxia (mSigDB)	Glycolysis (mSigDB)	Angiogenesis (mSigDB)	Cytoskeleton regulation(mSigDB)	TGFB Pathway(mSigDB)	Rho A Pathway (mSigDB)	Egfr Pathway (mSigDB)	Beta-Catenin Pathway (mSigDB)	PI3K PI3K (mSigDB)	Ras Pathway (mSigDB)	Src Pathway (mSigDB)	Demonstrated as regulator of breast cancer metastasis (reference)	Demonstrated as regulator of breast cancer metastatic features *in vitro* (reference)
Vegfa	1.888174	1.630074	A,B, LA, H2	Yes	Hypoxia	Glycolysis	Angiogenesis		Tgfb	RhoA	Egfr					[199]	
Hk2	0.465832	1.435557	H2	Yes	Hypoxia	Glycolysis			Tgfb	RhoA	Egfr						
Ldha	1.888174	1.131396	A, LB	Yes	Hypoxia	Glycolysis				RhoA							
Pgk1	0.773842	1.250938	A, B, LA	Yes	Hypoxia	Glycolysis				RhoA							
Aldoa	0.773842	1.217873	A, B, LA	Yes	Hypoxia	Glycolysis											
Pfkp	0	1.415108	A, B, LA	Yes	Hypoxia	Glycolysis											
Pkm2	1.22928	1.145071	H2	Yes	Hypoxia	Glycolysis											
Coro1c	2.892629	1.189088	A, LA	Yes	Hypoxia			Cytoskeleton regulation		RhoA							[211]
Myo1c	4.307465	1.179644	LB, H2	Yes	Hypoxia			Cytoskeleton regulation									
Slc2a1	0.773842	1.445333	A, B, LA, LB, H2	Yes	Hypoxia				Tgfb	RhoA							
Maff	0.465832	1.478268	A,B	Yes	Hypoxia				Tgfb		Egfr						
Adm	0.773842	2.203166	A,B,LA, LB	Yes	Hypoxia					RhoA	Egfr						
Tgm2	0	2.230846	LB	Yes	Hypoxia						Egfr						
Akap8	2.892629	1.08564	LB	Yes	Hypoxia												
Lama5	0	1.329392	A, LA, H2	Yes	Hypoxia												[212]
P4hb	4.307465	1.171495	A, H2	Yes	Hypoxia												
Hbegf	0	1.673486	LB	Yes			Angiogenesis		Tgfb	RhoA	Egfr		PI3K	Ras		[200]	
Hspb1	2.892629	1.251572	A,B, LA, LB, H2	Yes				Cytoskeleton regulation						Ras		[201]	
Fgf13	1.22928	1.476689	A,B, LB, H2	Yes				Cytoskeleton regulation									
Tpm4	1.888174	1.138223	A, LA, H2	Yes				Cytoskeleton regulation									
Pvr	0	1.30474	A, LB	Yes					Tgfb	RhoA	Egfr						
Slc20a1	0	2.388173	A, LA	Yes					Tgfb		Egfr						
Plk3	1.22928	1.44195	A, LA, LB, H2	Yes					Tgfb								
Il1rap	0	1.494084	B	Yes					Tgfb								
Tubb6	2.892629	1.434252	A,B, LA	Yes						RhoA	Egfr						
Hnrpdl	0.773842	1.182597	A, H2	Yes							Egfr						

TABLE 4.1 (cont'd)

Zwint	2.892629	1.216663	A, LA, LB	Yes							Egfr						
Aire	0	1.372985	A,B	Yes													
Atf4	1.888174	1.154385	H2	Yes													
Chac1	1.22928	1.654794	A,LA	Yes													
Mafk	1.888174	1.136931	H2	Yes													
Map1lc3b	2.892629	1.150717	A,B, LA	Yes													
Psmf1	4.307465	1.166417	LA	Yes													
Tead1	2.892629	1.24441	B, LB, H2	Yes													[210]
Flt1	0	1.288865	A,LA, H2		Hypoxia		Angiogenesis		Tgfb							[202]	
Gys1	1.22928	1.268834	A, LB		Hypoxia	Glycolysis											
L1cam	2.892629	1.523696	A, LA		Hypoxia											[203]	
Plaur	0.773842	1.235096	A		Hypoxia				Tgfb	RhoA	Egfr			Ras		[204]	
Gdpd3	0	1.572679	LA, H2		Hypoxia										Src		
Hspa1a	1.888174	1.429015	A,LA,LB		Hypoxia												
Acta1	0.773842	5.110624	A, LA, LB					Cytoskeleton regulation	Tgfb	RhoA							
Fgf7	2.892629	1.16499	B					Cytoskeleton regulation	Tgfb				PI3K				[214]
Ckm	1.888174	2.691264	LB														
Myh1	1.22928	6.52414	H2														
Myl1	1.22928	3.913239	LA														
Myog	1.22928	1.159692	LB														
Pvalb	4.307465	2.233215	LA, LB														
Ryr1	2.892629	2.02497	A, LA														
Sln	2.892629	1.110799	B, LB														
Alpk3	1.22928	1.25759	H2														
Areg	1.888174	1.704905	B								Egfr						[209]
Ckmt1	1.22928	1.388296	A, B, LA, LB, H2														
Mlf1	2.892629	1.441451	A, LA, LB														
Mt2	4.307465	1.55117	A, B, H2							RhoA							
Wnt8b	1.888174	1.132226	LB														

TABLE 4.2: A SUMMARY OF THE METASTATIC FUNCTIONS OF CELL SIGNALLING PATHWAYS WITH LOW ACTIVITY IN E2F1 -/- TUMORS

Pathway	Model	Description	Reference
Src	MDA-MB-231 cells in nude mice.	Src kinase dead MDA-MB-231 cells exhibited a reduction in bone metastasis following intracardiac injection. In addition, Src kinase dead MDA-MB-231 cells reduced metastasis to the lungs compared to controls following tail vein injection.	[206]
Tgfb	MMTV-PyMT Mice, MMTV-PyMT Tumor Cells in FVB mice.	Doxycycline induction of Tgfb1 increased lung metastases in the transgenic mice. Antisense-mediated inhibition of Tgfb1 reduced metastases following orthotopic injection of tumor cells..	[205]
Beta-catenin	Tumor cells from the Erbb2 knockin transgenic mouse model.	shRNA to knockdown beta-catenin expression in Errb2 knockin tumors cells impaired metastasis followig orthotopic injection.	[132]
RhoA	MDA-MB-231 cells in nude mice.	shRNA to knockdown rhoA expression in MDA-MB-231 cells reduced lung metastasis following tail vein injection.	[207]
Egfr	SUM149 cells in nude mice.	Treatment of mice with the Egfr inhibitor erlotinib inhibited metastasis following orthotopic injection.	[208]

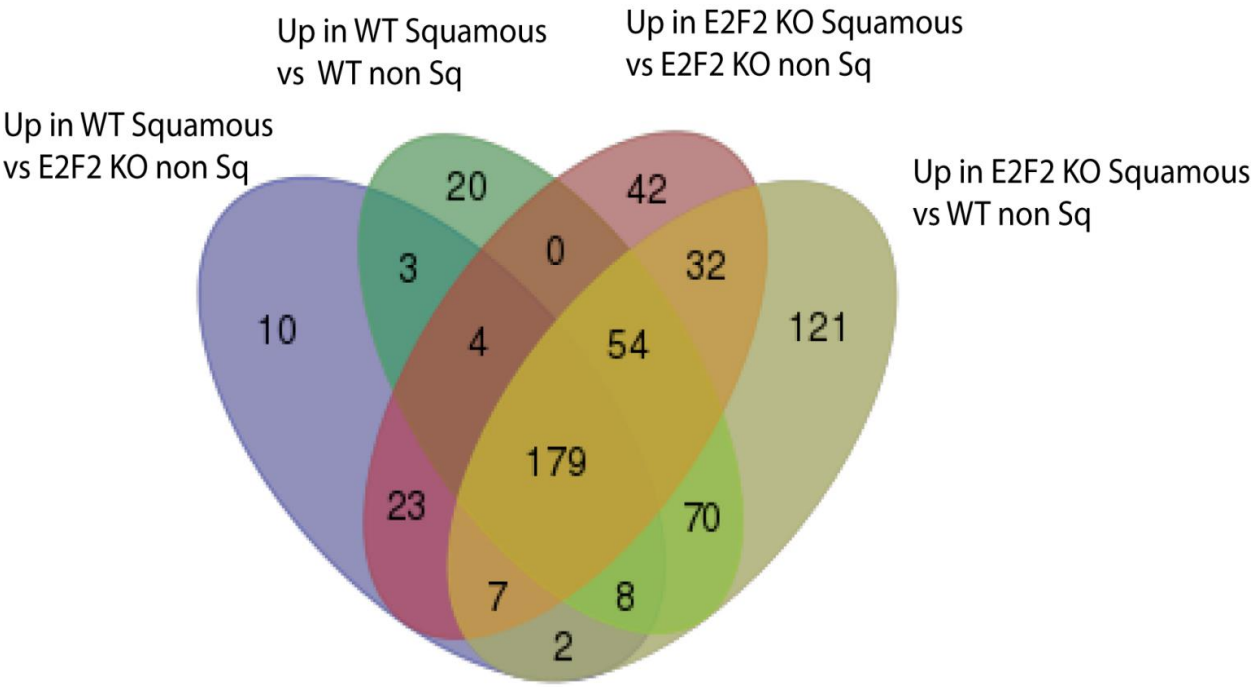
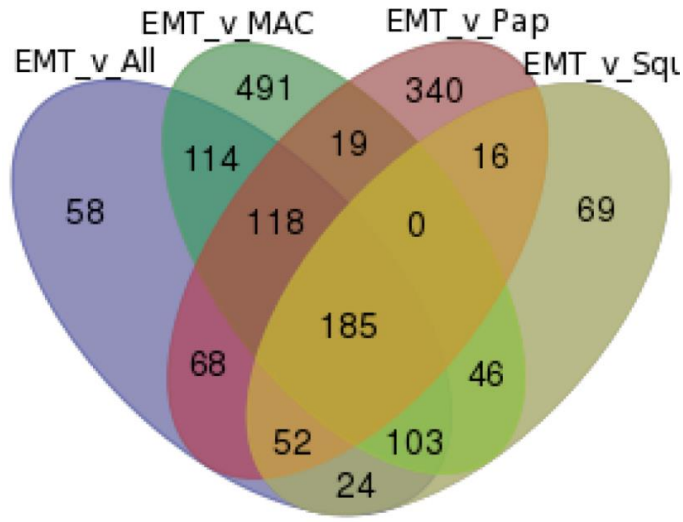


FIGURE 5.1: VENN DIAGRAM ILLUSTRATION OF THE IDENTIFICATION OF SQUAMOUS SIGNATURE GENES

To identify squamous genes we used significance analysis of microarrays to set up the following comparisons: E2F $^{WT/WT}$ squamous tumors compared to E2F $^{WT/WT}$ non-squamous tumors, E2F2 $^{-/-}$ squamous tumors compared to E2F2 $^{-/-}$ non-squamous tumors, E2F2 $^{-/-}$ squamous tumors compared to E2F $^{WT/WT}$ non-squamous tumors, and E2F $^{WT/WT}$ squamous tumors compared to E2F2 $^{-/-}$ non-squamous tumors. The Venn diagram shows the overlap of all of the comparisons, revealing 179 genes that are consistently upregulated in squamous tumors from MMTV-PyMT mice.

A



B

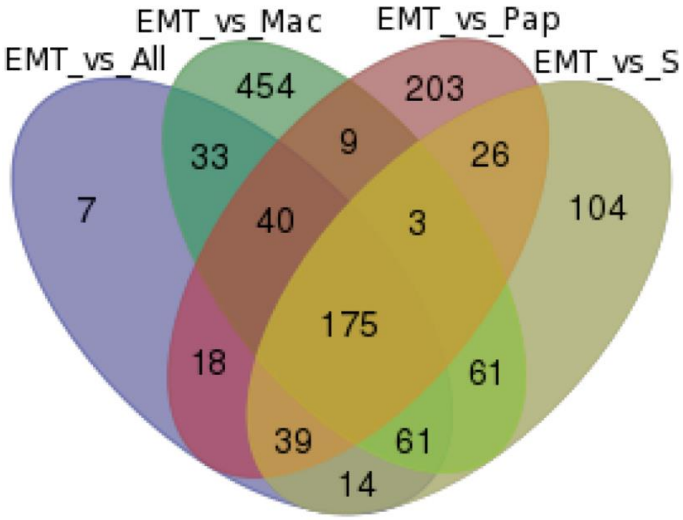


FIGURE 5.2: VENN DIAGRAM ILLUSTRATION OF THE IDENTIFICATION OF EMT-LIKE SIGNATURE GENES

To identify EMT-like genes we used significance analysis of microarrays to set up the following comparisons: EMT-like tumors compared to all non-EMT-like tumors, EMT-like tumors compared to squamous tumors, EMT-like tumors compared to papillary tumors, and EMT-like

FIGURE 5.2 (cont'd)

tumors compared to microacinar tumors. Signature genes focused on genes that were differentially regulated in each comparison. (A) The overlapping genes that were upregulated in each of the comparisons. In total, the analysis revealed 185 genes consistently upregulated in EMT-like tumors. (B) The overlapping genes that were downregulated in each of the comparisons, depicting 175 genes consistently downregulated in EMT-like tumors.

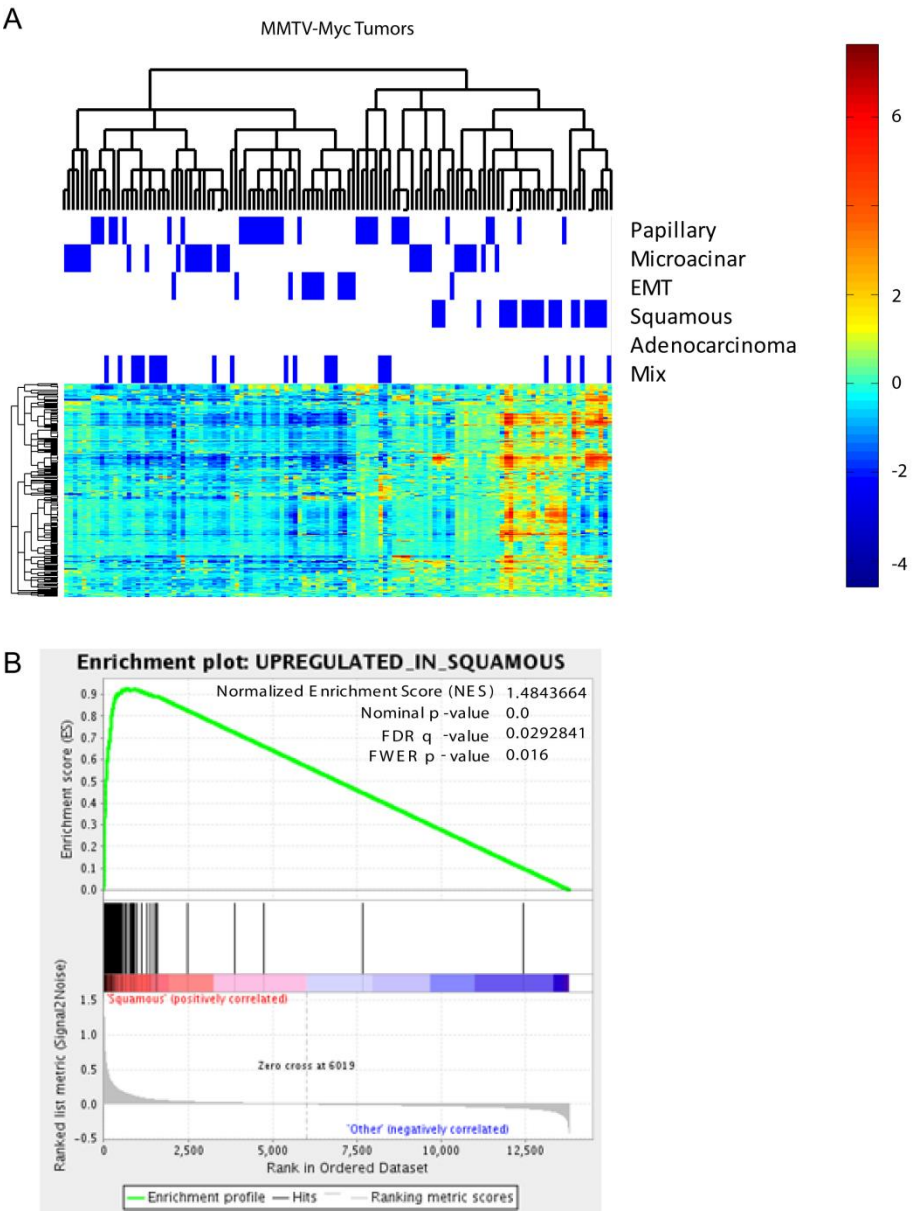
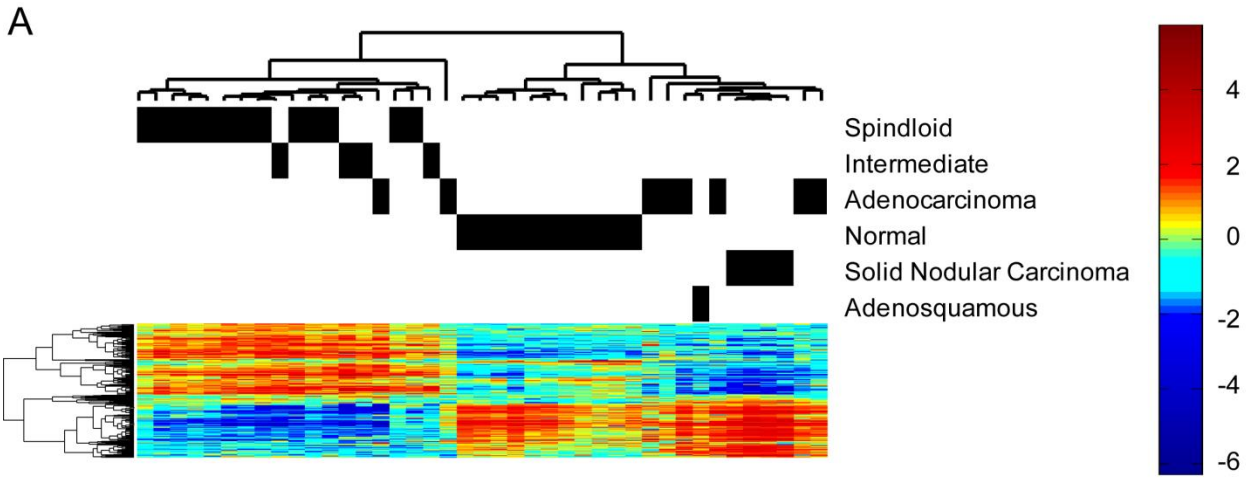
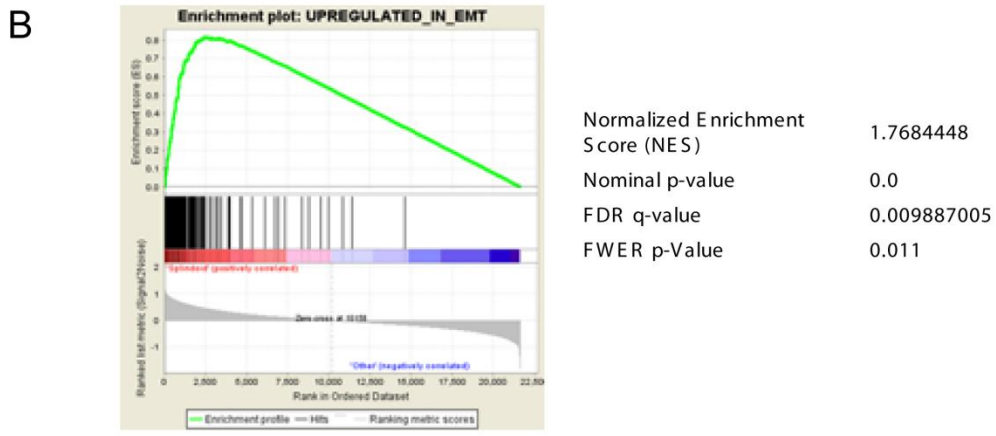


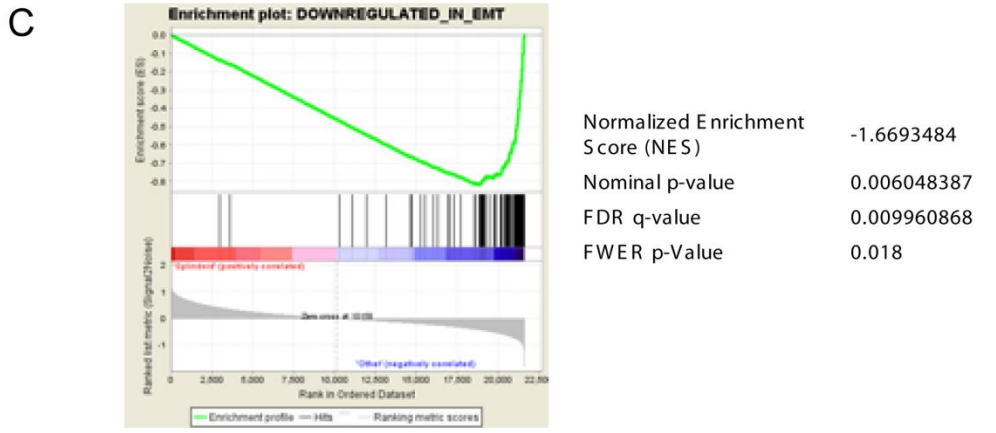
FIGURE 5.3: VALIDATION OF SQUAMOUS SIGNATURE GENES USING MMTV-MYC TUMORS

(A) Unsupervised hierarchical clustering using squamous signature genes accurately splits out squamous tumors from the other tumor histologies. (B) Gene set enrichment analysis (GSEA) comparing MMTV-Myc squamous tumors to all non-squamous tumors. The squamous signature genes derived from MMTV-PyMT tumors were significantly enriched in MMTV-Myc squamous tumors (NES 1.48, nominal p-value=0.0, FDR q-value =0.029, FWER p-value=0.016).





Normalized Enrichment Score (NES)	1.7684448
Nominal p-value	0.0
FDR q-value	0.009887005
FWER p-Value	0.011



Normalized Enrichment Score (NES)	-1.6693484
Nominal p-value	0.006048387
FDR q-value	0.009960868
FWER p-Value	0.018

FIGURE 5.4: VALIDATION OF EMT SIGNATURE GENES USING MMTV-MET TUMORS

FIGURE 5.4 (cont'd)

(A) Unsupervised hierarchical clustering using squamous signature genes accurately splits out the MMTV-Met Emt-like (splindoid) tumors from the other tumor histologies. (B) Testing for enrichment of the signature genes with GSEA illustrated a significant enrichment for upregulation of the upregulated EMT-like signature genes in Met-induced EMT-like tumors compared to non EMT-like tumors (FIGURE 5.4B, NES=1.76, nominal p-value=0.0, FDR q-value= 0.009, FWER p-value = 0.011). (B) GSEA found significant enrichment for downregulation of the downregulated EMT-like signature genes in Met-induced EMT-like tumors compared to non EMT-like tumors (FIGURE 5.4C, NES=-1.66, nominal p-value=0.006, FDR q-value= 0.009, FWER p-value = 0.018).

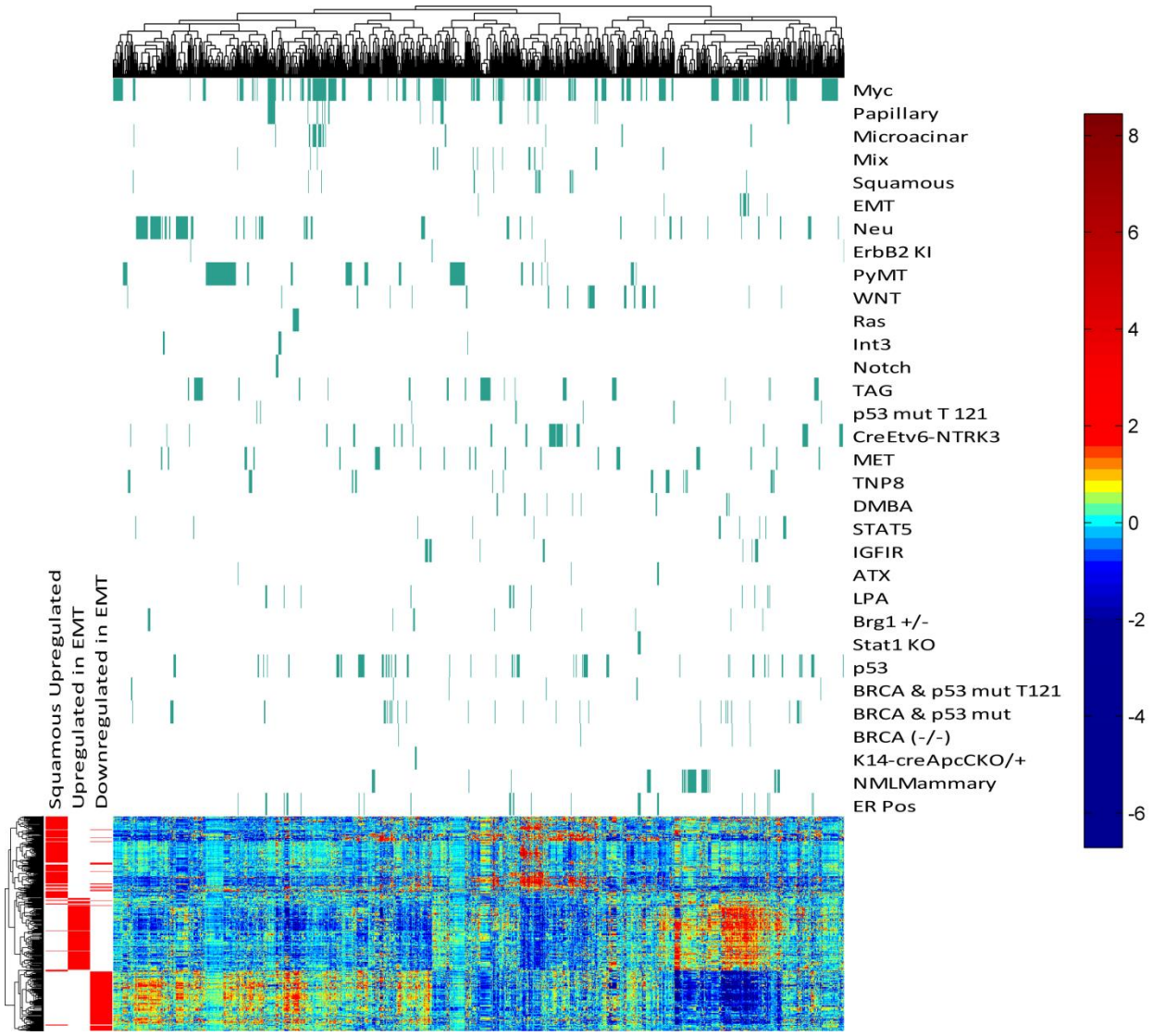


FIGURE 5.5 : UNSUPERVISED HIERARCHICAL CLUSTERING OF A MOUSE MAMMARY TUMOR MODEL GENE EXPRESSION DATABASE USING SQUAMOUS AND EMT-LIKE SIGNATURE GENES

Unsupervised hierarchical clustering of a mouse mammary tumor model gene expression database using squamous and Emt-like signature genes organizes tumors into clusters with Emt-like gene expression patterns, squamous gene expression patterns, clusters with tumors that have neither squamous or Emt-like gene expression patterns

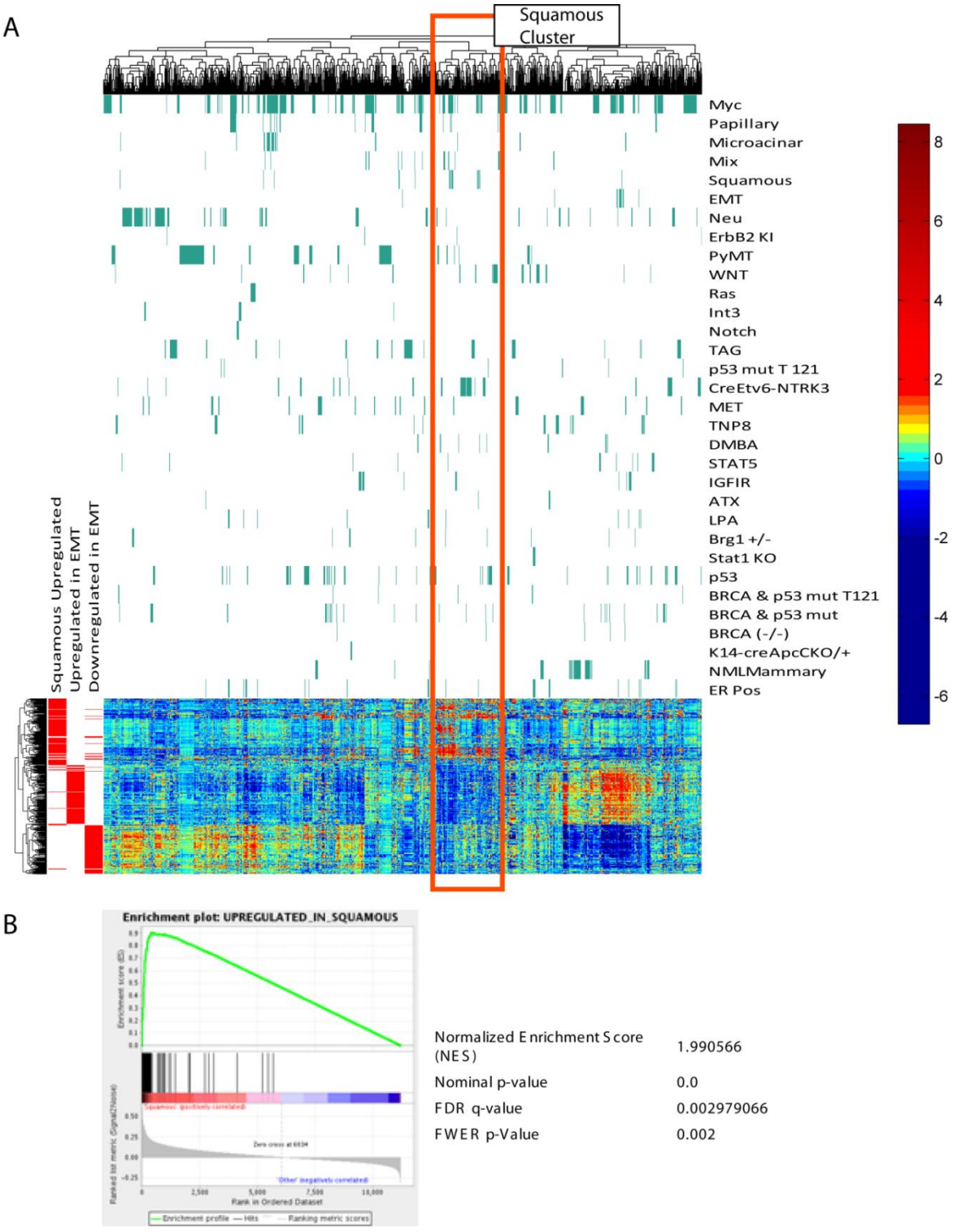


FIGURE 5.6: IDENTIFICATION OF SQUAMOUS MOUSE MAMMARY TUMORS

(A) Unsupervised hierarchical clustering of a mouse mammary tumor model gene expression database using squamous and Emt-like signature genes organizes tumors into clusters with Emt-

FIGURE 5.6 (cont'd)

like gene expression patterns, squamous gene expression patterns, clusters with tumors that have neither squamous or Emt-like gene expression patterns. The red box highlights the cluster of tumors with squamous gene expression features. (B) Gene set enrichment analysis shows tumors in the cluster highlighted in A, are significantly enriched for squamous gene expression features.

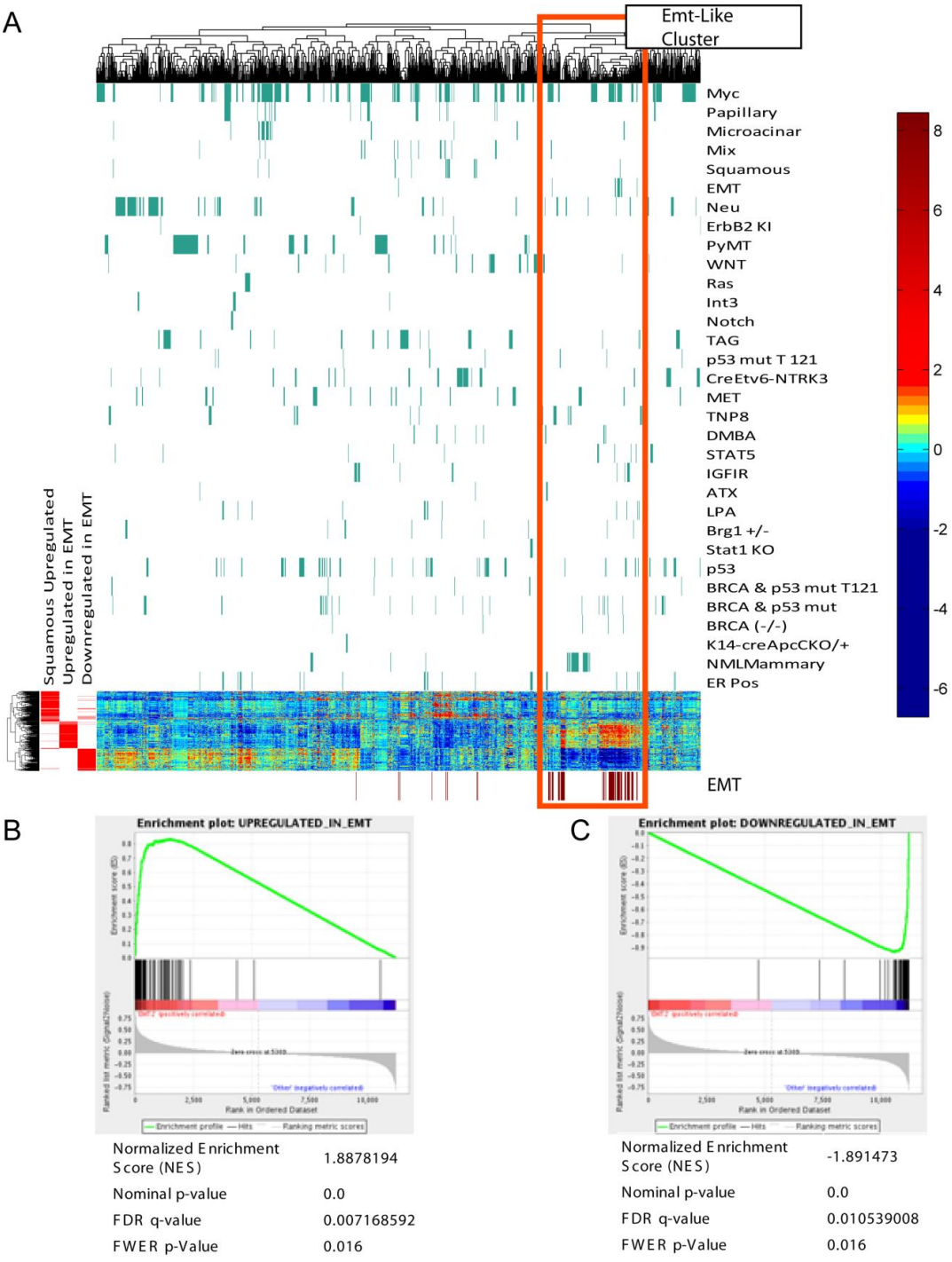


FIGURE 5.7: IDENTIFICATION OF EMT-LIKE MOUSE MAMMARY TUMORS

(A)Using unsupervised hierarchical clustering we observed expression patterns for the EMT-like signature genes. As marked by the red bar below the heatmap, the majority of the tumor samples

FIGURE 5.7 (cont'd)

with an annotation for an EMT-like histology had high expression of the genes upregulated in EMT-like tumors and low expression of genes downregulated in EMT-like tumors. As a result, we call the cluster highlighted with the red box the Emt-like cluster. (B) GSEA detected a significant enrichment for high expression of the genes upregulated in EMT-like tumors (FIGURE 5.7B, NES=1.88, nominal p-value = 0.0, FDR q-value=0.0071, FWER p-value = 0.016) and (C) low expression of the genes downregulated in EMT-like tumors (FIGURE 5.7C, NES=-1.89, nominal p-value = 0.0, FDR q-value=0.01, FWER p-value = 0.016).

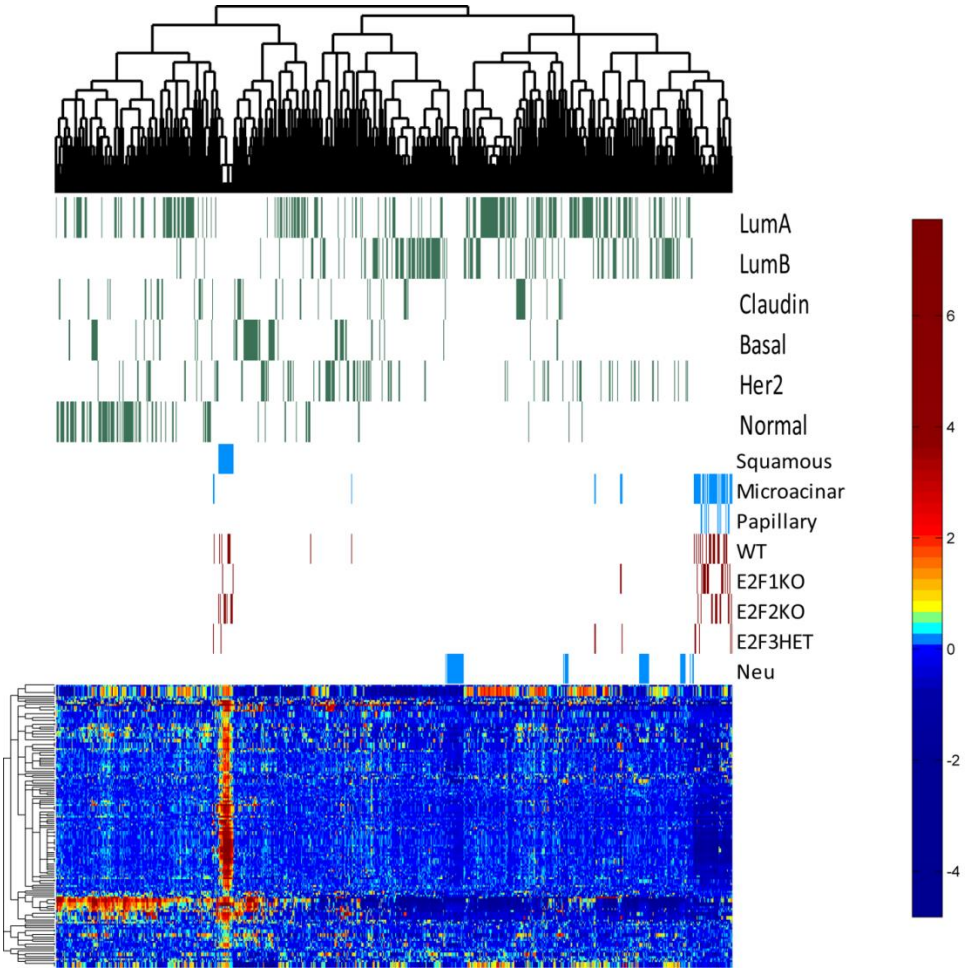


FIGURE 5.8: TESTING HUMAN BREAST CANCER EXPRESSION PROFILES OF SQUAMOUS SIGNATURE GENES

Unsupervised hierarchical clustering of human breast cancer samples and mouse mammary tumor samples on the basis of the squamous signature genes correctly identifies squamous tumors from the MMTV-PyMT mouse model and does not show high expression in human breast cancer.

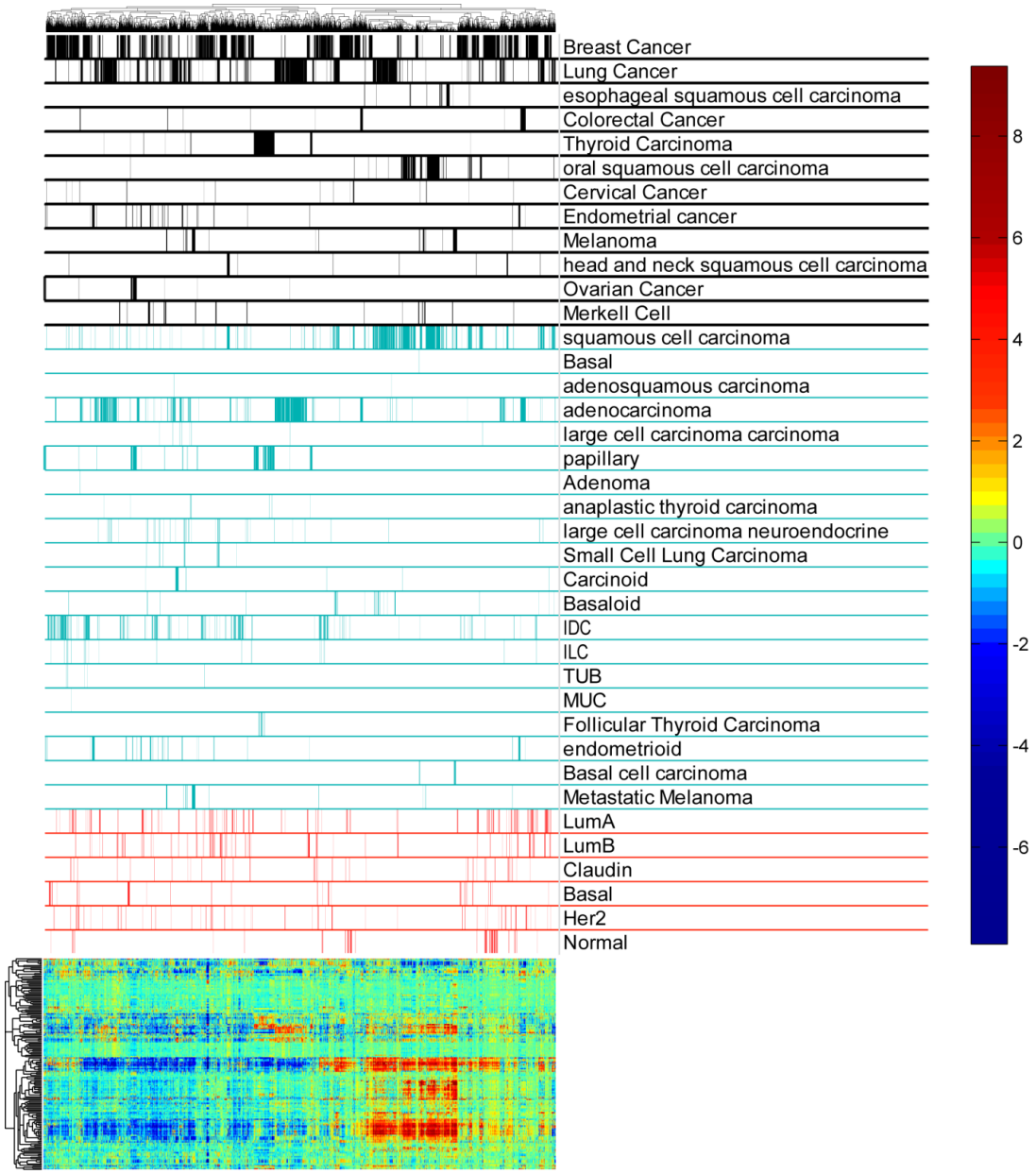


FIGURE 5.9: UNSUPERVISED HIERARCHICAL CLUSTERING OF HUMAN CANCER GENE EXPRESSION DATABASE USING SQUAMOUS SIGNATURE GENES

Unsupervised hierarchical clustering of human tumor sample (n=3186) gene expression data shows that squamous tumors from a variety of human cancers have high expression of squamous signature genes.

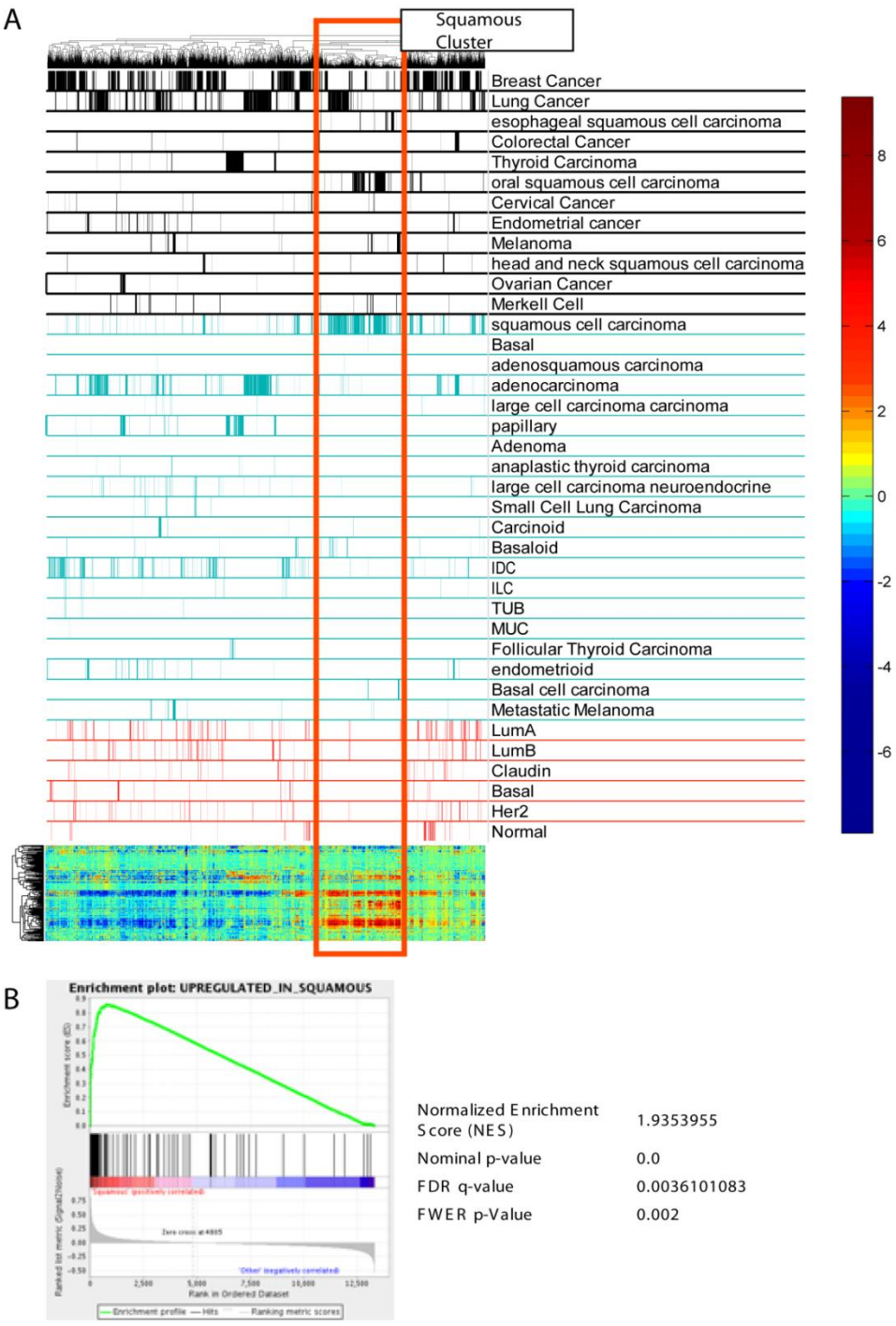


FIGURE 5.10: SQUAMOUS SIGNATURE GENES ARE HIGHLY EXPRESSED AND ARE ENRICHED IN A VARIETY OF HUMAN CANCERS OF SQUAMOUS HISTOLOGY

FIGURE 5.10 (cont'd)

(A) Using unsupervised hierarchical clustering we observed expression patterns for the squamous signature genes. As marked by the blue color bar above the heatmap, the majority of the tumor samples with an annotation for a squamous histology clustered together had high expression of the squamous signature genes. As a result, we call the cluster highlighted with the red box the squamous cluster. (B) Using GSEA to test for statistical enrichment of the squamous signatures in these samples indicted these samples demonstrated that this enrichment is significant (FIGURE 5.10B, NES= 1.93, nominal p-value 0.0, FDR q-value 0.003, FWER p-value =.002).

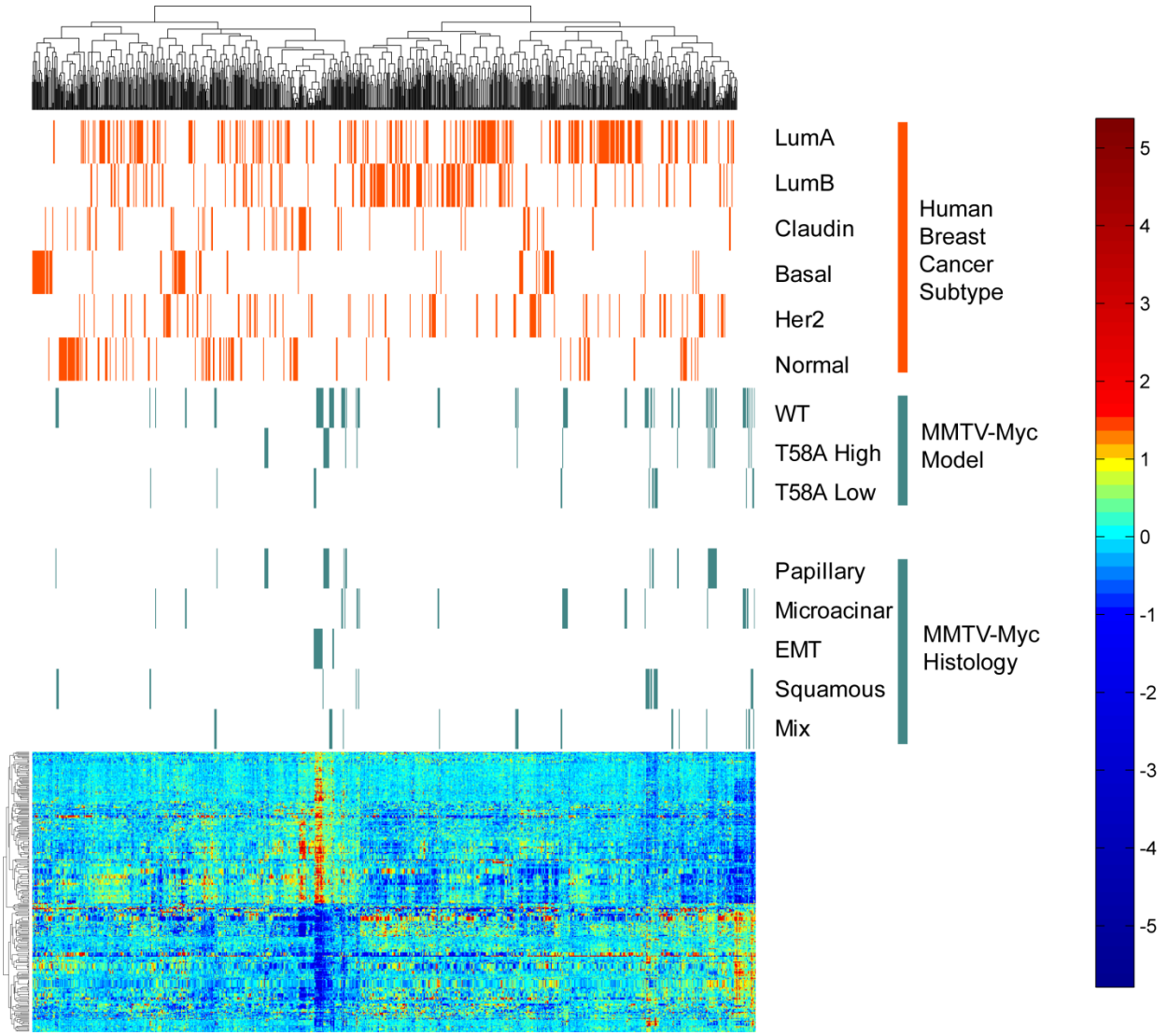
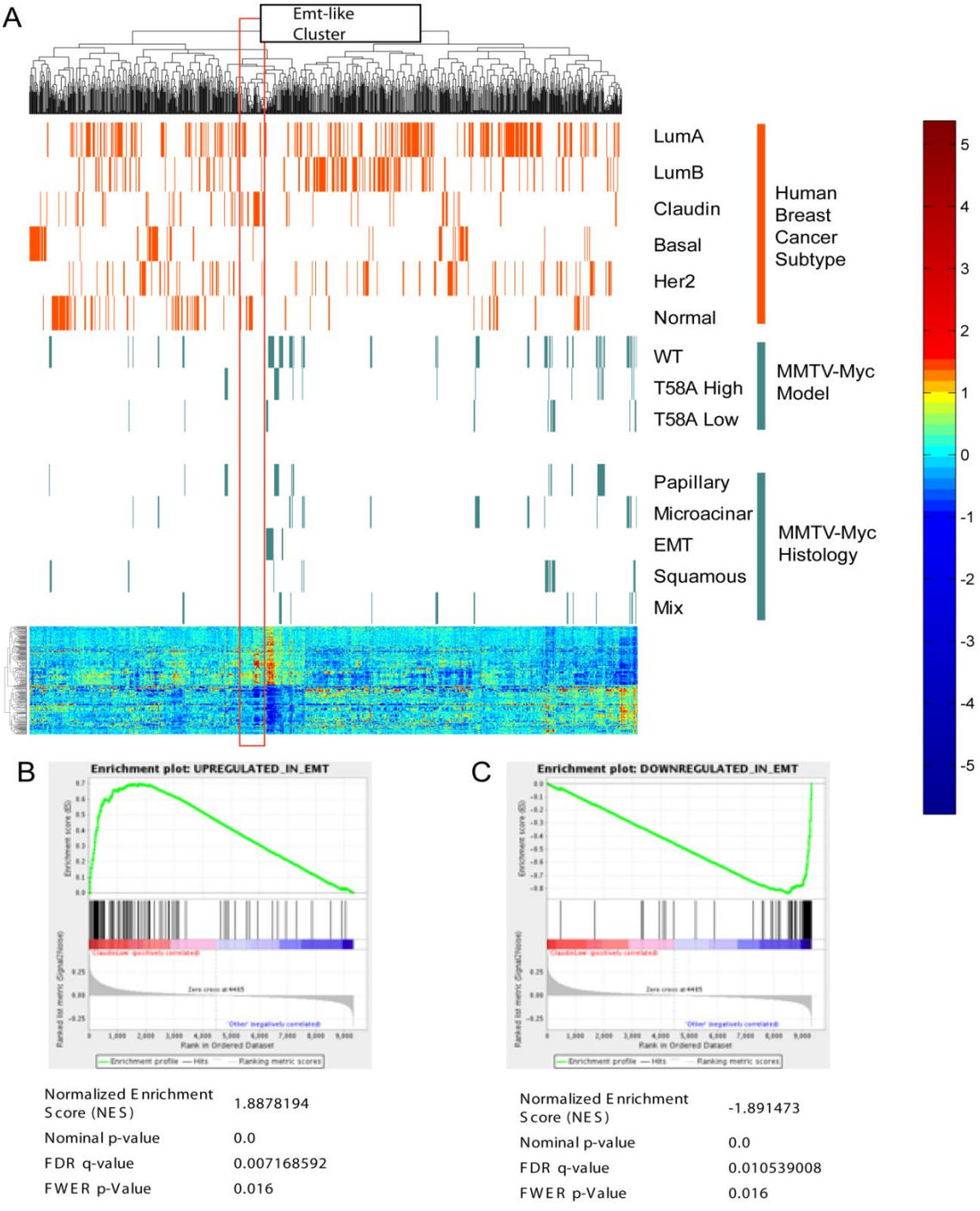


FIGURE 5.11: UNSUPERVISED HIERARCHICAL CLUSTERING OF HUMAN BREAST CANCER AND MMTV-MYC TUMORS USING THE EMT-LIKE GENE SIGNATURE

Unsupervised hierarchical clustering of human breast cancer and MMTV-Myc tumors using the Emt-like gene signature shows high expression of the Emt-like genes in a subset of claudin low human breast cancer and in the MMTV-Myc Emt-like tumors.



Normalized Enrichment Score (NES)	1.8878194
Nominal p-value	0.0
FDR q-value	0.007168592
FWER p-Value	0.016

Normalized Enrichment Score (NES)	-1.891473
Nominal p-value	0.0
FDR q-value	0.010539008
FWER p-Value	0.016

FIGURE 5.12: IDENTIFICATION OF EMT-LIKE SIGNATURE ENRICHMENT IN A SUBSET OF HUMAN CLAUDIN LOW BREAST CANCER

(A) Unsupervised hierarchical clustering revealed a subset of human claudin low breast cancer tumors that showed high expression of the EMT-like signature genes and clustered with EMT-

FIGURE 5.12 (cont'd)

like tumors from the MMTV-Myc mouse model. As result, we call the claudin low tumors in this cluster Emt-like. (B)Testing for significant enrichment for the Emt-like gene expression patterns with GSEA revealed that the claudin low tumors that clustered with mouse EMT-like tumors were significantly enriched for high expression of genes upregulated in EMT-like tumors (NES=1.87, nominal p-value =0.0m FDR q-value=.002, FWER p-value = 0.004) and (C)enriched for low expression of genes downregulated in EMT-like tumors (NES=-2.11, nominal p-value=0.0, FDR q-value =0.0, FWER p-value 0.0).

REFERENCES

REFERENCES

1. Hughes, T.R., et al., *Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.* Nature biotechnology, 2001. **19**(4): p. 342-347.

2. Miller, M.B. and Y.-W. Tang, *Basic concepts of microarrays and potential applications in clinical microbiology.* Clinical microbiology reviews, 2009. **22**(4): p. 611-633.

3. Agilent, *Two-Color Microarray-Based Gene Expression Analysis (Quick Amp Labeling) with Tecan HS Pro Hybridization*, 2008, Agilent.

4. Goryachev, A.B., P.F. Macgregor, and A.M. Edwards, *Unfolding of microarray data.* Journal of Computational Biology, 2001. **8**(4): p. 443-461.

5. Ideker, T., et al., *Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data.* Journal of Computational Biology, 2000. **7**(6): p. 805-817.

6. Kerr, M.K., M. Martin, and G.A. Churchill, *Analysis of variance for gene expression microarray data.* J Comput Biol, 2000. **7**(6): p. 819-37.

7. Wang, X., S. Ghosh, and S.W. Guo, *Quantitative quality control in microarray image processing and data acquisition.* Nucleic Acids Res, 2001. **29**(15): p. E75-5.

8. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic acids research, 2015: p. gkv007.

9. Schulze, A. and J. Downward, *Navigating gene expression using microarrays—a technology review.* Nature cell biology, 2001. **3**(8): p. E190-E195.

10. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic acids research, 2003. **31**(4): p. e15-e15.

11. Affymetrix, *Statistical Algorithms Reference Guide.* Affymetrix, 2007(Rev 2).

12. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-264.

13. Parrish, R.S. and H.J. Spencer III, *Effect of normalization on significance testing for oligonucleotide microarrays.* Journal of biopharmaceutical statistics, 2004. **14**(3): p. 575-589.

14. Holder, D., et al. *Statistical analysis of high density oligonucleotide arrays: a SAFER approach.* in *GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data*. 2001.

15. Affymetrix, *Affymetrix Expression Console™ Software.* 2011. **Rev 2**.

16. Ringnér, M., *What is principal component analysis?* Nature biotechnology, 2008. **26**(3): p. 303-304.

17. Jolliffe, I., *Principal component analysis*2002: Wiley Online Library.

18. Alter, O., P.O. Brown, and D. Botstein, *Singular value decomposition for genome-wide expression data processing and modeling.* Proceedings of the National Academy of Sciences, 2000. **97**(18): p. 10101-10106.

19. Nielsen, T.O., et al., *Molecular characterisation of soft tissue tumours: a gene expression study.* Lancet, 2002. **359**(9314): p. 1301-7.

20. Chung, C.H., et al., *Gene expression profiles identify epithelial-to-mesenchymal transition and activation of nuclear factor-κB signaling as characteristics of a high-risk head and neck squamous cell carcinoma.* Cancer research, 2006. **66**(16): p. 8210-8218.

21. Benito, M., et al., *Adjustment of systematic microarray data biases.* Bioinformatics, 2004. **20**(1): p. 105-14.

22. Furey, T.S., et al., *Support vector machine classification and validation of cancer tissue samples using microarray expression data.* Bioinformatics, 2000. **16**(10): p. 906-914.

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

24. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-27.

25. Carvalho, C.M., et al., *High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics.* J Am Stat Assoc, 2008. **103**(484): p. 1438-1456.

26. Hastie, T., et al., *The elements of statistical learning*. Vol. 2. 2009: Springer.

27. Greene, D., P. Cunningham, and R. Mayer, *Unsupervised learning and clustering*, in *Machine learning techniques for multimedia*2008, Springer. p. 51-90.

28. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.

29. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature genetics, 2000. **25**(1): p. 25-29.

30. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature protocols, 2008. **4**(1): p. 44-57.

31. Chang, J.T. and J.R. Nevins, *GATHER: a systems approach to interpreting genomic signatures.* Bioinformatics, 2006. **22**(23): p. 2926-33.

32. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.

33. Matys, V., et al., *TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes.* Nucleic acids research, 2006. **34**(suppl 1): p. D108-D110.

34. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

35. Huang, E., et al., *Gene expression phenotypic models that predict the activity of oncogenic pathways.* Nat Genet, 2003. **34**(2): p. 226-30.

36. West, M., et al., *Predicting the clinical status of human breast cancer by using gene expression profiles.* Proc Natl Acad Sci U S A, 2001. **98**(20): p. 11462-7.

37. Bild, A.H., et al., *Oncogenic pathway signatures in human cancers as a guide to targeted therapies.* Nature, 2006. **439**(7074): p. 353-7.

38. Peterka, V., *Bayesian system identification.* Automatica, 1981. **17**(1): p. 41-53.

39. Gilks, W.R., *Markov chain monte carlo*2005: Wiley Online Library.

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

41. Hawkins, D.M., *The problem of overfitting.* Journal of chemical information and computer sciences, 2004. **44**(1): p. 1-12.

42. Babyak, M.A., *What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models.* Psychosomatic medicine, 2004. **66**(3): p. 411-421.

43. Dudoit, S., J.P. Shaffer, and J.C. Boldrick, *Multiple hypothesis testing in microarray experiments.* Statistical Science, 2003: p. 71-103.

44. Talloen, W., et al., *I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data.* Bioinformatics, 2007. **23**(21): p. 2897-2902.

45. Bellman, *Adaptive Control Processes*1961, Princeton, NJ: Princeton University Press.

46. Varshavsky, R., et al., *Novel unsupervised feature filtering of biological data.* Bioinformatics, 2006. **22**(14): p. e507-e513.

47. Gatza, M.L., et al., *A pathway-based classification of human breast cancer.* Proc Natl Acad Sci U S A, 2010. **107**(15): p. 6994-9.

48. Simon, R., et al., *Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification.* Journal of the National Cancer Institute, 2003. **95**(1): p. 14-18.

49. Li, C.I., et al., *Trends in incidence rates of invasive lobular and ductal breast carcinoma*. Jama, 2003. **289**(11): p. 1421-1424.

50. DeSantis, C., et al., *Breast cancer statistics, 2013.* CA: a cancer journal for clinicians, 2014. **64**(1): p. 52-62.

51. Antoniou, A., et al., *Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies.* The American Journal of Human Genetics, 2003. **72**(5): p. 1117-1130.

52. Slamon, D.J., et al., *Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene.* Science, 1987. **235**(4785): p. 177-82.

53. Slamon, D.J., et al., *Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer.* Science, 1989. **244**(4905): p. 707-12.

54. Slamon, D.J., et al., *Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2.* N Engl J Med, 2001. **344**(11): p. 783-92.

55. Paik, S., et al., *Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor–positive breast cancer*. Journal of clinical oncology, 2006. **24**(23): p. 3726-3734.

56. Albain, K.S., et al., *Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial.* The lancet oncology, 2010. **11**(1): p. 55-65.

57. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.

58. Prat, A., et al., *Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer.* Breast Cancer Res, 2010. **12**(5): p. R68.

59. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.

60. Carey, L.A., et al., *Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study.* JAMA, 2006. **295**(21): p. 2492-502.

61. Network, T.C.G.A., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.

62. Gatza, M.L., et al., *An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer.* Nat Genet, 2014. **46**(10): p. 1051-9.

63. Nguyen, D.X. and J. Massagué, *Genetic determinants of cancer metastasis.* Nature Reviews Genetics, 2007. **8**(5): p. 341-352.

64. Nguyen, D.X., P.D. Bos, and J. Massague, *Metastasis: from dissemination to organ-specific colonization.* Nat Rev Cancer, 2009. **9**(4): p. 274-84.

65. Weigelt, B., J.L. Peterse, and L.J. van 't Veer, *Breast cancer metastasis: markers and models.* Nat Rev Cancer, 2005. **5**(8): p. 591-602.

66. Cancer, A.J.C.o., *Breast Cancer Staging.* American Joint Committee on Cancer, 2009. **7th Edition**.

67. Program, S.R., *SEER 18 2004-2010, All Races, Females by SEER Summary Stage 2000.* National Cancer Institute, 2011.

68. Bravo-Cordero, J.J., L. Hodgson, and J. Condeelis, *Directed cell invasion and migration during metastasis.* Current opinion in cell biology, 2012. **24**(2): p. 277-283.

69. Saharinen, P., et al., *VEGF and angiopoietin signaling in tumor angiogenesis and metastasis.* Trends in molecular medicine, 2011. **17**(7): p. 347-362.

70. Moss, L.A.S., S. Jensen-Taubman, and W.G. Stetler-Stevenson, *Matrix metalloproteinases: changing roles in tumor progression and metastasis.* The American journal of pathology, 2012. **181**(6): p. 1895-1899.

71. Chaffer, C.L., et al., *Poised chromatin at the ZEB1 promoter enables breast cancer cell plasticity and enhances tumorigenicity.* Cell, 2013. **154**(1): p. 61-74.

72. Marusyk, A., et al., *Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity.* Nature, 2014. **514**(7520): p. 54-58.

73. Bernards, R. and R.A. Weinberg, *Metastasis genes: a progression puzzle.* Nature, 2002. **418**(6900): p. 823-823.

74. Gilkes, D.M., G.L. Semenza, and D. Wirtz, *Hypoxia and the extracellular matrix: drivers of tumour metastasis.* Nature Reviews Cancer, 2014. **14**(6): p. 430-439.

75. Qian, B.-Z. and J.W. Pollard, *Macrophage diversity enhances tumor progression and metastasis.* Cell, 2010. **141**(1): p. 39-51.

76. DeNardo, D.G., P. Andreu, and L.M. Coussens, *Interactions between lymphocytes and myeloid cells regulate pro-versus anti-tumor immunity.* Cancer and Metastasis Reviews, 2010. **29**(2): p. 309-316.

77. Kennecke, H., et al., *Metastatic behavior of breast cancer subtypes.* Journal of clinical oncology, 2010. **28**(20): p. 3271-3277.

78. Fong, M.Y., et al., *Breast-cancer-secreted miR-122 reprograms glucose metabolism in premetastatic niche to promote metastasis.* Nature cell biology, 2015. **17**(2): p. 183-194.

79. Olkhanud, P.B., et al., *Breast cancer lung metastasis requires expression of chemokine receptor CCR4 and regulatory T cells.* Cancer research, 2009. **69**(14): p. 5996-6004.

80. Gay, L.J. and B. Felding-Habermann, *Contribution of platelets to tumour metastasis.* Nature Reviews Cancer, 2011. **11**(2): p. 123-134.

81. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-6.

82. Aigner, K., et al., *The transcription factor ZEB1 (δEF1) promotes tumour cell dedifferentiation by repressing master regulators of epithelial polarity.* Oncogene, 2007. **26**(49): p. 6979-6988.

83. Wang, Y., et al., *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* Lancet, 2005. **365**(9460): p. 671-9.

84. Kang, Y., et al., *A multigenic program mediating breast cancer metastasis to bone.* Cancer Cell, 2003. **3**(6): p. 537-49.

85. Minn, A.J., et al., *Genes that mediate breast cancer metastasis to lung.* Nature, 2005. **436**(7050): p. 518-24.

86. Gyorffy, B., et al., *Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer.* PLoS One, 2013. **8**(12): p. e82241.

87. Neve, R.M., et al., *A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes.* Cancer Cell, 2006. **10**(6): p. 515-27.

88. Taxman, D.J., et al., *Short hairpin RNA (shRNA): design, delivery, and assessment of gene knockdown*, in *RNA Therapeutics*2010, Springer. p. 139-156.

89. Zhang, F., Y. Wen, and X. Guo, *CRISPR/Cas9 for genome editing: progress, implications and challenges.* Human molecular genetics, 2014: p. ddu125.

90. Pagès, F., et al., *Effector memory T cells, early metastasis, and survival in colorectal cancer.* New England journal of medicine, 2005. **353**(25): p. 2654-2666.

91. Callahan, R. and G.H. Smith, *MMTV-induced mammary tumorigenesis: gene discovery, progression to malignancy and cellular pathways.* Oncogene, 2000. **19**(8): p. 992-1001.

92. Nusse, R. and H.E. Varmus, *Many tumors induced by the mouse mammary tumor virus contain a provirus integrated in the same region of the host genome.* Cell, 1982. **31**(1): p. 99-109.

93. Taneja, P., et al., *MMTV mouse models and the diagnostic values of MMTV-like sequences in human breast cancer.* Expert Rev Mol Diagn, 2009. **9**(5): p. 423-40.

94. Xu, X., et al., *Conditional mutation of Brca1 in mammary epithelial cells results in blunted ductal morphogenesis and tumour formation.* Nat Genet, 1999. **22**(1): p. 37-43.

95. Jonkers, J., et al., *Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer.* Nat Genet, 2001. **29**(4): p. 418-25.

96. Nass, S.J. and R.B. Dickson, *Detection of cyclin messenger RNAs by nonradioactive ribonuclease protection assay: a comparison of four detection methods.* Biotechniques, 1995. **19**(5): p. 772-6, 778.

97. Donehower, L.A., et al., *Deficiency of p53 accelerates mammary tumorigenesis in Wnt-1 transgenic mice and promotes chromosomal instability.* Genes Dev, 1995. **9**(7): p. 882-95.

98. Bearss, D.J., et al., *Differential effects of p21(WAF1/CIP1) deficiency on MMTV-ras and MMTV-myc mammary tumor properties.* Cancer Res, 2002. **62**(7): p. 2077-84.

99. Adams, J.R., et al., *Cooperation between Pik3ca and p53 mutations in mouse mammary tumor formation.* Cancer Res, 2011. **71**(7): p. 2706-17.

100. Boxer, R.B., et al., *Lack of sustained regression of c-MYC-induced mammary adenocarcinomas following brief or prolonged MYC inactivation.* Cancer Cell, 2004. **6**(6): p. 577-86.

101. D'Cruz, C.M., et al., *c-MYC induces mammary tumorigenesis by means of a preferred pathway involving spontaneous Kras2 mutations.* Nat Med, 2001. **7**(2): p. 235-9.

102. Leung, J.Y., et al., *Heterogeneity in MYC-induced mammary tumors contributes to escape from oncogene dependence.* Oncogene, 2012. **31**(20): p. 2545-54.

103. Podsypanina, K., et al., *Oncogene cooperation in tumor maintenance and tumor recurrence in mouse mammary tumors induced by Myc and mutant Kras.* Proc Natl Acad Sci U S A, 2008. **105**(13): p. 5242-7.

104. Andrechek, E.R. and J.R. Nevins, *Mouse models of cancers: opportunities to address heterogeneity of human cancer and evaluate therapeutic strategies.* J Mol Med, 2010. **DOI: 10.1007/s00109-010-0644-z**.

105. Herschkowitz, J.I., et al., *Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors.* Genome Biol, 2007. **8**(5): p. R76.

106. Andrechek, E.R., et al., *Genetic heterogeneity of Myc-induced mammary tumors reflecting diverse phenotypes including metastatic potential.* Proc Natl Acad Sci U S A, 2009. **106**(38): p. 16387-92.

107. Ponzo, M.G., et al., *Met induces mammary tumors with diverse histologies and is associated with poor outcome and human basal breast cancer.* Proc Natl Acad Sci U S A, 2009. **106**(31): p. 12903-8.

108. Fujiwara, K., et al., *Prediction and Genetic Demonstration of a Role for Activator E2Fs in Myc-Induced Tumors.* Cancer Res, 2011. **71**(5): p. 1924-32.

109. Van Den Heuvel, S. and N.J. Dyson, *Conserved functions of the pRB and E2F families.* Nature reviews Molecular cell biology, 2008. **9**(9): p. 713-724.

110. Guy, C.T., R.D. Cardiff, and W.J. Muller, *Induction of mammary tumors by expression of polyomavirus middle T oncogene: a transgenic mouse model for metastatic disease.* Mol Cell Biol, 1992. **12**(3): p. 954-61.

111. Landskroner-Eiger, S., et al., *Proangiogenic contribution of adiponectin toward mammary tumor growth in vivo.* Clin Cancer Res, 2009. **15**(10): p. 3265-76.

112. Duggan, C., et al., *Associations of insulin resistance and adiponectin with mortality in women with breast cancer.* J Clin Oncol, 2010. **29**(1): p. 32-9.

113. Wertheim, G.B., et al., *The Snf1-related kinase, Hunk, is essential for mammary tumor metastasis.* Proc Natl Acad Sci U S A, 2009. **106**(37): p. 15855-60.

114. Quintela-Fandino, M., et al., *HUNK suppresses metastasis of basal type breast cancers by disrupting the interaction between PP2A and cofilin-1.* Proc Natl Acad Sci U S A, 2010. **107**(6): p. 2622-7.

115. Schade, B., et al., *PTEN deficiency in a luminal ErbB-2 mouse model results in dramatic acceleration of mammary tumorigenesis and metastasis.* J Biol Chem, 2009. **284**(28): p. 19018-26.

116. Sorlie, T., et al., *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8418-23.

117. Sotiriou, C., et al., *Breast cancer classification and prognosis based on gene expression profiles from a population-based study.* Proc Natl Acad Sci U S A, 2003. **100**(18): p. 10393-8.

118. Deming, S.L., et al., *C-myc amplification in breast cancer: a meta-analysis of its occurrence and prognostic relevance.* Br J Cancer, 2000. **83**(12): p. 1688-95.

119. Blancato, J., et al., *Correlation of amplification and overexpression of the c-myc oncogene in high-grade breast cancer: FISH, in situ hybridisation and immunohistochemical analyses.* Br J Cancer, 2004. **90**(8): p. 1612-9.

120. Nesbit, C.E., J.M. Tersak, and E.V. Prochownik, *MYC oncogenes and human neoplastic disease.* Oncogene, 1999. **18**(19): p. 3004-16.

121. Chrzan, P., et al., *Amplification of c-myc gene and overexpression of c-Myc protein in breast cancer and adjacent non-neoplastic tissue.* Clin Biochem, 2001. **34**(7): p. 557-62.

122. Sears, R., et al., *Ras enhances Myc protein stability.* Mol Cell, 1999. **3**(2): p. 169-79.

123. Sears, R., et al., *Multiple Ras-dependent phosphorylation pathways regulate Myc protein stability.* Genes Dev, 2000. **14**(19): p. 2501-14.

124. Amati, B., et al., *Transcriptional activation by the human c-Myc oncoprotein in yeast requires interaction with Max.* Nature, 1992. **359**(6394): p. 423-6.

125. Kim, J., et al., *A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs.* Cell, 2010. **143**(2): p. 313-24.

126. Stewart, T.A., P.K. Pattengale, and P. Leder, *Spontaneous mammary adenocarcinomas in transgenic mice that carry and express MTV/myc fusion genes.* Cell, 1984. **38**(3): p. 627-37.

127. Sinn, E., et al., *Coexpression of MMTV/v-Ha-ras and MMTV/c-myc genes in transgenic mice: synergistic action of oncogenes in vivo.* Cell, 1987. **49**(4): p. 465-75.

128. Wang, X., et al., *Phosphorylation regulates c-Myc's oncogenic activity in the mammary gland.* Cancer Res, 2011. **71**(3): p. 925-36.

129. Herschkowitz, J.I., et al., *Comparative oncogenomics identifies breast tumors enriched in functional tumor-initiating cells.* Proc Natl Acad Sci U S A, 2011.

130. Lim, E., et al., *Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways.* Breast Cancer Res, 2010. **12**(2): p. R21.

131. Smith, R., G. Peters, and C. Dickson, *Genomic organization of the mouse cyclin D1 gene (Cyl-1).* Genomics, 1995. **25**(1): p. 85-92.

132. McCormack, S.J., et al., *Myc/p53 interactions in transgenic mouse mammary development, tumorigenesis and chromosomal instability.* Oncogene, 1998. **16**(21): p. 2755-66.

133. Andrechek, E.R., et al., *Patterns of cell signaling pathway activation that characterize mammary development.* Development, 2008. **135**(14): p. 2403-13.

134. Muller, W.J., et al., *Single-step induction of mammary adenocarcinoma in transgenic mice bearing the activated c-neu oncogene.* Cell, 1988. **54**(1): p. 105-15.

135. Andrechek, E.R., et al., *Amplification of the neu/erbB-2 oncogene in a mouse model of mammary tumorigenesis.* Proc Natl Acad Sci U S A, 2000. **97**(7): p. 3444-9.

136. Fluck, M.M. and B.S. Schaffhausen, *Lessons in signaling and tumorigenesis from polyomavirus middle T antigen.* Microbiol Mol Biol Rev, 2009. **73**(3): p. 542-63, Table of Contents.

137. Hunter, K.W., et al., *Predisposition to efficient mammary tumor metastatic progression is linked to the breast cancer metastasis suppressor gene Brms1.* Cancer Res, 2001. **61**(24): p. 8866-72.

138. Andrechek, E.R., et al., *Gene expression profiling of neu-induced mammary tumors from transgenic mice reveals genetic and morphological similarities to ErbB2-expressing human breast cancers.* Cancer Res, 2003. **63**(16): p. 4920-6.

139. Rosner, A., et al., *Pathway pathology: histological differences between ErbB/Ras and Wnt pathway transgenic mammary tumors.* Am J Pathol, 2002. **161**(3): p. 1087-97.

140. Knight, J.F., et al., *Met synergizes with p53 loss to induce mammary tumors that possess features of claudin-low breast cancer.* Proc Natl Acad Sci U S A, 2013. **110**(14): p. E1301-10.

141. Hollern, D.P., I. Yuwanita, and E.R. Andrechek, *A mouse model with T58A mutations in Myc reduces the dependence on KRas mutations and has similarities to claudin-low human breast cancer.* Oncogene, 2012.

142. Schade, B., et al., *beta-Catenin Signaling Is a Critical Event in ErbB2-Mediated Mammary Tumor Progression.* Cancer Res, 2013. **73**(14): p. 4474-4487.

143. Dourdin, N., et al., *Phosphatase and tensin homologue deleted on chromosome 10 deficiency accelerates tumor induction in a mouse model of ErbB-2 mammary tumorigenesis.* Cancer Res, 2008. **68**(7): p. 2122-31.

144. Khalil, S., et al., *Activation status of Wnt/ss-catenin signaling in normal and neoplastic breast tissues: relationship to HER2/neu expression in human and mouse.* PLoS One, 2012. **7**(3): p. e33421.

145. Wang, J., et al., *Notch1 is involved in migration and invasion of human breast cancer cells.* Oncol Rep, 2011. **26**(5): p. 1295-303.

146. Muraoka, R.S., et al., *Blockade of TGF-beta inhibits mammary tumor cell viability, migration, and metastases.* J Clin Invest, 2002. **109**(12): p. 1551-9.

147. Kumar, R., D. Medina, and S. Sukumar, *Activation of H-ras oncogenes in preneoplastic mouse mammary tissues.* Oncogene, 1990. **5**(8): p. 1271-7.

148. Kim, S., A. Roopra, and C.M. Alexander, *A phenotypic mouse model of basaloid breast tumors.* PLoS One, 2012. **7**(2): p. e30979.

149. Schade, B., et al., *beta-catenin signaling is a critical event in ErbB2-mediated mammary tumor progression.* Cancer Res, 2013.

150. Guy, C.T., et al., *Activation of the c-Src tyrosine kinase is required for the induction of mammary tumors in transgenic mice.* Genes Dev, 1994. **8**(1): p. 23-32.

151. Visvader, J.E., *Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis.* Genes Dev, 2009. **23**(22): p. 2563-77.

152. Bos, P.D., et al., *Genes that mediate breast cancer metastasis to the brain.* Nature, 2009. **459**(7249): p. 1005-9.

153. Lin, E.Y., et al., *Colony-stimulating factor 1 promotes progression of mammary tumors to malignancy.* J Exp Med, 2001. **193**(6): p. 727-40.

154. Forrester, E., et al., *Effect of conditional knockout of the type II TGF-beta receptor gene in mammary epithelia on mammary gland development and polyomavirus middle T antigen induced tumor formation and metastasis.* Cancer Res, 2005. **65**(6): p. 2296-302.

155. Dillon, R.L., et al., *Akt1 and akt2 play distinct roles in the initiation and metastatic phases of mammary tumor progression.* Cancer Res, 2009. **69**(12): p. 5057-64.

156. Nevins, J.R., *The Rb/E2F pathway and cancer.* Hum Mol Genet, 2001. **10**(7): p. 699-703.

157. Trimarchi, J.M. and J.A. Lees, *Sibling rivalry in the E2F family.* Nat Rev Mol Cell Biol, 2002. **3**(1): p. 11-20.

158. Attwooll, C., E. Lazzerini Denchi, and K. Helin, *The E2F family: specific functions and overlapping interests.* EMBO J, 2004. **23**(24): p. 4709-16.

159. Hallstrom, T.C., S. Mori, and J.R. Nevins, *An E2F1-dependent gene expression program that determines the balance between proliferation and cell death.* Cancer Cell, 2008. **13**(1): p. 11-22.

160. Chen, H.Z., S.Y. Tsai, and G. Leone, *Emerging roles of E2Fs in cancer: an exit from cell cycle control.* Nat Rev Cancer, 2009. **9**(11): p. 785-97.

161. Alla, V., et al., *E2F1 in melanoma progression and metastasis.* J Natl Cancer Inst, 2010. **102**(2): p. 127-33.

162. Ebihara, Y., et al., *Over-expression of E2F-1 in esophageal squamous cell carcinoma correlates with tumor progression.* Dis Esophagus, 2004. **17**(2): p. 150-4.

163. Foster, C.S., et al., *Transcription factor E2F3 overexpressed in prostate cancer independently predicts clinical outcome.* Oncogene, 2004. **23**(35): p. 5871-9.

164. Franci, C., et al., *Biomarkers of residual disease, disseminated tumor cells, and metastases in the MMTV-PyMT breast cancer model.* PLoS One, 2013. **8**(3): p. e58183.

165. Leone, G., et al., *Myc requires distinct E2F activities to induce S phase and apoptosis.* Mol Cell, 2001. **8**(1): p. 105-13.

166. Kong, L.J., et al., *Compensation and specificity of function within the E2F family.* Oncogene, 2007. **26**(3): p. 321-7.

167. Danielian, P.S., et al., *E2f3a and E2f3b make overlapping but different contributions to total E2f3 activity.* Oncogene, 2008. **27**(51): p. 6561-70.

168. Humbert, P.O., et al., *E2f3 is critical for normal cellular proliferation.* Genes Dev, 2000. **14**(6): p. 690-703.

169. Zheng, N., et al., *Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP.* Genes Dev, 1999. **13**(6): p. 666-74.

170. Freedman, J.A., et al., *A combinatorial mechanism for determining the specificity of E2F activation and repression.* Oncogene, 2009. **28**(32): p. 2873-81.

171. Merdzhanova, G., et al., *The transcription factor E2F1 and the SR protein SC35 control the ratio of pro-angiogenic versus antiangiogenic isoforms of vascular endothelial growth factor-A to inhibit neovascularization in vivo.* Oncogene, 2010. **29**(39): p. 5392-403.

172. He, T., et al., *MicroRNA-542-3p inhibits tumour angiogenesis by targeting angiopoietin-2.* J Pathol, 2014. **232**(5): p. 499-508.

173. Zhang, Z.L., et al., *Suppression of angiogenesis and tumor growth and using an anti-angiopoietin-2 single-chain antibody.* Exp Ther Med, 2014. **7**(3): p. 543-552.

174. Imanishi, Y., et al., *Angiopoietin-2 stimulates breast cancer metastasis through the alpha(5)beta(1) integrin-mediated pathway.* Cancer Res, 2007. **67**(9): p. 4254-63.

175. Lin, J., et al., *A novel anti-Cyr61 antibody inhibits breast cancer growth and metastasis in vivo.* Cancer Immunol Immunother, 2011. **61**(5): p. 677-87.

176. Harris, L.G., et al., *Increased vascularity and spontaneous metastasis of breast cancer by hedgehog signaling mediated upregulation of cyr61.* Oncogene, 2012. **31**(28): p. 3370-80.

177. Fidler, I.J., *The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited.* Nat Rev Cancer, 2003. **3**(6): p. 453-8.

178. Ricciardelli, C., et al., *The ADAMTS1 protease gene is required for mammary tumor growth and metastasis.* Am J Pathol, 2011. **179**(6): p. 3075-85.

179. Shintani, Y., et al., *Collagen I promotes metastasis in pancreatic cancer by activating c-Jun NH(2)-terminal kinase 1 and up-regulating N-cadherin expression.* Cancer Res, 2006. **66**(24): p. 11745-53.

180. Gilkes, D.M., et al., *Collagen prolyl hydroxylases are essential for breast cancer metastasis.* Cancer Res, 2013. **73**(11): p. 3285-96.

181. Gilkes, D.M., et al., *Procollagen lysyl hydroxylase 2 is essential for hypoxia-induced breast cancer metastasis.* Mol Cancer Res, 2013. **11**(5): p. 456-66.

182. Kretzschmar, M., et al., *The TGF-beta family mediator Smad1 is phosphorylated directly and activated functionally by the BMP receptor kinase.* Genes Dev, 1997. **11**(8): p. 984-95.

183. Park, E.S., D.C. Woods, and J.L. Tilly, *Bone morphogenetic protein 4 promotes mammalian oogonial stem cell differentiation via Smad1/5/8 signaling.* Fertil Steril, 2013. **100**(5): p. 1468-75.

184. Lai, D. and X. Yang, *BMP4 is a novel transcriptional target and mediator of mammary cell migration downstream of the Hippo pathway component TAZ.* Cell Signal, 2013. **25**(8): p. 1720-8.

185. Guo, D., J. Huang, and J. Gong, *Bone morphogenetic protein 4 (BMP4) is required for migration and invasion of breast cancer.* Mol Cell Biochem, 2011. **363**(1-2): p. 179-90.

186. Garcia-Montero, A.C., et al., *Transforming growth factor beta-1 enhances Smad transcriptional activity through activation of p8 gene expression.* Biochem J, 2001. **357**(Pt 1): p. 249-53.

187. Ree, A.H., et al., *Expression of a novel factor, com1, in early tumor progression of breast cancer.* Clin Cancer Res, 2000. **6**(5): p. 1778-83.

188. Newlaczyl, A.U. and L.G. Yu, *Galectin-3--a jack-of-all-trades in cancer.* Cancer Lett, 2011. **313**(2): p. 123-8.

189. Warfield, P.R., et al., *Adhesion of human breast carcinoma to extracellular matrix proteins is modulated by galectin-3.* Invasion Metastasis, 1997. **17**(2): p. 101-12.

190. Ochieng, J., V. Furtak, and P. Lukyanov, *Extracellular functions of galectin-3.* Glycoconj J, 2004. **19**(7-9): p. 527-35.

191. Shekhar, M.P., et al., *Alterations in galectin-3 expression and distribution correlate with breast cancer progression: functional analysis of galectin-3 in breast epithelial-endothelial interactions.* Am J Pathol, 2004. **165**(6): p. 1931-41.

192. Zou, J., et al., *Peptides specific to the galectin-3 carbohydrate recognition domain inhibit metastasis-associated cancer cell adhesion.* Carcinogenesis, 2005. **26**(2): p. 309-18.

193. Newton-Northup, J.R., et al., *Inhibition of metastatic tumor formation in vivo by a bacteriophage display-derived galectin-3 targeting peptide.* Clin Exp Metastasis, 2012. **30**(2): p. 119-32.

194. Kim, J., et al., *An extended transcriptional network for pluripotency of embryonic stem cells.* Cell, 2008. **132**(6): p. 1049-61.

195. Asp, P., et al., *E2f3b plays an essential role in myogenic differentiation through isoform-specific gene regulation.* Genes Dev, 2009. **23**(1): p. 37-53.

196. Ernst, J., et al., *Integrating multiple evidence sources to predict transcription factor binding in the human genome.* Genome Res, 2010. **20**(4): p. 526-36.

197. Xu, X., et al., *A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members.* Genome Res, 2007. **17**(11): p. 1550-61.

198. Gyorffy, B., et al., *An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients.* Breast Cancer Res Treat, 2010. **123**(3): p. 725-31.

199. Field, S.J., et al., *E2F-1 functions in mice to promote apoptosis and suppress proliferation.* Cell, 1996. **85**(4): p. 549-61.

200. Murga, M., et al., *Mutation of E2F2 in mice causes enhanced T lymphocyte proliferation, leading to the development of autoimmunity.* Immunity, 2001. **15**(6): p. 959-70.

201. Wang, S., et al., *Disruption of the SRC-1 gene in mice suppresses breast cancer metastasis without affecting primary tumor formation.* Proc Natl Acad Sci U S A, 2009. **106**(1): p. 151-6.

202. Borowsky, A.D., et al., *Syngeneic mouse mammary carcinoma cell lines: two closely related cell lines with divergent metastatic behavior.* Clin Exp Metastasis, 2005. **22**(1): p. 47-59.

203. Selvaraj, N., et al., *Prostate cancer ETS rearrangements switch a cell migration gene expression program from RAS/ERK to PI3K/AKT regulation.* Mol Cancer, 2014. **13**: p. 61.

204. Hollern, D.P., et al., *The E2F transcription factors regulate tumor development and metastasis in a mouse model of metastatic breast cancer.* Mol Cell Biol, 2014.

205. Bieda, M., et al., *Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome.* Genome research, 2006. **16**(5): p. 595-605.

206. Yang, J.-H., et al., *ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data.* Nucleic acids research, 2013. **41**(D1): p. D177-D187.

207. Riese, D.J., et al., *The epidermal growth factor receptor couples transforming growth factor-α, heparin-binding epidermal growth factor-like factor, and amphiregulin to Neu, ErbB-3, and ErbB-4.* Journal of Biological chemistry, 1996. **271**(33): p. 20047-20052.

208. Franceschini, A., et al., *STRING v9. 1: protein-protein interaction networks, with increased coverage and integration.* Nucleic acids research, 2013. **41**(D1): p. D808-D815.

209. Schoeffner, D.J., et al., *VEGF contributes to mammary tumor growth in transgenic mice through paracrine and autocrine mechanisms.* Laboratory investigation, 2005. **85**(5): p. 608-623.

210. Zhou, Z., et al., *Autocrine HBEGF expression promotes breast cancer intravasation, metastasis and macrophage-independent invasion in vivo.* Oncogene, 2014. **33**(29): p. 3784-3793.

211. Gibert, B., et al., *Targeting heat shock protein 27 (HspB1) interferes with bone metastasis and tumour formation in vivo.* British journal of cancer, 2012. **107**(1): p. 63-70.

212. Taylor, A.P. and D.M. Goldenberg, *Role of placenta growth factor in malignancy and evidence that an antagonistic PlGF/Flt-1 peptide inhibits the growth and metastasis of human breast cancer xenografts.* Molecular cancer therapeutics, 2007. **6**(2): p. 524-531.

213. Zhang, H., et al., *HIF-1-dependent expression of angiopoietin-like 4 and L1CAM mediates vascular metastasis of hypoxic breast cancer cells to the lungs.* Oncogene, 2012. **31**(14): p. 1757-1770.

214. Xing, R.H. and S.A. Rabbani, *Overexpression of urokinase receptor in breast cancer cells results in increased tumor invasion, growth and metastasis.* International journal of cancer, 1996. **67**(3): p. 423-429.

215. Muraoka-Cook, R.S., et al., *Conditional overexpression of active transforming growth factor β1 in vivo accelerates metastases of transgenic mammary tumors.* Cancer research, 2004. **64**(24): p. 9002-9011.

216. Myoui, A., et al., *C-SRC tyrosine kinase activity is associated with tumor colonization in bone and lung in an animal model of human breast cancer metastasis.* Cancer Research, 2003. **63**(16): p. 5028-5033.

217. Valastyan, S., et al., *Concurrent suppression of integrin α5, radixin, and RhoA phenocopies the effects of miR-31 on metastasis.* Cancer research, 2010. **70**(12): p. 5147-5154.

218. Zhang, D., et al., *Epidermal growth factor receptor tyrosine kinase inhibitor reverses mesenchymal to epithelial phenotype and inhibits metastasis in inflammatory breast cancer.* Clinical Cancer Research, 2009. **15**(21): p. 6639-6648.

219. Higginbotham, J.N., et al., *Amphiregulin exosomes increase cancer cell invasion.* Current Biology, 2011. **21**(9): p. 779-786.

220. Zhang, H., et al., *TEAD transcription factors mediate the function of TAZ in cell growth and epithelial-mesenchymal transition.* Journal of biological chemistry, 2009. **284**(20): p. 13355-13362.

221. Wang, J., et al., *miR-206 inhibits cell migration through direct targeting of the actin-binding protein Coronin 1C in triple-negative breast cancer.* Molecular oncology, 2014. **8**(8): p. 1690-1702.

222. Giannelli, G., et al., *Induction of cell migration by matrix metalloprotease-2 cleavage of laminin-5.* Science, 1997. **277**(5323): p. 225-228.

223. Mangala, L., et al., *Tissue transglutaminase expression promotes cell attachment, invasion and survival in breast cancer cells.* Oncogene, 2007. **26**(17): p. 2459-2470.

224. Zang, X.P. and J.T. Pento, *Keratinocyte growth factor-induced motility of breast cancer cells.* Clinical & experimental metastasis, 2000. **18**(7): p. 573-580.

225. Ma, J., et al., *Characterization of mammary cancer stem cells in the MMTV-PyMT mouse model.* Tumor Biology, 2012. **33**(6): p. 1983-1996.

226. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC bioinformatics, 2008. **9**(1): p. 559.

227. Errico, M., et al., *Identification of placenta growth factor determinants for binding and activation of Flt-1 receptor.* Journal of Biological Chemistry, 2004. **279**(42): p. 43929-43939.

228. Martínez, A., et al., *The effects of adrenomedullin overexpression in breast tumor cells.* Journal of the National Cancer Institute, 2002. **94**(16): p. 1226-1237.

229. Oehler, M., et al., *Tissue and plasma expression of the angiogenic peptide adrenomedullin in breast cancer.* British journal of cancer, 2003. **89**(10): p. 1927-1933.

230. Ribatti, D., et al., *The role of adrenomedullin in angiogenesis.* Peptides, 2005. **26**(9): p. 1670-1675.

231. Lu, X. and Y. Kang, *Hypoxia and hypoxia-inducible factors: master regulators of metastasis.* Clinical cancer research, 2010. **16**(24): p. 5928-5935.

232. Krishnamachary, B., et al., *Hypoxia-inducible factor-1-dependent repression of E-cadherin in von Hippel-Lindau tumor suppressor–null renal cell carcinoma mediated by TCF3, ZFHX1A, and ZFHX1B.* Cancer research, 2006. **66**(5): p. 2725-2731.

233. Fujiwara, S., et al., *Silencing hypoxia-inducible factor-1α inhibits cell migration and invasion under hypoxic environment in malignant gliomas.* International journal of oncology, 2007. **30**(4): p. 793-802.

234. Steinbrech, D.S., et al., *Fibroblast response to hypoxia: the relationship between angiogenesis and matrix regulation.* Journal of surgical Research, 1999. **84**(2): p. 127-133.

235. Sullivan, R. and C.H. Graham, *Hypoxia-driven selection of the metastatic phenotype.* Cancer and Metastasis Reviews, 2007. **26**(2): p. 319-331.

236. Danø, K., et al., *The urokinase receptor. Protein structure and role in plasminogen activation and cancer invasion.* Fibrinolysis, 1994. **8**: p. 189-203.

237. Bailly, K., et al., *RhoA activation by hypoxia in pulmonary arterial smooth muscle cells is age and site specific.* Circulation research, 2004. **94**(10): p. 1383-1391.

238. Hayashi, M., et al., *Hypoxia up-regulates hypoxia-inducible factor-1α expression through RhoA activation in trophoblast cells.* The Journal of Clinical Endocrinology & Metabolism, 2005. **90**(3): p. 1712-1719.

239. Franovic, A., et al., *Translational up-regulation of the EGFR by tumor hypoxia provides a nonmutational explanation for its overexpression in human cancer.* Proceedings of the National Academy of Sciences, 2007. **104**(32): p. 13092-13097.

240. O'Reilly, S.M., et al., *Hypoxia induces epithelial amphiregulin gene expression in a CREB-dependent manner.* American Journal of Physiology-Cell Physiology, 2006. **290**(2): p. C592-C600.

241. Mateus, A.R., et al., *EGFR regulates RhoA-GTP dependent cell motility in E-cadherin mutant cells.* Human molecular genetics, 2007. **16**(13): p. 1639-1647.

242. Pío, R., et al., *Complement factor H is a serum-binding protein for adrenomedullin, and the resulting complex modulates the bioactivities of both partners.* Journal of Biological Chemistry, 2001. **276**(15): p. 12292-12300.

243. Siclari, V.A., et al., *Tumor-expressed adrenomedullin accelerates breast cancer bone metastasis.* Breast Cancer Research, 2014. **16**(6): p. 458.

244. Wu, Q.-F., et al., *Fibroblast growth factor 13 is a microtubule-stabilizing protein regulating neuronal polarization and migration.* Cell, 2012. **149**(7): p. 1549-1564.

245. Vial, E., E. Sahai, and C.J. Marshall, *ERK-MAPK signaling coordinately regulates activity of Rac1 and RhoA for tumor cell motility.* Cancer cell, 2003. **4**(1): p. 67-79.

246. Takemura, R., et al., *Increased microtubule stability and alpha tubulin acetylation in cells transfected with microtubule-associated proteins MAP1B, MAP2 or tau.* Journal of Cell Science, 1992. **103**(4): p. 953-964.

247. Waterman-Storer, C.M. and E. Salmon, *Positive feedback interactions between microtubule and actin dynamics during cell motility.* Current opinion in cell biology, 1999. **11**(1): p. 61-67.

248. Jaffe, A.B. and A. Hall, *Rho GTPases in transformation and metastasis.* Advances in cancer research, 2002. **84**: p. 57-80.

249. Hollern, D.P. and E. Andrechek, *A genomic analysis of mouse models of breast cancer reveals molecular features of mouse models and relationships to human breast cancer.* Breast Cancer Research, 2014. **16**(R59).

250. Kent, W.J., et al., *The human genome browser at UCSC.* Genome research, 2002. **12**(6): p. 996-1006.

251. Lukacher, A.E., et al., *Susceptibility to tumors induced by polyoma virus is conferred by an endogenous mouse mammary tumor virus superantigen.* The Journal of experimental medicine, 1995. **181**(5): p. 1683-1692.

252. Qiu, T.H., et al., *Global Expression Profiling Identifies Signatures of Tumor Virulence in MMTV-PyMT-Transgenic Mice Correlation to Human Disease.* Cancer research, 2004. **64**(17): p. 5973-5981.

253. Ellis, I., et al., *Pathological prognostic factors in breast cancer. II. Histological type. Relationship with survival in a large study with long-term follow-up.* Histopathology, 1992. **20**(6): p. 479-489.

254. Weigelt, B., F.C. Geyer, and J.S. Reis-Filho, *Histological types of breast cancer: how special are they?* Molecular oncology, 2010. **4**(3): p. 192-208.

255. Ellis, I., et al., *Invasive breast carcinoma.* World Health Organization Classification of Tumours. Tumours of the Breast and Female Genital Organs, 2003: p. 13-59.

256. Pfefferle, A.D., et al., *Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts.* Genome Biol, 2013. **14**(11): p. R125.

257. Cardiff, R.D., et al., *The mammary pathology of genetically engineered mice: the consensus report and recommendations from the Annapolis meeting.* Oncogene, 2000. **19**(8): p. 968-88.

258. Smith, B.A., et al., *Targeting the PyMT Oncogene to Diverse Mammary Cell Populations Enhances Tumor Heterogeneity and Generates Rare Breast Cancer Subtypes.* Genes Cancer, 2013. **3**(9-10): p. 550-63.

259. Miyoshi, K., et al., *Activation of different Wnt/b-catenin signaling components in mammary epithelium induces transdifferentiation and the formation of pilar tumors.* Oncogene, 2002. **21**: p. 5548-55556.

260. Li, Z., et al., *ETV6-NTRK3 fusion oncogene initiates breast cancer from committed mammary progenitors via activation of AP1 complex.* Cancer Cell, 2007. **12**(6): p. 542-58.

261. Currier, N., et al., *Oncogenic signaling pathways activated in DMBA-induced mouse mammary tumors.* Toxicol Pathol, 2005. **33**(6): p. 726-37.

262. Lin, S.-C.J., et al., *Somatic mutation of p53 leads to estrogen receptor α-positive and-negative mouse mammary tumors with high frequency of metastasis.* Cancer research, 2004. **64**(10): p. 3525-3532.

263. Jiang, Z., et al., *Rb deletion in mouse mammary progenitors induces luminal-B or basal-like/EMT tumor subtypes depending on p53 status.* J Clin Invest, 2010. **120**(9): p. 3296-309.

264. Hu, G., et al., *MTDH activation by 8q22 genomic gain promotes chemoresistance and metastasis of poor-prognosis breast cancer.* Cancer Cell, 2009. **15**(1): p. 9-20.

265. Sandgren, E.P., et al., *Inhibition of mammary gland involution is associated with transforming growth factor alpha but not c-myc-induced tumorigenesis in transgenic mice.* Cancer Res, 1995. **55**(17): p. 3915-27.

266. Zhu, M., et al., *Integrated miRNA and mRNA expression profiling of mouse mammary tumor models identifies miRNA signatures associated with mammary tumor lineage.* Genome Biol, 2011. **12**(8): p. R77.

267. Ranger, J.J., et al., *Identification of a Stat3-dependent transcription regulatory network involved in metastatic progression.* Cancer Res, 2009. **69**(17): p. 6823-30.

268. Guy, C.T., et al., *Expression of the neu protooncogene in the mammary epithelium of transgenic mice induces metastatic disease.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10578-82.

269. Schade, B., et al., *Distinct ErbB-2 coupled signaling pathways promote mammary tumors with unique pathologic and transcriptional profiles.* Cancer Res, 2007. **67**(16): p. 7579-88.

270. Ursini-Siegel, J., et al., *Receptor tyrosine kinase signaling favors a protumorigenic state in breast cancer cells by inhibiting the adaptive immune response.* Cancer Res, 2010. **70**(20): p. 7776-87.

271. Schoenherr, R.M., et al., *Proteome and transcriptome profiles of a Her2/Neu-driven mouse model of breast cancer.* Proteomics Clin Appl, 2011. **5**(3-4): p. 179-88.

272. Hu, Y., et al., *Integrated cross-species transcriptional network analysis of metastatic susceptibility.* Proc Natl Acad Sci U S A, 2012. **109**(8): p. 3184-9.

273. Bu, W., et al., *Keratin 6a marks mammary bipotential progenitor cells that can give rise to a unique tumor model resembling human normal-like breast cancer.* Oncogene, 2011. **30**(43): p. 4399-409.

274. Flowers, M., et al., *Pilot study on the effects of dietary conjugated linoleic acid on tumorigenesis and gene expression in PyMT transgenic mice.* Carcinogenesis, 2010. **31**(9): p. 1642-9.

275. Klein, A., et al., *Comparison of gene expression data from human and mouse breast cancers: identification of a conserved breast tumor gene set.* Int J Cancer, 2007. **121**(3): p. 683-8.

276. Maroulakou, I.G., et al., *Prostate and mammary adenocarcinoma in transgenic mice carrying a rat C3(1) simian virus 40 large tumor antigen fusion gene.* Proc Natl Acad Sci U S A, 1994. **91**(23): p. 11236-40.

277. Kretschmer, C., et al., *Identification of early molecular markers for breast cancer.* Mol Cancer, 2011. **10**(1): p. 15.

278. Zhang, M., et al., *Identification of tumor-initiating cells in a p53-null mouse model of breast cancer.* Cancer Res, 2008. **68**(12): p. 4674-82.

279. Backlund, M.G., et al., *Impact of ionizing radiation and genetic background on mammary tumorigenesis in p53-deficient mice.* Cancer Res, 2001. **61**(17): p. 6577-82.

280. Cho, R.W., et al., *Isolation and molecular characterization of cancer stem cells in MMTV-Wnt-1 murine breast tumors.* Stem Cells, 2008. **26**(2): p. 364-71.

281. Tsukamoto, A.S., et al., *Expression of the int-1 gene in transgenic mice is associated with mammary gland hyperplasia and adenocarcinomas in male and female mice.* Cell, 1988. **55**(4): p. 619-25.

282. Pond, A.C., et al., *Fibroblast growth factor receptor signaling dramatically accelerates tumorigenesis and enhances oncoprotein translation in the mouse mammary tumor virus-Wnt-1 mouse model of breast cancer.* Cancer Res, 2010. **70**(12): p. 4868-79.

283. Franks, S.E., et al., *Transgenic IGF-IR overexpression induces mammary tumors with basal-like characteristics, whereas IGF-IR-independent mammary tumors express a claudin-low gene signature.* Oncogene, 2011. **31**(27): p. 3298-309.

284. Liu, S., et al., *Expression of autotaxin and lysophosphatidic acid receptors increases mammary tumorigenesis, invasion, and metastases.* Cancer Cell, 2009. **15**(6): p. 539-50.

285. Bultman, S.J., et al., *Characterization of mammary tumors from Brg1 heterozygous mice.* Oncogene, 2008. **27**(4): p. 460-8.

286. Yin, Y., et al., *Characterization of medroxyprogesterone and DMBA-induced multilineage mammary tumors by gene expression profiling.* Mol Carcinog, 2005. **44**(1): p. 42-50.

287. Gallahan, D., et al., *Expression of a truncated Int3 gene in developing secretory mammary epithelium specifically retards lobular differentiation resulting in tumorigenesis.* Cancer Res, 1996. **56**(8): p. 1775-85.

288. Kuraguchi, M., et al., *Genetic mechanisms in Apc-mediated mammary tumorigenesis.* PLoS Genet, 2009. **5**(2): p. e1000367.

289. Chan, S.R., et al., *STAT1-deficient mice spontaneously develop estrogen receptor alpha-positive luminal mammary carcinomas.* Breast Cancer Res, 2012. **14**(1): p. R16.

290. Pollock, C.B., et al., *PPARdelta activation acts cooperatively with 3-phosphoinositide-dependent protein kinase-1 to enhance mammary tumorigenesis.* PLoS One, 2011. **6**(1): p. e16215.

291. McCarthy, A., et al., *A mouse model of basal-like breast carcinoma with metaplastic elements.* J Pathol, 2007. **211**(4): p. 389-98.

292. Jin, W., et al., *Cellular transformation and activation of the phosphoinositide-3-kinase-Akt cascade by the ETV6-NTRK3 chimeric tyrosine kinase requires c-Src.* Cancer Res, 2007. **67**(7): p. 3192-200.

293. Maroulakou, I.G., et al., *Akt1 ablation inhibits, whereas Akt2 ablation accelerates, the development of mammary adenocarcinomas in mouse mammary tumor virus (MMTV)-ErbB2/neu and MMTV-polyoma middle T transgenic mice.* Cancer Res, 2007. **67**(1): p. 167-77.

294. Liu, M.L., et al., *Amplification of Ki-ras and elevation of MAP kinase activity during mammary tumor progression in C3(1)/SV40 Tag transgenic mice.* Oncogene, 1998. **17**(18): p. 2403-2411.

295. Taneja, P., et al., *Transgenic and knockout mice models to reveal the functions of tumor suppressor genes.* Clin Med Insights Oncol, 2011. **5**: p. 235-57.

296. Jackson, J.G., et al., *p53-mediated senescence impairs the apoptotic response to chemotherapy and clinical outcome in breast cancer.* Cancer Cell, 2012. **21**(6): p. 793-806.