CLINICAL INSIGHTS FROM MOUSE MODELS OF BREAST CANCER

By

Jonathan Paul Rennhack

A DISSERTATION

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

Physiology – Doctor of Philosophy

2018

ABSTRACT

CLINICAL INSIGHTS FROM MOUSE MODELS OF BREAST CANCER

By

Jonathan Paul Rennhack

Breast cancer presents an enormous public health concern. One out of eight women will experience breast cancer in her lifetime. To understand the root of breast cancer initiation and progression, many multi "-omic" projects have been undertaken. This effort has been extremely fruitful in in the discovery of many new genomic events in breast cancer. However, what the studies lack is the functional impact of the genomic events discovered. To understand this, researchers must use *in vitro* and *in vivo* models of the disease. One common *in vivo* model used is the genetically engineered mouse model. Despite their widespread use, there is no integrative database to capture the similarity and differences between human tumors and genetically engineered mouse models.

To begin to address this critical need, we started by identifying genomic copy number alterations (CNAs) in 600 tumors across 27 major mouse models of breast cancer through the application of a predictive algorithm to publicly available gene expression data. It was found that despite the presence of strong oncogenic drivers in most mouse models, CNAs are extremely common but heterogeneous both between models and within models.

Due to the predictive nature of the previous study, we have completed whole genome sequencing and transcriptome profiling of two widely used mouse models of breast cancer, MMTV-Neu and MMTV-PyMT. This genomic information was integrated with phenotypic data and CRISPR/Cas9 studies to understand the impact of key events on tumor biology

To functionalize this data, we followed up on one key amplification event that we found on chromosomes 11D. We identified this event to be associated with worse distant metastasis free survival due to the presence of Co11a1 and CHAD within the amplification event. This was identified through the use of a wound healing assay, tail vein injection, and mammary fat pad injection of CRISPR-Cas9 generated knockout cell lines for Col1a1 and CHAD. In all assays the reduction of metastatic potential was seen. Importantly, we are also able to identify the vulnerability of tumors with the 17q21.33 amplicon to AKT targeted therapy. This was predicted through a number of high throughput genomic and drug compound screens in which unique vulnerabilities were identified in those cell lines containing the 17q21.33 amplicon.

Here we also identified a conserved mutation in phosphotyrosine receptor phosphatase type H (*Ptprh*). The mutation is highly conserved in mouse models of breast cancer and is identified to be mutant in 81% of MMTV-PyMT tumors. A key finding is that *Ptprh* mutations are associated with high EGFR activity, lower latency and more aggressive tumors in a variety of cancer types. Importantly when cell lines with the *Ptprh* mutation were compared against those without the mutation we identified an increased sensitivity to EGFR targeted therapy such as erlotinib associated with *Ptprh* mutation.

Through these studies we have identified key genomic alterations within mouse models of breast cancer. Both of the explored events serve as biomarkers of treatment response and could change the course of therapy for patients. We believe that these events are just case studies and many other events exist within mouse models. Taken together this shows the critical need to increase the depth and breadth of full characterization of mouse models of breast cancer.

This work is dedicated to:
My parents: Dennis and Rhonda.
My brother: Aaron.
My grandparents: Gene and Lorraine Rennhack and Donald and Shirley Hughes.

ACKNOWLEDGEMENTS

The work presented within this dissertation could not have been possible without all of the amazing people that have helped to guide me as a person and as a scientist. Without each and every one of you I would not be the person I am today. While there are far too many to name, I would like to start by thanking the faculty of Lakeshore High School Math and Science Center and Albion College Biology Department for nurturing a love of science and equipping me with the tools to pursue it.

First of all, my family and friends have been an incredible support network. Thank you to my parents Dennis and Rhonda, who have been there for me every step of the way by supporting me and loving me unconditionally. I would not have made it this far without you. Thank you to my grandparents, cousins, and brother for the lifetime of love, support, and friendship. Thank you for my lifelong friends, James Marsh, Nick Arend, Nick and Abbey Herrman, Mark and Elise Miller, Nick Abbey, Jeremy Covell, and Briana To who have been a constant support system.

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 could not have picked a better mentor to start my scientific career. You have given me the balance of freedom and support that is so critical as a graduate student. I believe you have given me the skills and support to succeed as a cancer researcher.

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 past and present. Thank you for all of the shared reagents, protocols, advice, good times and teamwork. Thank you to the MSU histopathology core, high performance compute cluster, and RTSF staff for their constant research support.

A special thank you to my committee: Dr. Bill Henry, Dr. Kathy Gallo, Dr. Hua Xiao, and Dr. Laura McCabe for the letters of recommendation, time spent listening to presentations and the valuable feedback on my scientific directions. Similarly, I'd like to acknowledge members of the cancer research network for their support and guidance especially Dr. Sophia Lunt. Finally, I would like to thank the Physiology program for funding and providing me with the opportunity to earn my degree at Michigan State University.

TABLE OF CONTENTS

LIST OF FIGURES

KEY TO SYMBOLS AND ABBREVIATIONS

ACE - Analysis of CNAs by Expression data

aCGH – Array Comparative genomic hybridization

ADCY33 - Adenylate cyclase 33

AFF2 - AF4/FMR2 Family Member 2

BFB - Breakage-Fusion-Bridge

BRCA1 – Breast Cancer 1

BRCA2 – Breast Cancer 2

CBFb - Core-Binding Factor Beta Subunit

CCND3 – Cyclin D3

CCT4 - Chaperonin Containing TCP1 Subunit 4

CDH1 – Cadherin 1

CDKN1B - Cyclin Dependent Kinase Inhibitor 1B

CDKN2 - Cyclin-Dependent Kinase Inhibitor

cDNA – Complementary DNA

Cenpo - Centromere Protein O

Chad - Chondroadherin

CNA – Copy Number Alteration

CNVs – Copy Number Variants

Col1a1 – Type 1 Collagen Alpha 1

COSMIC – Catalog of Somatic Mutations in Cancer

Cp - Ceruloplasmin

CRISPR - Clustered Regularly Interspaced Short Palindromic Repeats

DMBA - 7,12-Dimethylbenz(a)anthracene

DMN2 - Dynamin 2

DRCR - double rolling-circle replication

EGFR – Epidermal Growth Factor Receptor

EMT – Epithelial-Mesenchymal transition

ER – Estrogen Receptor

FoSTes - replication fork stalling and template switching

FOXA1 - Forkhead Box A1

GATA3 - GATA Binding Protein 3

GEMMs – Genetically Engineered Mouse Models

HER2 - Human Epidermal Growth Factor Receptor 2

Hnrnpab - Heterogeneous Nuclear Ribonucleoprotein A/B

ICGC - The International Cancer Genome Consortium

IVIS – *In Vivo* Imaging System

MAP2K4 - Mitogen-Activated Protein Kinase Kinase 4

MAP3K1 - Mitogen-Activated Protein Kinase Kinase Kinase 1

Matn2 - Matrilin 2

MLL3 - Mixed-Lineage Leukemia Protein 3

MMP23 - Matrix Metalloproteinase 23

MMTV – Mouse Mammary Tumor Virus

MSH6 - MutS Homolog 6

Muc4 – Mucin 4

Muc6 – Mucin 6

NBN - Nibrin

NDL2-5 - Neu Deletion 2-5

NF1 - Neurofibromatosis type 1

NGS – Next Generation Sequencing

PALB2 - Partner and Localizer of BRCA2

PCR – Polymerase Chain Reaction

PDX – Patient Derived Xenograft

PHB - Prohibitin

PIK3CA - Phosphoinositide-3-Kinase Subunit CA

PIK3R1 - Phosphoinositide-3-Kinase Regulatory Subunit 1

Plekhm1 - Pleckstrin Homology And RUN Domain Containing M1

PMS2 - PMS1 Homolog 2, Mismatch Repair System

PR – Progestin Receptor

PTEN - Phosphatase And Tensin Homolog

PTPN22 - Protein Tyrosine Phosphatase, Non-Receptor Type 22

Ptpr – Phosphotyrosine Phosphatase Receptor

Ptprd - Phosphotyrosine Phosphatase Receptor Type D

Ptprh - Phosphotyrosine Phosphatase Receptor Type D

PyMT – Polyoma Middle T

qPCR – Quantitative Polymerase Chain Reaction

RAD51C - RAD51 homolog C

RB1 - Retinoblastoma Protein

RNA-Seq – Ribonucleic Acid Sequencing

RPPA – Reverse Phase Protein Array

RTKs – Receptor Tyrosine Kinases

RUNX1 - Runt Related Transcription Factor 1

SF3B1 - Splicing factor 3B subunit 1

siRNA – Small Interfering Ribonucleic Acid

SNVs – Single Nucleotide Variant

Sumo2 - Small Ubiquitin-Like Modifier 2

TALEN - Transcription Activator-Like Effector Nucleases

TBX3 - T-Box 3

TCGA – The Cancer Genome Atlas

TP53 – Tumor Protein 53

WAP – Whey Acidic Protein

INTRODUCTION

BREAST CANCER

Breast cancer is an extremely prevalent disease. At its most basic definition it is uncontrolled cellular proliferation of tissue in the breast. In fact, it is estimated that there will be approximately a quarter million new diagnoses of breast cancer in the US in the year 2018[1]. Compounded over the life span of a women, 1 in 8 women, will experience breast cancer in their lifespan[2]. There is also a rate of 1 in 1000 for male breast cancer[3]. The sheer number of cases causes breast cancer to be a huge public health concern.

Breast cancer largely begins in two areas. These are the duct of the mammary gland and the milk producing lobule. Ductal breast cancer is the most common with almost 90% of all cases resulting from this cell type[4]. Lobular breast cancer is much rarer. Both subtypes are divided into two classes: invasive and *in situ*. A breast cancer is classified as *in situ* if it has not invaded out of its original structure (duct or lobule) to the surrounding tissue. If this is the case, the survival rate of the patient is very high[4]. Survival rates are worse for invasive disease if they started in the duct or lobule[4]. Other rarer types of breast cancer such as inflammatory breast cancer are less understood and thus have fewer treatment options and worse outcomes.

For women, breast cancer is the second leading cause of death behind lung cancer[5]. The average five-year survival rate of women with breast cancer is 90%[6]. However, this is closely linked to staging of the disease. Stage 0 or 1 breast cancer, local disease, has a five-year survival rate of close to 100%. However, as the disease progresses to stage 2, 3, or 4 the survival rate drops dramatically. Stage 4 has a five-year survival rate of only 22% [6]. The poor survival associated with stage 4 breast cancer is because this staging of breast cancer represents systemic disease. At this disease stage, the tumor has invaded beyond the breast and colonized different organs in a process called metastasis.

Once the disease is metastatic it is systemic. This causes several unique challenges in treatments with regards to treatment delivery and efficacy which are explored in later chapters of this work. Ultimately the colonization of a vital distant organ, not the primary tumor in the breast, results in the death of the patient. To improve patient outcomes researchers must develop therapeutic regimens which not only treat the primary tumor but also prevent the development of new and kill metastatic lesions existing at the time of diagnosis.

TUMOR METASTASIS

The process of tumor metastasis involves a number of complex stages each of which has its own challenges that the tumor cell must overcoming. Each of these stages have been profiled from a gene expression point of view[7,8], but there is still much more work to do to understand all of the genes involved with tumor metastasis.

The first stage of metastasis is the tumor cell must develop invasive properties. The cause of this remains unknown, but a number of factors have been implicated in the initiation of the metastatic cascade. These factors include hypoxia[9] and paracrine signaling[10] from the tumor. Once the cell has acquired an invasive phenotype, it can begin to migrate to the surrounding tissue, or enter the lymphatic system or circulatory system. Key proteins have been identified in the extracellular matrix[11] that encourage migration and the perpetration of the invasive phenotype.

The process of entering the blood vasculature is call intravasation. Once in the blood stream the cell has a number of factors it must evade to survive. This includes an unnatural pH, surviving in a liquid environment rather than a solid tissue, and nutrient restraints. The mechanisms that a tumor cell uses to survive in the vasculature is still unknown. However, it has been shown that there is increased survival of cells which migrate in clusters as well as those who are able to attract platelets as a level of protection[12].

The collection of tumor cells and platelets are relatively bulky and tend to get caught in capillary beds. Once stuck in the capillary bed the tumor must leave the vasculature in a process called extravasation and colonize the distant organ. A still relatively unstudied part of the metastatic cascade is how tumor cells begin to proliferate again in the distant organ. In some cases, it has been shown that the tumor cell in the distant organs can remain dormant for decades before becoming active and forming a metastatic lesion[13].

The most common site of breast cancer metastasis are capillary rich organs including the bone, brain, liver and lung[14]. Interestingly, each of these metastatic sites show unique transcriptomic profiles. Work by the Massague group identified that cells with a unique transcriptional profile will preferentially colonize the lung[15], bone[16], and brain[17]. The causes behind the identified transcriptional changes are largely unknown. Surprisingly, whole genome sequencing has not identified characteristic changes in the genome between metastatic locations[18]. This indicates that the changes are not hardwired into the tumor cell but are of a more transient nature. Emerging evidence has pinpointed that these changes may be epigenetic in nature and are flexible throughout the lifespan of the tumor and metastatic cascade[19]. The epigenetic nature of the changes reveals an interesting therapeutic avenue which is currently being explored for treatment of the metastatic lesions.

GENOMIC STABILITY

Tumor metastasis is not the sole process involved in the progression of tumors. In Hanahan and Weinberg's seminal paper stating the changes that must occur for cancer to occur they outlined the six hallmarks of cancer as: sustained proliferative signaling, evasion of growth suppressors, activation of invasion and metastasis, enabling replicative immortality, introducing angiogenesis, and resisting cell death[20]. While the mechanism of how each of these processes is activated is still under debate, Hanahan and Weinberg agree that an enabling characteristic underlying each hallmark is the presence of genomic instability.

Genomic instability causes changes in the genome of a cancer cell. These changes come in three varieties: Single nucleotide variants or short indels, gene copy number changes, and gene translocations. Each change alters the behavior of a critical protein in each process and leads to tumor progression.

A common type of variant, single nucleotide variants and indels are frequent in breast cancer. In these types of variants, a single nucleotide is changed, or a small section of nucleotides are lost or gained. This in turn will alter the protein structure and eventual function. Across all subtypes of breast cancer, it was found that PIK3CA, PTEN, AKT1, TP53, GATA3, CDH1, RB1, MLL3, MAP3K1, CDKN1B, TBX3, RUNX1, CBFB, AFF2, PIK3R1, PTPN22, PTPRD, NF1, SF3B1 and CCND3 were all mutated [21]. Many of these mutations have been identified to have a pro tumorigenic effect through activating oncogenic changes others through loss of function mutations of tumor suppressors.

Copy number alterations are extremely common in breast cancer. They present as either gene deletions or gene amplifications. These alterations can span either short regions encompassing just a few genes to large regions spanning entire chromosomal arms and cause changes to many genes. Gene amplifications typically present as either chromosomal tandem repeat regions or extrachromosomal double minute events. In tandem repeats a region of the chromosome in duplicated in series head to tail. For double minute events the amplification event has broken from chromosome and has formed a small chromosome like structure containing the amplified region.

The process of gene amplification is still debated however the prevailing hypothesis involves mistakes in the process of DNA replication during mitosis. There have been four main models proposed for the mechanism of gene amplification. These include: extrareplication and recombination, the breakage-fusion-bridge (BFB) cycle, double rolling-circle replication (DRCR), and replication fork stalling and template switching (FoSTes)[22].

In extrareplication and recombination a secondary replication fork is formed during DNA replication. This creates an extra copy of the region being replicated. After completion of the replication the newly created “replication bubble” will break and fuse forming classic double minute structures. After the double minutes are formed they are able to stay as extrachromosomal structures or be embedded in chromosomes through break and repair events [23–26].

The breakage-fusion-bridge (BFB) model supports the generation of both tandem repeats and double minutes. In this model, a break occurs in the chromosome. During replication, due to the lack of telomere the broken ends of a chromosome are fused together creating head to tail repeats. This can continue for many cycles and create many head to tail repeats or during subsequent cell divisions homologous recombination can occur. If recombination occurs between replicated intrachromosomic events, the resulting structure is a double minute[27–30].

The last two models, DRCR and FoSTes are relatively young models and do not have the same type of support from a molecular biology point of view and from a community acceptance standpoint. DRCR has been described as an important experimental system and is widely utilized in the Cre-Lox system; however, there is no evidence that this occurs in cancer[31]. FoSTes also has little support to no evidence of occurring in cancer but has been shown to be involved in other diseases including Pelizaeus-Merzbacher and Charcot-Marie-Tooth disease[32]. For the sake of brevity and the lack of involvement of these models in cancer they will not be discussed further here

Common copy number alterations in breast cancer were identified by the TCGA group. Across breast cancer regardless of subtype TCGA identified PIK3CA, EGFR, FOXA1 and HER2 as focally amplified and MLL3, PTEN, RB1 and MAP2K4 as focally deleted[21]. However, it is likely there are many more gene amplification and deletions that are influential on tumor behavior or are subtype specific.

The last type of instability event, translocations, are not well classified in breast cancer. In this event one part of a chromosome breaks and is fused to a new location either on the same chromosome or a different chromosome. The resulting structure can have a number of effects. It can link a gene with the wrong regulatory regions. It can cause premature ending or elongation of the translation of a protein. In rare cases, the resulting structure can cause the production of a fusion protein in which two functional parts of a protein are fused together causing the creation of a new protein with oncogenic properties. Classical translocation events such as BCR-ABL are extremely rare and it was believed until recently that there were very few defining breast cancer translocation. However recent work has begun to show that subsets of patients do have oncogenic translocations including various *NRG1 translocations and a recurrent* MAGI3–AKT3 translocation[33].

The events described above are all somatic events. Thus, they arise in somatic cells in the patient and are not passed on to the offspring. This is the case in 90% of breast cancer patients. However around 10% of patients have disease which is influenced by germline mutations. The most common of these are mutations in BRCA1 and BRCA2. However there have also been inherited mutations in PALB2, PTEN, NBN, RAD51C, RAD51D, MSH6, and PMS2 as well[34].

BREAST CANCER HETEROGENEITY

Another compounding factor in the poor outcomes associated with breast cancer is the heterogeneity associated with the disease. Due to difference in selective pressure and random chance compounded with the inherent genomic instability in cancer each breast cancer is unique. There are differences in growth rate, metastatic ability, and response to treatment.

There is also heterogeneity within a single tumor. Different regions are under different selective pressures. Due to the inherent genomic instability in tumor cells, this lead to the tumor being under Darwinian principles. In a survival of the fittest type model, different regions of the tumor will have a different mutational profile. The most apparent indicator of the heterogeneity found within a tumor is the presence of multiple cell morphologies or histological subtypes in different regions of the tumor. There are also genomic studies that have confirmed this diversity including aCGH as well as single cell sequencing studies genomic and transcriptomic level. Recent studies have shown differences in key oncogenic pathways such as TP53 and PI3KCA[35].

The type of diversity noted above is spatial heterogeneity. However, throughout the development of the tumor there has also been evidence of temporal heterogeneity. The temporal heterogeneity typically presents itself in two unique situations. In the case of metastasis, a group of primary tumor cell seed a distant organ. This distant organ has distinct selective pressures and shifts the evolution of the metastatic site in a unique direction from the primary tumor. This increases the diversity of the tumor.

A large selective pressure that a tumor undergoes during the course of disease is treatment. This pressure leads to a bottleneck in evolution. In this case, those clonal population that are sensitive to the treatment are killed but the resistant populations can survive and repopulate the tumor and or metastatic lesions.

This has obvious implication in the patient treatment and survival and indicates a specific model for how cancer care might be handled in the future. One could envision a situation where a patient that is diagnosed with go through various rounds of biopsy, genetic profiling, treatment, relapse, and subsequent biopsy to understand the new dominant clonal population. The understanding of each genetic change of the dominant clone and tailoring treatment accordingly has spawned the field of research call precision medicine.

PRECISION MEDICINE

Precision medicine is the tailoring of medical treatment to a particular patient group based upon genomic or imaging analysis of a disease. The goal of the field is to stratify diseases in such a way that treatment can be tailored to a particular disease subtype and improve patient outcomes through identifying therapeutic vulnerabilities or by reducing treatment side effects. In this pursuit scientists and clinicians hope to advance the field in such a way to provide the right treatment to the right patient at the right time.

The field of precision medicine is not a new field. It has long been common practice to tailor disease treatment to specific patient populations. Notably in cancer, in the early 1980's it was discovered to give patients positive for Estrogen Receptor competitive estrogen binding molecules like tamoxifen[36] significantly improved patient survival. However, there has been renewed interest in the field since the establishment of the Precision Medicine Initiative by President Barack Obama in the 2015 State of the Union address. The call to action and founding of the initiative have been ignited by recent advances in high throughput technology to understand disease on a molecular level in a large-scale manner.

Due to the genetic nature of the disease, the field of cancer research and treatment has embraced the use of precision treatment more quickly than other fields. Breast cancer, has seen large advances in patient outcomes due to the use of tailored therapeutics and genetic classification of the disease. Historically, breast cancer has been classified on the status of Estrogen Receptor (ER), Progesterone Receptor (PR), Human Epidermal Growth Factor Receptor 2 (HER2) as well as proliferative markers such as Ki67.

ER is the single most important marker for clinical classification of the disease. 75% of breast cancers are classified as ER positive and are shown to be response to estrogen targeting therapy. The other clinically relevant endocrine marker in breast cancer, PR, is found in approximately 70% of breast cancer patients. While ER and PR status do not differ frequently they are all markers of responsiveness to endocrine targeted therapy[37].

HER2 is an extremely important marker for clinical decision making regarding treatment. It has been identified that approximately 20% of breast cancers have amplification and/or overexpression of the HER2 gene[38,39]. This has been shown to be a member of the epidermal growth factor receptor protein family and alteration of these family members have been shown to cause uncontrolled cell proliferation. Importantly breast cancer patient's classified as HER2 positive are responsive to HER2 targeted therapy.

Ki67 is a proliferative marker. It identifies highly proliferative breast cancer patients. Due to the proliferative status of these tumors, it is an important marker of response to adjuvant chemotherapy [40]. A combination of ER, PR, HER2, and Ki67 status has largely driven the course of treatment for the majority of recent breast cancer patients.

However, with the advancement of transcriptomics and the advent of DNA microarray technology breast cancers began to be profiled on their global genomic profile. Based upon hierarchical clustering of breast cancer patients four subtypes were identified Luminal A, Luminal B, Basal, and HER2 positive [41]. More recently a new subtype, Claudin Low has been established [42]. Luminal A and B tumor cells resemble cells that start in the lumin of the mammary duct. These tumor types tend to be ER or PR positive with luminal A being slower proliferating and thus not responsive to traditional chemotherapy. The vast majority of basal tumors are ER, PR, and HER2 negative, also known as triple negative, and resemble cells that are outside the mammary duct. Basal tumors do not respond to endocrine therapy or HER2 targeted therapy. The HER2 molecular subtype is not the same as HER2 positive tumors but like basal and triple negative they correlate closely. The HER2 molecular subtype of tumors are responsive to HER2 targeted therapy. Despite the PAM50 subtype's potential for driving therapy at this point it does not add much information to ER, PR, HER2, and Ki67 markers in terms of clinical decision making.

However, there are other gene expression-based assays which are routinely used in making treatment decisions. These are most commonly Mammaprint[43] and Oncotype DX[44]. Oncotype DX is 21 gene signature used in ER+ patients. This is used to predict recurrence and identify a subset of ER+ tumors which are responsive to chemotherapy. Mammaprint is a larger 70 gene assay which is used regardless of ER status. Mammaprint helps to identify tumors that are likely to metastasize and a class to respond to chemotherapy.

With the drastic drop in next generation sequencing pricing, the development of sequencing-based panels was accelerated. The most common of these, the Foundation One ("FoundationOne - Foundation Medicine," 2018) panel, include 61 genes that are tested for sequence or copy number variants. Based upon the status of these genes the report matches therapeutics and clinical trials that the patient might benefit from.

The advances in transcriptomic and next generation sequencing have rapidly brought down the price of understanding the tumor genetics and advancing precision medicine. However, the problem has now shifted from generating the data to understand the genomic landscape but correlating that data with clinical outcomes.

"-OMIC" PROFILING OF THE DISEASE

The advances in precision medicine have mirrored the advances in high throughput "-omic" technologies. There have been three major advancements that have caused a dynamic shift in the field of precision medicine. These are the ability to profile the entire transcriptome through the use of microarray technology, advancements in next generation sequencing to profile the transcriptome and genome, and most recently the advancement of single cell profiling. Each of these technologies have revealed a deeper understanding the large amount of heterogeneity in breast cancer.

The earliest profile of the transcriptome in a high throughput manner was made possible through the invention of cDNA Microarrays. In short, the microarrays are chips with tens of thousands of oligos fused to them which are unique to a specific transcript. To analyze transcription levels, RNA is reverse transcribed to cDNA, labeled with a fluorescent probe and allowed to flow across the chip. The amount of transcript level in interpreted through, after a number of normalization steps, the intensity of bound fluorescent cDNA at each fused oligo location. There are two major types of chips - Agilent and Affymetrix.

Though the outcome of both types of microarrays are the same, the transcriptomic profile of the tumor, there are key differences. Agilent assays use unique 60-mer probe IDs. Furthermore Agilent uses a Cy3, Cy5 dye system to calculate expression values[46].

Affymetrix microarrays use a 25-mer system which contains exact matches to the gene of interest and probes with a mismatched nucleotide. This match/mismatch system allows the calculation of real signal to noise to be completed with just a single dye system[47]. This is in contrast to the two dye Agilent system described earlier.

While both of these systems fill similar niches of transcription, they produce slightly different results. In head to head comparisons Affymetrix seems to be able to pick up more minor differences in transcriptional levels. However, this is at the cost of more false negative gene calls. Agilent has less false positives but transcriptional differences must be relatively large to be picked up on this technology[48].

Microarray technology has several advantages. The major advantage that microarray had over other types of platforms is the cost. There is also an advantage with the speed and ease that data can be collected from microarrays. Microarray experiments can be completely quickly and with no advanced computing power needed. Furthermore, once the data is normalized the file sizes are small and can be analyzed using simple scripts or even commercially available spreadsheet management software such as Microsoft Excel. However, there are key deficiencies in microarray technology. The most glaring deficiency that microarrays have is their limited ability to determine the frequency at with SNPs occur in a transcript, the abundance of splice forms, and the presence of rare transcript.

Another issue with microarray technology is the variability of microarray data. There will be large differences in the data returned based upon each experimental run. These are known as batching effects and can largely be removed through mathematical manipulation to remove variance. However, with the use of this software one can remove technical artifacts that have been introduced but also remove some biological variance. This mutes any biological differences seen and makes it difficult to identify small differences in the biological variation.

With the dramatic decrease in the cost of RNA sequencing (RNA-seq) has become more common than microarray because it addresses some of these key deficiencies. RNA-seq also has the first step of conversion to cDNA. However, next it involves the shattering of cDNA into smaller chunks and the fusing of adapter sequences to one or both sides of the fragment. These known adapter sequences are fused to a plate to allow for the fluorescence based sequencing of the unknown regions. Each of these unknown sequences, called reads, are then mapped to their transcriptomic location. The relative abundance of each transcript is calculated from the number of reads that map to a particular transcript after size of the transcript is adjusted for.

The largest benefit of RNA-seq is the quantity of data generated. Nucleotide variants, alternative splicing, as well as the identification of rare transcripts can easily be identified using RNA-seq. However, the drawback of the RNA-seq is the accessibility of the data. Due to the large file sizes and scale of the data, an enormous amount of processing power must be used to analyze the data in a time friendly manner. This usually requires the use of a high performance compute cluster either locally or in the cloud. Also due to the relatively young age of the field the pipelines for data analysis are not well worked out and many times software is not user friendly. This makes the burden for entry to dealing with NGS data relatively high.

Coupled with the advance in RNA-seq came the ability to sequence the entire genome. Using a workflow similar to RNA-seq, a researcher can now determine the entire genomic landscape of a tumor. This includes nucleotide variants in both coding and non-coding regions of the genome, copy number changes, and translocations. The amount of information the one is able to obtain from next generation sequencing (NGS) is directly proportional to the depth of coverage (the theoretical amount of times that the entire genome has been sequenced). With deep enough sequencing, one can also understand the frequency at which a variant occurs throughout the tumor and give some insight into the clonal heterogeneity within the tumor.

Further information about the intratumoral heterogeneity can be obtained through the use of single cell sequencing technology. This technology is emerging and has become widespread in the last two years. With the advance of DNA and RNA amplification technology as well as single cell isolation technologies it is now possible to perform NGS of RNA and DNA on single cells. After the single cell isolation and amplification of DNA or RNA, the process of data generation and analysis is extremely similar to RNA-Seq and DNA-Seq. The major drawback of this type of data is similar to the RNA-seq and DNA-seq where there is a large amount of computational power required to analyze the data. Also, there is controversy in the field as to the utility and translatability of findings using single cell technology.

CELL LINES

A key tool in the study of what each genomic event does in cancer is through the use of *in vitro* cell line experimentation. The first cell line was originally established in the early 1950's by George Gey[49]. This cell line, named HeLa, was a cervical cancer line and it revolutionized the way that cancer was studied. With this line, it was now possible to culture a cancer line with normal cell culture media and perform a host of *in vitro* experiments.

Shortly after the derivation of the HeLa line, many other lines were developed including the first breast cancer line, BT-20 in 1958. Since that time a variety of lines have been developed representing various subtypes of breast cancer. With the establishment of various cell lines researches have virtually unlimited access to a relatively homogenous population of tumor cells for experimentation.

This has been paired with various genomic techniques such as siRNA, TALEN, and CRISPR to manipulate the genomic and expression landscape of many lines to tease out hypotheses and make translational findings. Another key strength of the model system that complements nicely with the ease of genetic engineering is the ability of high throughput screening. The cell line system allows for the ease of global siRNA or CRISPR and drug screens to identify key dependencies in various tumor types

Two major studies through the Broad institute and Sanger Institute have helped to define vulnerabilities of cell lines in regard to genomics and chemical compounds. With the Broad project, named cell dependency map, a panel of cell lines has been carefully gnomically characterized and then screened with a genomic siRNA[51] and CRISPRi screen[52]. This analysis allowed for correlation of genomic events including, single nucleotide variants (SNVs), Copy number variants (CNVs), and translocations with genetic dependencies[53]. The Sanger group took a similar approach where they identified well characterized cell lines and subjected them to a high throughput screen. However, instead of using genomic purturbins they used chemical compounds with the hope of identifying compounds that will target specific genomic events[54]. The goal is that these compounds would also target tumors in patients with similar genomic changes as those identified in the cell lines.

However, there has been debate about the translatability of findings with cell lines to the clinic. This largely is rooted in the environment which they are grown. The majority of cell lines are grown on a 2D plane in plastic dishes. This is very different from the three-dimensional complex tissue setting that a tumor is typically found in. The differences have been shown to cause major differences in drug response between a 2D cell culture system and a 3D complex environment.

Furthermore, many cell lines, especially in breast cancer, are not derived from the primary tumor. Instead they are derived from distant metastatic lesions, pleural effusions, and ascites. It is predicted that due to the large transcriptional differences present in each site it is predicted that cell lines derived from a metastatic lesion may not represent a primary tumor[56].

Many of these differences are represented when a cell line is transplanted into a mouse host. Importantly the largest change is that many cell lines derived from metastatic tumors have lost their ability to metastasize in the mouse. This fact along limits the utility of breast cancer cell line *in vivo*.

PATIENT DERIVED XENOGRAPH MODELS

In order to improve on cell line research, researchers developed patient derived xenograft (PDX) models. In these models, tumor biopsies are taken directly from the patient and implanted into a mouse host within hours. As of 2016 over 500 different PDX models have been created from breast cancer patients[57]. These models represent a variety of histological and molecular subtypes of cancer. These models have been shown (at least initially) to reflect their parent tumor with genomic alterations, gene expression, histology, and treatment response[58].

The obvious benefit to using PDX mouse models is the quick translatability to the clinic. It has been shown that in a number of studies and drug designs that a response in PDX is indicative of a response in humans[58]. This not only allows for the potential impact on the clinic immediately, but it also provides an import resource for moving treatments to clinical trials. In fact, they serve as an early indicator as to if a novel treatment will pass a clinical trial.

However, despite these advantages a number of flaws exist with the PDX models. The models only represent the most aggressive cancers and only represent a relatively small amount of the diversity found in breast cancer patients. In one such study where novel PDX's were derived, 113 tumors were implanted with an overall take rate of 27.4% (31/113). These were highly skewed towards basal breast cancer subtype. Specifically, the take rate was 51.3% (20/39) in basal cancer, 26.5% (9/34) in HER2+, 5.0% (2/40) in luminal B and 0% (0/3) in luminal A. Furthermore, through multiple passages the PDX models undergo selective pressures and develop mouse specific genomic changes.

The biggest challenge to the PDX models is the presence of human tissue in mouse. This causes unnatural interactions between the human tumor and the stroma from the mouse. Furthermore, the mouse must be immunocompromised in order to prevent immune rejection of the foreign tissue. With the renewed interest in the tumor and its relationship to the immune system this has become an increasingly large flaw with PDX models. To combat this, researchers have developed humanized mouse models, but these lines are extremely expensive to create and maintain. Thus, the model is cost prohibitive for large scale studies such as those needed for high throughput drug treatment studies. Due to these limitations, genetically engineered mouse models have become increasingly popular to use.

TRANSGENIC MOUSE MODELS

Mouse models are an extremely important tool in understanding basic tumor biology. To initiate tumorigenesis in mice, researchers have employed several strategies including chemical carcinogen treatment, viral infection, and genetically engineered mouse models (GEMMs). Using these models, key oncogenes and tumor suppressors have been characterized. Furthermore, many models have been created to model different subtypes of the disease and various characteristics such as tumor metastasis or genomic instability.

The most common chemically induced breast cancer tumor is generated by giving 7,12-dimethylbenz[a]anthracene (DMBA) orally[59]. This model has been shown to create mammary tumors in 75% of mice with an average latency of 22 weeks. Importantly this model has helped reveal the full process of exposure to a carcinogen, to mutational impact, to tumorigenesis (Currier 2005). Also, these tumors are shown to have a variety of histological subtypes and pathways active which is reflective of the human disease. Despite these similarities to the human disease criticisms of the model remain due to the limited scope of tumorigenesis in this model: Many human breast cancers are not due to chemical carcinogen exposure.

The most early studies of most models began last century with the identification of a number of inbred strains which had a tendency to develop breast cancer[60]. To begin to understand why these specific strains developed tumors more frequently than other strains, crosses were performed between high incidence strains and low incidence strains. It was shown that there was a high influence of the mother on the incidence of tumorigenesis in the offspring. This indicated that there was a sex linked or epigenetic factor driving the tumors in the high incidence strains. Early work by Bitner suggested that this influence was in fact milk derived[61] and follow-up work showed that it was in fact a virus, named the mouse mammary tumor virus (MMTV), which caused tumors. Despite the fact that this virus was poorly infective, and the model was much maligned about its relevance it was found that the viral protomer could be used to drive gene expression in the mammary gland.

To have an in-depth view of each discovered oncogene researchers have utilized GEMMs. In many models a tissue specific promoter such as the MMTV or WAP (whey acid protein) promoters are used to drive tumorigenic in a mammary specific manner with some leaky expression noted in other endothelial cells. These models allow for the identification of specific changes or dependencies related to a specific oncogene or tumor suppressor of interest. Two key models, the MMTV-Neu and MMTV-PyMT, have had a remarkable impact on the discovery of basic tumor biology.

The MMTV-Neu mouse model was meant to mimic HER2 amplified subtype of cancer. In this model, a non-activated form of Neu is overexpressed in the mammary gland. This causes the development of mammary tumors with a relatively long latency. Tumors develop in 50% of mice with a median latency of 202 days[63]. Importantly, these tumors were shown to carry the phosphorylated form of Neu; indicating that the tumors were in fact driven by the Neu transgene. Beyond this, the tumors showed a singular histological subtype of adendocarcinoma. Like human HER2 positive tumors, 72% of tumors formed in the mouse model were shown to be metastatic to the lung. A further finding which increased the translatability of the model is that it has been shown to have similar oncogenic signaling pathways as HER2 positive tumors. Notably this includes AKT and the E2F family of transcription factors[64].

The other model discussed at length in this thesis is the MMTV-PyMT tumor model. The MMTV-PyMT model uses MMTV promoter to drive expression of the middle T antigen and subsequently generate mammary tumors[65]. This tumor model is highly aggressive and metastatic. Tumors developed with an average latency of 45 days in this model and the majority (94%) of mice had metastatic disease to the lung. The molecular pathways associated with this model are also consistent with human tumors[66]. Furthermore, the model like the human disease produces a variety of histological features.

RATIONALE FOR DISSERTATION

To improve patient outcomes, I believe that we must continue to advance the field of precision medicine with an eye towards treatment of tumor metastasis. We must develop therapeutic regimens that match the right therapy with the right patient, at the right time. To pursue this goal, we must identify biomarkers present in the tumor that are associated with various tumor behaviors and ultimately treatment response.

In order to study this, I chose to use genetically engineered mouse model systems. I believe that GEMMs are an ideal model system for the discovery of new biomarkers for a number of reasons. Importantly, tumors formed in many of these models are metastatic. This is in contrast to cell lines and xenograft models where metastases are rare after orthotopic injection. Furthermore, mouse models undergo evolutionary selection throughout their development. Despite the presence of oncogenic initiating events, the mouse model is still subjected to Darwinian pressures. Other models such as PDX and cell lines are already transformed and do not have the same type of pressures. This limits their utility for discovering new tumor influencing genomic event. Finally, genetically engineered mouse models have been shown to be heterogeneous and capture a wide range of human tumor diversity. The diversity present in mouse models allow findings in this system to be readily translated to the clinical setting. On the other hand, PDX and cell line models capture a relatively small patient population.

Despite the wide spread use of mouse models, there remains a critical need to understand which models resemble which subtypes of human breast cancer. Historically, researchers chose their model based upon the initiating oncogenic event. That is if a study is on HER2 positive tumors, a researcher might choose to use the MMTV-Neu mouse model. Or if one was interested in basal tumors an MMTV-Myc mouse model may be used. However, these may not be the best option. Recent work has shown the MMTV-Neu tumors to be much more similar to Luminal A or B tumors than HER2 positive tumors from a transcriptional viewpoint. Furthermore, MMTV-Myc tumors have been shown to be extremely heterogeneous with some tumors resembling basal tumors but many others resemble other human breast cancer subtypes. In fact, it was shown that many mouse models are extremely heterogeneous and the adage “One oncogene, one tumor” is false when it comes to mouse models.

This heterogeneity has been profiled in a few papers from a transcriptional viewpoint. However, what causes this heterogeneity is unknown. I set out with the central hypothesis that genetically engineered mouse models have the same underlying genomic instability as human cancers and are subjected to the same evolutionary forces. I first wanted to pursue this at a copy number level. I designed an experiment to identify the gene copy number changes present in mouse models of cancer.

To pursue this, I assembled a large database of 27 different mouse model of breast cancer and 600 tumors spread across the models. This database was of publicly available transcriptomic data from the various mouse models. From this transcriptomic data I used a previously established predictive algorithm to identify gene copy number variants. After calling copy number variants I worked to identify key variants in each model and how these variants were reflective of human breast cancer.

This analysis revealed, as predicted, mouse models had a large number of copy number changes in them. Furthermore, the copy number variants reflected the heterogeneity found in the model. Importantly I identified conserved copy number variants in both mouse and human. However, these variants were not in traditional tumor suppressors or oncogenes. From these I hypothesized that the events were causing other evolutionary advantages for the tumor and maybe involved in other aspects of tumor progression such as angiogenesis or metastasis.

To follow-up on the prediction that copy number variants were involved with secondary tumor characteristics and due to the predictive nature of the previous study I sought to perform an integrative copy number analysis. In this experiment I performed next generation whole genome sequencing on two highly utilized mouse models (MMTV-PyMT and MMTV-Neu). With this data I identified single nucleotide variants, copy number variants, and translocations. I integrated these variant calls with transcriptomic data to get a full understanding of the genomic and transcriptomic landscape of the tumors. Furthermore, I integrated these changes with tumor phenotypes with an emphasis on tumor metastasis. In the final stage of this study I used CRISPR-Cas9 experiments to confirm phenotypic impacts of each variant through the use of *in vitro* and *in vivo* studies.

Specifically, in the MMTV-Neu model I identified an amplified region associated with metastasis. Importantly this region is amplified in 25% of human HER2+ve breast cancer and is linked with metastatic progression. Knocking out genes in this region resulted in reduced migration and metastasis in both mouse and human breast cancer cell lines. Likewise, in the MMTV-PyMT model we identified a mutation in PTPRH, a protein tyrosine phosphatase, which was conserved in over 80% of tumors. PTPRH normally dephosphorylates EGFR, and the PTPRH mutation that we identified is associated with an ~15-fold increase in phosphorylated EGFR levels. Critically, we found that human lung cancers had mutations in PTPRH that were mutually exclusive with amplified or mutated EGFR. Furthermore, cell lines derived from PTPRH mutant tumors were responsive to tyrosine kinase inhibitor therapy while wild type PTPRH tumors were not responsive.

While my studies are exciting, they are not exhaustive and there is much more work to be done. The most immediate work to be done is to expand the study beyond the MMTV-Neu and MMTV-PyMT models. More models must have comprehensive transcriptomic and genomic profiling, so the research community can make educated decisions about which model to use for a particular study. Furthermore, the analysis presented in this thesis is a first pass analysis. More work should be performed to characterize the translocations as well as the non-coding variants. Last, further biochemical work needs to be performed to fully characterize the mechanism between the amplification event and metastasis as well as PTPRH mutations and EGFR signaling dependence.

My work will have an immediate impact to the mouse modelling, cancer research, and clinical communities. This is the study, the first of its kind, profiling the MMTV-Neu and MMTV-PyMT mouse models at the sequence level. This shows the importance of secondary genomic events in driving tumor progression and heterogeneity in mouse models which were previously thought to be largely driven by the initiating engineered event. We expect this manuscript to have translational findings to other models of breast and other cancer types and to inspire similar studies in other models. I have uncovered and functionally characterized found two novel alterations in breast cancer, these findings and the other alterations we describe will have an immediate impact on the basic cancer research community. The genomic events we observed include a copy number alteration which drives tumor metastasis and a single nucleotide variant which modifies EGFR signaling. The events are described more in detail below. Finally, I have identified a mutation which will be directly relevant in a clinical setting and our immediate next steps are geared towards translation to the clinic. Tumors with the uncovered mutation readily respond to EGFR targeted therapy and this mutation has strong potential to be a deciding biomarker in the course of patient therapy.

WORKS CITED

WORKS CITED

1. ACS. Breast Cancer Facts and Figures. (2018).

2. DeSantis, C., Ma, J., Bryan, L. & Jemal, A. Breast cancer statistics, 2013. *CA. Cancer J. Clin.* **64,** 52–62 (2014).

3. Giordano, S. H. A review of the diagnosis and management of male breast cancer. *Oncologist* **10,** 471–9 (2005).

4. Li, C. I., Anderson, B. O., Daling, J. R. & Moe, R. E. Trends in incidence rates of invasive lobular and ductal breast carcinoma. *JAMA* **289,** 1421–4 (2003).

5. ACS. Cancer Facts and Figures. (2017).

6. Iqbal, J., Ginsburg, O., Rochon, P. A., Sun, P. & Narod, S. A. Differences in Breast Cancer Stage at Diagnosis and Cancer-Specific Survival by Race and Ethnicity in the United States. *JAMA* **313,** 165 (2015).

7. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 10869–74 (2001).

8. Aktas, B. *et al.* Stem cell and epithelial-mesenchymal transition markers are frequently overexpressed in circulating tumor cells of metastatic breast cancer patients. *Breast Cancer Res.* **11,** R46 (2009).

9. Rausch, L. K., Netzer, N. C., Hoegel, J. & Pramsohler, S. The Linkage between Breast Cancer, Hypoxia, and Adipose Tissue. *Front. Oncol.* **7,** 211 (2017).

10. Patsialou, A. *et al.* Invasion of human breast cancer cells in vivo requires both paracrine and autocrine loops involving the colony-stimulating factor-1 receptor. *Cancer Res.* **69,** 9498–506 (2009).

11. Lochter, A. & Bissell, M. J. Involvement of extracellular matrix constituents in breast cancer. *Semin. Cancer Biol.* **6,** 165–173 (1995).

12. Aceto, N. *et al.* Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell* **158,** 1110–1122 (2014).

13. Giancotti, F. G. Mechanisms governing metastatic dormancy and reactivation. *Cell* **155,** 750–64 (2013).

14. Kennecke, H. *et al.* Metastatic behavior of breast cancer subtypes. *J. Clin. Oncol.* **28,** 3271–7 (2010).

15. Minn, A. J. *et al.* Genes that mediate breast cancer metastasis to lung.

16. Kang, Y. *et al.* A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell* **3,** 537–49 (2003).

17. Bos, P. D. *et al.* Genes that mediate breast cancer metastasis to the brain. *Nature* **459,** 1005–9 (2009).

18. Brastianos, P. K. *et al.* Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets. *Cancer Discov.* **5,** 1164–1177 (2015).

19. McDonald, O. G. *et al.* Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis. *Nat. Genet.* **49,** 367–376 (2017).

20. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144,** 646–674 (2011).

21. TCGA. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

22. Matsui, A., Ihara, T., Suda, H., Mikami, H. & Semba, K. Gene amplification: mechanisms and involvement in cancer. *Biomol. Concepts* **4,** 567–582 (2013).

23. Osheim, Y. N. & Miller, O. L. Novel amplification and transcriptional activity of chorion genes in Drosophila melanogaster follicle cells. *Cell* **33,** 543–553 (1983).

24. Brodeur, G. M. *et al.* Cytogenetic features of human neuroblastomas and cell lines. *Cancer Res.* **41,** 4678–86 (1981).

25. Woodcock, D. M. & Cooper, I. A. Evidence for double replication of chromosomal DNA segments as a general consequence of DNA replication inhibition. *Cancer Res.* **41,** 2483–90 (1981).

26. Smith, C. A. & Vinograd, J. Small polydisperse circular DNA of HeLa cells. *J. Mol. Biol.* **69,** 163–178 (1972).

27. Selvarajah, S. *et al.* The breakage–fusion–bridge (BFB) cycle as a mechanism for generating genetic heterogeneity in osteosarcoma. *Chromosoma* **115,** 459–467 (2006).

28. Coquelle, A., Pipiras, E., Toledo, F., Buttin, G. & Debatisse, M. Expression of Fragile Sites Triggers Intrachromosomal Mammalian Gene Amplification and Sets Boundaries to Early Amplicons. *Cell* **89,** 215–225 (1997).

29. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–52 (2012).

30. McCLINTOCK, B. Chromosome organization and genic expression. *Cold Spring Harb. Symp. Quant. Biol.* **16,** 13–47 (1951).

31. Watanabe, T., Tanabe, H. & Horiuchi, T. Gene amplification system based on double rolling-circle replication as a model for oncogene-type amplification. *Nucleic Acids Res.* **39,** e106–e106 (2011).

32. Zhang, F. *et al.* The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat. Genet.* **41,** 849–853 (2009).

33. Howarth, K. *et al.* Chromosome translocations in breast cancer. *Breast Cancer Res.* **10,** P6 (2008).

34. Sun, J. *et al.* Germline Mutations in Cancer Susceptibility Genes in a Large Series of Unselected Breast Cancer Patients. *Clin. Cancer Res.* **23,** 6113–6119 (2017).

35. Martelotto, L. G., Ng, C. K., Piscuoglio, S., Weigelt, B. & Reis-Filho, J. S. Breast cancer intra-tumor heterogeneity.

36. Furr, B. J. A. & Jordan, V. C. The pharmacology and clinical uses of tamoxifen. *Pharmacol. Ther.* **25,** 127–205 (1984).

37. Schneeweiss, A. *et al.* Update Breast Cancer 2018 (Part 2) – Advanced Breast Cancer, Quality of Life and Prevention. *Geburtshilfe Frauenheilkd.* **78,** 246–259 (2018).

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

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

40. Ellis, M. J. *et al.* Ki67 Proliferation Index as a Tool for Chemotherapy Decisions During and After Neoadjuvant Aromatase Inhibitor Treatment of Breast Cancer: Results From the American College of Surgeons Oncology Group Z1031 Trial (Alliance). *J. Clin. Oncol.* **35,** 1061–1069 (2017).

41. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–52 (2000).

42. Wallden, B. *et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med. Genomics* **8,** 54 (2015).

43. Agendia. Agendia. (2017).

44. McVeigh, T. P. *et al.* The impact of Oncotype DX testing on breast cancer management and chemotherapy prescribing patterns in a tertiary referral centre. *Eur. J. Cancer* **50,**

2763–70 (2014).

45. FoundationOne - Foundation Medicine. *2018* Available at: https://www.foundationmedicine.com/genomic-testing/foundation-one. (Accessed: 27th March 2018)

46. Agilent. No Title. (2018).

47. Affymetrix. No Title. (2018).

48. Yauk, C. L., Berndt, M. L., Williams, A. & Douglas, G. R. Comprehensive comparison of six microarray technologies. *Nucleic Acids Res.* **32,** e124 (2004).

49. Gey GO, Coffman WD, K. M. Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. *Cancer Res.* **12,** 264–265 (1952).

50. LASFARGUES, E. Y. & OZZELLO, L. Cultivation of human breast carcinomas. *J. Natl. Cancer Inst.* **21,** 1131–47 (1958).

51. Cowley, G. S. *et al.* Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context- specific genetic dependencies. (2014). doi:10.1038/sdata.2014.35

52. Evers, B. *et al.* CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nat. Biotechnol.* **34,** 631–633 (2016).

53. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170,** 564–576.e16 (2017).

54. Haverty, P. M. *et al.* Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **533,** 333–7 (2016).

55. IMAMURA, Y. *et al.* Comparison of 2D- and 3D-culture models as drug-testing platforms in breast cancer. *Oncol. Rep.* **33,** 1837–1843 (2015).

56. Holliday, D. L. & Speirs, V. Choosing the right cell line for breast cancer research. *Breast Cancer Res.* **13,** 215 (2011).

57. Pompili, L., Porru, M., Caruso, C., Biroccio, A. & Leonetti, C. Patient-derived xenografts: a relevant preclinical model for drug development. *J. Exp. Clin. Cancer Res.* **35,** 189 (2016).

58. Whittle, J. R., Lewis, M. T., Lindeman, G. J. & Visvader, J. E. Patient-derived xenograft models of breast cancer and their predictive power. *Breast Cancer Res.* **17,** 17 (2015).

59. Medina, D. Mammary Tumorigenesis in Chemical Carcinogen-Treated Mice. I. Incidence in BALB/c and C57BL Mice2. *JNCI J. Natl. Cancer Inst.* **53,** 213–221 (1974).

60. Callahan, R. & Smith, G. H. MMTV-induced mammary tumorigenesis: gene discovery,

progression to malignancy and cellular pathways. *Oncogene* **19,** 992–1001 (2000).

61. Bittner, J. J. SOME POSSIBLE EFFECTS OF NURSING ON THE MAMMARY GLAND TUMOR INCIDENCE IN MICE. *Science (80-.).* **84,** 162–162 (1936).

62. M Hilgers, J. S. Mammary Tumors in the Mouse. *Elsevier/North Holl. Bimedical Press* (1981).

63. Guy, C. T. *et al.* Expression of the neu protooncogene in the mammary epithelium of transgenic mice induces metastatic disease. *Proc. Natl. Acad. Sci.* **89,** 10578–10582 (1992).

64. Andrechek, E. R. HER2/Neu tumorigenesis and metastasis is regulated by E2F activator transcription factors. *Oncogene* (2013). doi:10.1038/onc.2013.540

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

66. Hollern, D. P. & Andrechek, E. R. A genomic analysis of mouse models of breast cancer reveals molecular features of mouse models and relationships to human breast cancer. *Breast Cancer Res.* **16,** R59 (2014).

CHAPTER 1

CONSERVED E2F MEDIATED METASTASIS IN MOUSE MODELS OF BREAST CANCER AND HER2 POSITIVE PATIENTS

PREFACE

This chapter while not directly related to the work in the thesis was included due to the fact that it is a case study for the power of integrating informatics and traditional laboratory science. It has been previously been published in Oncoscience as:

Rennhack, J. & Andrechek, E. Conserved E2F mediated metastasis in mouse models of breast cancer and HER2 positive patients. *Oncoscience* **2,** 867 (2015).

ABSTRACT

To improve breast cancer patient outcome work must be done to understand and block tumor metastasis. This study leverages bioinformatics techniques and traditional genetic screens to create a novel method of discovering potential contributors of tumor progression with a focus on tumor metastasis. A database of 1172 of expression data from a variety of mouse models of breast cancer was assembled and queried using previously defined oncogenic activity signatures. This analysis revealed high activity of the E2F family of transcription factors in the MMTV-Neu mouse model. A genetic cross of MMTV-Neu mice into an E2F1 null, E2F2 null, or E2F3 heterozygous background revealed significant changes in tumor progression specifically reductions in tumor latency and metastasis with E2F1 or E2F2 loss. These findings were found to be conserved in human HER2 positive patients. Patients with high E2F1 activity were shown to have worse outcomes such as relapse free survival and distant metastasis free survival. This study shows conserved mechanisms of tumor progression in human breast cancer subtypes and analogous mouse models and underlies the importance of increased research into the characterization of and comparisons between mouse and human tumors to identify which mouse models resemble each subtype of human breast cancer.

MAIN TEXT

BREAST CANCER AS A HETEROGENEOUS DISEASE

Breast cancer is an extremely common and deadly disease. With over 200,000 new cases and 40,000 deaths in the United States annually contributed to the cancer, it is the second leading cause of cancer deaths in women. The main cause of these deaths is the ability of the tumor to metastasize to the lungs, liver, bone, and brain[1]. This is reflected in the survival rates of patients diagnosed with or without tumor metastasis. The five year survival rate of a patient without tumor metastasis is over 90% in contrast to a patient with tumor metastasis who only has approximately a 20% five year survival rate[2]. In order to improve patient outcomes, significant research effort must be placed on treating and preventing tumor metastasis.

A defining characteristic of breast cancer is heterogeneity. Tumors from different patients will have a wide variety of tumor growth rates, response to treatment, and metastatic potential. In order to understand the mechanism behind the diversity of characteristics from one tumor to another many multi "-omic" studies such as TCGA and Metabric have begun to profile tumors from a molecular standpoint[3,4]. Gene expression data has classified tumors into six main subgroups: Luminal A, Luminal B, Basal, Claudin Low, Normal, and HER2 positive[5]. Each subtype has key driving events such as basal breast cancer being largely associated with p53 mutations or Myc amplification, while HER2+ breast cancer is characterized by the amplification/overexpression of the HER2 protein.

The HER2 subtype has been of special interest due to its clinical relevance. Approximately 25% of breast cancer patients have a HER2 amplification event[6,7]. This causes the upregulation of HER2, a growth factor receptor, on the cell surface leading to uncontrolled cell growth and increased metastatic capability. Despite the aggressive nature of the subtype, there has been success in developing treatment targeted against the HER2 protein. However, these treatments, such as Herceptin[8] and Lapatinib[9],are not effective in all HER2 positive patients. This indicates that there is heterogeneity in the subgroups as well as redundant oncogenic signaling allowing for survival of the cancer cell without the HER2 signaling cascade.

To better understand and predict the activation of key signaling pathways, oncogenic activation signatures were created. These signatures, developed through Bayesian regression analysis and induced expression of a specific oncogenic driver[10–12], have shown key signaling pathways involved in each molecular subtype. As expected, the basal subgroup has low activation of ER and PR while HER2 positive subtypes have HER2 activation. However it is also seen that subsets of each tumor subtype have a specific oncogenic signaling pattern including a subset of Luminal A tumors with high Src activity. The high Src signaling indicates that a subgroup of Luminal A tumors is dependent upon the Src signaling pathway.

MOUSE MODELS OF BREAST CANCER

Mouse models have been created to mimic specific oncogenic drivers, such as Src, in hopes to mirror different types of breast cancer to better understand tumor progression that is dependent on a specific signaling pathway. Induction of breast cancer in a mouse model can be accomplished in a number of different manners. These methods include leveraging tissue specific promoters such as MMTV or WAP to drive expression of an oncogene such as Neu[13], or the use of a tissue specific Cre[14] or inducible drug system to create conditional knockouts of tumor suppressors. Other models use a carcinogen induced model such as DMBA treatment.

Models have also been created to investigate specific aspects of breast cancer progression including genomic instability through the loss of key checkpoint or repair proteins like p53[15] or BRCA[16] or tumor metastasis through induction of PyMT[17]. Given the variety of methods to induce tumors as well activation of unique tumor driving pathways, the transcriptional program in each model would be expected to be unique.

GENE EXPRESSION PROFILING OF MOUSE MODELS OF BREAST CANCER

To profile this diversity, a database consisting of 1172 tumors from a variety of mouse models was generated[18]. As expected, there was a significant amount of diversity between samples from different models and also within each model (Figure 1.1A). Despite these differences it was found through unsupervised hierarchical clustering that mouse models of breast cancer clustered into four distinct clusters. These clusters contain transcriptional profiles which regulate different tumor characteristics and are associated with histological patterns such as epithelial to mesenchymal transition (EMT). As expected each of the clusters also had unique oncogenic pathway activation.

Oncogene activation signatures were calculated for each sample in the manner described above, and hierarchical clustering was performed. It was seen that within models sets of tumors had the same signature profile. A key example being the Myc induced tumor models. Tumors derived from these models were extremely heterogeneous[19] and subsets of tumors contained the same oncogenic signaling pattern as tumors from each of the human subclasses of breast cancer[18,20].

HIGH E2F ACTIVITY IN MMTV-NEU MOUSE MODEL

Surprisingly it was noted that the activator subclass of the E2F family of transcription family was seen to be highly active in MMTV-Neu tumor samples (Figure 1.1A) [21]. The E2F family, classically known to regulate cell cycle[22,23], has recently been shown to regulate a number of tumor characteristics beyond proliferation such as DNA repair, angiogenesis, and immune-evasion[24,25]. When oncogenic signatures were applied to a group of human breast cancer patients it was seen that a subset of HER2+ patients with unique E2F signaling had worse outcomes, including relapse free survival[21]. This indicates that the E2F family of transcription factors play an important role in HER2 positive tumor progression.

LOSS OF E2FS IMPACT TUMOR PROGRESSION MMTV-NEU MOUSE MODEL

To test the hypothesis that the E2Fs are critical in HER2 tumor progression, MMTV-Neu tumors were crossed into an E2F1 null, E2F2 null, and E2F3 heterozygous background (Figure1.1B)[21]. The E2Fs have been shown to be redundant in their binding sites and function, so as expected there was compensation by other E2F family members with the loss of individual E2Fs[26]. Despite the apparent compensation of the E2F knockouts, significant differences were identified in tumor progression between the E2F wildtype and E2F null background indicating specificity in the functions of each E2F family member in regards to tumor progression. There was a significant delay in tumor latency associate with E2F1, E2F2 and E2F3 loss. Furthermore, there was a reduction in tumor burden showing a decrease from an average of 2.5 tumors per mouse in wildtype E2Fs to 1.5 tumors per mouse in the E2F1 null background. The growth rate of the tumors was not affected with E2F2 and E2F3; however, there was a significant increase in the growth rate of E2F1 null tumors. This is likely due to the role of E2F1 in tumor apoptosis.

Striking differences were seen in tumor metastasis. There was a significant reduction in the number of mice with metastasis with the loss of specific E2Fs[21]. In a wildtype MMTV-Neu background it was seen that 73% of mice with tumors develop metastasis to the lungs (Figure 1.1B). This number is reduced to 40% and 35% with the loss of E2F1 and E2F2 respectively (Figure 1.1B). It was also seen that the E2Fs affect both early and late stages on metastasis in a cell independent manner. A colony formation assay from circulating tumor cells showed a reduction in the amount of colonies formed in the E2F2 null background indicating a block in the early stages of tumor metastasis. However, the E2F1 null tumors did not show a significant reduction in the amount of colonies formed indicating a block in the late stages of metastasis. The metastasis effects were seen to be background independent with E2F1 null tumors still being non-metastatic when transplanted into a wildtype host.

CONSERVATION OF THE E2FS ROLE IN METASTASIS OF HUMAN BREAST CANCER

A dataset of gene expression data from human HER2 breast cancer patients was assembled and E2F activity was assessed. It was shown that patients with high E2F1 activity compared to those with relatively lower E2F1 activity had worse metastasis free survival[21]. Furthermore patients were separated on the basis of low and high E2F1 activity regardless or subtype[27], and it was shown that patients with high E2F1 levels had worse distant metastasis free survival (Figure 1.1C).

FUTURE DIRECTIONS

With the establishment of the role of the E2Fs in tumor metastasis, the next goal is to leverage them as a therapeutic target to block metastasis and reduce the mortality associated with breast cancer. It is not predicted that the E2Fs themselves will be good targets for therapy due to their involvement in a myriad of normal cell processes. However, one might predict that there are specific downstream targets of E2F1 or E2F2 that mediate discreet steps in the development of tumor metastasis. As these genes are identified and characterized they may provide opportunities for development as therapeutic targets.

The description of the role of E2Fs in Neu mediated tumors is an example of how an integrative approach can be used to uncover genes that regulate metastasis. As such, this study demonstrates the need for increased basic research into mouse models. In this study we have taken a bioinformatics prediction in a mouse model about the essential nature of the E2Fs in a model, MMTV-Neu. This was investigated and validated through traditional genetic studies, and the role of E2F1 and E2F2 was shown in tumor metastasis. The finding was consistent in HER2 positive patients leading to a potential new therapeutic avenue to block tumor metastasis. To continue studies of this kind, more work must be completed to understand mouse models from a molecular standpoint and to understand which mouse models represent which classes of human tumors. Leveraging advances in bioinformatics and applying them to mouse models of breast cancer therefore presents a unique opportunity to develop and test hypotheses for how metastatic breast cancer progresses.

APPENDIX

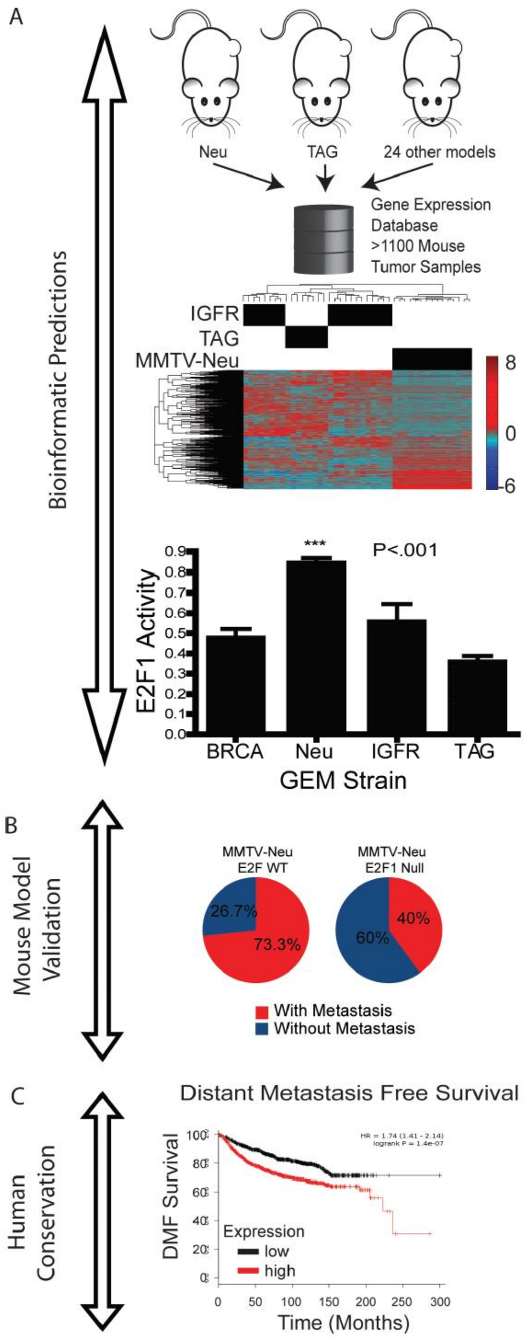

Figure 1.1: An integration of traditional genetics and bioinformatics to understand the role of E2F1 in breast cancer

Identification and validation of conserved mechanism of tumor metastasis in mouse models and human breast cancer patients

WORKS CITED

WORKS CITED

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

2. Kohler, B. A. *et al.* Annual Report to the Nation on the Status of Cancer, 1975-2011, Featuring Incidence of Breast Cancer Subtypes by Race/Ethnicity, Poverty, and State. *J. Natl. Cancer Inst.* **107,** djv048 (2015).

3. TCGA. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

4. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–52 (2012).

5. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–52 (2000).

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

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

8. De Laurentiis, M. *et al.* Targeting HER2 as a therapeutic strategy for breast cancer: a paradigmatic shift of drug development in oncology. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **16 Suppl 4,** iv7-13 (2005).

9. Geyer, C. E. *et al.* Lapatinib plus capecitabine for HER2-positive advanced breast cancer. *N. Engl. J. Med.* **355,** 2733–43 (2006).

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

11. Bild, A. H. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439,** 353–7 (2006).

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

13. Muller, W. J., Sinn, E., Pattengale, P. K., Wallace, R. & Leder, P. Single-step induction of mammary adenocarcinoma in transgenic mice bearing the activated c-neu oncogene. *Cell* **54,** 105–15 (1988).

14. Andrechek, E. R. *et al.* Amplification of the neu/erbB-2 oncogene in a mouse model of mammary tumorigenesis. *Proc. Natl. Acad. Sci.* **97,** 3444–3449 (2000).

15. Backlund, M. G. *et al.* Impact of ionizing radiation and genetic background on mammary tumorigenesis in p53-deficient mice. *Cancer Res.* **61,** 6577–82 (2001).

16. Xu, X. *et al.* Conditional mutation of Brca1 in mammary epithelial cells resultsin blunted ductal morphogenesis and tumour formation. *Nat. Genet.* **22,** 37–43 (1999).

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

18. Hollern, D. P. & Andrechek, E. R. A genomic analysis of mouse models of breast cancer reveals molecular features of mouse models and relationships to human breast cancer. *Breast Cancer Res.* **16,** R59 (2014).

19. 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.* **106,** 16387–92 (2009).

20. Hollern, D. P., Yuwanita, I. & Andrechek, E. R. A mouse model with T58A mutations in Myc reduces the dependence on KRas mutations and has similarities to claudin-low human breast cancer. *Oncogene* **32,** 1296–304 (2013).

21. Andrechek, E. R. HER2/Neu tumorigenesis and metastasis is regulated by E2F activator transcription factors. *Oncogene* (2013). doi:10.1038/onc.2013.540

22. Nevins, J. R. The Rb/E2F pathway and cancer. *Hum. Mol. Genet.* **10,** 699–703 (2001).

23. Trimarchi, J. M. & Lees, J. A. Sibling rivalry in the E2F family. *Nat. Rev. Mol. Cell Biol.* **3,** 11–20 (2002).

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

25. Chen, H.-Z., Tsai, S.-Y. & Leone, G. Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat. Rev. Cancer* **9,** 785–797 (2009).

26. Kong, L.-J., Chang, J. T., Bild, A. H. & Nevins, J. R. Compensation and specificity of function within the E2F family. *Oncogene* **26,** 321–327 (2007).

27. Györffy, 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.* **123,** 725–31 (2010).

CHAPTER 2

MOUSE MODELS OF BREAST CANCER SHARE AMPLIFICATION AND DELETION EVENTS WITH HUMAN BREAST CANCER

PREFACE

This chapter has previously been published as:

Rennhack, J., To, B., Wermuth, H., & Andrechek, E. R. (2017). Mouse Models of Breast Cancer Share Amplification and Deletion Events with Human Breast Cancer. *Journal of Mammary Gland Biology and Neoplasia*, *22*(1), 71–84.

ABSTRACT

Breast tumor heterogeneity has been well documented through the use of multiplatform –omic studies in human tumors. However, there is no integrative database to capture the heterogeneity within mouse models of breast cancer. This project identifies genomic copy number alterations (CNAs) in 600 tumors across 27 major mouse models of breast cancer through the application of a predictive algorithm to publicly available gene expression data. It was found that despite the presence of strong oncogenic drivers in most mouse models, CNAs are extremely common but heterogeneous both between models and within models. Many mouse CNA events are largely conserved in human tumors and in the mouse we show that they are associated with secondary tumor characteristics such as tumor histology, metastasis, as well as enhanced oncogenic signaling. These data serve as an important resource in guiding investigators when choosing a mouse model to understand the gene copy number changes relevant to human breast cancer.

INTRODUCTION

Genomic instability, including point mutations, translocation, and gene copy number alteration in key oncogenic signaling genes, is an underlying driver of breast cancer development and progression. Gene copy number changes, containing amplifications such as *HER2*[1] and *MYC*[2] or deletion events such as *PTEN*[3], are key biomarkers of tumor onset[4–7], histology[8], metastatic potential[9], and treatment response[10,11]. The amplification of *HER2* in 20-30% of breast cancer patients results in significantly more aggressive tumors and is an important prognostic marker in a patient's ability to respond to anti-HER2 therapy such as trastuzumab. Despite the success of HER2 therapy, the majority of patients with the amplification event will have primary or acquired resistance to trastuzumab[11], indicating potential heterogeneity in these tumors.

To investigate tumor heterogeneity, large multiplatform studies such as The Cancer Genome Atlas (TCGA)[12] and Metabric[13] projects have begun to integrate transcriptional data and genomic data, as well other data platforms. Traditionally, breast cancer diversity has been classified at the transcriptional level into six basic subtypes: basal, luminal A, luminal B, HER2 positive, claudin-low, and normal like[14,15]. The integration of gene copy number showed that there are a number of classical gene copy number alterations that are associated with each tumor subtype. For instance, this revealed that *MYC* amplification occurs in all breast cancer subtypes, but *MYC* is only transcriptionally active in the basal subtype[12]. This study underscores the importance of integration of multiple platforms to understand tumor heterogeneity.

In order to understand the function of oncogenic drivers, research has employed mouse models. There are a variety of methods to induce breast cancer in a mouse model. These range from tissue specific overexpression using promoters including mouse mammary tumor virus (MMTV) or Whey acidic protein (WAP) promoters to drive oncogene expression[16–19], to conditional knockouts of tumor suppressive genes through the use of a tissue specific Cre[20–22] or inducible system[23,24] and carcinogen induced model such as DMBA treatment. Furthermore, models have been used to investigate particular aspects of tumor development such as metastasis via MMTV-PyMT[25],[26] or genomic instability with loss of p53[27]. Given these varied methods and drivers of tumor formation, the transcriptional program in each model would be expected to be unique. Importantly the recent advent of patient derived xenografts (PDX) models has given a new option. These models have been shown to reflect their human counterpart at a genomic[28] and transcriptional level; however, other common mouse models of breast cancer have not been described in such a manner.

Recent work has captured the gene expression diversity between models and within tumors of the same model[29–31]. However, these works do not describe multiple levels of genomic diversity in mouse models like the TCGA and Metabric projects do in human tumors. Largely this is due to a lack of multi-platform omic studies present for the mouse models. Small scale studies that integrate CNA with expression data across species have identified a unique CNA in Basal like breast cancer[32]. The lack of such profiling on a large scale leaves researchers relatively uninformed about genomic changes present in a mouse tumor when choosing a mouse model which is representative of a specific subtype of human breast cancer.

Here we describe a large scale investigation of copy number changes in 600 tumors across 27 mouse models of breast cancer for the use of the ACE algorithm[33]. In short, the ACE algorithm predicts CNA from gene expression data through the use of a weighted mean of gene expression across a given genomic region. Due to its reliance on consistent regulation across an entire genomic region it has been shown to accurately predict copy number variants and has been shown to have consistent results to traditional genetic predictors of gene copy number tumors[33]. This predicted CNA across our dataset demonstrated wide heterogeneity across mouse models of breast cancer. Interestingly, consistent CNA changes were noted in the microacinar histological subtypes of breast cancer indicating a role in copy number changes and a tumor's histological phenotype. Moreover, in an important observation we noted that CNA was associated with breast cancer metastasis and enhanced oncogenic signaling in both mouse models and human breast cancer.

RESULTS

IDENTIFICATION OF GENE COPY NUMBER ALTERATION IN MOUSE MODELS OF BREAST CANCER

A large number of mouse model tumors have been examined by microarray for gene expression but few have been assessed for genome wide copy number alteration. To computationally predict CNA from gene expression data, we applied the ACE algorithm[33]. The ACE algorithm is shown to be consistent with traditional means of determining CNA utilized by the TCGA breast cancer study. To validate the algorithm, we used the entire TCGA breast cancer dataset in which both gene expression and copy number data is available. A random selection of three samples from the dataset show a high degree of similarity between the ACE calls and the TCGA CNA calls (Figure 2.1A)[12,34]. Application of the algorithm across the entire dataset shows a false positive rate of 25.4% and 24.3% for amplifications and deletions respectively for the ACE algorithm (Figure 2.1B). However, the false negative rate was higher at greater than 90% (Figure 2.1B). This includes both amplification of whole arms of the chromosome and very small amplification events. For the false negatives, the algorithm also shows a dependence upon gene expression being coordinated with CNA. Importantly, in the case of amplification events with a high transcriptional impact, noted by a z-score of greater than 4, we found a success rate of 31.1% of TCGA events called by the ACE algorithm.

We have also shown the translational effects of copy number gains and losses in human tumors through investigation of Reverse Phase Protein Array (RPPA) data associated with *EGFR* (Figure 2.3A) and *FOXO3* (Figure 2.3B) amplification events. This analysis shows that in both EGFR and FOXO3 the protein level is directly correlated with gene copy number. This validation of the ACE software across the TCGA dataset shows that the events predicted in this manuscript are an understatement of the events in the tumor. However, the false low false positive rate shows that the events called in the manuscript can be used for predictive purposes and begin to show the copy number profile in mouse models of breast cancer.

To investigate the presence of copy number alterations in mouse models we applied the ACE algorithm to gene expression data from a normal mammary gland from an FVB Wildtype mouse (Figure 2.1C) and an MMTV-Neu derived tumor from the same background (Figure 2.1D). As expected, no CNA was identified in the control mammary gland. In contrast, the MMTV-Neu sample is characterized by a large amplification event on chromosome 3 as well as a large deletion on chromosome 4. This deletion is consistent with previously published findings of chromosome 4 loss in MMTV-Neu mouse models[35]. In addition to these major CNA events, there are a number of smaller CNAs throughout of the genome of the sample.

We then hypothesized as a further check that unstable models would have significantly more CNAs than oncogene induced models. To investigate this hypothesis, we tested for CNA in multiple samples from unique mouse models of breast cancer. The ACE algorithm was applied to tumor samples from MMTV-Myc, MMTV-PyMT, MMTV-Neu, TAG, and DMBA treated models derived from the FVB/NJ mouse background (Figure 2.1E). ACE analysis showed genomic stability in the MMTV-Neu driven mouse models. This model had significantly fewer amplification or deletion events than more classically unstable models such as the TAG (p<.05) and DMBA (p<.05). This is mirrored in human cancer where certain tumors such as Basal tumors are shown to be more unstable than other subtypes especially Luminal A[36]. It was also noted that mice with the same oncogenic initiation event such as PyMT had a difference in copy number based upon the background of the mouse model (Figure 2.4). We noted that the FVB background was the most unstable when compared to the AKXD background in PyMT driven tumors and the Balb/C background in the TAG or various p53 driven models.

CONSERVATION OF CNA VARIABILITY IN MOUSE MODELS

To investigate the extent of CNA in mouse models and to determine if there were common alterations across various oncogenic driver genes, expression data was downloaded from a previously assembled mouse model database[29]. Specifically, for copy number variability we used 600 tumor samples from 27 mouse models that had been analyzed through the use of Affymetrix microarrays. ACE analysis was run and the percent of samples, regardless of model, amplified or deleted with CNA at a specific locus across the genome was calculated (Figure 2.2A).

This data shows the vast majority of genes across the genome are amplified or deleted in less than 5% of the total samples with a few distinct regions being amplified or deleted in a larger fraction of samples. However, we did identify regions of instability that were conserved across models. We identified a number of genes that were both amplified and deleted in greater than 10% of mouse models. Specifically, we identified *Gsn* is amplified 10.9% of samples and deleted in 11.4% of samples. Other genes that were amplified and deleted at a high level included *Cct4*, *Hnrnpab*, *Cp*, *Cklf*, *Cenpo*, and *Dnm2*. These genes are all located at regions previously described in the mouse genome to be unstable[37].

Given the extensive heterogeneity in breast cancer, we sought to test the hypothesis that heterogeneity was present at the level of gene copy number within individual mouse models of cancer. The extent of heterogeneity of CNAs within a tumor model was analyzed by examining the fraction of mice within a model with a given amplification or deletion event at a particular locus (Figure 2.2B). This analysis revealed a large degree of heterogeneity within models with most models having the majority of loci amplified or deleted in less than 50% of the sample within a model. This is despite that fact that many tumor samples are driven by the same oncogenic driver and are biological replicates. Some of the genes that were amplified in greater than 50% of samples within a given tumor model represent key genes in tumor development, progression and metastasis including well known genes such as *Cdkn2*, *Mmp23*, *Sumo2*, and *Adcy33*. Interestingly, conserved CNA events were not seen to span models, reinforcing the genomic diversity both within each model system and between the model systems. In addition, we noted some models with more copy number events, such as the p53 induced models.

These CNA changes were then divided into amplification events (Figure 2.5A) or deletion events (Figure 2.5B) to reflect the copy number diversity in each model. This revealed that mouse tumor models largely fall into three categories. First, we observed unstable models with a high degree of amplification or deletion in a large number of genes but with low levels of conservation, this including many models with a *p53* mutation. Secondly, we noted models that are relatively stable with no amplification or deletion at the vast majority of genes, including models such as MMTV-PyMT. Lastly there are models with a few highly conserved amplification or deletion events. These conserved events were noted in more than 25% of samples in lines such as the *Erbb2* knock-in model or the WAP TAG model.

CONSERVED ROLE OF CNAS IN HUMAN AND MOUSE TUMORS

A key feature of the mouse tumors is their ability to model human cancer. To test the hypothesis that there were conserved CNAs in both species, we began by identifying the fraction of tumors within each model with an amplification or a deletion at genes prone to copy number alteration in human breast cancer such as *ERBB2*, *MYC*, *PTEN*, *RB,* and others identified in the TCGA study. Driver genes were found to be amplified or deleted in specific mouse models. Specifically, this occurred in the BRCA/p53 modified models, which have a fraction of samples with amplification in common oncogenes such as *CCNE* or deletion of common tumor suppressors like *CMTM3*.

To identify genes that were amplified or deleted at a high level in both mouse and human that are not traditional drivers of tumorigenesis, we used unsupervised hierarchical clustering of CNAs in human tumors of various subtypes and mouse models of breast cancer (Figure 2.6A). As expected, human tumors and mouse models showed diverse copy number profiles and in general showed similarity in copy number profiles. This was illustrated through co-clustering in an unsupervised hierarchical clustering of gene loci amplified or deleted in more than 5% of mouse and human tumors (Figure 2.7A). Through this analysis we identified three tightly clustered groups of human and mouse samples. The sub-cluster indicated by the purple portion of the dendrogram was largely dominated by human Luminal A and normal like tumors while the yellow and green portions of the dendrogram had clusters characterized by the presence of Luminal B tumors. Largely absent in this analysis were the HER2 positive tumors. This finding indicates the lack of mouse tumor models with the HER2 amplification event or associated copy number changes. Of interest we noted that a fraction of MMTV-PyMT samples were present in each of the clusters, indicating considerable diversity within the MMTV-PyMT tumor model from a gene copy number perspective. To show that within these clusters there is similarity between mouse and human samples, we showed a significant increase in the Jaccard index (a similarity metric of mouse to human samples) within the purple cluster, and a low Jaccard index between the mouse samples of the purple cluster and the human samples of the other two defined clusters (Figure 2.7B). This shows that each cluster of tumors has a significantly different copy number profile.

While we noted model to model heterogeneity, our previous analysis revealed within model heterogeneity. It was hypothesized that the heterogeneity was due to differences in histological subtypes. To test this, unsupervised hierarchical clustering of CNA events from the MMTV-Myc tumor model by the top 4118 commonly amplified or deleted genes, filtered by standard deviation of neighborhood score, was performed (Figure 2.8A). The major clusters demonstrated that distinct copy number changes are associated specific histological subtypes within the *Myc* driven model. Specifically, there is a cluster enriched for the microacinary subtype. The mircoacinar subtype shows the majority of samples containing amplification of many genes located on chromosomes 11 and 15. The EMT subtype is characterized by having few amplification or deletion events. Tumors of this histological subtype have previously been noted to have activating mutations in *Kras* that contribute to the EMT histological subtype. These activating mutations may result in a reduced requirement for other copy number events.

The genes amplified in mice with a microacinar histological subtype, located on mouse chromosome 11, are also conserved in humans. Specifically, we found a region of fourteen genes which mapped to chromosome 17q25.1 in humans. These genes are amplified shown to be amplified in a subset of breast tumors as identified by cBio portal. An assessment of mouse (Figure 2.8B), and human (Figure 2.8C) tumors, from the MMTV-Myc mouse model and TCGA breast cancer dataset respectively, showed conserved histology across species.

Human tumor samples were divided into those containing the microacinar specific genes through the use of a 14 gene signature (Figure 2.8D). Tumors with at least two of the 14 genes amplified were considered to be in the amplified subgroup. This produced a subgroup for 28 human tumors and was compared against 30 randomly chosen tumors that contained none of the 14 amplified microacinar associated genes. Histology for each of the groups of samples was determined and it was found that the microacinar associated gene amplified subgroup contained an enrichment of tumors with a microacinar histology (Figure 2.8E). This indicated a conserved role in gene copy number across mouse and human breast cancer in determining tumor histology specifically with respect to the microacinar subtype.

To investigate the role of the CNA on tumor progression, we examined tumor metastasis by integrating gene expression data with gene copy number data. Specifically we used a previously identified lung metastasis gene signature[38] and correlated gene copy number events with the sample's metastasis score through Spearman's Rank correlation (Figure 2.9A) to reveal amplification and deletion events associated with highly metastatic samples. This was applied to the human TCGA breast dataset (Figure 2.9B – top) as well as the mouse model dataset (Figure 2.9B – bottom). When differences in gene location were taken into account, there were 132 gene copy number alterations highly correlated with metastasis that were conserved in the TCGA breast cancer and mouse datasets (Figure 2.9B). To provide validation of the 132 genes and their association with tumor metastasis we leveraged the KM-plot human dataset to show that over express or decreased expression of amplified or deleted genes was associated with worse distant metastasis free survival. We showed that 55% of these regions had significantly increased or decreased metastasis free survival depending upon the transcript level of the gene. To further investigate the metastasis related genes we used the Metabric data with associated overall survival data. Of the samples with an identified amplification or deletion event in a predicted metastasis region, we showed 28% of the events resulted in a decrease in overall survival of the patients. A closer examination of mouse chromosome 3 (Figure 2.9C) revealed a number of genes in the 3F region where amplification is associated with a high metastasis score. It was seen that some regions, such as the 3F region, associated with the metastasis signature. It was seen that chromosome 3F amplification was also associated with high RAS activity revealing a potential mechanism of metastasis (Figure 2.9C).

To test the role of chromosome 3F on metastasis, we examined mouse tumor samples where metastasis data and pathway activity predictions were available. In particular, we used the MMTV-Myc, MMTV-Neu, and MMTV-PyMT models. As predicted those samples with the 3F amplification had much higher predicted *Ras* activity and number of lung metastases than those with a deletion. (Figure 2.9D). The 3F amplification event is also conserved in human Luminal A tumors. When tumors were split on the basis of amplification of the analogous human region it was seen that they exhibited higher *Ras* pathway activity (Figure 2.9E) and had worse metastasis free survival (Figure 2.9F).

Given that CNA was associated with metastatic progression we then hypothesized that CNA would also impact key cell signaling pathways. To test this hypothesis we examined the role of CNAs on major oncogenic pathways including *BCAT*, *SRC*, *E2F*, and others (38, 39). This experiment then used the same workflow to coordinate CNA and pathway activation status as was used to coordinate amplification and deletion events with the tumor metastasis signature (Figure 2.10). This analysis revealed amplification and deletion regions associated with each major oncogenic signature. The regulation of signaling pathways can occur through amplifying key genes within the signaling pathway. An example of this was observed when specific amplified genes associated with high *AKT* activity located on chromosome 4 or the specific amplified genes associated located on chromosome 14 associated with high *E2F2* activity were tested. When these genes are displayed in an interaction network, the vast majority of the genes can be found to be located either up or downstream of their respective key signaling protein such as or RB/E2F2 (Figure 2.10B). This suggests the chromosome 14 region is associated with Rb/E2F signaling.

DISCUSSION

Here we have described the copy number alteration across the genome of 27 mouse models of breast cancer. This has been completed through the use of an algorithm (ACE) to infer gene copy number profile from gene expression data. The ACE algorithm was identified to have a high rate of false negatives and a relatively low rate of false positive calls. When the algorithm was run across the TCGA dataset, it was seen to have a moderate rate of concurrence with the TCGA copy number calls. Due to this, it is important to note the predictive nature of this database. While the copy number calls found in this dataset have not been validated using traditional means, the dataset begins to identify potential copy number variants in the mouse models of breast cancer. This is an important step in understanding tumorigenesis in these models specifically from a copy number point of view until a more robust and accurate profiling of the tumors can be completed.

The copy number profiles have been examined in a number of ways to classify inter and intra model heterogeneity as well as the similarities between copy number profiles in mouse models and human breast cancer. This study makes important contributions in understanding CNA in mouse models. Beyond this the CNAs are profiled for their contribution to tumor progression and the conservation of this role in human tumors.

Despite the presence of strong oncogenic signals, gene copy number alterations are still extremely common in mouse models of breast cancer. Common human drivers of breast cancer such as *HER2*, *MYC* or *PTEN* were not observed to be amplified or deleted at a high level across mouse models. This is unsurprising due to the lack of selective pressure for CNAs in these oncogenes or tumor suppressors because of the presence of a strong oncogenic signal.

The exceptions to this are the p53/BRCA induced models which do not have a strong oncogenic signal but instead induce genomic instability. Amplification or deletion events in common human oncogenes are more frequent in these models.

A key finding of this manuscript is the heterogeneity of copy number alterations both within a model and between models. The within model heterogeneity is surprising due to the fact that each tumor is a biological replicate with the same driving oncogenic event. We identified that most events that occur within a given tumor model are not shared among even 50% of tumors from that same model. This finding underscores the importance of gene expression and genomic characterization of tumor studies when dealing with mouse models due to the inherent genomic variability. It further emphasizes the need for a large enough cohort to capture the heterogeneity of all tumor models.

During preparation of this manuscript, a complementary study was published examining CNA in mouse models of breast cancer[39]. This publication uses a different algorithm to predict CNAs, one that predicts resolution on the whole chromosome scale while the ACE algorithm provides finer resolution. Due to the predictive nature of defining gene amplification events from gene expression data, we believe that it is important to compare their manuscript with the data herein. This demonstrates that multiple algorithms call the same dataset with overlapping findings, resulting in a comprehensive view of CNA in mouse model tumors. The two manuscripts agree on a number of findings including the stability of mouse models with rapid latency, the p53 KO model being the most unstable, within model heterogeneity, and the association of CNA changes with the microacinar subtype. The increased precision of the ACE method has allowed us to identify small focal events in many of the models including PyMT that the published paper did not uncover. Indeed, the data we present here allows one to search for a mouse model with amplification or deletion of particular genes. We have also leveraged human data through the use of the TCGA, Metabric, and KMplotter datasets to provide a comprehensive comparison of mouse models and the five main subtypes of human breast cancer tumors. In addition, we have also shown the conservation of regions between the two species to predict a number of new metastasis related copy number changes.

We noted that the amplification or deletion events are associated with secondary tumor characteristics such as tumor histology, enhanced oncogenic signaling, and tumor metastasis. Specifically, we observed unique copy number profiles for the microacinar tumor histology including the amplification of fourteen genes on chromosome 17q25.1. However other histological subtypes did not have characteristic copy number profiles. Surprisingly, we noted that EMT tumors were stable in regards to copy number change, likely due to activation of *Kras* in MYC tumors[23,40,41]. The lack of pattern of amplification or deletion of other histological subtypes indicates that there are other factors such as point mutations or transcriptional changes associate with these subtypes. This can also be said for oncogenic signaling pathways and tumor metastasis. While CNAs contribute to each of these, there are also contributions of single nucleotide variants (SNVs) and transcriptional changes. For this reason, it is important to integrate multiple platforms to understand tumor heterogeneity.

There is conservation between mouse and human subtypes in regards to tumor metastasis and oncogenic signaling. 132 genes that were amplified or deleted in mouse and human contributed to increased metastasis. Furthermore, these genes were located in the same oncogenic signaling pathways indicating conserved mechanisms of metastasis in human and mouse.

When comparing heterogeneity of breast cancer, we found a large degree of heterogeneity both between models and within specific models. Given these findings, it is therefore critical to understand copy number profile when choosing a strain to model human breast cancer. For example, if one is interested in the *HER2* oncogene there are a number of mouse models including the MMTV-Neu[16,25], *Erbb2* Knock-in[20], NDL[42] and others with conditional activation[24]. Each of these models has completely different CNA and transcriptional profiles leading to different oncogenic signaling and subsequently different tumor properties.

This heterogeneity also exists in other common models such as MMTV-Myc. This strain has previously been identified to be heterogeneous from a transcriptional viewpoint[40] and therefore it is unsurprising that it is also heterogeneous from a copy number standpoint. Due to the heterogeneity present at a gene expression and copy number level, investigators must take care when choosing tumor models of breast cancer to ensure that the chosen model reflects all aspects of the human breast cancer subtype they wish to model. We have also noted strain specific differences for some of the models. It was seen that the FVB model was found to be more unstable when compared to tumors derived from other backgrounds. This finding emphasizes the importance of researchers understanding the background of their mouse strain when choosing mouse models for their study.

Projects such as TCGA have profiled human tumors at multiple levels. This allows researchers to stratify human tumors by gene expression, copy number profile, as well as SNVs and epigenetic markings to find a tumor population that is relevant for their study. However, there is not a mouse model equivalent to this dataset, so researchers are unable to choose mouse models which represent their specific tumor subtype at multiple levels. Recent studies such as this and others have begun to make strides in this area by profiling tumors at a CNA and expression levels. However, there is still a need to continue to profile mouse models through the use of whole genome sequencing as well as epigenetic markings. This information needs to be available to researchers in order to design studies that accurately represent the human subtypes of breast cancer.

This study clearly illustrates the importance of gene copy number alterations in tumor progression even in the presence of strong oncogenic drivers. Many mouse models contain a high degree of gene copy number alterations. These copy number alterations are highly heterogeneous both between models and within a model of breast cancer. Despite this heterogeneity, it was seen that the CNAs found in mice are conserved in humans. Conserved variants were associated with tumor progression and potentially play a role in enhanced oncogenic signaling, histological appearance, and the tumor's metastatic potential in both human and mouse tumors.

Beyond the profiling of mouse tumors and the conserved roles of CNAs in mouse and human tumors this study has a broader impact on the field of cancer research. It, when used in combination with gene expression studies, begins to create a comprehensive molecular portrait of tumors derived from mouse models of breast cancer. These studies could be significantly enhanced if outcome, pathology, metastasis and other clinical data was included when publishing tumor data from mouse models. However, this current study provides an essential resource to researchers to contemplate as they choose a model system to mimic a specific subtype of human breast cancer.

MATERIALS AND METHODS

DATASET AND ACE ANALYSIS

A comprehensive mouse dataset was downloaded and assembled as previous described[29] including GSE15263, GSE3165, GSE37954, GSE32152, GSE10450, GSE22406, GSE42533, GSE15904, GSE8836, GSE27101, GSE30864, GSE20416, GSE10193, GSE23938, GSE15119, GSE16110, GSE25488, GSE21444, GSE8828, E-TABM-684. ACE analysis was run as previously described[33] comparing each individual sample to a wildtype mouse of the same strain (FVB/NJ = GSE25488, Balb/C = GSE21444, C57BL/6 = GSE14753) with a significance threshold p and q value of .05 with any size of the event.

Z-SCORE CALCULATION

The microarray based expression data was downloaded from the TCGA dataset. Z score was calculated for each gene in each sample and each event was classified based on the Z-score for validation analysis.

HUMAN DATASET AND ANALYSIS

The TCGA breast cancer and KMplot.com datasets were used for human copy number analysis[12] and validation of results. Specific tumor breast cancer subtype and copy number calls were used from the TCGA dataset[12,34]. To run ACE the gene symbols were replaced with their Affymetrix U133A_2 probe ID. This was queried using the cbio portal visualization tool. Distant metastasis free survival results were obtained using the KMplot.com dataset[43]. ACE analysis for the human analysis compared expression to normal HMEC gene expression (n=10) data gathered from GSE24468.

MOUSE METASTASIS DATASET

A dataset with known lung metastasis from MMTV-Neu[44], MMTV-Myc[45,46], and MMTV-PyMT[47] was compiled for metastasis free survival of mouse models.

MOUSE AND HUMAN GENE LOCATION CONVERSION

Locations of mouse and human genes were taken from the Affymetrix array annotation files from mouse 430A_2 and human U133A_2 array. These locations were merged by common gene symbol to provide a conversion table between the two species for the location of a particular gene.

CLUSTERING OF HUMAN AND MOUSE TUMORS

ACE analysis was performed as previously described on the TCGA breast cancer human dataset as well as the mouse dataset. Significant CNAs were mapped onto the mouse genome for clustering. For human to mouse comparisons genes were filtered to those genes that were amplified or deleted in at least 5% of human and mouse tumors (n=594). Unsupervised hierarchical clustering was performed using cluster 3.0 and Java Tree View. For tumor histology the MMTV-Myc dataset with histological annotations, GSE15904, was used. To cluster this dataset we filtered the genes to 4118 genes through the use of standard deviation of the neighborhood score. This removed all genes that were unaltered across the dataset. For all clustering analysis Euclidian distance, complete linkage was used for the similarity metric and clustering method respectively.

JACCARD INDEX

Jaccard index was calculated between clusters through use of the R package “sets” through use of the similarity function.

MOUSE AND HUMAN HISTOLOGICAL COMPARISONS

Histological annotations were analyzed from a group of MMTV-Myc mouse tumors[40], GSE15904, as well as the human TCGA tumors[12]. We identified the genes within mouse chromosome 11 which were also amplified in a subset of human tumors using cbio portal. These fourteen identified genes mapped to the 17q25.1 region in humans and are referred to as the microacinar associated event. For overrepresentation analysis we compared the number of human tumors with the microacinar subtype from a group with the microacinar associated amplification event found in mice against a random set of tumors not containing that amplification event through the use of a 2x2 contingency table.

ONCOGENIC SIGNATURE APPLICATION

Predefined oncogenic signatures were applied to the dataset. Briefly, the training data was merged with the full dataset and batch effects removed through the use of COMBAT. These samples were then subjected to binary regression analysis with a predefined gene list and conditions for each individual signature[40,48–51].

COORDINATION OF CNA WITH ONCOGENIC SIGNATURE

Oncogenic signatures[40,48,49,51] and lung metastasis signatures[38] were applied to mouse and human datasets as previously described. These scores were coordinated to neighborhood score through the use of a Spearman rank correlation applied through R. A significance threshold of $P<.01$ was applied and the results were visualized using MATLAB.

GENE NETWORK INTERACTION

Interaction networks were visualized through the use of STRING-DB[52]. Input nodes were those genes significantly correlated with the particular pathway in a specific region as well as key signaling proteins for the pathway (Rb/E2F2). Twenty additional white nodes were added to complete the network.

APPENDIX

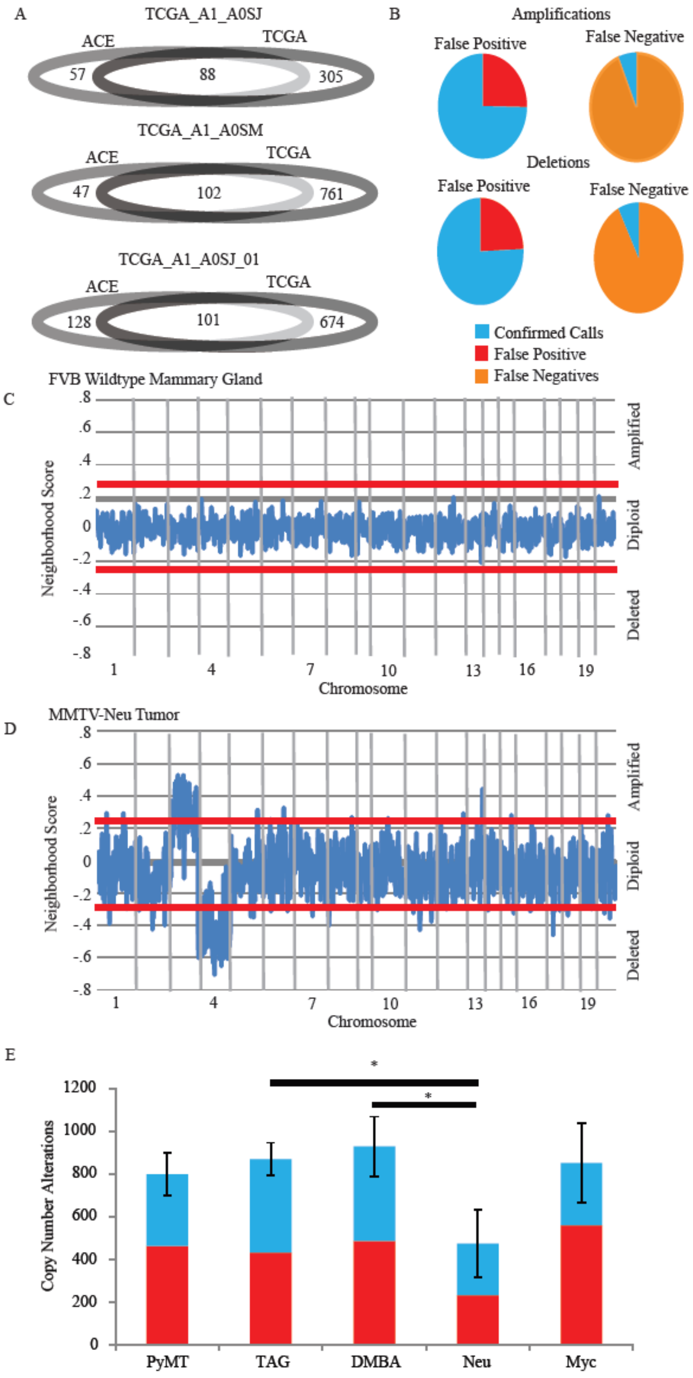

Figure 2.1: Identification of copy number alteration through gene expression data from mouse models of breast cancer

Figure 2.1 (cont'd)

(a) The Venn diagram illustrates the consistency of ACE copy number calls with traditional copy number calls of three randomly chose TCGA breast cancer samples. The ACE algorithm was applied to a three TCGA sample and genes that were significantly amplified ($p<.05$, $q<.05$) were compared against the TCGA copy number predictions for the same sample. (b) When applied to the entire TCGA breast cancer dataset with microarray and complete CNV calls (N=478) the ACE algorithm is able to identify amplification (top) and deletion (bottom) events with a false discovery rate of 25.4% and 23.3% respectively (left) and a false negative rate of over 90% (right) (c) The ACE algorithm predicts no copy number alteration in an FVB control mouse. The graph shows neighborhood score, the weighted mean expression value used by the algorithm to predict copy number, at each genomic location (blue line). Significant regions of amplification or deletion are identified when the blue line falls outside of the red bounds ($p<.05$, $q<.05$). (d) Copy number predictions across the genome of a FVB MMTV-Neu tumor reveals notable amplification in chromosome 3 and deletion in chromosome 4 in addition to several smaller amplification and deletion events. (e) When the ACE algorithm is applied to a set of mouse tumors made up of MMTV-Myc tumors (N=12), MMTV-Neu tumors (N=15), MMTV-PyMT tumors (N=26) and DMBA (N=14) and TAG models (N=37), all on the FVB/NJ background, the MMTV-Neu model has significantly ($P<.05$) less amplified or deleted genes compared to the TAG or DMBA models. The MMTV-Neu model was significantly more stable ($P<.05$) and showed on average less copy number changes than all other models.

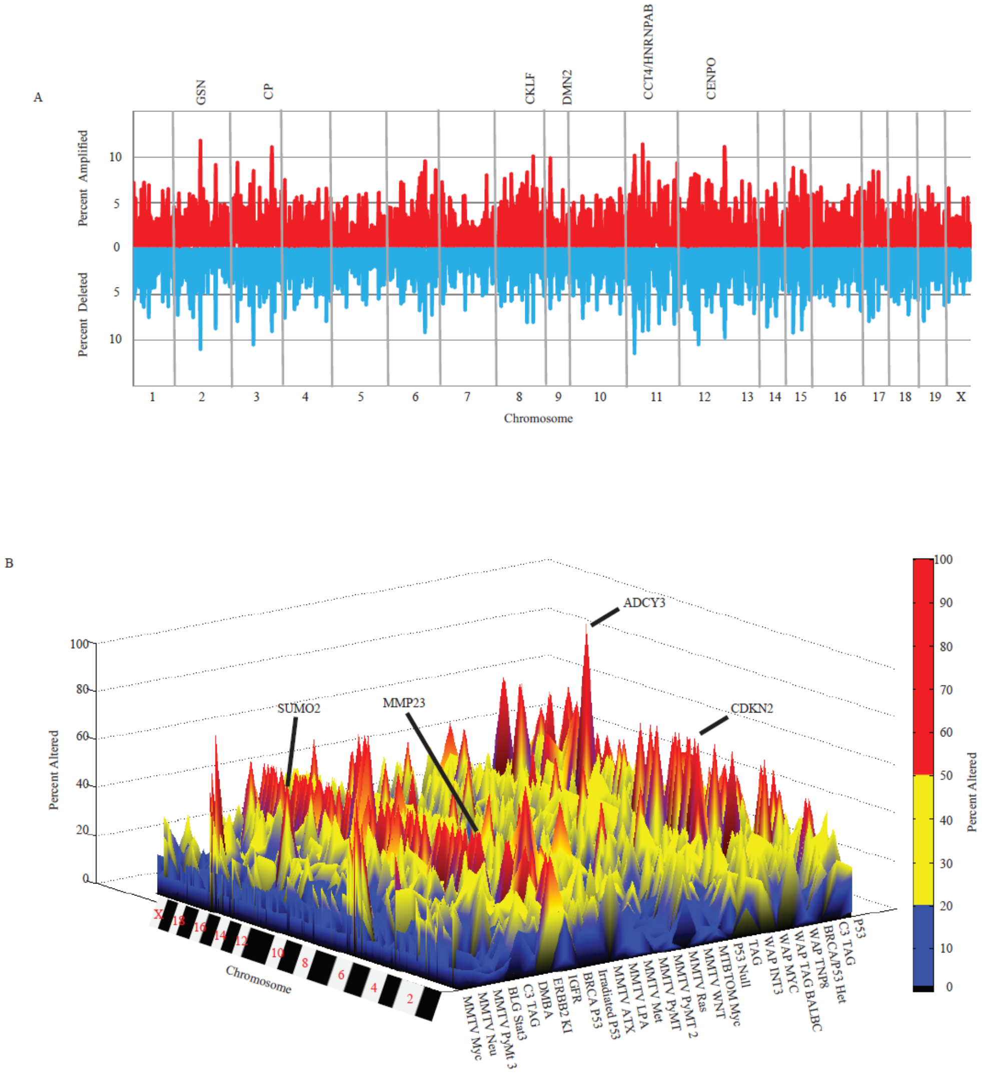

Figure 2.2: Landscape of CNA heterogeneity across mouse models of breast cancer

(a) ACE analysis to predict CNA was applied to 600 mouse model tumors arising from 27 major models of breast cancer. The percentage of mice with amplification (red) or deletion (blue) regardless of mouse model at a particular locus across the genome is shown as identified by ACE (q<.05). All genes present on the microarray were graphed. Genes that are amplified or deleted in more than 10% of mice are identified. (b) The fraction of mice with amplification or

Figure 2.2 (cont'd)

deletion event at individual loci is shown across all chromosomes for individual mouse models. The bar height and color illustrates the percent of samples with an amplification or deletion at the individual gene locus as indicated by the legend.

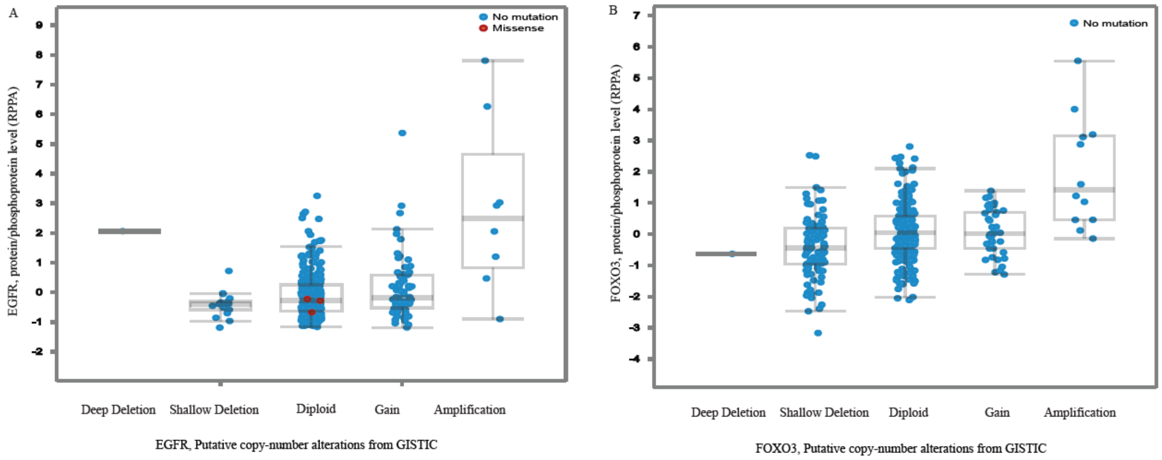

Figure 2.3: Correlation between copy number alterations and gene expression data

The TGCA data was queried for copy number alterations and protein levels in EGFR (a) and FOXO3 (b). These samples were separated in to five categories, Deep deletion (homozygous deletion), Shallow deletion (heterozygous deletion), diploid, Gain (low level amplification), and amplification (high level amplification). A positive correlation between increased copy number and protein level was identified.

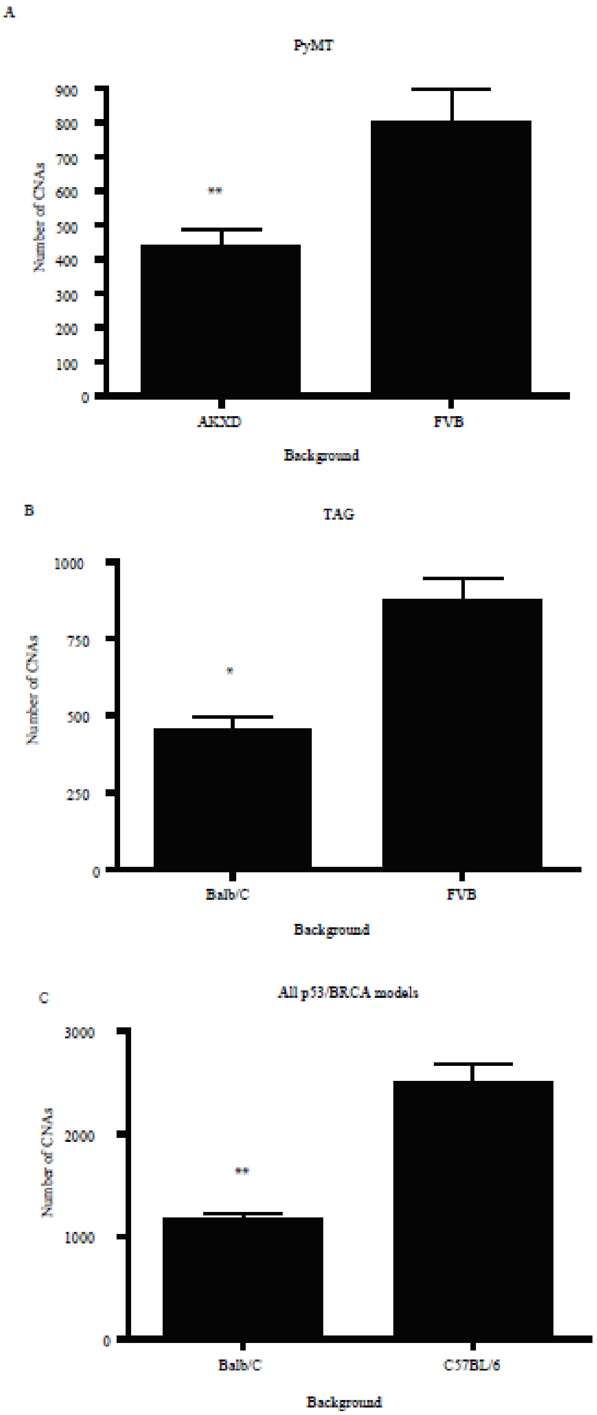

Figure 2.4: Mouse genetic background and number of copy number alterations

To identify the effect of mouse strain on the stability of a mouse model we used mouse models with the same oncogenic driver on different mouse model backgrounds. This was done with

Figure 2.4 (cont'd)

the MMTV-PyMT (a), TAG (b), and p53/BRCA (c) models. It was found that in the PyMT model significantly more alterations were found in the FVB background (N=66) when compared to the AKXD model (N=55) (P<.01). A similar result was noted with the TAG model where the FVB background (N=37) had significantly more alterations than TAG driven tumors in a Balb/C background (N=3) (P<.05). In the BRCA/p53 models we found the C5Bl/6 model (N=12) to be more unstable compared to the Balb/c background (N=73) (P<.01).

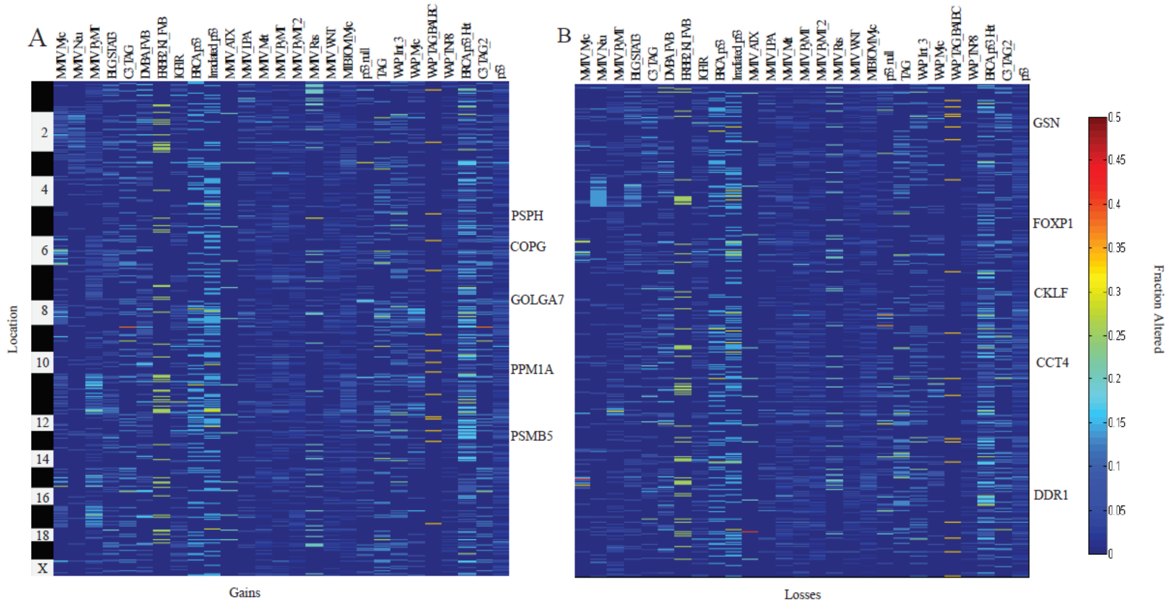

Figure 2.5: Amplification or Deletion in specific mouse models

Heatmap representation of the data in Figure 2B. Containing amplification or deletion percentages in specific mouse models. Percentages are displayed as a value between 0 (blue) and 100% (red). The figure is split into amplifications (left) and deletions (right)

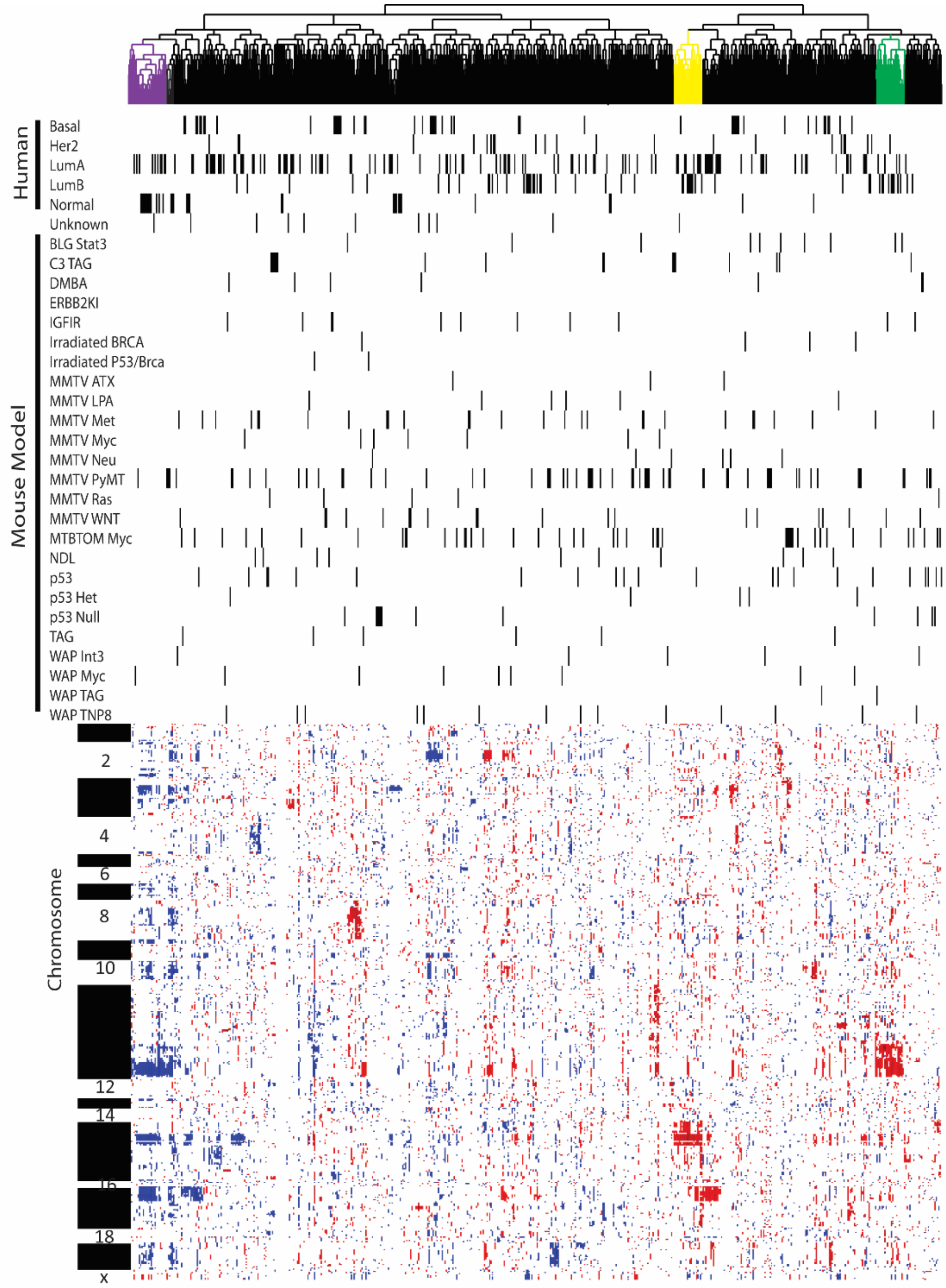

Figure 2.6: Full heatmap associated with Figure 2.7A

(a)To assess the conservation of CNAs in mouse models and human patients unsupervised hierarchical complete linkage clustering of samples across human and mouse tumors were clustered by recurrent CNA events (N=597) that were amplified or deleted in greater than 5% of mouse and human tumors. The dataset used the complete mouse models dataset of 27 mouse

Figure 2.6 (cont'd)

models (N=600) and randomly chosen TCGA breast cancer tumors across all five major subtypes of breast cancer (N=559). The clustering revealed three tight clusters composed of human and mouse samples as indicated by the purple, yellow, and green clusters.

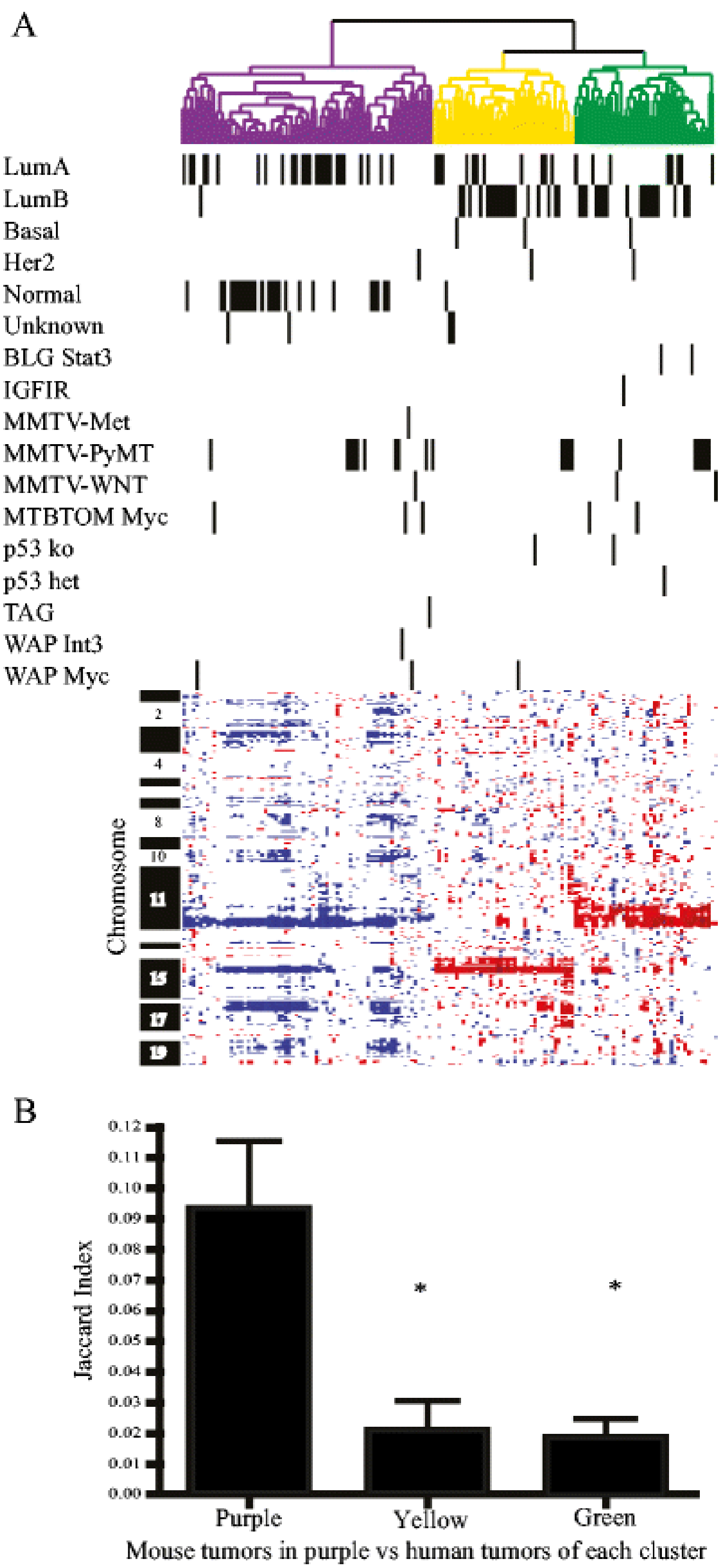

Figure 2.7: Conservation of common human CNAs in specific mouse models

Figure 2.7 (cont'd)

(a) To assess the conservation of CNAs in mouse models and human patients unsupervised hierarchical complete linkage clustering of samples across human and mouse tumors were clustered by recurrent CNA events (N=597) that were amplified or deleted in greater than 5% of mouse and human tumors. The dataset used the complete mouse models dataset of 27 mouse models (N=600) and randomly chosen TCGA breast cancer tumors across all five major subtypes of breast cancer (N=559). The clustering revealed three tight clusters composed of human and mouse samples as indicated by the purple, yellow, and green clusters. A highlighted version of this figure is seen (a) while the full version can be found in the supplemental materials. The analysis showed a large fraction of PyMT tumors sorted into each major cluster indicating there are shared copy number alterations between the MMTV-PyMT mouse model and human tumors particularly the Lumina A, Lumina B and normal like subtypes of breast cancer.(b) Through the use of a Jaccard index we showed within cluster similarity. This reveals a significant decrease of Jaccard Index score when comparing mouse samples from the purple cluster to human samples of the purple, yellow (P<.05) and green cluster (p<.05).

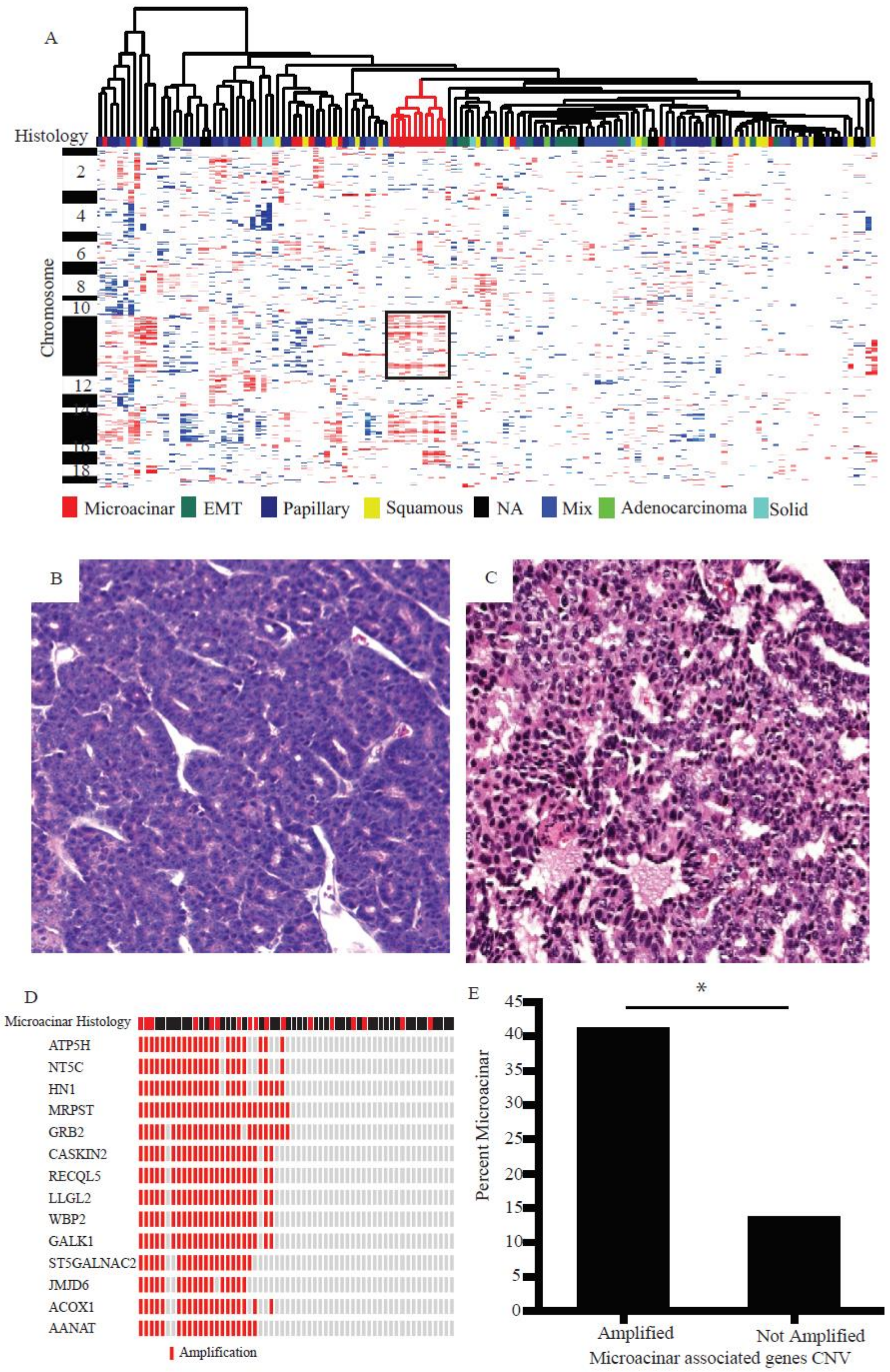

Figure 2.8: Within model CNA heterogeneity associates with tumor histology

Figure 2.8 (cont'd)

(a) For the MMTV-Myc tumor model (n=105), individual tumor samples were clustered by their copy number profile and separated largely into histological subtypes. Histological subtype of each sample is indicated by the color of the bar below the dendrogram as specified by the legend. Vertically the genes (n=4118) are ordered by their chromosomal location from 1 to X. The relationship of the samples is indicated by the dendrogram. The heatmap indicates amplification (red) or deletion (blue) at a particular locus (p<.05, q<.05). Chromosome 11 amplification was noted in a large fraction of mouse tumor with microacinar histology (Boxed). (b) Mouse tumors with chromosome 11 amplification display a distinct microacinar histological subtype. (c) Human tumors with analogous region (17q25.1) amplified exhibit similar microacinar-like histological patterns. (d) The cbio oncoprint of the microacinar associated genes across 58 samples. The genes were identified as amplified on mouse chromosome 11 as well as human chromosome region 17q25.1 and total 14 genes in all and identify the core genes associated with microacinar like tumor histology(28 control samples with amplification events and 30 control samples) (e) Across the TCGA breast cancer dataset, patients with a consistent amplification pattern of chromosome 17q23.1 have a microacinar like histological subtype significantly (P=.01) more often than those without the amplification event (N= 28 for amplified, N=30 for non-amplified) indicating a role in this region in defining tumor histology.

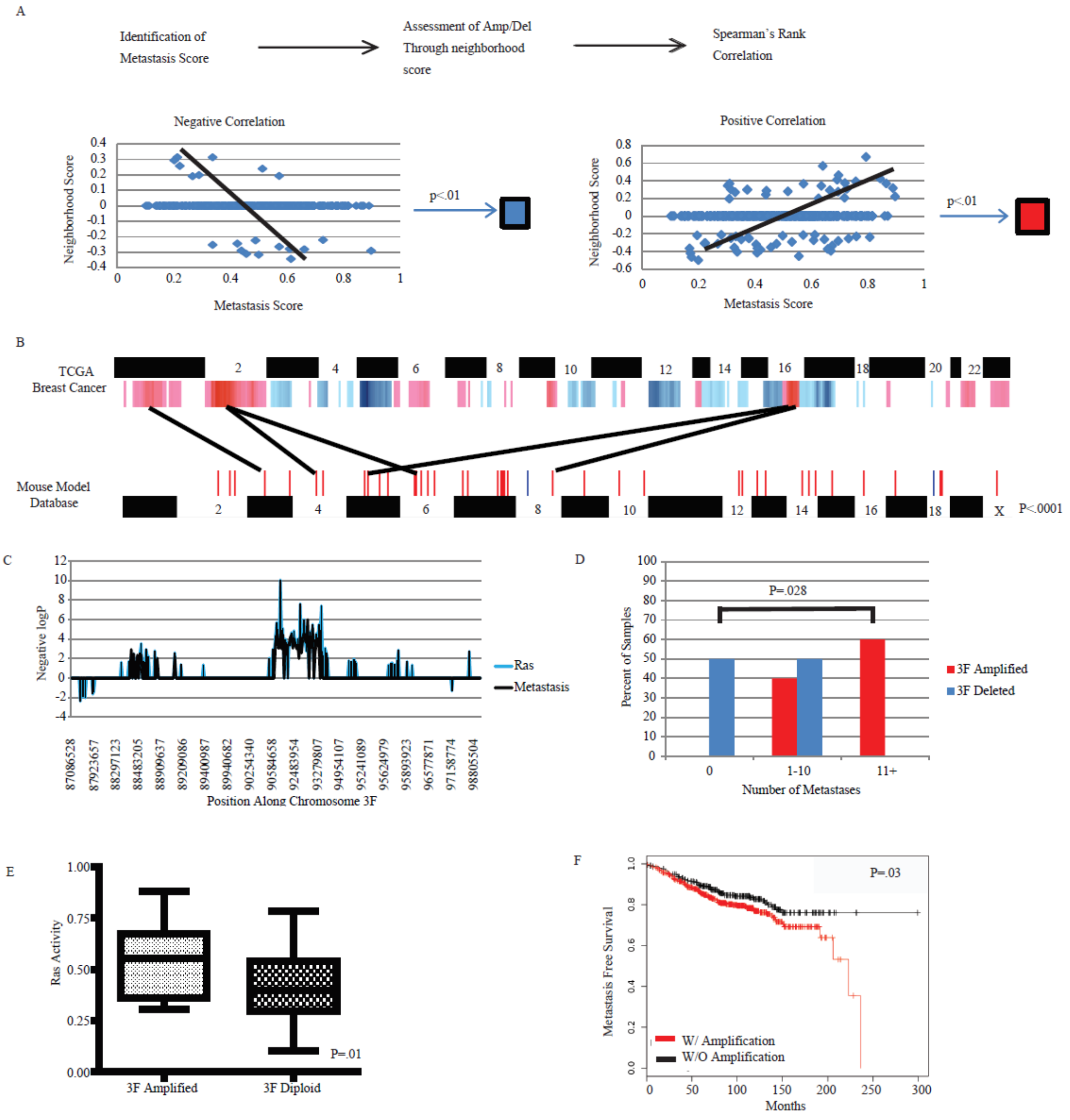

Figure 2.9: Role of CNA in tumor metastasis

(a) The schematic outlining the strategy to associate copy number alteration with metastasis score is shown where the metastasis score was correlated with the neighborhood score to assess CNA. Negative correlation (blue) and positive correlation (red) examples are shown. (b) Metastasis associated copy number gains (red) or losses (blue) in the TCGA human breast

Figure 2.9 (cont'd)

cancer dataset are shown by human chromosome (B - top). The location of the homologous loci and their amplification status are shown in the mouse model database (B - bottom). Dark black lines indicate example conserved human locations and their associated mouse chromosomal location. Conserved amplification and deletion events between both species are overrepresented when compared to a random set of genes (P<.0001) An identified region is conserved between humans chromosome 1 and mouse chromosome 3F indicating a role in tumor metastasis by this region in mouse and human tumors and is more completely explored in panel c. (c) To identify a potential pathway through which metastasis was being mediated we coordinated Ras activity with each amplification or deletion event and looked to identify regions that were associated with metastasis and high ras activity. This is graphed for mouse chromosome 3F by the negative log of P value for the association of amplification (positive) and deletion (negative) of mouse chromosome 3 where amplification is associated with both metastasis and Ras activity in the 3F region. (d) When tumors are split on the basis of chromosome 3F status those mice, from and MMTV-Myc, MMTV Neu, or MMTV-PyMT background, with amplification (Red) (n=5) are shown to have significantly more metastases than those with a deletion at the same locus (Blue) (n=6) (e). When the region is identified in the kmplot.com dataset this event is shown to be conserved in human Luminal A tumors when the analogous human region is amplified there is significantly (P=.03) higher Ras activity and lower (P<.01) metastasis free survival (f).

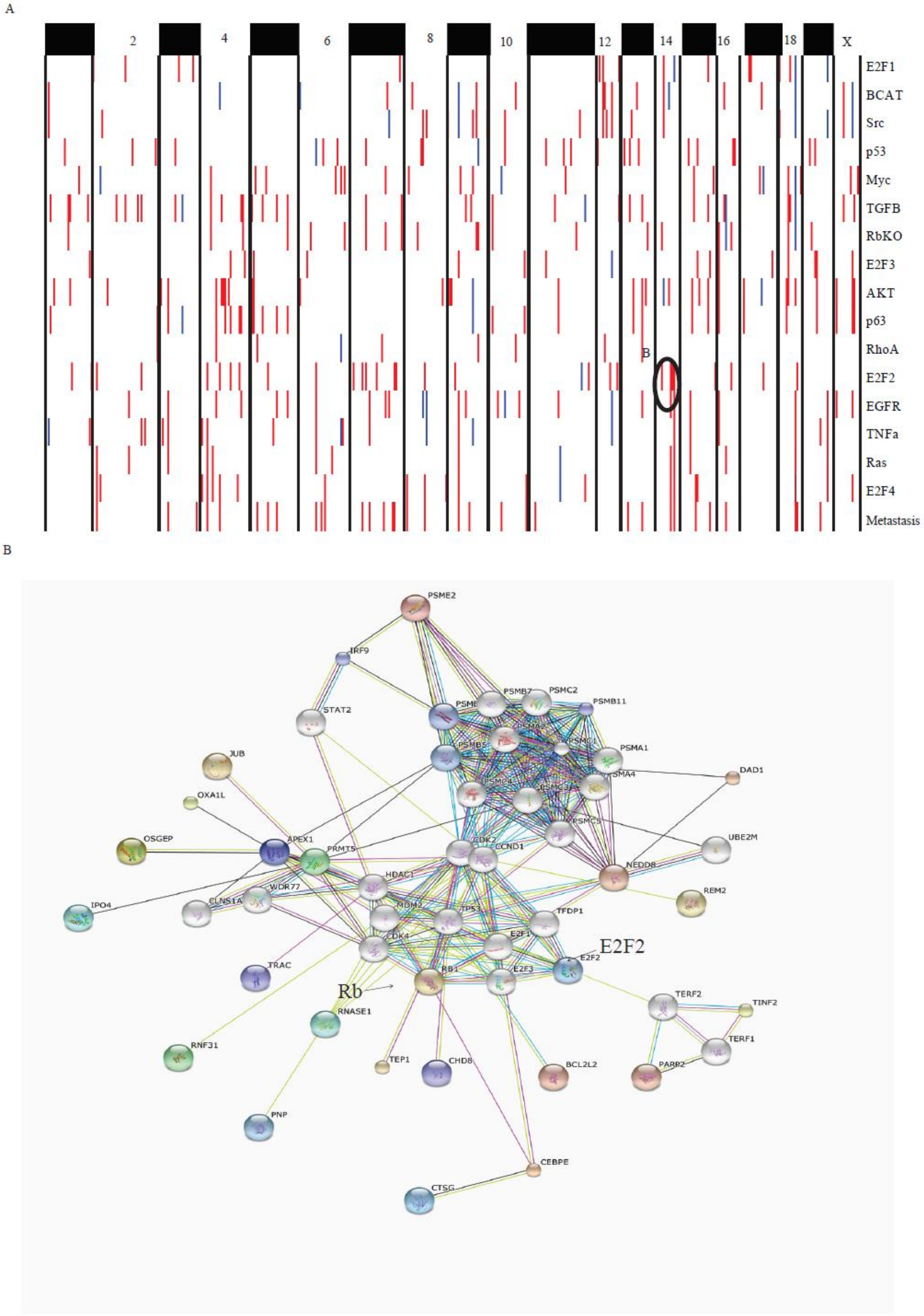

Figure 2.10: Role of CNA in oncogenic signaling pathways

Figure 2.10 (cont'd)

(a) Spearman's rank correlation of amplification (red) or deletion (blue) events with high activity of oncogenic signaling pathways is shown. Events are arranged by chromosomal location as indicated at the top for the pathways indicated at the right. The String-DB derived connectivity map of RB-E2F (B) networks is depicted. Rb and E2F2 are denoted by black arrows. All other colored nodes are genes which have a copy number alteration significantly correlated with a particular signaling pathway indicated by black circles, with the exception of Rb and E2F2.

WORKS CITED

WORKS CITED

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

2. Escot, C. *et al.* Genetic alteration of the c-myc protooncogene (MYC) in human primary breast carcinomas. *Proc. Natl. Acad. Sci. U. S. A.* **83,** 4834–8 (1986).

3. Li, J. *et al.* PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275,** 1943–7 (1997).

4. Deming, S. L., Nass, S. J., Dickson, R. B. & Trock, B. J. C-myc amplification in breast cancer: a meta-analysis of its occurrence and prognostic relevance. *Br. J. Cancer* **83,** 1688–95 (2000).

5. Ménard, S., Fortis, S., Castiglioni, F., Agresti, R. & Balsari, A. HER2 as a Prognostic Factor in Breast Cancer. *Oncology* **61,** 67–72 (2001).

6. Hall, J. M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250,** 1684–9 (1990).

7. Feilotter, H. E. *et al.* Analysis of the 10q23 chromosomal region and the PTEN gene in human sporadic breast carcinoma. *Br. J. Cancer* **79,** 718–23 (1999).

8. Shimada, H. *et al.* Identification of subsets of neuroblastomas by combined histopathologic and N-myc analysis. *J. Natl. Cancer Inst.* **87,** 1470–6 (1995).

9. Wang, S. *et al.* Prostate-specific deletion of the murine Pten tumor suppressor gene leads to metastatic prostate cancer. *Cancer Cell* **4,** 209–21 (2003).

10. Fujita, T. *et al.* PTEN activity could be a predictive marker of trastuzumab efficacy in the treatment of ErbB2-overexpressing breast cancer. *Br. J. Cancer* **94,** 247–252 (2006).

11. Vogel, C. L. *et al.* Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J. Clin. Oncol.* **20,** 719–26 (2002).

12. TCGA. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

13. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–52 (2012).

14. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor

subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 10869–74 (2001).

15. Prat, A. *et al.* Research-based PAM50 subtype predictor identifies higher responses and improved survival outcomes in HER2-positive breast cancer in the NOAH study. *Clin. Cancer Res.* **20,** 511–21 (2014).

16. Muller, W. J., Sinn, E., Pattengale, P. K., Wallace, R. & Leder, P. Single-step induction of mammary adenocarcinoma in transgenic mice bearing the activated c-neu oncogene. *Cell* **54,** 105–15 (1988).

17. 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* **49,** 465–75 (1987).

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

19. Cardiff, R. D. *et al.* The mammary pathology of genetically engineered mice: the consensus report and recommendations from the Annapolis meeting. *Oncogene* **19,** 968–88 (2000).

20. Andrechek, E. R. *et al.* Amplification of the neu/erbB-2 oncogene in a mouse model of mammary tumorigenesis. *Proc. Natl. Acad. Sci.* **97,** 3444–3449 (2000).

21. McCarthy, A. *et al.* A mouse model of basal-like breast carcinoma with metaplastic elements. *J. Pathol.* **211,** 389–398 (2007).

22. Xu, X. *et al.* Conditional mutation of Brca1 in mammary epithelial cells resultsin blunted ductal morphogenesis and tumour formation. *Nat. Genet.* **22,** 37–43 (1999).

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

24. Moody, S. E. *et al.* Conditional activation of Neu in the mammary epithelium of transgenic mice results in reversible pulmonary metastasis. *Cancer Cell* **2,** 451–61 (2002).

25. Guy, C. T. *et al.* Expression of the neu protooncogene in the mammary epithelium of transgenic mice induces metastatic disease. *Proc. Natl. Acad. Sci.* **89,** 10578–10582 (1992).

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

27. Herschkowitz, J. I. *et al.* Comparative oncogenomics identifies breast tumors enriched in functional tumor-initiating cells. *Proc. Natl. Acad. Sci.* **109,** 2778–2783 (2012).

28. DeRose, Y. S. *et al.* Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nat. Med.* **17,** 1514–20 (2011).

29. Hollern, D. P. & Andrechek, E. R. A genomic analysis of mouse models of breast cancer reveals molecular features of mouse models and relationships to human breast cancer. *Breast Cancer Res.* **16,** R59 (2014).

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

31. Pfefferle, A. D. *et al.* Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. *Genome Biol.* **14,** R125 (2013).

32. Silva, G. O. *et al.* Cross-species DNA copy number analyses identifies multiple 1q21-q23 subtype-specific driver genes for breast cancer. *Breast Cancer Res. Treat.* **152,** 347–56 (2015).

33. Hu, G. *et al.* MTDH activation by 8q22 genomic gain promotes chemoresistance and metastasis of poor-prognosis breast cancer. *Cancer Cell* **15,** 9–20 (2009).

34. Ciriello, G. *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163,** 506–519 (2015).

35. Hodgson, J. G. *et al.* Copy Number Aberrations in Mouse Breast Tumors Reveal Loci and Genes Important in Tumorigenic Receptor Tyrosine Kinase Signaling. *Cancer Res.* **65,** 9695–9704 (2005).

36. Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* **121,** 2750–67 (2011).

37. Helmrich, A., Stout-Weider, K., Hermann, K., Schrock, E. & Heiden, T. Common fragile sites are conserved features of human and mouse chromosomes and relate to large active genes. *Genome Res.* **16,** 1222–1230 (2006).

38. Minn, A. J. *et al.* Genes that mediate breast cancer metastasis to lung.

39. Ben-David, U. *et al.* The landscape of chromosomal aberrations in breast cancer mouse models reveals driver-specific routes to tumorigenesis. *Nat. Commun.* **7,** 12160 (2016).

40. 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.* **106,** 16387–92 (2009).

41. Boxer, R. B., Jang, J. W., Sintasath, L. & Chodosh, L. A. Lack of sustained regression of c-MYC-induced mammary adenocarcinomas following brief or prolonged MYC inactivation. *Cancer Cell* **6,** 577–586 (2004).

42. Siegel, P. M., Ryan, E. D., Cardiff, R. D. & Muller, W. J. Elevated expression of activated forms of Neu/ErbB-2 and ErbB-3 are involved in the induction of mammary tumors in transgenic mice: implications for human breast cancer. *EMBO J.* **18,** 2149–64 (1999).

43. Györffy, 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.* **123,** 725–31 (2010).

44. Andrechek, E. R. HER2/Neu tumorigenesis and metastasis is regulated by E2F activator transcription factors. *Oncogene* (2013). doi:10.1038/onc.2013.540

45. Fujiwara, K., Yuwanita, I., Hollern, D. P. & Andrechek, E. R. Prediction and genetic demonstration of a role for activator E2Fs in Myc-induced tumors. *Cancer Res.* **71,** 1924–32 (2011).

46. Yuwanita, I., Barnes, D., Monterey, M. D., O'Reilly, S. & Andrechek, E. R. Increased metastasis with loss of E2F2 in Myc-driven tumors. *Oncotarget* **6,** 38210–24 (2015).

47. Hollern, D. P., Honeysett, J., Cardiff, R. D. & Andrechek, E. R. The E2F transcription factors regulate tumor development and metastasis in a mouse model of metastatic breast cancer. *Mol. Cell. Biol.* (2014). doi:10.1128/MCB.00737-14

48. Bild, A. H. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439,** 353–7 (2006).

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

50. Andrechek, E. R., Mori, S., Rempel, R. E., Chang, J. T. & Nevins, J. R. Patterns of cell signaling pathway activation that characterize mammary development. *Development* **135,** 2403–13 (2008).

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

52. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39,** D561-8 (2011).

CHAPTER 3

GENOMIC LANDSCAPE OF MMTV-NEU AND MMTV-PYMT TUMORS

ABSTRACT

Mouse models have an essential role in cancer research, yet little is known about how various models resemble human cancer at a genomic level. However, the shared genomic alterations in each model and corresponding human cancer are critical for translating findings in mice to the clinic. We have completed whole genome sequencing and transcriptome profiling of two widely used mouse models of breast cancer, MMTV-Neu and MMTV-PyMT. This genomic information was integrated with phenotypic data and CRISPR/Cas9 studies to understand the impact of key events on tumor biology. Despite the engineered initiating transgenic event in these mouse models, they contain similar copy number alterations, single nucleotide variants, and translocation events as human breast cancer. A key finding showed that both models had similar mutational processes. Despite this similarity, it was shown that the MMTV-Neu mouse model was significantly more unstable than the MMTV-PyMT tumor model. A larger panel of tumors for each model showed a large amount of diversity in both model systems. This diversity was shown in key genes associated with tumor progression and potentially modify the behavior of the tumors in both mouse and human. These findings underscore the importance of understanding the complete genomic landscape of a mouse model and illustrate the utility this has in understanding human cancers.

INTRODUCTION

Breast cancer is a large public health concern with an estimated one in eight women experiencing the disease during her lifespan[1]. Compounded with this, breast cancer is the second leading cause of cancer related deaths in women[2]. Largely, the cause of death in patients is the heterogeneity of the disease. This heterogeneity is shown at all level of genetic regulation including genome[3], transcriptome[4], and proteome[5]. Due to the diversity in the disease it becomes very difficult to match the correct patient with the correct treatment.

There have been international efforts including, TCGA[3,6], ICGC[7], and Metabric[8] to understand the underlying causes and diversity in breast cancer. These studies have been extremely productive and have identified a number of subtypes of breast cancer each with unique genomic profile. A number of candidate oncogenes and tumor suppressors have been potentially identified through these studies; however, in order to validate these hypotheses researchers must turn to *in vitro* and *in vivo* models of the disease.

A key *in vivo* tool that researchers use is the genetically engineered mouse model. In this system, a mouse has been altered in such a way that it is predisposed to tumors in the mammary gland. Many times, this involves the overexpression of an oncogene through the use of a tissue specific manner. In the case of breast cancer this is typically the mouse mammary tumor virus (MMTV)[9,10] or whey acidic protein (WAP)[11,12] promoter system.

Two key models using this system to model breast cancer are the MMTV-Neu and MMTV-PyMT mouse models. The MMTV-Neu[13] mouse model is meant to model HER2 positive breast cancer through the forced over expression of the HER2 homolog, Neu, in the mammary gland. These tumors have a moderate latency of 202 days and the majority of the tumors metastasize to the lung. The MMTV-PyMT[14] mouse model, formed through the over expression of the polyoma middle T antigen in the mammary gland, is meant to model highly aggressive endstage disease. The model has a latency of 45 days and is highly metastatic.

Despite the frequency of these models there is debate about the findings in them and how well they translate to the human patient setting. This is largely due to the fact it is unknown how well the models actually represent the human disease from a genomic viewpoint. There have been preliminary bioinformatic studies profiling the gene expression profile[15–17] and gene copy number status[18,19] of mouse models of breast cancer. These studies have revealed that mouse models on a whole are extremely heterogeneous like human cancer. Surprisingly, many mouse models produce tumors which represent multiple subtypes of human cancer.

To understand this further, and to confirm genomic events underlying the transcriptional diversity in mouse models of cancer researchers have turned to next generation sequencing. To date, there have been studies profiling the genome of lung cancer mouse models[20,21], models of melanoma[22] as well as tumor suppressor driven[23] breast cancer mouse models. However, to our knowledge the MMTV-Neu and MMTV-PyMT models have not been profiled in such a way.

Here we present an integrative genomic profile of the MMTV-Neu and MMTV-PyMT mouse models. In this pursuit we have performed whole genome sequencing as well transcriptomic profiling of the two models. We have carefully integrated this with tumor phenotypes including latency, metastatic ability, as well as the histological subtype.

Through this analysis we further confirmed this heterogeneity found in mouse models. Despite the heterogeneity, we were able to identify a number of defining events in each model including a copy number variant which modulates metastatic potential and a single nucleotide variant that modifies EGFR signaling.

This study is an important proof of concept study and underscores the importance of understanding the entire genomic landscape of mouse models. Furthermore, this type of study has immediate clinical benefits beyond the impact in basic research.

RESULTS

INTER AND INTRA TUMOR DIVERSITY IN MMTV-NEU AND MMTV-PYMT

To characterize the genomic landscape of the MMTV-Neu and MMTV-PyMT tumors, we created a tumor database with complete phenotypic characterization including tumor latency, histology, and metastatic burden. Representative tumors from this database were selected for whole genome sequencing and whole transcriptome profiling by microarray. The analysis pipeline then correlated phenotypic changes with molecular profiling, including transcriptomics and sequence alterations. The resulting genes were then filtered through human breast cancer datasets to ensure relevance to human breast cancer and confirmed with *in vitro* and *in vivo* experiments (Figure 3.1A). A high degree of transcriptomic diversity both between and within each model was observed in hierarchical clustering (Figure 3.1B). As expected, this heterogeneity correlated with tumor histological subtype rather than tumor model, consistent with recent studies[15,24]. It was hypothesized that these differences in expression were driven by genomic changes.

Following standard informatic pipelines, the whole genome sequencing data was analyzed. To validate bioinformatic calls of SNVs and CNVs we used PCR and qPCR, observing a validation rate of 85%. Whole genome sequencing revealed large differences in the genomic landscape of the MMTV-Neu (Figure 3.2A) and MMTV-PyMT (Figure3.2B) tumors. The two tumor models had similar numbers of SNVs (Figure 3.2C), however both models were ~20X more stable than human breast tumors with 0.049 mutations/megabase in the mouse models in comparison to an average of approximately 1 mutation/megabase in breast cancer[25]. Copy number alterations (Figure 3.2D) and translocations (Figure 3.2E) were more frequent in the MMTV-Neu model relative to MMTV-PyMT.

To understand the specific role of copy number alterations within the two models we compared copy number variants present in the mouse models with those also in the human breast cancer. This analysis identified 11 candidate genes which were highly altered in breast cancer (Figure 3.4) and predicted to impact tumor biology based upon a literature screen. qPCR gene copy number analysis across an extended tumor panel (15 MMTV-PyMT, 10 MMTV-Neu) identified the rate at which each copy number variant occurred throughout the model (Figure 3.3A). This analysis showed that while each of the copy number variants predicted through bioinformatic means were valid, the depth of the amplification was largely around 1.5 fold indicating shallow amplification events (Figure 3.3B). Interestingly we identified the largest diversity of copy number profiles in the 11D locus.

MUTATIONAL PROCESS OF GENETICALLY ENGINEERED MOUSE MODELS

In addition to copy number alterations, the whole genome sequence data resulted in the identification of numerous mutations (Figure 3.2C). When COSMIC mutational signatures[26,27] were applied to the models, it was observed that the tumor models had similar mutational processes (Figure 3.5). The MMTV-Neu and MMTV-PyMT tumors both contain the same trinucleotide context of their mutation spectrum. The mutation spectrum shows all nucleotide substitutions present with a slight bias towards C/T and T/C transitions. When compared to the human mutational signatures, the mutational processes present in both mouse models closely resembles COSMIC signature 5 (Figure 3.5C). This signature has been shown to be present in breast cancer patients with disease associated with late onset[28], indicating a similar mutational process in both the human disease and mouse models

SINGLE NUCLEOTIDE VARIANTS IN GEMMS

The distribution of SNVs reflected patterns seen in the transcriptional data (Figure 3.1B) with some events shared between Neu and PyMT tumors while others were unique to the models. Considerable SNV diversity within a model was also prevalent. For instance, the MMTV-Neu model had no genes with shared mutations in all samples and only five genes containing a coding, non-synonymous mutation in more than one sample (Figure 3.6). Notably we identified mutations within Mucin 4 (*Muc4*) which are potentially impactful due to *Muc4*'s emerging roles in HER2 positive cancer and metastasis[29].

Interestingly, we observed that PyMT induced tumors had more SNVs in the coding regions of the genome despite not having significantly more mutations overall. Specifically, these mapped to 34 genes, 9 of which overlapped with Neu tumors. A number of genes with coding mutations specifically in PyMT tumors, including *Matn2, Plekhm1, Muc6* and *Ptprh* were observed. *Matn2*[30]*, Plekhm1*[31] and *Muc6*[32] have all been demonstrated to have roles in tumor progression and metastasis and may contribute to the high metastatic capacity of the MMTV-PyMT model. *Ptprh* is a phoso-tyrosine receptor phosphatase that is shown to regulate EGFR signaling[33]. This gives it obvious implications in tumor signaling and progression.

To test the frequency of these coding mutations in the models as a whole we selected a population of 10 MMTV-Neu tumors and 15 MMTV-PyMT tumors for targeted resequencing. From these tumors we extracted genomic DNA and performed PCR based amplification followed by Sanger sequencing of *Matn2, Plekhm1* and *Ptprh*. While *Matn2* and *Plekhm1* confirmed the whole genome sequencing variant calls in the sequenced tumors, additional mutations were not found. Strikingly, *Ptprh* was found to be mutated in 81% of MMTV-PyMT tumors. This indicates a conserved and perhaps necessary role of *Ptprh* mutation within MMTV-PyMT tumors which was explored in greater depth in future chapters.

DISCUSSION

This data is to our knowledge the first comprehensive multi-omic profile of mouse models of breast cancer specifically the MMTV-Neu and MMTV-PyMT models. These models provide key insights into HER2 positive breast cancer and the process of tumor metastasis and are critical for the advancement of their respective research fields. This research fills the critical need of understanding translatability of mouse models to human cancer.

We found that mouse models reflect the heterogeneity found in breast cancer. In both models, there is a large amount of diversity both with the transcriptome as well as the genome. Interestingly it was found that these differences largely correlate with the histological subtype of the tumor. This finding should change the way that researchers choose mouse models to use for study. Instead of choosing a mouse model based upon the oncogenic driver, researchers should instead choose the tumors within the model that capture their disease context better. We have identified tumor histological subtype as an indicator of genomic profile and a good proxy for determining what class of tumor that a given tumor from a mouse model belongs to.

Interestingly despite the presence of heterogeneity in the models, all of the sequenced tumors, regardless of model, had a similar mutational profile, referred to as mutagenic signature. In human cancer, specific mutagenic signatures correlate with specific mutational processes such as BRCA loss or exposure to UV light. The presence of a consistent mutational signature indicates the same driver of instability underlying both models. Strikingly the mutational signature found in MMTV-Neu and MMTV-PyMT is highly consistent with human mutational signature number 5. Human signature number 5 is found in breast cancer patients that present at and old age; however, the driver of the signature is unknown. Study of the mouse models may reveal insight into the driver of signature 5. However, more work must be done to understand the drivers of instability in mouse models. The most obvious limitation of this study is with the number of samples. More samples must be sequenced to see if the consistency of mutational profile extends beyond the individual tumors presented in this study. Also, it will be important to expand this type of analysis to other tumor models with different oncogenic drivers including those with loss of tumor suppressors or carcinogen induced model.

A key opportunity of this study was to examine the “two hit hypothesis.” In human cancers, the prevailing hypothesis is that it takes multiple genomic alterations (hits) to cause the development of tumors. Some mouse models require multiple hits to initiate tumors including models such as the BRCA/p53 knockout models. However, this is not the case for the MMTV-Neu or MMTV-PyMT tumor models. In these models, the only engineered oncogenic event is the force overexpression of Neu or the Middle T antigen in each model respectively. It is possible that in order for the tumor to develop and progress additional oncogenic drivers must be affected.

Surprisingly, in neither model were traditional human drivers of cancer identified as altered in the sequenced tumors. In the MMTV-PyMT tumors, we saw an extremely prevalent mutation in the protein *Ptprh*. Indicating that loss of *Ptprh* is required for PyMT tumors to progress, or certainly it is highly advantageous for the development of tumors. We do not identify analogous mutation of this protein in breast cancer patients; however, it is present in other tumor types such as lung cancer. With regards to MMTV-Neu tumors we did not identify any highly conserved mutations. This made the identification of additional oncogenic hits difficult. With this study, we cannot conclude rather this model requires additional oncogenic promotion in addition to the overexpression of Neu. To pursue this question, more individuals from the model must be examined.

We did identify a number of copy number variants that were in traditional oncogenic or tumor suppressive tumors. However, when these were validated through qPCR they were identified to be low level copy number variants with only 1.5 to 3 fold extra copies. To identify if this is an artifact of the assay additional assays such as fish must be performed. If the copy number variants are determined to be low level, next will be to understand if the upregulation is significant enough to cause impacts in cell signaling. It has recently been shown that there is a subclass of human tumors with the low level copy number variants. A number of extra experiments must be performed to determine if and how mouse models of breast cancer resemble low level copy number variant driven human tumors.

As noted in earlier chapters, the identified copy number and single nucleotide variants were associated with tumor progression, not initiation. Specifically, we identified a number of genes that are implicated in tumor metastasis. These are explored in more detail in additional chapters.

A key finding is that mouse models have significantly less mutations than human tumors. This is unsurprising due to the lack of evolutionary pressure on these tumors. This finding as impacts that both improve and hinder the utility of mouse models. An advantage to using mouse models is that there are relatively few alterations. This makes the system relatively clean without the same magnitude of genomic noise and passenger events found in human tumors. The lack of genomic variants improves the ability of researchers to find novel impactful events such as the variation of *Ptprh* found in this work. However, the lack of genomic events limits of the utility of traditional mouse models for tumor immunology research. It has recently been shown that tumors with a high mutation burden have a high number of neo-peptides and have evolved ways to evade immune detection. This is not the case with the MMTV-Neu and MMTV-PyMT tumors. The low mutation burden causes low neo-peptide production and thus few immune evasion techniques. To study these aspects of tumor biology new genetically engineered mouse models with a higher mutational load must be developed.

This project is an important proof of concept study. It provides important information about the heterogeneity in and progression of MMTV-PyMT and MMTV-Neu tumors. Beyond this, the study shows the utility of sequencing specific mouse models. In these relatively small studies we identified a number of events with direct translatability to the clinic. We expect that with an expansion of the number of samples in the study the amount of clinically relevant findings will grow exponentially. It is critical for the advancement of basic cancer research as well as translational findings to have a comprehensive understanding of genetically engineered mouse models and the genomic changes present in them.

MATERIALS AND METHODS

ANIMAL STUDIES

All animal husbandry and use was conducted according to local, national and institutional guidelines. The MMTV-Neu[13] and MMTV-PyMT[14] mice were in the FVB background. MMTV-PyMT634 and MMTV-Neu mice were obtained from The Jackson Laboratory. Mice were monitored twice weekly for tumor initiation and growth. At a 2000 mm^3 endpoint, mice were necropsied. For mice with multiple tumors the endpoint was established when the primary tumor was at 2000 mm^3.Tumors and lungs were collected for genomic analysis, hematoxylin and eosin staining for histological subtyping and presence of pulmonary metastases. The number of metastasis was quantified using a single cut through the lung and count of the number of micro-metastases in that plane. Masson's trichrome staining was used to examine tumors for collagen deposition using standard methods.

WHOLE GENOME SEQUENCING

Flash frozen tumor pieces were ground and DNA was extracted with the Qiagen Genomic-tip 20/G with the manufacturer's protocol. DNA was sequenced to a depth of 40X with paired end 150 base pair reads on an Illumina HiSeq 2500 using the Illumina TruSeq Nano DNA library preparation.

TRANSCRIPTOMIC PROFILING

Transcriptome data for this study was previously published[34–36]. Data was downloaded from GSE42533 (MMTV-Neu) and GSE104397 (MMTV-PyMT) as .cel files. Affymetrix expression console was used to normalize each individual dataset using RMA normalization. To remove batch effects between datasets BRFM normalization[37] was performed with standard parameters.

CLUSTERING

Unsupervised hierarchical clustering was performed using Cluster 3.0 and heatmaps were created using the MATLAB imagesc function.

VARIANT CALLING

Generated .fastq files were assessed for quality control using FASTQC analysis http://www.bioinformatics.babraham.ac.uk/projects/fastqc. Reads were trimmed for quality using Trimmomatic[38]. After trimming, data was reassessed for quality using FASTQC. Then reads were aligned to the mm10 mouse reference genome using BWA-mem. After alignment, base recalibration and pcr induced biases were removed using PICARD tools (http://broadinstitute.github.io/picard). For variant calling we utilized four software packages, GATK[39], Mutect2[40], Strelka[41], and SomaticSniper[42]. To be a legitimate variant we filtered to only those variants called by 3 of the 4 packages. To control for differences in the FVB strain and the mm10 reference genome we used previously published normal FVB tissue (ERR046395)[43]. To call copy number and structural variants we used Delly[44]. For copy number we used default quality control settings and only analyzed those copy number events which had precise boundaries and were larger than 100KB. For translocations we used default quality control setting and precise breakpoints.

VARIANT VERIFICATION AND EXTENDED TUMOR PANEL SEQUENCING

For verification of SNVs we used PCR based amplification followed by Sanger sequencing. For validation of CNVs we used qPCR on the genomic DNA with the Quantabio PerfeCTa SYBR green kit under the manufacturer's specifications.

CIRCOS VISUALIZATION

Representative MMTV-Neu and MMTV-PyMT samples were chosen to be displayed as CIRCOS[33] plots. CIRCOS plots were generated using CIRCOS v 0.69 and SNVs, CNVs, and translocations were mapped according to their location on the mm10 genome.

MUTATION SIGNATURES

Due to the low mutational burden of MMTV-Neu and MMTV-PyMT tumors, mutations were combined into a signal analysis for each model. These samples were processed with MutSpec-NMF[45] for trinucleotide context and comparison to the known human mutation signatures.

APPENDIX

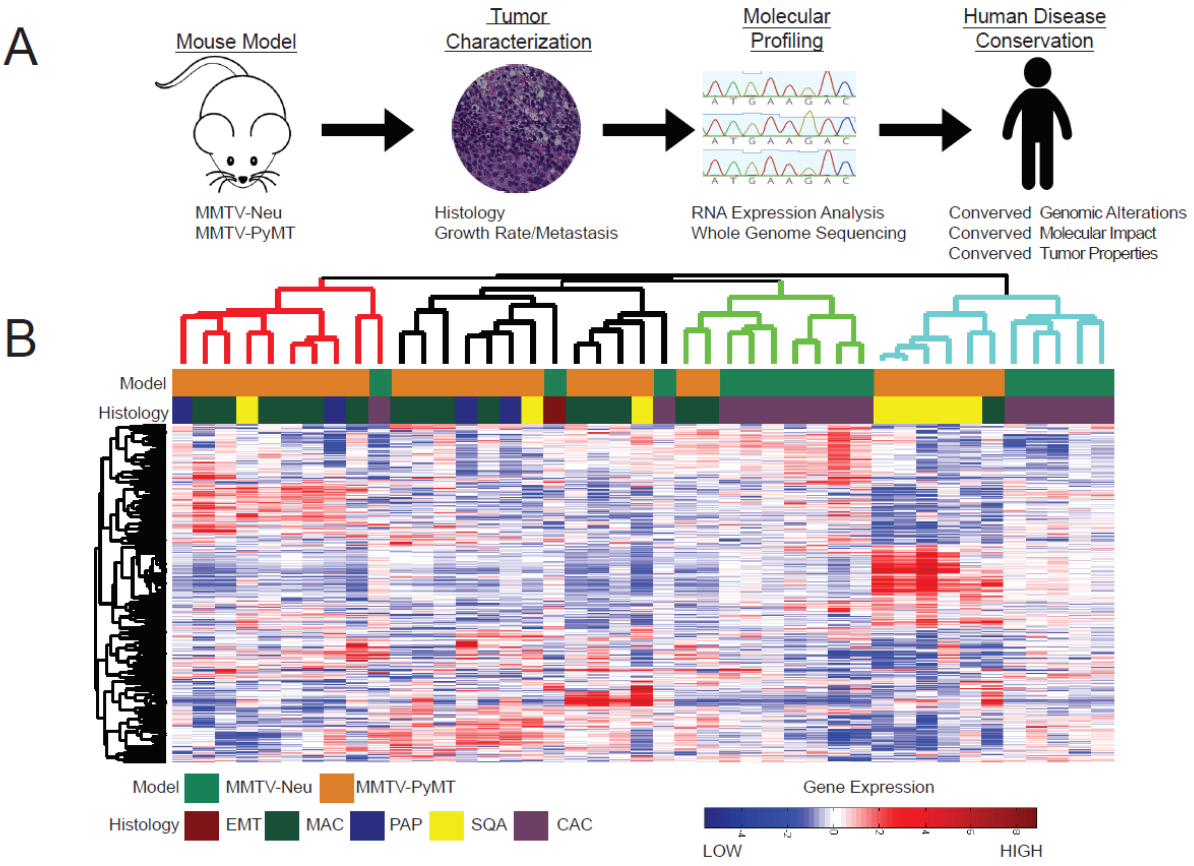

Figure 3.1: Genomic landscape of MMTV-Neu and MMTV-PyMT tumors

The schematic representation of the project workflow is depicted (A), where mammary tumors from two major mouse models are completely characterized through histological, molecular, sequence and transcriptomic methods. After data integration and analysis, the tumors were compared to human cancers at both genomic and phenotypic levels. Gene expression patterns from MMTV-Neu and MMTV-PyMT tumors were compared by unsupervised clustering, revealing substantial heterogeneity both between and within models. Tumors clustered largely based on histological subtype and not simply genotype. SQU – squamous, MAC – microacinar, PAP – papillary, and CAC – comedo-adenocarcinoma (n=15 for MMTV-Neu, n=25 for MMTV-PyMT) (B).

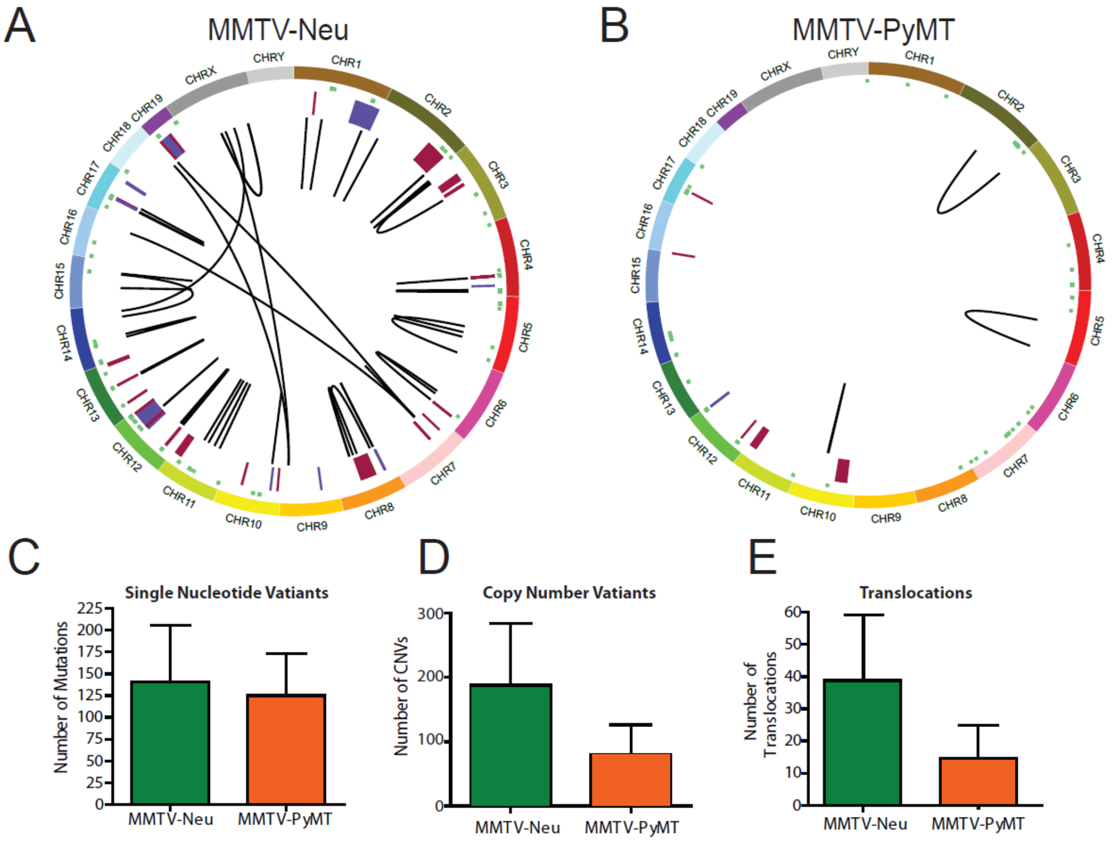

Figure 3.2: MMTV-Neu tumors are significantly more unstable that MMTV-PyMT tumors

Circos plots from whole genome sequencing results for MMTV-Neu (C) and MMTV-PyMT (D) tumors revealed differences between the strains for genomic alterations. Plots display from outside in; Chromosomal location (Each chromosome is unique color), SNVs (green), copy number alterations (Amplification – Red and Deletions – Blue), and translocations (black lines). Variation from multiple tumors is shown for Single Nucleotide Variants (E), Copy Number Variants (F) and translocations (G) (n=3 for each).

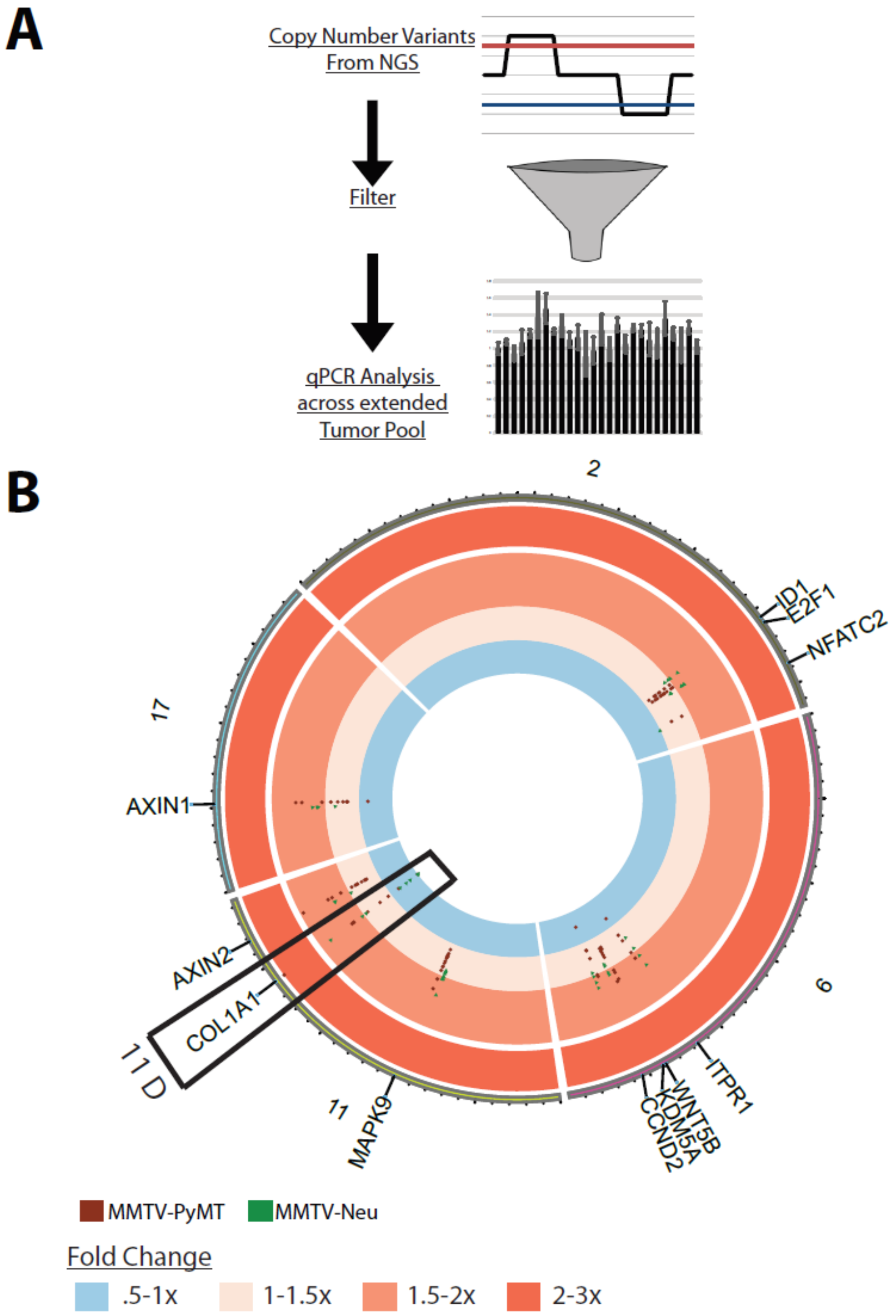

Figure 3.3: Copy number alterations in mouse models of breast cancer

Schematic representation of filtering of copy number variants (A) where copy number variants were detected from NGS and filtered to the top 11 variants by selecting those genes that

Figure 3.3 (cont'd)

encompassed both mouse models and were conserved in human breast cancer. These genes were then assayed using qPCR analysis across a larger panel of MMTV-Neu (n=10) and MMTV-PyMT (n=15) tumors and depicted using a circos plot for genes in chromosomes 2, 6, 11 and 17 (B). Deletion on the interior of the plot to a 3 fold amplification on the exterior of the plot is shown. A key copy number alteration in the 11D region

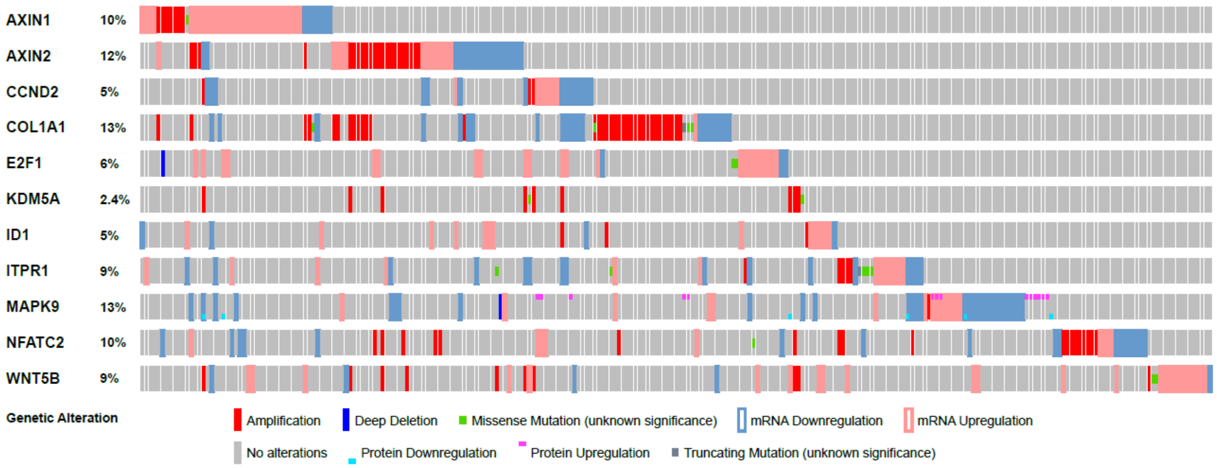

Figure 3.4: Copy number alterations TCGA Breast Cancer Oncoprint

Oncoprint of the human TCGA Breast cancer cohort (Nature 2012) displaying the alteration of genes altered at a high rate in mouse models with regards to copy number

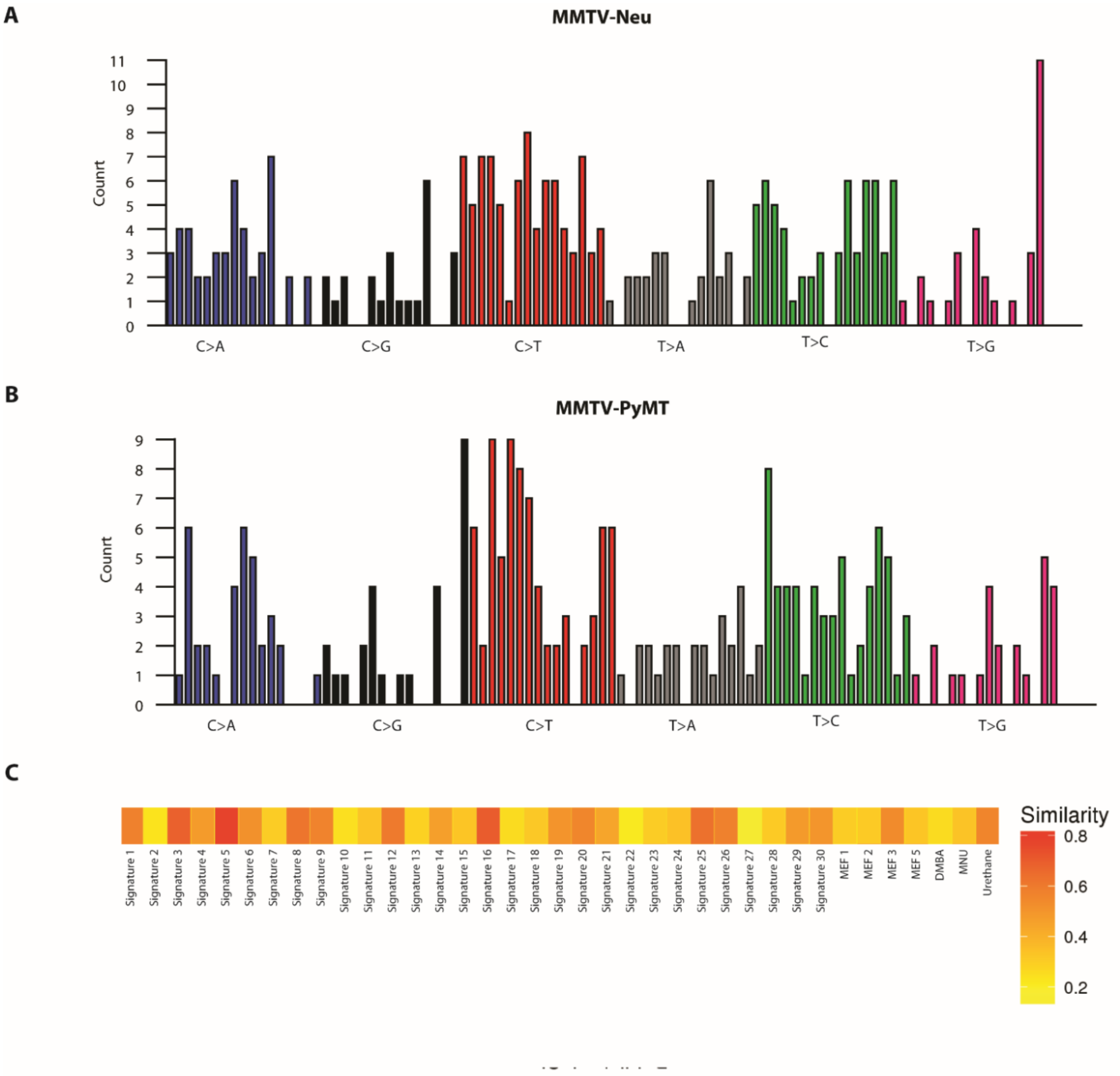

Figure 3.5: Mutational Signatures of MMTV-Neu and MMTV-PyMT Models

The trinucleotide context of MMTV-Neu (A) and MMTV-PyMT (B) samples are similar. They show the presence of every mutation possibility with the overrepresentation of the C>T and T>C transitions. These trinucleotide signatures were compared with human mutational signatures through the use of a Baysian model high similarity (Red) and low similarity (yellow) were identified through the use of a heat map. Signature 5 represented the highest similarity score.

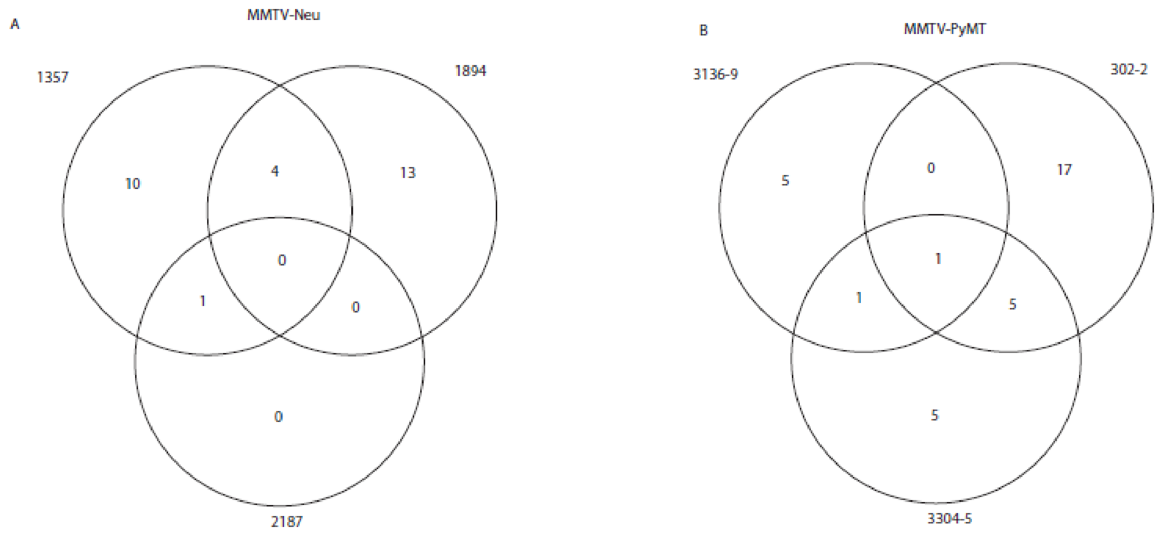

Figure 3.6: Heterogeneity of SNVs in mouse models of breast cancer

The MMTV-Neu (A) and MMTV-PyMT (B) models have considerable diversity in regards to SNVs. Samples were analyzed for overlap in SNV calls through the use of a Venn Diagram

WORKS CITED

WORKS CITED

1. ACS. Breast Cancer Facts and Figures. (2018).

2. ACS. Cancer Facts and Figures. (2017).

3. TCGA. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

4. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 10869–74 (2001).

5. Tyanova, S. *et al.* Proteomic maps of breast cancer subtypes. *Nat. Commun.* **7,** 10259 (2016).

6. Ciriello, G. *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163,** 506–519 (2015).

7. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534,** 47–54 (2016).

8. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–52 (2012).

9. 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* **49,** 465–75 (1987).

10. Muller, W. J., Sinn, E., Pattengale, P. K., Wallace, R. & Leder, P. Single-step induction of mammary adenocarcinoma in transgenic mice bearing the activated c-neu oncogene. *Cell* **54,** 105–15 (1988).

11. 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.* **55,** 3915–27 (1995).

12. Schoenenberger, C. A., Zuk, A., Groner, B., Jones, W. & Andres, A. C. Induction of the endogenous whey acidic protein (Wap) gene and a Wap-myc hybrid gene in primary murine mammary organoids. *Dev. Biol.* **139,** 327–37 (1990).

13. Guy, C. T. *et al.* Expression of the neu protooncogene in the mammary epithelium of transgenic mice induces metastatic disease. *Proc. Natl. Acad. Sci.* **89,** 10578–10582 (1992).

14. Guy, C. T., Cardiff, R. D. & Muller, W. J. Induction of mammary tumors by expression of

polyomavirus middle T oncogene: a transgenic mouse model for metastatic disease. *Mol. Cell. Biol.* **12,** 954–61 (1992).

15. Hollern, D. P. & Andrechek, E. R. A genomic analysis of mouse models of breast cancer reveals molecular features of mouse models and relationships to human breast cancer. *Breast Cancer Res.* **16,** R59 (2014).

16. Pfefferle, A. D. *et al.* Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. *Genome Biol.* **14,** R125 (2013).

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

18. Rennhack, J., To, B., Wermuth, H. & Andrechek, E. R. Mouse Models of Breast Cancer Share Amplification and Deletion Events with Human Breast Cancer. *J. Mammary Gland Biol. Neoplasia* **22,** 71–84 (2017).

19. Ben-David, U. *et al.* The landscape of chromosomal aberrations in breast cancer mouse models reveals driver-specific routes to tumorigenesis. *Nat. Commun.* **7,** 12160 (2016).

20. Westcott, P. M. K. *et al.* The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. *Nature* **517,** 489–492 (2015).

21. McFadden, D. G. *et al.* Mutational landscape of *EGFR-* , *MYC-* , and *Kras-* driven genetically engineered mouse models of lung adenocarcinoma. *Proc. Natl. Acad. Sci.* **113,** E6409–E6417 (2016).

22. Kwong, L. N. *et al.* Modeling Genomic Instability and Selection Pressure in a Mouse Model of Melanoma. *Cell Rep.* **19,** 1304–1312 (2017).

23. Francis, J. C. *et al.* Whole-exome DNA sequence analysis of *Brca2* - and *Trp53* -deficient mouse mammary gland tumours. *J. Pathol.* **236,** 186–200 (2015).

24. 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.* **106,** 16387–92 (2009).

25. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

26. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149,** 979–993 (2012).

27. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,**

415–421 (2013).

28. Nik-Zainal, S. & Morganella, S. Mutational Signatures in Breast Cancer: The Problem at the DNA Level. *Clin. Cancer Res.* **23,** 2617–2629 (2017).

29. Rowson-Hodel, A. R. *et al.* Membrane Mucin Muc4 promotes blood cell association with tumor cells and mediates efficient metastasis in a mouse model of breast cancer. *Oncogene* **37,** 197–207 (2018).

30. Minn, A. J. *et al.* Genes that mediate breast cancer metastasis to lung.

31. Permuth-Wey, J. *et al.* Identification and molecular characterization of a new ovarian cancer susceptibility locus at 17q21.31. *Nat. Commun.* **4,** 1627 (2013).

32. Leir, S.-H. & Harris, A. MUC6 mucin expression inhibits tumor cell invasion. *Exp. Cell Res.* **317,** 2408–19 (2011).

33. Yao, Z. *et al.* A Global Analysis of the Receptor Tyrosine Kinase-Protein Phosphatase Interactome. *Mol. Cell* **65,** 347–360 (2017).

34. Hollern, D. P. & Andrechek, E. R. A genomic analysis of mouse models of breast cancer reveals molecular features of mouse models and relationships to human breast cancer. *Breast Cancer Res.* **16,** R59 (2014).

35. Hollern, D. P., Honeysett, J., Cardiff, R. D. & Andrechek, E. R. The E2F transcription factors regulate tumor development and metastasis in a mouse model of metastatic breast cancer. *Mol. Cell. Biol.* (2014). doi:10.1128/MCB.00737-14

36. Andrechek, E. R. HER2/Neu tumorigenesis and metastasis is regulated by E2F activator transcription factors. *Oncogene* (2013). doi:10.1038/onc.2013.540

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

38. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* btu170- (2014). doi:10.1093/bioinformatics/btu170

39. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).

40. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–9 (2013).

41. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28,** 1811–1817 (2012).

42. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28,** 311–7 (2012).

43. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477,** 289–94 (2011).

44. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012).

45. Ardin, M. *et al.* MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. *BMC Bioinformatics* **17,** 170 (2016).

CHAPTER 4

CHARACTERIZATION OF 17Q21.33 AMPLIFICATION

ABSTRACT

The HER2 positive subtype of breast cancer presents a unique clinical challenge. It is present in approximately 20% of HER2 positive breast cancer and it is shown to be particularly aggressive with regard to metastatic potential compared to breast cancer patients of other subtypes. A unique challenge of the subtype is the presence of multiple amplicons located on chromosome 17 that lead to differential gene signaling and potentially different tumor phenotypes. Here we use an integrative *in silico*, *in vitro* and *in vivo* approach we present here characterization of one such event, 17q21.33. The 17q21.33 amplicon is present in 25% of HER2 positive breast cancer patients and 8% of patients regardless of subtype. We identified this event to be associated with worse distant metastasis free survival due to the presence of Co11a1 and CHAD within the amplification event. This was identified through the use of a wound healing assay, tail vein injection, and mammary fat pad injection of CRISPR-Cas9 generated knockout cell lines for Col1a1 and CHAD. In all assays the reduction of metastatic potential was seen. Importantly, we are also able to identify the vulnerability of tumors with the 17q21.33 amplicon to AKT targeted therapy. This was predicted through a number of high throughput genomic and drug compound screens in which unique vulnerabilities were identified in those cell lines containing the 17q21.33 amplicon. This study underscores the importance of understanding the diversity within HER2 positive cancer. Furthermore, the work presented here has immediate clinical impact due to its translatability with screening for metastatic lesions as well as AKT targeted therapy.

INTRODUCTION

Approximately 20% of breast cancers are classified as HER2 positive[1,2]. This subtype is defined by the amplification and/or overexpression of the epidermal growth factor receptor family protein HER2. This subclass is associated with a high metastatic burden with an overrepresentation of metastases to the brain[3]. The highly metastatic potential of this tumor subclass makes it particularly difficult to treat.

The process of tumor metastasis involves a number of complex steps as a tumor proceeds from its primary location in the breast, through the lymph and vasculature, to a distant organ. While many stages of this process are unknown it has been shown that a tumor microevironment plays an important role in both the initiation of metastasis[4] and progression of the process. Specifically, it has been shown that the extracellular matrix (ECM) must be in a certain formation to allow for cell migration[5]. This involves the remodeling of the ECM through the production of collagen fibers[6], matrix metalloproteinase[7] and adherens[8] proteins.

Despite the metastatic nature of the disease, there has been success in developing targeted therapy for this disease subtype. Specifically, these therapies target the HER2 signaling cascade through binding to HER2 directly or indirectly stopping signaling[9–11]. These therapies have successful in improving patients outcomes; however, a significant portion of patients either fail to respond initially or acquire resistance to the therapeutic agent[12,13].

Underlying the diversity of response to HER2 targeted therapy is diversity within HER2 positive breast cancer[14]. A characteristic of the HER2 subtype is not just the amplification of the HER2 coding region but amplification in other regions of the genome as well. In fact, many HER2 positive tumors present with a “firestorm” amplification pattern along chromosome 17[15].

In these tumors many regions along chromosome 17 are amplified. It is hypothesized that each of these amplification events have impacts on cell signaling and eventual tumor characteristics including metastasis and treatment response.

Here we present an in depth muti-platform analysis of one of the firestorm events on chromosome 17 specifically on band 21.33 (17q21.33). This event was shown to be found in approximately 25% of breast cancer patients. This patient population was also shown to have lower distant metastasis free survival. Through the use of CRIPSR-Cas9 mediated studies we identified Collagen Type 1 Alpha 1 (Col1a1) and Chondroadherin (CHAD) as contributing genes to the metastatic phenotype. We also used bioinformatic and high throughput screening studies to predict therapeutic impacts of the amplification event. While we were unable to identify an association with HER2 targeted therapy, we did identify an association of tumors with the 17q21.33 amplification event and response to AKT targeted therapy. This result indicates that patients with the HER2 amplicon and co-amplification of 17q21.33 may benefit from addition AKT targeted therapy.

RESULTS

IDENTIFICATION OF 11D COPY NUMBER VARIATIONS

Through the resequencing approach discussed in the previous chapter, we identified low number gene copy number alterations in the mouse models. Interestingly we identified the largest diversity of copy number profiles in the 11D locus. This locus includes a total of 40 genes, 19 with transcriptomic differences. Depending on the presence or absence of the locus, the tumors exhibited striking differences in structure and behavior. We identified dramatic differences in the tumors with the presence of an 11D amplification with regards to collagen content through a Mason's trichrome (Figure 4.1A) stain and the presence of metastatic lesions in the lungs (Figure 4.1 B).

To identify the driving genes of the metastatic phenotype associated with 11D amplification, we examined human breast cancer for distant metastasis free survival outcomes and then created CRISPR-Cas9 generated knockouts of two potential metastasis related proteins within the region, Collagen type 1 alpha 1 (*Col1a1*) and Chondroadherin (*Chad*). Knockouts were generated in two mouse driven tumor cell lines NDL2-5[16] and PyMT 419[17]. NDL2-5 is an 11D amplified Neu driven line, while the 419 line is diploid for the 11D locus and is driven by PyMT expression. Knockouts of each gene in both cell lines revealed defects in the ability to migrate in a wound healing assay (Figure 4.2A and Figure 4.2B). Defects in lung colonization in a tail vein injection were also observed with the *Col1a1* and *Chad* knockout cell lines (Figure 4.2C and Figure 4.2D). The differences in migratory ability may be contributed by the completentess of the knowout/knockdown in each line (Figure 4.4). Migration was partially rescued with addback of wildtype *Col1a1* or *Chad*, demonstrating that migration defects were not due to off target effects (Figure 4.5).

11D AMPLIFICATION IS ANALOGOUS TO 17Q21.33 AMPLIFICATION IN HUMANS

Mouse chromosome 11D is conserved in humans and is analogous to chromosomal region 17q21.33. There is similar amplification event at 17q21.33, including *COL1A1* and *CHAD* that occurs in 8% of breast cancer patients. Array CGH from the TCGA data[18,19] demonstrates that *COL1A1/CHAD* amplification was distinct from HER2 amplification (Figure 4.3A). Importantly, this amplification is subtype specific; 25% of HER2+ breast cancers have a co-amplification of the 17q21.33 region along with the HER2 amplicon while only 6% of Luminal A, 7% of Luminal B, and 1.2% of Basal breast cancers have amplification (Figure 4.3B). To investigate the transcriptional impact of the amplification event we used weighted gene correlation network analysis[20]. This identified a robust transcriptional signature that differentiated *COL1A1*/*CHAD*, HER2 positive tumors from HER2 positive tumors without the amplification event. Unsupervised hierarchical clustering readily identified separation of the two HER2 positive subtypes based on this signature (Figure 4.3C). These correlated genes were used in a predictive signature to correlate patient outcome with predictive amplification status (Figure 4.6) revealing that metastasis was associated with the amplification event (Figure 4.3D).

To test whether *COL1A1* and *CHAD* were driving the metastasis phenotype in human breast cancer, we used CRISPRi[21] to knockdown *COL1A1* and *CHAD* in the HER2 amplified, *COL1A1*/*CHAD* amplified breast cancer line BT-474. These knockdowns showed a decreased ability to migrate in a wound healing assay (Figure 4.3E and 4.3F). Importantly the knockdown lines also were unable to metastasize to the lung after being injected into the mammary fat pad (Figure 4.3G and 4.3H). Together these data underscore the importance of identifying copy number variation in mouse models of cancer.

TREATING THE 17Q21.33 AMPLIFICATION EVENT

Many other genes were also noted to be co-amplified with the presence of *Col1a1* and *Chad* (Figure 4.7) With the presence of the 17q21.33 we identified a consistent transcriptional profile indicating that the tumors had a unique profile from other HER2 amplified tumors (Figure 4.8). It was hypothesized that the differences in transcription identified between the HER2 positive 17q21.33 amplified tumors and the HER2 positive tumors without the 17q21.33 amplification event may introduce new genetic dependencies and potential therapeutic targets. To pursue this hypothesis, we first identified differences in key oncogenic pathways. We identified slight increases in AKT, E2F2, and Myc (Figure 4.9) signaling associated with 17q21.33 indicating reliance upon these pathways.

To identify specific gene upregulations associated with the 17q21.33 amplicon we first defined the gene amplification region and identified those genes within the region that had coordinated RNA upregulation associated with gene amplification (Figure 4.7). We identified this through the use of cbio portal and the TCGA breast cancer dataset. If is also possible that genes may be differentially regulated as downstream effects of the amplification event and not within genomic region of the event. To identify this we used cbioportal to correlate genes that are overexpressed or underexpressed in the HER2 positive patients with 17q21.33 amplification (Figure 4.10). This identified a number of interesting genes.

To find genes that might contribute to the higher AKT signaling that was identified previously in 17q21.33 amplified patients, we used a stringdb[22] approach. This approach interestingly identified three of the correlated proteins, PHB, Beclin, and SLC45A to be directed interacting with AKT (Figure 4.10). Because of their known roles in apoptosis and autophagy PHB and Beclin were selected to be tested for downstream analysis.

Interestingly, when PHB was knocked down in a 17q21.33 amplified cell line it showed that the cells performed better. That is, they had a higher rate of proliferation than those cells that did not have amplification of 17q21.33 when PHB was knocked down (Figure 4.12) this is not the case for any other gene within the amplicon event as shown by the random growth effects for *Ankrd40*. This indicates that PHB has a tumor suppressive role in the tumors in which it is amplified.

To overcome the high levels of PHB and continue to proliferate the tumor must inactivate or evade the PHB growth signals. According to recent literature, AKT is directly responsible for the phosphorylation or PHB and its inhibition. This drove us to hypothesize that 17q21.33 are dependent upon high AKT signaling. To pursue this, we used a number of publicly available high-throughput screening database.

We first used the Achilles[23] database to identify differentially lethal siRNA in cell lines with co-amplification of the HER2 amplicon and the 17q21.33 amplification as well as those with only HER2 amplification. Unsurprisingly, a number of differentially lethal siRNA's were identified between the two cell groups (Figure 4.11). In support of our hypothesis of AKT addiction, it was shown that the vast majority of the lethal siRNA's were associated with AKT signaling as identified through a string-DB network. This indicates that any amount of AKT signaling perturbation will cause the death of the cell line.

To confirm this using a chemical compound we identified a PDX[24] highthroughput drug screening dataset[25–27]. Similar to the Achilles dataset we dived the PDX samples into HER2 amplified and HER2 amplified with the presence of the 17q21.33 amplification event. Then we identified the response of the tumor to AKT targeted therapy. In particular we identified two compounds which inhibited PI3K signaling (an important activator of AKT). In both cases the samples with the 17q21.33 amplification event was shown to be more sensitive to this line of therapy (Figure 4.11).

DISCUSSION

Here we identify an important subclass of HER2 positive cancer. This subclass is defined by the presence of the 17q21.33 amplification event. Furthermore, we have shown this subclass to be directly impactful clinically with these patients having more metastatic tumors. These patients should have a different course of care than other HER2 positive patients with the additional screening to identify metastatic lesions early. Furthermore, these patients may benefit from the addition of AKT targeted therapy.

The identification of Col1a1 and CHAD as drivers of metastasis presents a unique opportunity to intervene in the metastatic cascade. The data presented in this manuscript indicates that Col1a1 and CHAD play a role in both early and late stages of metastasis. If a therapy was designed to either inhibit the production of these proteins or the ability of the tumor cell to migrate along them, it would inhibit the formation of new metastatic lesions. One could envision such a therapy being concurrent to cytotoxic therapies. This type of therapy would greatly reduce the metastatic potential of the tumors and greatly improve patient outcomes.

The identified AKT addiction presents a key proof of concept type study for the identification of therapeutic avenues. Here we see the amplification of a tumor suppressor, PHB, introduce a unique vulnerability in 17q21.33 tumors. We believe that PHB is amplified as part of a passenger event due to its proximity to other tumor promoting proteins such as KAT7. It is not selected against, due to the high AKT activity present due to the co-occurring HER2 amplification event[28]. However, in the absence or reduction of AKT signaling, PHB is able to return to its natural tumor suppressive role[29] and kill the tumor. This manuscript exposes a unique vulnerability to 17q21.33 amplified tumors and it shows the importance of oncogenic addiction. Furthermore, this study shows the special notice that should be taken when tumors are identified with the amplification of traditional tumor-suppressors due to the therapeutic opportunity that they present.

It is critical to understand the diversity within the HER2 subtype to improve patient care. This has never been more evident than with AKT targeted therapy. Currently, there are no AKT targeting agents that are approved in breast cancer. However, there have been a number of clinical trials presenting AKT targeted therapy in various contexts[30]. However, these have all failed in either phase II or phase III settings due to efficacy. We believe that these studies were fundamentally flawed in their design by not accounting for the heterogeneity within the HER2 subtype. Based on the data presented in this manuscript we believe that those patients noted in each trial to have partial or complete response may be 17q21.33 amplified. It is predicted that the other copy number events associated with HER2 amplicon may have similar impact on tumor behaviors and treatment response. To design and advance clinical trials in the era of precision medicine researchers much account for the diversity within the HER2 positive subtype.

MATERIALS AND METHODS

CELL LINES

The PyMT 419 cell lines were a gracious gift from Dr. Stuart Sell and Dr. Ian Guess[17]. The NDL2-5 cells lines were obtained as a gift from Dr. Peter Siegel[16]. The BT-474 cell line was obtained from Dr. Kathy Gallo and validated using fingerprinting analysis performed at Michigan State University.

CRISPR GENERATED KNOCKOUTS OF PYMT 419 AND NDL2-5

CRISPR/Cas9 constructs were created to knockout *Col1a1* and *Chad* in PyMT 419 and NDL2-5. Guides were designed and inserted into Px458, obtained from addgene (Addgene #48138) as a gift from Feng Zhang, as previously described[31]. Cells were sorted using FACS technology into single cells and grown into clonal population, then screened for the presence of INDELs using Sanger sequencing. Knockouts were further confirmed for the NDL2-5 lines using western blot.

CRISPRI GENERATED KNOCKDOWNS IN BT-474

Knockdowns of Col1a1 and CHAD were created in the BT-474 line using CRISPRi technology. gRNA were cloned into a plasmid containing the gRNA under the control of the U6 promoter (Addgene plasmid #60955)[32]. Lenti virus was created for stable expression of this plasmid and the stable expression of KRAB-Cas9 fusion protein (Addgene plasmid #60954)[32]. Cells were infected with KRAB-Cas9 expression virus first and selected for uptake by puromycin treatment. The stable KRAB-Cas9, BT474 line was then infected with the virus for stable selection of the gRNA for CHAD or COL1A1. These were then sorted using flow cytometery for RFP expression into a pooled population and validated knockdown through western blot. The plasmids used in the part of the project were obtained through Addgene as a gift from Jonathan Weissman.

WOUND HEALING ASSAY

Wound healing assays were performed similarly for all cell lines in the manuscript. Cells were grown to 100% confluence in a six well plate then a wound was created in the middle of the plate. Cells were allowed to close the wound for 24 hours in the presence of Mitomycin C growth inhibitor then the cells were imaged. Images were quantified for the amount of migration into the wound using ImageJ.

TAIL VEIN INJECTION

NDL2-5 *Chad* and *Col1a1* knockout cell lines were injected into the tail vein of syngeneic FVB/NJ mice. Cell were suspended in PBS in a single cell population and injected in a single bolus of 500x10^5 cells in 50uL. Mice were monitored for 9 weeks then euthanized. At this point, lungs were collected and stained with Hematoxylin and Eosin to identify the presence of pulmonary metastases.

MAMMARY FAT PAD INJECTION

NDL2-5 WT and cell lines were suspended in PBS and injected into mammary gland number four in syngeneic FVB/NJ mice as a single bolus of 1x10^6 cells. The mice were monitored twice weekly until tumors reached an endpoint of 2000 mm^3 in diameter.

BT474 wildype and CHAD/COL1A1 knockout lines were suspended in a 1:1 concentration of matrigel:PBS mixture and injecting into the mammary gland number four in a single bolus of 1x10^6 cells. Balb/C nude mice were used for these studies. Tumors were monitored until a size of 1000 mm^3 in diameter. Tumors were then resected, and mice were monitored for an additional four weeks. At necropsy lungs were imaged for RFP using the IVIS imaging system and then processed for hematoxylin and eosin staining.

HUMAN DATASET USAGE

All human datasets used in this study are publicly available and noted as used in the manuscript. For genomic alteration frequency the TCGA Breast cancer[18,19] and the TCGA-pan Lung cancer[33] datasets were used. For the expression based survival data the KMPlot.com dataset[34] was used.

WESTERN BLOTTING

Western blots in this manuscript were completed under manufacturer's specifications. Blocking was performed for 1 hour by incubation at room temperature with the LiCor blocking reagents. Western blots were imaged using the LiCor system. The following antibodies were used: COL1A1 (Origene TA309096), CHAD (Abcam ab104757), HSP90 (CST 4874S), Beta-tubulin (CST 2128S), anti-rabbit secondary (Licor 926-32211), anti-mouse secondary (Licor 926-68070)

AKT SENSITIVITY EXPERIMENTS

CCLE[35] data was download and separated into HER2+ lines with 17q21.33 amplification and those without. The Achilles[23] data was filtered to those samples and the top 200 differentially lethal siRNA's were identified between the two subgroups. For the PDX[24] analysis a similar approach was taken with the division of HER2+ tumors into the two subgroups and differentially lethal compounds were identified.

APPENDIX

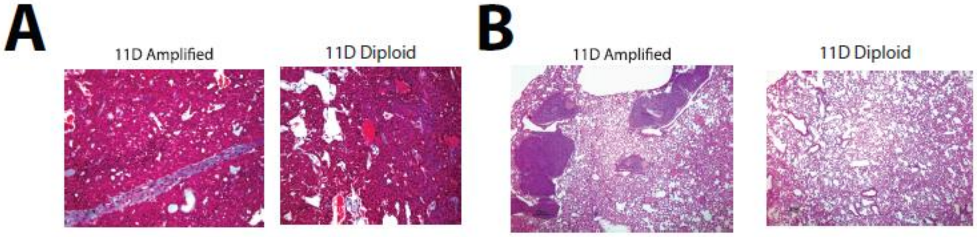

Figure 4.1: Association of ECM changes and metastatic potential changes with 11D CNV

A key copy number alteration in the 11D region encompassing the *Col1a1* gene was observed to correlate with reduction and lack of collagen alignment in Masons trichrome staining (C) and an increase in metastases in the lungs of mice with *Col1a1* amplification in the primary tumors (D).

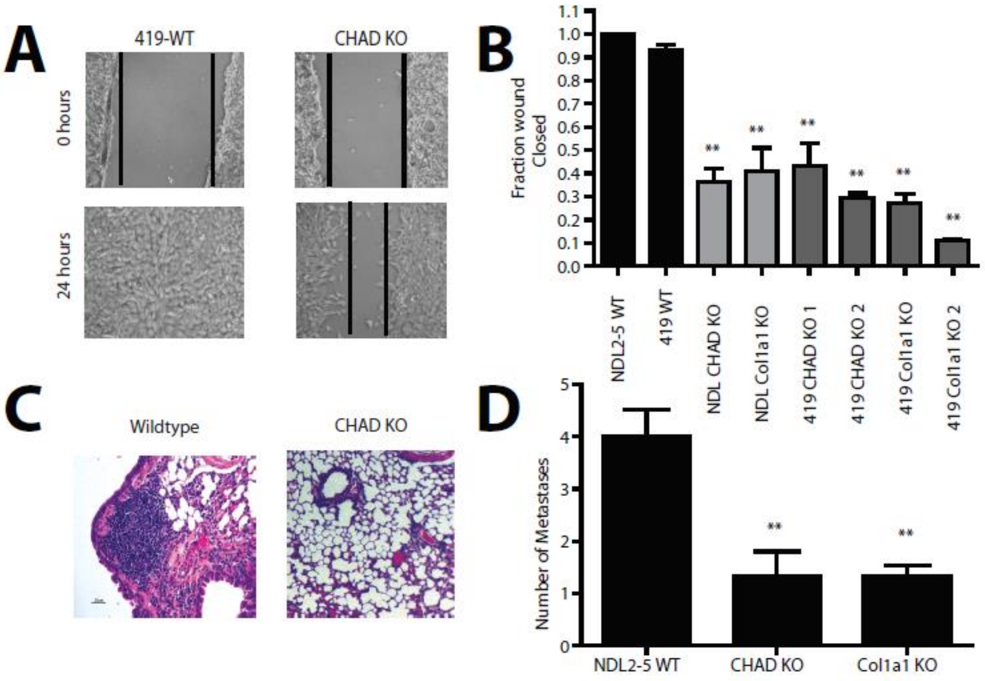

Figure 4.2: Loss of Col1a1 or CHAD effects migration in vitro and lung colonization in vivo in mouse mammary tumor cell line

CRISPR-Cas9 mediated knockout of two key genes within this region, *Col1a1* and *Chad*, show defects in wound healing (A, B) (n=9). Knockout also impaired the ability to colonize the lung through a tail vein injection (C, D) (WT n=12, Chad KO n=9, Col1a1 KO n=6). (**=P<.01 for D)

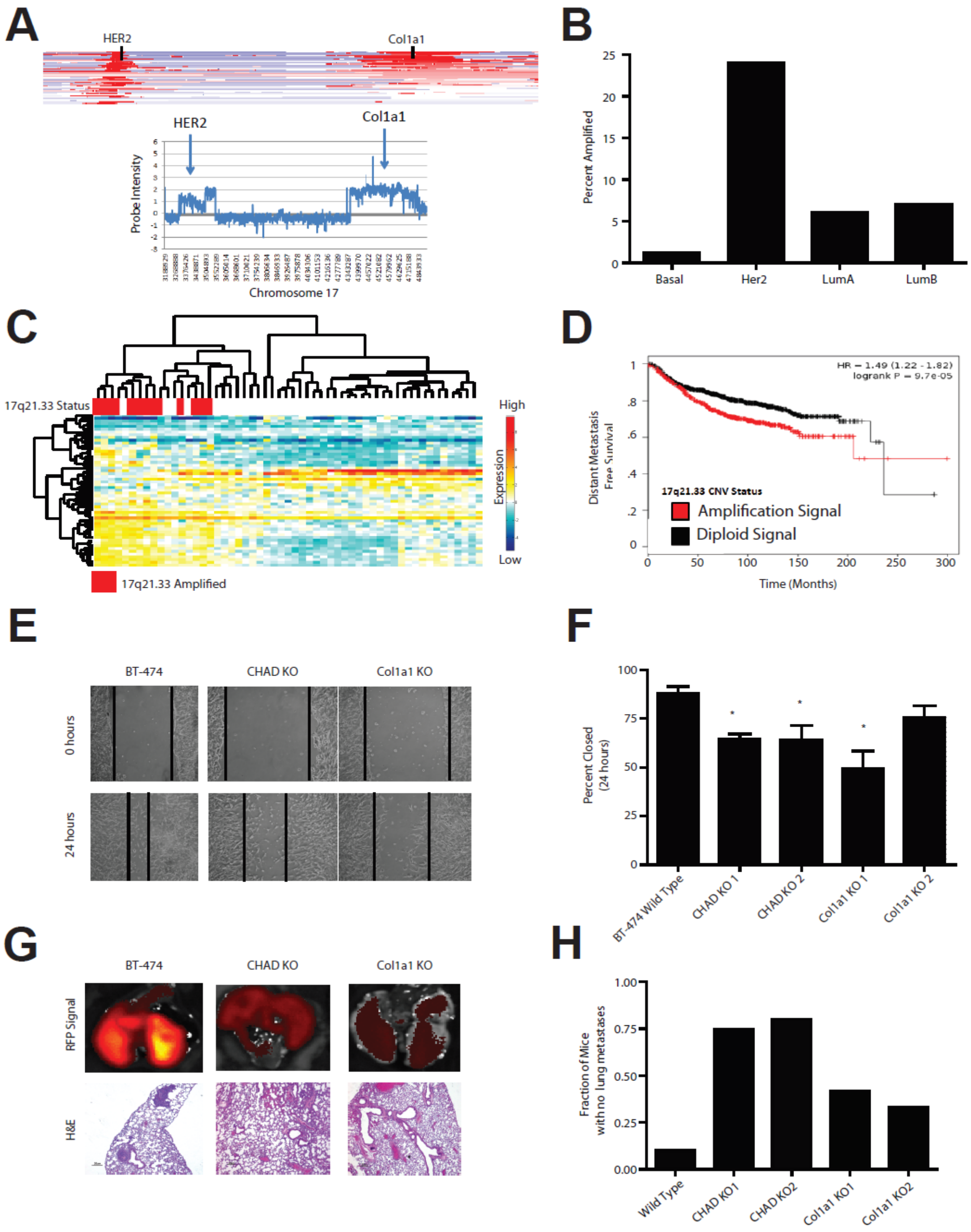

Figure 4.3: 11D amplicon presence and function is conserved in human breast cancer

Figure 4.3 (cont'd)

TCGA breast cancer copy number dataset analysis revealed co-amplification of HER2 and the COL1A1 locus through a heatmap across chromosome 17 of multiple samples with each row representing an independent patient sample (A-top) with red representing amplification and blue representing deletion. The COL1A1 amplification event occurred independently of HER2 (A-bottom) as identified by probe intensity of aCGH data of a single TCGA breast cancer patient. The COL1A1/CHAD amplification event was disproportionately found in HER2 positive tumors and is present in approximately 25% of HER2 positive tumors (B). Gene expression of HER2+ samples with and without the Col1A1 17q21.33 amplification demonstrated a unique gene expression profile as identified by unsupervised hierarchical clustering (C) and overall survival within the KMplotter dataset (P<.001) (D). CRISPRi mediated knockdown of CHAD and COL1A1 in human cell line BT-474 resulted in defects in wound healing (F, G) (*=p<.05, n=9) and distant metastasis to the lung after orthotopic injections (H, I n=10, for WT, n= 4 for CHAD KO1, n = 5 for CHAD KO2 n= 12 for Col1a1 KO1, n=6 for Col1a1 KO2).

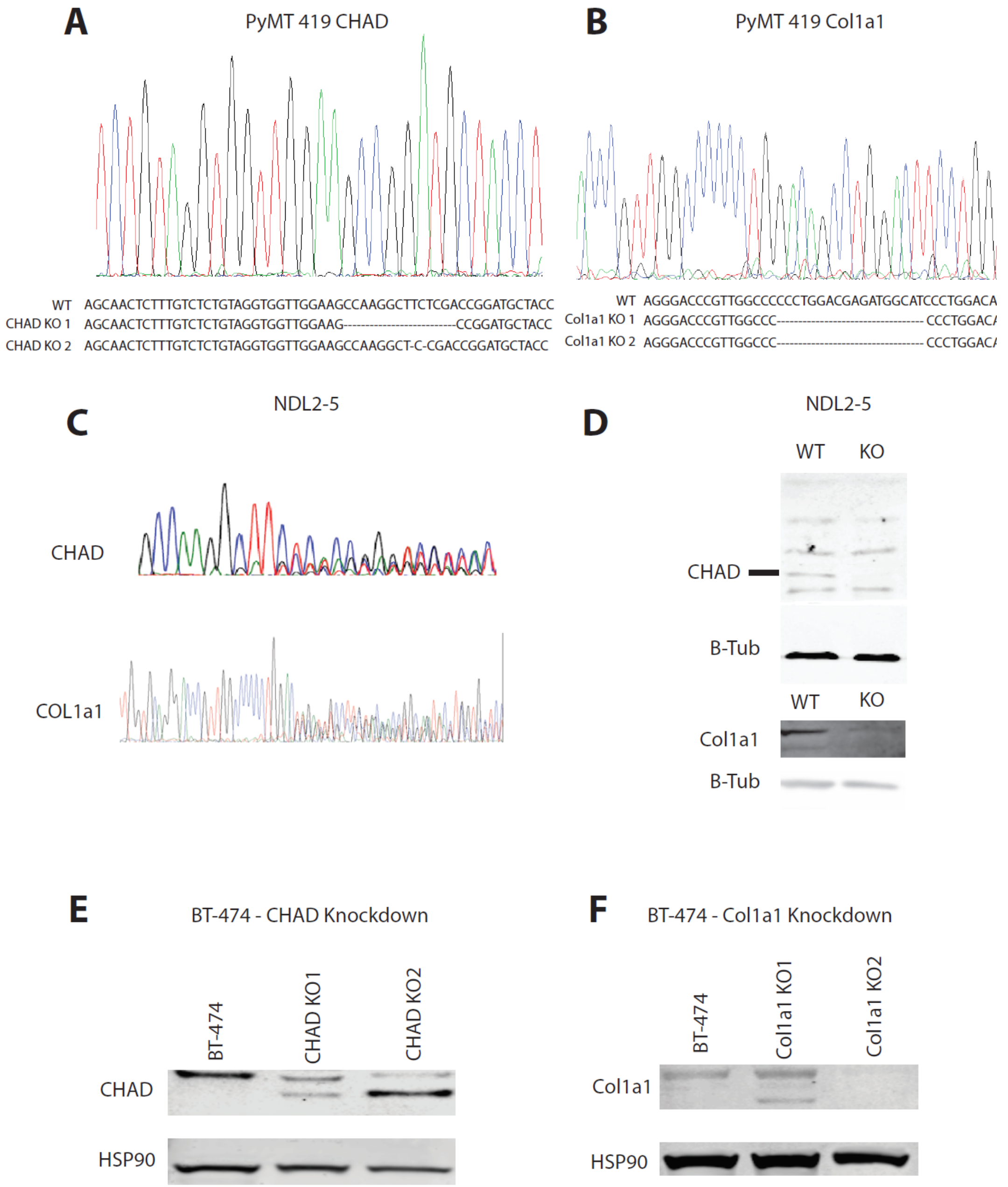

Figure 4.4: Confirmation of Col1a1 and CHAD knockout in PyMT 419, NDL2-5, and BT-474 cell lines

Sanger sequencing of CHAD (A) and Col1a1 (B) KO clones revealed the production of indels within the coding sequence of each protein within the PyMT 419 line. This is also the case

Figure 4.4 (cont'd)

where multiple different indels were shown in the Col1a1/CHAD amplified cell line NDL2-5 (C). The confirmation of knockdown was completed through western blot for CHAD (top) and Col1a1 (Bottom). The CRISPRi system with guides against early exons of CHAD and Col1a1 was used to generate knockdowns of the respective genes in the human HER2 positive, COL1A1/CHAD amplified line BT474. The efficiency of knockdown in the pooled population was assessed through western blot for CHAD (E) and COL1A1 (F).

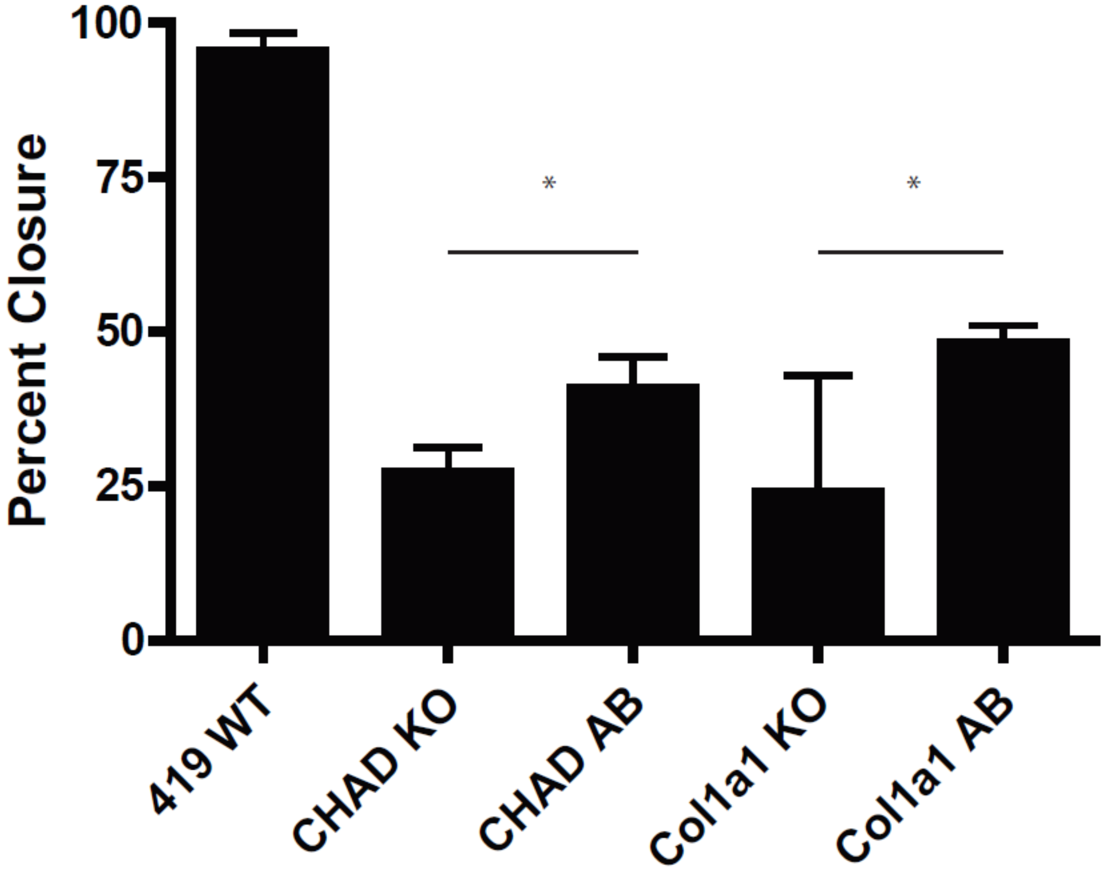

Figure 4.5: Addback of Col1a1 and CHAD in PyMT 419 cell lines

The addback of wildype *Col1a1* and *Chad* into the CRISPR generated knockout lines showed partial recovery of movement in a scratch assay (*=P<.05).

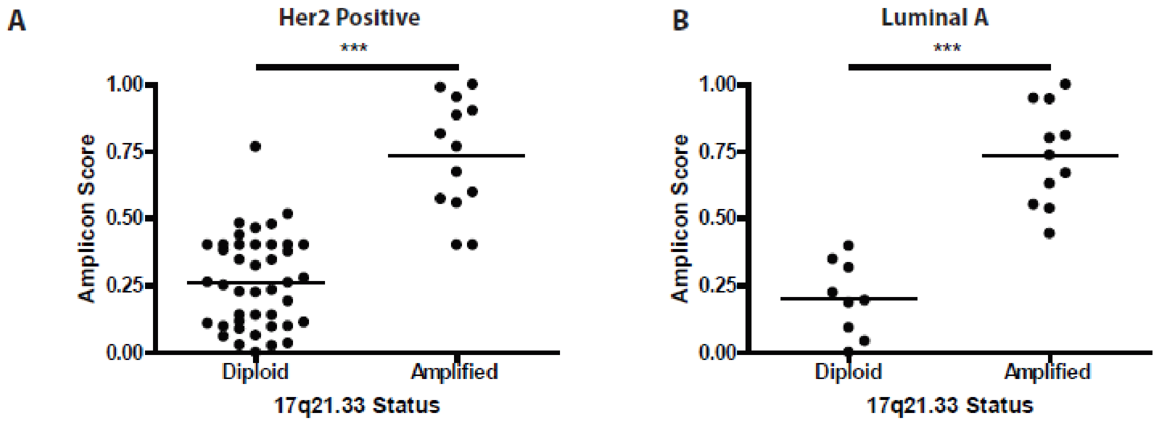

Figure 4.6: Validation of COL1A1/CHAD amplicon gene expression signature

A score between 0 (diploid) and 1 (amplified) was generated for the predicted presence of the COL1A1/CHAD amplification event based upon a weighted gene expression data. This signature showed a robust prediction of the amplification event in both the training HER2 positive dataset (A) and the Luminal A validation cohort (B)

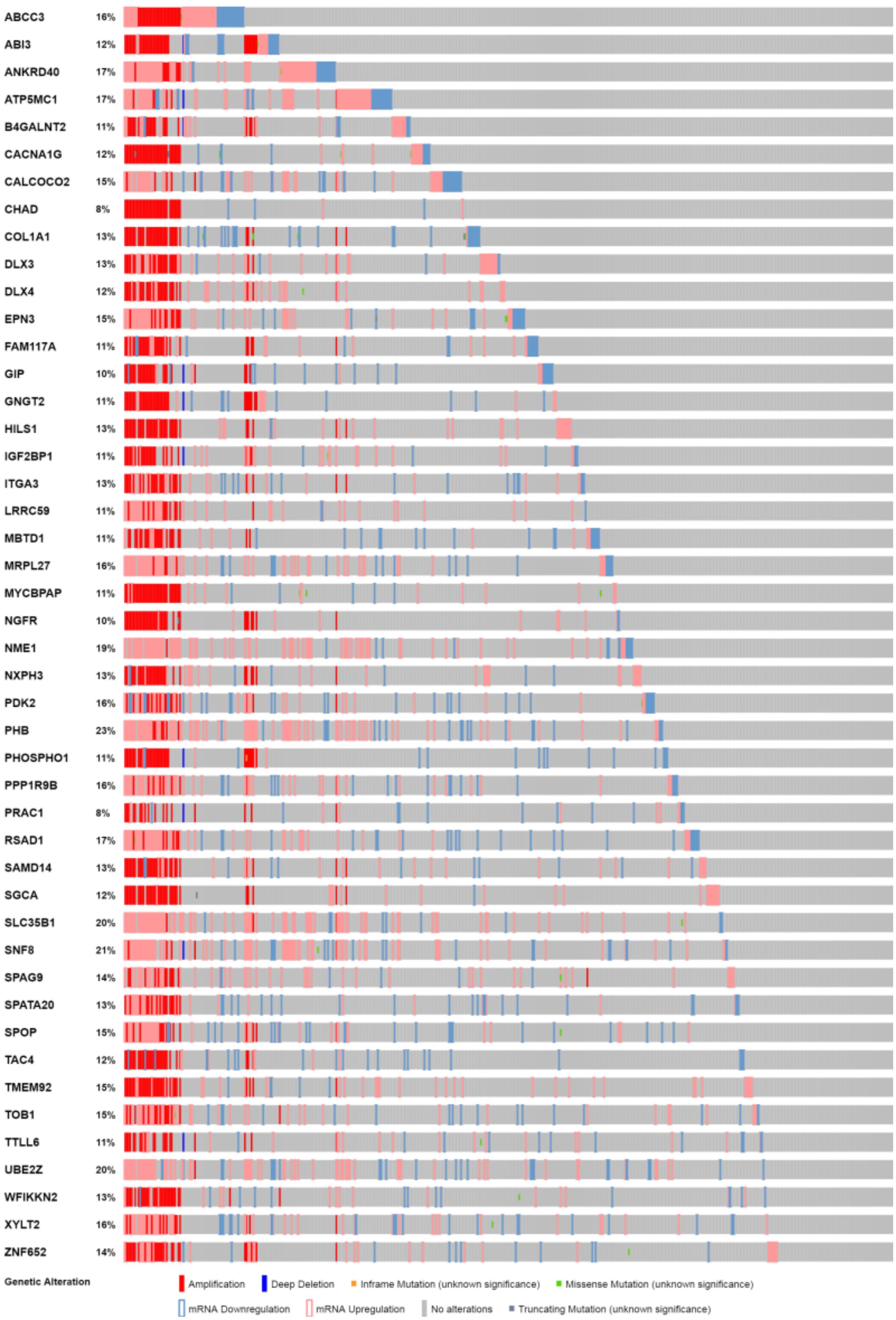

Figure 4.7: 17q21.33 Oncoprint

Figure 4.7 (cont'd)

An oncoprint showing all the genes within the amplification event shows co-amplification of approximately 40 genes. Many of these genes show an upregulation of RNA to accompany the amplification at the DNA level

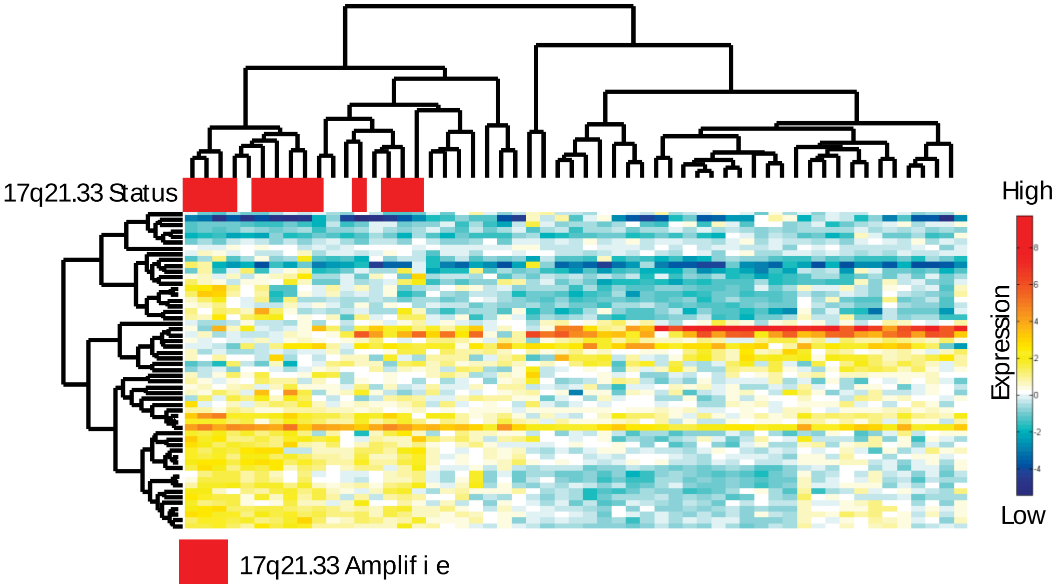

Figure 4.8: Gene expression impact of 17q21.33

Consistent gene expression changes were identified with the presence of the 17q21.33 amplification event (Red) through unsupervised hierarchical clustering. Gene expression data is color coded as shown in the color bar with high being red and low being blue.

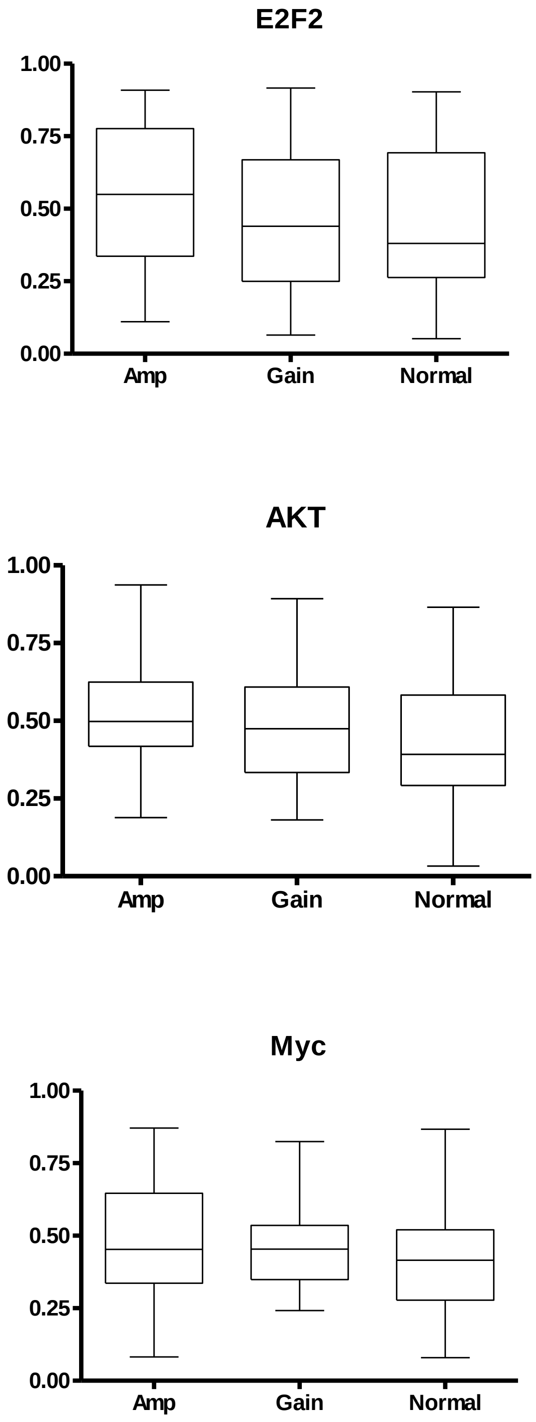

Figure 4.9: Key signaling changes

Significant differences were seen between amplified and diploid pathway activity signatures in E2F2 (Top), AKT (Middle), and Myc (Bottom) (P<.05)

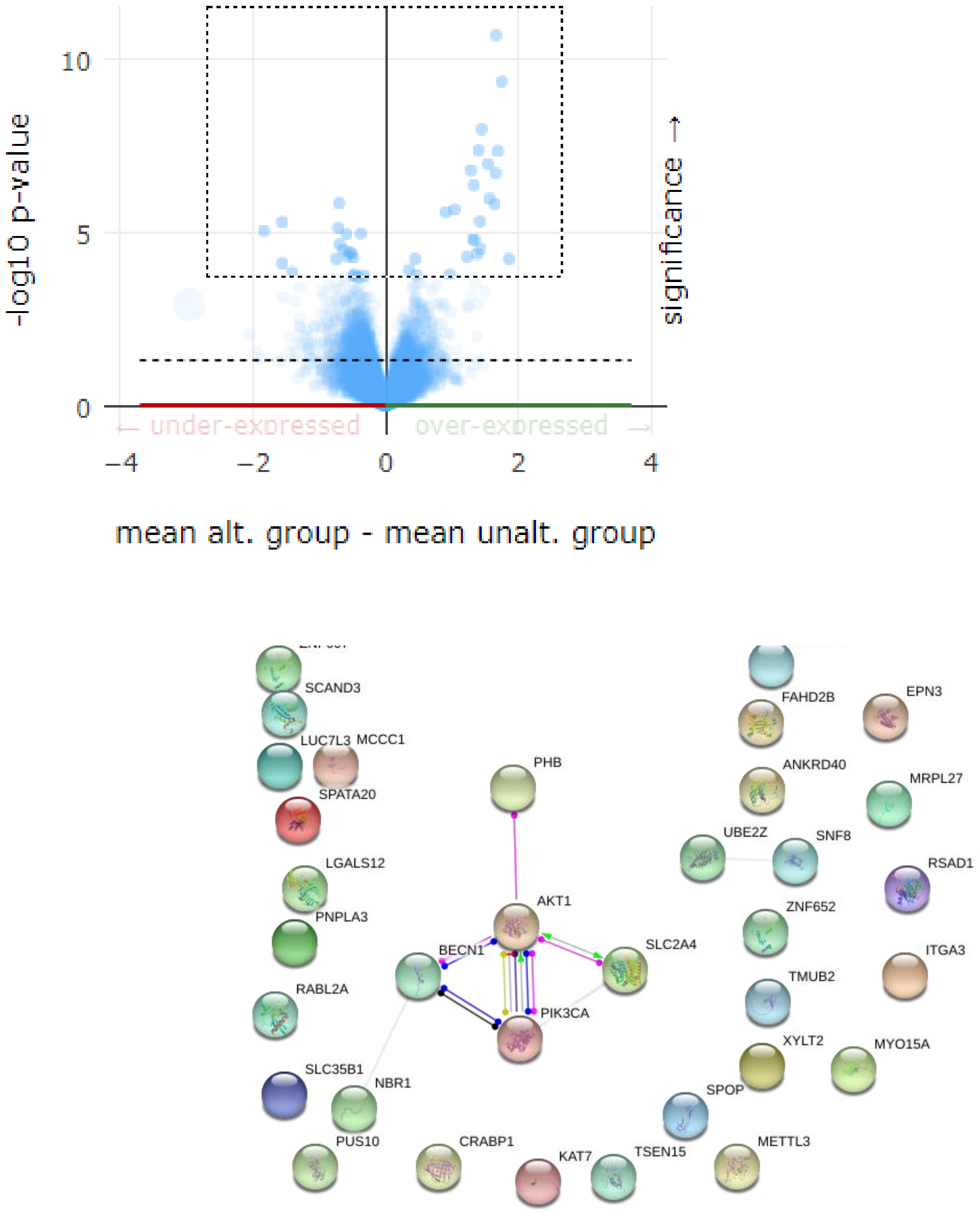

Figure 4.10: 17q21.33 Amplified gene expression correlation

Genes which had expression correlated with Col1A1/CHAD amplification within the TCGA breast cancer cohort were identified through the use of cbio portal (top). Many of these genes

Figure 4.10 (cont'd)

were shown to be unrelated; however, a key network of genes revolving around AKT was identified including PHB, SLC2A4, and Beclin (bottom)

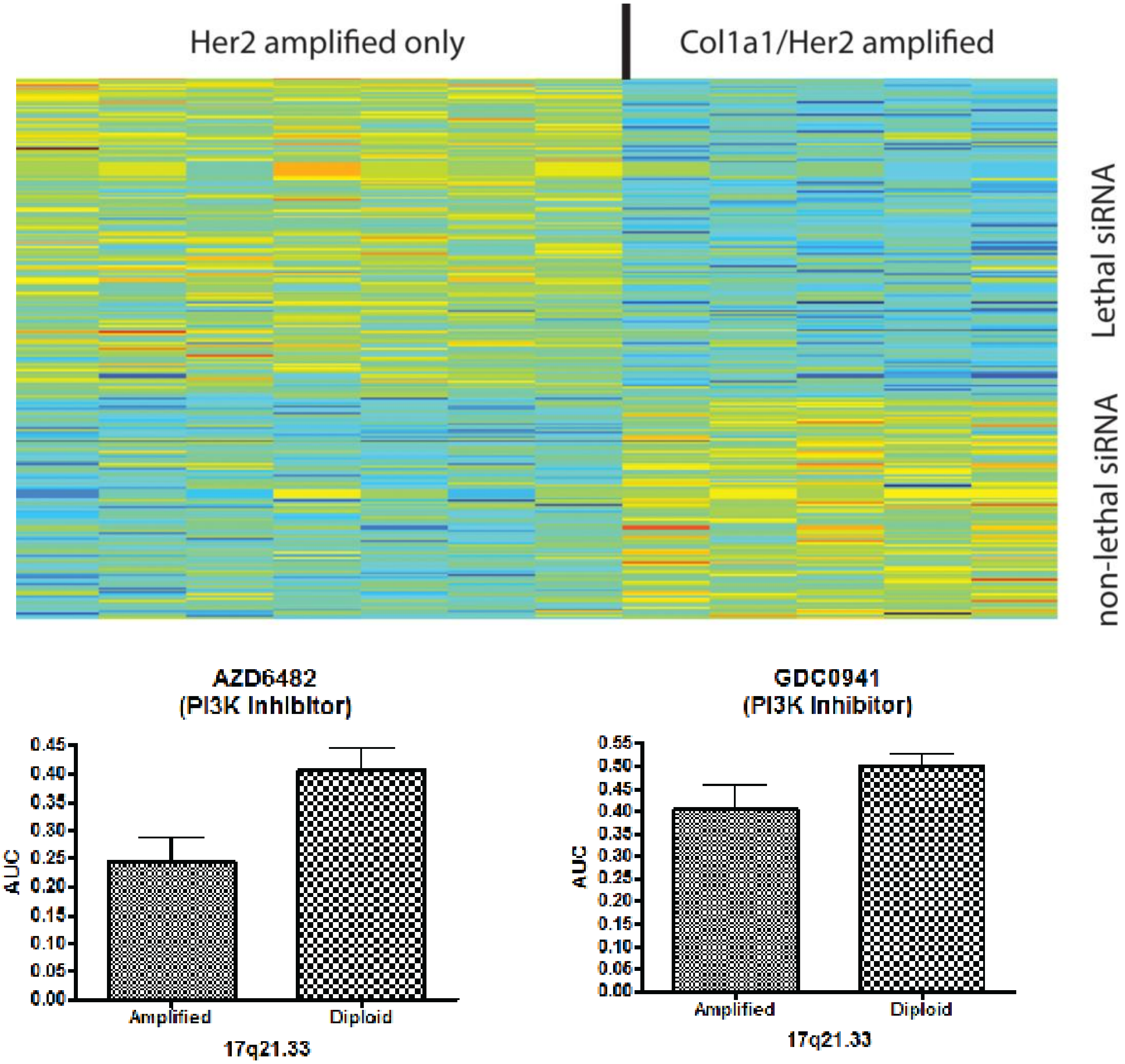

Figure 4.11: Targeting of Col1a1/CHAD amplicon

Use of the Achilles data identified preferentially lethal siRNA target genes between HER2 amplified and HER2/Col1a1 Amplified cell lines (top). The proliferation score is color coded with Red being highly proliferative (non-lethal) and Blue being lowly proliferative (Lethal). This data was followed up on in PDX drug studies where it was identified that 17q21.33 amplified lines were more sensitive to PI3K inhibition (bottom)

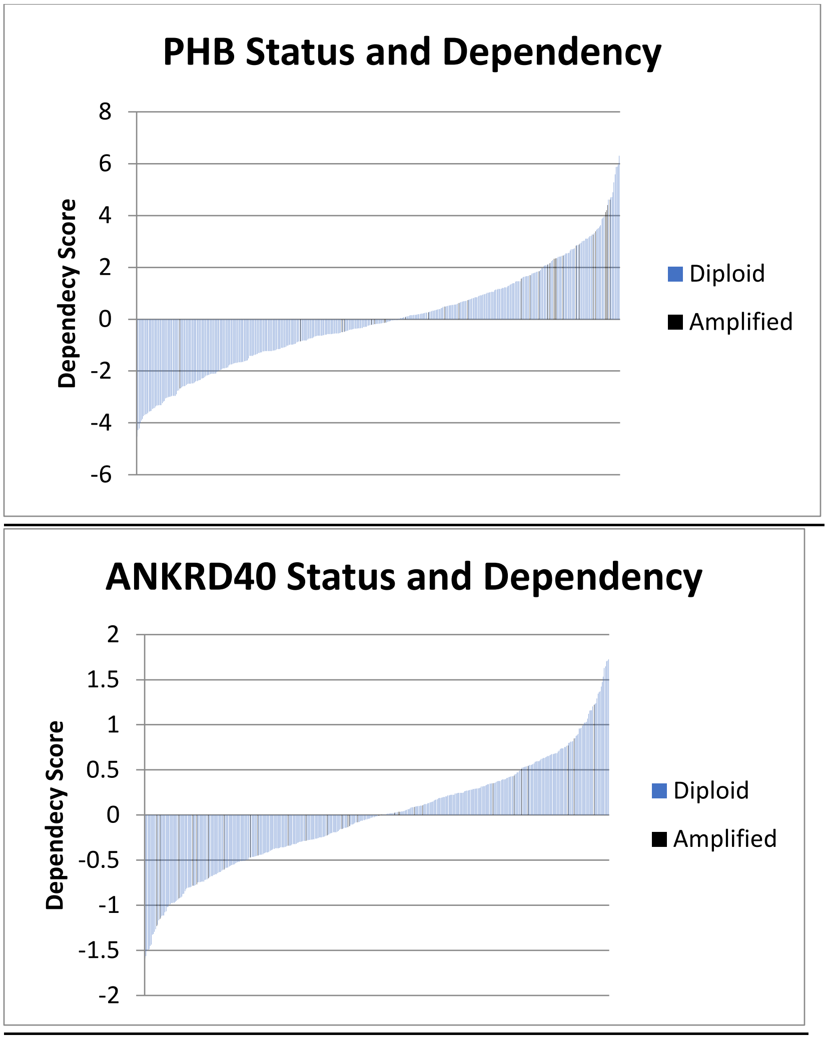

Figure 4.12: 17q21.33 correlated genes and dependency

A ranking of CCLE cell lines from low to high of dependency score shows differential survival depending upon amplification status of 17q21.33. PHB knockdown shows a large increase in cell viability when in an amplified setting (Red) but not in a diploid (Blue) setting (top). This is not the case for other 17q21.33 correlated genes such as ANKRD40 (bottom)

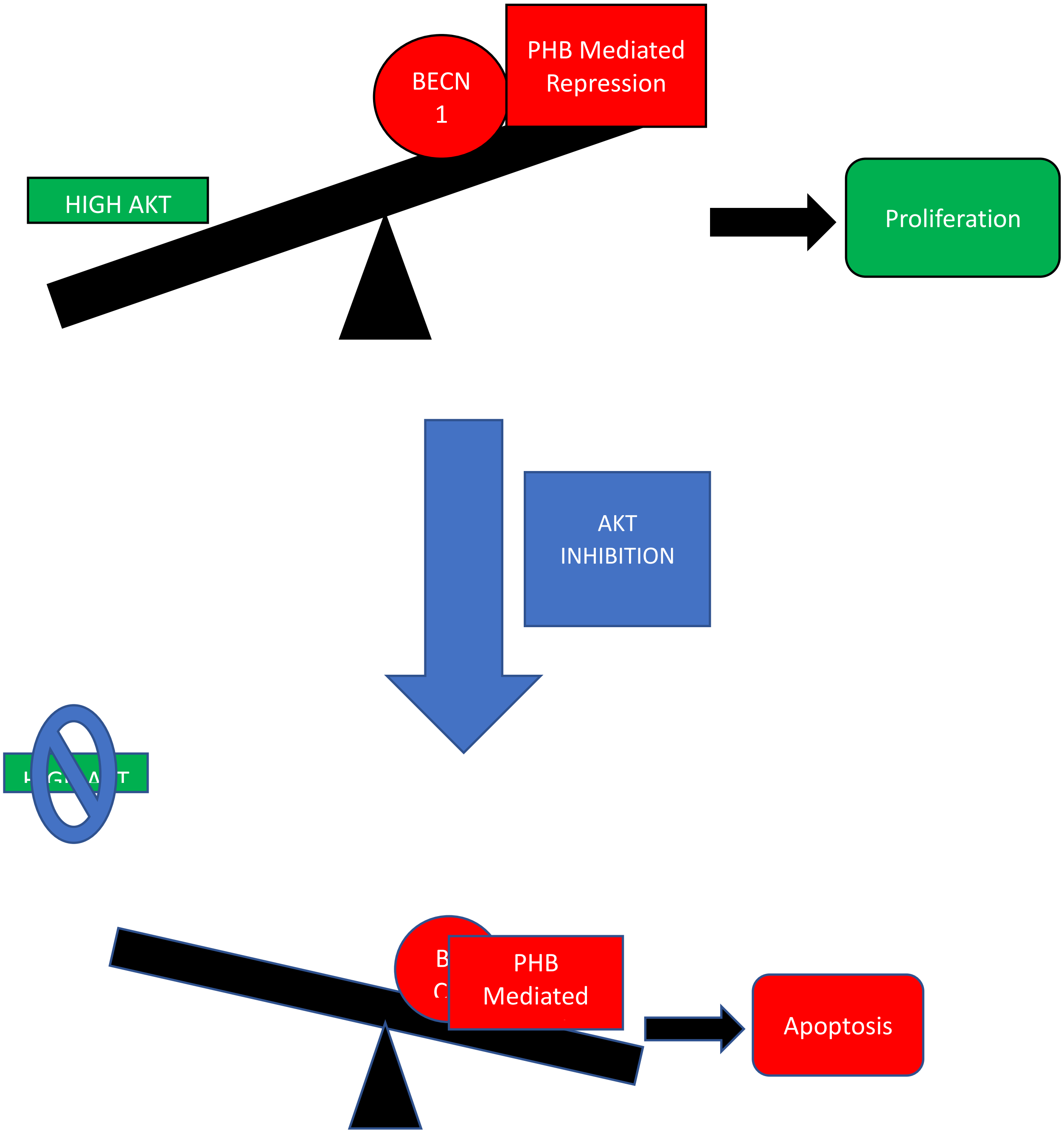

Figure 4.13: Working model of AKT sensitivity

Figure 4.13 (cont'd)

In a HER2/Col1a1 amplified setting High AKT signaling is able to overcome PHB and Beclin mediated apoptosis (top). However, in the presence of AKT inhibition PHB mediated apoptosis occurs and kills the tumor cells (bottom).

WORKS CITED

WORKS CITED

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

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

3. Lin, N. U. & Winer, E. P. Brain metastases: the HER2 paradigm. *Clin. Cancer Res.* **13,** 1648–55 (2007).

4. Stetler-Stevenson, W. G., Aznavoorian, S. & Liotta, L. A. Tumor Cell Interactions with the Extracellular Matrix During Invasion and Metastasis. *Annu. Rev. Cell Biol.* **9,** 541–573 (1993).

5. Raines, E. W. The extracellular matrix can regulate vascular cell migration, proliferation, and survival: relationships to vascular disease. *Int. J. Exp. Pathol.* **81,** 173–182 (2001).

6. Levental, K. R. *et al.* Matrix crosslinking forces tumor progression by enhancing integrin signaling. *Cell* **139,** 891–906 (2009).

7. Deryugina, E. I. & Quigley, J. P. Matrix metalloproteinases and tumor metastasis. *Cancer Metastasis Rev.* **25,** 9–34 (2006).

8. Tsukita, S., Tsukita, S., Nagafuchi, A. & Yonemura, S. Possible involvement of adherens junction plaque proteins in tumorigenesis and metastasis. *Princess Takamatsu Symp.* **24,** 38–50 (1994).

9. Gennari, R. *et al.* Pilot study of the mechanism of action of preoperative trastuzumab in patients with primary operable breast tumors overexpressing HER2. *Clin. Cancer Res.* **10,** 5650–5 (2004).

10. Cuello, M. *et al.* Down-regulation of the erbB-2 receptor by trastuzumab (herceptin) enhances tumor necrosis factor-related apoptosis-inducing ligand-mediated apoptosis in breast and ovarian cancer cell lines that overexpress erbB-2. *Cancer Res.* **61,** 4892–900 (2001).

11. Cobleigh, M. A. *et al.* Multinational study of the efficacy and safety of humanized anti-HER2 monoclonal antibody in women who have HER2-overexpressing metastatic breast cancer that has progressed after chemotherapy for metastatic disease. *J. Clin. Oncol.* **17,** 2639–48 (1999).

12. Ahmad, S., Gupta, S., Kumar, R., Varshney, G. C. & Raghava, G. P. S. Herceptin resistance database for understanding mechanism of resistance in breast cancer patients. *Sci. Rep.*

4, 4483 (2014).

13. Luque-Cabal, M., García-Teijido, P., Fernández-Pérez, Y., Sánchez-Lorenzo, L. & Palacio-Vázquez, I. Mechanisms Behind the Resistance to Trastuzumab in HER2-Amplified Breast Cancer and Strategies to Overcome It. *Clin. Med. Insights. Oncol.* **10,** 21–30 (2016).

14. Seol, H. *et al.* Intratumoral heterogeneity of HER2 gene amplification in breast cancer: its clinicopathological significance. *Mod. Pathol.* **25,** 938–948 (2012).

15. Staaf, J. *et al.* High-resolution genomic and expression analyses of copy number alterations in HER2-amplified breast cancer. *Breast Cancer Res.* **12,** R25 (2010).

16. Siegel, P. M., Ryan, E. D., Cardiff, R. D. & Muller, W. J. Elevated expression of activated forms of Neu/ErbB-2 and ErbB-3 are involved in the induction of mammary tumors in transgenic mice: implications for human breast cancer. *EMBO J.* **18,** 2149–64 (1999).

17. Ma, J. *et al.* Characterization of mammary cancer stem cells in the MMTV-PyMT mouse model. *Tumor Biol.* **33,** 1983–1996 (2012).

18. TCGA. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

19. Ciriello, G. *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163,** 506–519 (2015).

20. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).

21. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* **8,** 2180–2196 (2013).

22. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39,** D561–D568 (2011).

23. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170,** 564–576.e16 (2017).

24. Pauli, C. *et al.* Personalized *In Vitro* and *In Vivo* Cancer Models to Guide Precision Medicine. *Cancer Discov.* **7,** 462–477 (2017).

25. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483,** 570–5 (2012).

26. Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41,** D955-61 (2013).

27. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166,** 740–754

(2016).

28. Tokunaga, E. *et al.* Akt is frequently activated in HER2/neu-positive breast cancers and associated with poor prognosis among hormone-treated patients. *Int. J. Cancer* **118,** 284–289 (2006).

29. Fan, W. *et al.* Prohibitin 1 suppresses liver cancer tumorigenesis in mice and human hepatocellular and cholangiocarcinoma cells. *Hepatology* **65,** 1249–1266 (2017).

30. Hudis, C. *et al.* A phase 1 study evaluating the combination of an allosteric AKT inhibitor (MK-2206) and trastuzumab in patients with HER2-positive solid tumors. *Breast Cancer Res.* **15,** R110 (2013).

31. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8,** 2281–2308 (2013).

32. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159,** 647–661 (2014).

33. Campbell, J. D. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48,** 607–616 (2016).

34. Györffy, 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.* **123,** 725–31 (2010).

35. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–7 (2012).

CHAPTER 5

CONSERVED MUTATIONS OF *PTPRH* IN MMTV-PYMT TUMORS

ABSTRACT

Receptor tyrosine kinases play a critical role in the development and progression of human cancers. A key protein in many cancers of this type is EGFR where uncontrolled activation of the protein results in uncontrolled cell proliferation. The activity of EGFR is tightly regulated by phosotyrosine receptor phosphatases. Here we identify a conserved mutation in one such protein, phosphotyrosine receptor phosphatase type H (*Ptprh*). The mutation is highly conserved in mouse models of breast cancer and is identified to be mutant in 81% of MMTV-PyMT tumors. Also, we identified the mutation to be present in a number of human tumors including Ovarian, Head and Neck, and lung tumors. A key finding is that *Ptprh* mutations are associated with high EGFR activity, lower latency and more aggressive tumors in a variety of cancer types. Importantly when cell lines with the *Ptprh* mutation were compared against those without the mutation we identified an increased sensitivity to EGFR targeted therapy such as erlotinib associated with *Ptprh* mutation. We believe that this mutation acts as a dominant negative mutation and leads to uncontrolled EGFR signaling. An important impact of this finding is its immediate clinical relevance. The presence of a *Ptprh* mutation in a patient's tumor may lead to sensitivity towards EGFR targeted therapy. This would have immediate impact on patient care and potential provide a new therapeutic option for many patients. The study underscores the importance of understanding mouse models of cancer. We believe that *Ptprh* is just one of many conserved events present in mouse models and with the increase in number and breadth of models we will identify other key tumor promoting genomic events.

INTRODUCTION

A hallmark of cancer is uncontrolled growth through the self-sufficiency in growth signals[1]. This allows the replication of a tumor cell without the presence of a growth factor. There have many different types of oncogenic, proliferative modifications identified that drive replication in cancer[2,3]. Many of these are receptor tyrosine kinases (RTKs) in the epidermal growth factor receptor family.

The epidermal growth factor receptor family has four family members EGFR, HER2/Neu, Her3, Her4[4]. The family members work together to promote cell proliferation. EGFR, Her3 and Her4 bind growth factors. Upon binding of their ligand, they will homodimerize or heterodimerize with other family members to lead to uncontrolled cell proliferation. Mutations or amplifications of these family members have been identified that cause uncontrolled cell proliferation in the absence of the growth factor and is the underlying cause of many cancer[5].

EGFR modifications have shown to play a key role in the development of lung cancers[6], anal cancers, glioblastoma[7], and tumors of the head and neck[8]. This dysregulation is caused by mutation or amplification of the EGFR protein in most cases. Upon binding of the growth factor, like other family members, EGFR with homo or hetero dimerize. This leads to of several intercellular tyrosine residues. In the human these are specifically Y992, Y1045, Y1068, Y1148 and Y1173[9]. The phosphorylation of these residues leads to activation of the AKT/PI3K[10], MAPK[11], or JNK[12] signaling cascades and leads to cellular division. Emerging research has identified that phosphorylation of specific residues leads to specific signaling cascades[13].

Due to its frequency of modulation and important role in cancer a number of targeted therapies have been developed to block EGFR signaling[14]. There are two main classes of EGFR targeted therapy. The first type are monoclonal antibodies raised against EGFR. These antibodies function in a number of ways by blocking binding to EGF, preventing dimerization and mediating antibody dependent cytotoxicity. A common drug from this class used to treat patients is pertuzumab[15]. The other main class of compounds are EGFR inhibitors. These are small molecule compounds which prevent the tyrosine kinase activity of the receptor. By inhibiting phosphorylation, the molecules prevent downstream signaling from occurring and stop cell growth. Some common small molecule inhibitors of EGFR are erolotinib[16], gefitinib[17], and afatinib[18].

An important class of regulatory proteins of EGFR and its family members are phosphotyrosine receptor (ptpr) proteins. This protein family has 21 members and plays a critical role in regulating a number of cellular processes including cell growth and differentiation[19]. Recent work has shown an emerging role of PTPRs in cancers where misregulation of the protein is present in the tumor but not the surrounding normal tissue[20]. The Ptprs vary in structure and thus vary in function with each one having a unique profile of EGFR family member that it is responsible to bind to, and tyrosine residues to de-phosphorylates. Ptprs work in a dimer system in which homo or heterodimer must form in order for the phosphatase activity to occur[21].

As expected, many mouse models of breast cancer that rely on EGFR signaling have been developed. Most common dysregulation of EGFR results in mouse models of lung cancer[22]. However, in a highly aggressive mouse model of breast cancer, MMTV-PyMT, it has been shown that EGFR is highly active and the model is dependent upon EGFR signaling for progression[23]. This is surprising due to the fact that breast cancer is not traditionally EGFR driven. Furthermore, the mechanism of how EGFR signaling is upregulated in this model is unknown. To date, EGFR amplification or mutations have not been found in this model.

Here we present an integrative whole genome and transcriptome analysis of MMTV-PyMT tumors to identify underlying drivers of tumorigeneisis in this model. Strikingly we identified a previously unknown mutation in *PTPRH* in 81% of PyMT tumors. Importantly we showed that similar mutations were present in around 5% of lung cancer patients. The tumors with *PTPRH* mutation were shown to be more aggressive than those with wildtype tumors with increased EGFR signaling and lower latency in the mouse as well as worse outcomes in the human. It was also found that tumors with *PTPRH* mutations with *Ptprh* mutations are more responsive to EGFR targeted therapy like erolotinib.

Taken together these findings have impacts in both the research community and also in the clinic. In the research community, this work gives an increased understanding of the tumorigenesis of the MMTV-PyMT tumors. This will help to translate the findings from the MMTV-PyMT into the appropriate patient population. The most immediate impact of this work is that the identification of a *Ptprh* mutation will lead to new treatment options for a number of lung cancer patients. These patients would otherwise be faced with traditional chemotherapy and would have worse outcomes. We believe that this study underscores the importance of understanding mouse models of cancer and shows the importance of increase the number and diversity of mouse models that have been profiled.

RESULTS

IDENTIFICATION OF *PTPRH* MUTATIONS

Ptprh was found to be mutated in 81% of MMTV-PyMT tumors. Furthermore, the *Ptprh* mutation was shown to be homozygously mutated in 21% of PyMT tumors and heterozygously mutated in 60% of PyMT tumors (Figure 4.1A and 4.1B). Surprisingly, an identical C to T mutation was observed in each tumor resulting in a valine residue being converted to a methionine at amino acid 483 (V483M). To test for the conservation of mutations of *Ptprh* in mouse strains beyond FVB/NJ, we sequenced *Ptprh* of MMTV-PyMT models in a C57/Bl6, C57/Bl10, CAST, and MOLF backgrounds as well as a different inbred MMTV-PyMT FVB /NJ line. This analysis showed consistent mutation in the structural fibronectin domains (FN3) and the phosphatase domain of *Ptprh* (Figure 4.1C). Interestingly we found that the two FVB models contained different mutational patterns indicating an impact of environmental and potential epigenetic causes of mutational hotspots.

PTPRH MUTATION MODULATES EGFR SIGNALING

Given that recent work identified the target of *PTPRH* as EGFR[20], we hypothesized that EGFR was not dephosphorylated with *Ptprh* mutation. Testing this, we observed that the V483M mutation correlated with pEGFR levels (Figure 4.2A and Figure 4.2B). With the resulting increase in EGFR activity, we also observed a significant decrease in tumor latency (Figure 4.2C). With an increase in EGFR activity, it was possible that tumors with mutant *Ptprh* would be dependent upon EGFR signaling. To test this prediction, cell lines derived from *Ptprh* wildtype and mutant tumors were treated with EGFR targeted therapy. After 48 hours, tumors containing *Ptprh* mutations were shown to be more sensitive to erlotinib treatment (Figure 4.2D).

Given the role of EGFR in lung cancer, we next sought to determine if there was a non-EGFR mutant patient population within lung cancer that could benefit from EGFR inhibition. Examination of the pan–lung TCGA data revealed 5% of patients with a mutation in *PTPRH* (Figure 4.3A). Importantly, these mutations were shown to be mutually exclusive from EGFR, indicating that patients were likely not treated with EGFR tyrosine kinase inhibitors. To confirm the impact of *PTPRH* mutations on EGFR activity in human lung tumors we used gene set enrichment analysis to predict EGFR activity of each mutant *PTPRH* sample. This analysis revealed four key hotspots of mutations driving high EGFR activity, including three in the FN3 domains and one in the phosphatase domain of *PTPRH* (Figure 4.3B).

PAN-CANCER IMPACTS OF *PTPRH* LOSS

To identify if *PTPRH* is altered in other tumor types beside lung we utilized the TCGA data from various tumor types (Figure 4.4A). It was seen that lung cancer had the highest number of samples with *Ptprh* mutation. Surprisingly, it was shown that a high degree of tumor samples had a heterogeneous loss of *Ptprh*. This includes a significant portion of ovarian and head and neck tumors. Interestingly both of these tumor types have been shown to have subsets of tumors which respond to EGFR targeted therapy. We also identified a protective role of high expression of *Ptprh* in breast (Figure 4.4B), ovarian (Figure 4.4C), stomach (Figure 4.4D) and liver (Figure 4.4E) cancer with regards to overall and relapse free survival. This confirms a pan-cancer protective role of *Ptprh*. Bioinfomatic analysis of the ovarian tumors and head and neck tumors however, did not reveal activation of the Egfr pathway. However, this could be due to the fact that the gene sets did not fully represent the specific type of Egfr signaling induced by *Ptprh* loss or other activators of Egfr in those tumors.

To identify if *Ptprh* mutation lead to specific Egfr signaling we used the breast cancer dataset. Despite the small number of samples with *Ptprh* mutations in breast cancer we utilized this dataset for the informatic extensive informatic tools which have been validated in it. Specifically, there are pathway activity signatures that would give insight into the downstream pathways of Egfr which are active. We hypothesized that different alterations in different Ptprs will lead to different downstream pathway activation. We pursued this hypothesis through the use of unsupervised hierarchical clustering of pathway activity and identified specific alterations with specific pathway activity associated (Figure 5.5A). With regards to *Ptprh* mutation we identified that the AKT/PI3K pathway was activated downstream of Egfr indicating a specific role of *Ptprh* in de-phosorylation and subsequent signaling (Figure 5.5B).

DISCUSSION

Here we have identified a conserved mutation in *Ptprh* in 81% of the MMTV-PyMT mouse model which is conserved in approximately 5% of lung cancer patients. We confirmed that mutations of *Ptprh* is associated with higher EGFR activity and more aggressive tumors in a variety of cancer types. The most immediate clinical impact is the identification of a new patient population that may respond to EGFR targeted therapy such as erlotitinib. We identified this through the isolation of *Ptprh* mutant and wildtype cell lines.

We identified specific pathway regulation associated with *Ptprh* mutation this indicates that *Ptprh* is responsible for de-phosphorylation of specific tyrosine residues that lead to downstream AKT/PI3K signaling. Preliminary evidence states that other phosphor-tyrosine receptor phosphatases regulate additional signaling cascades and could potentially be markers of other therapeutic response much like *Ptprh* is of erlotinib sensitivity.

A key remaining question is if mutations in *Ptprh* work in a haploinsuffienct manner or in a dominant negative manner. Based upon our copy number data we identify that a one copy number loss of *Ptprh* does not lead to changes in EGFR activity. However, a heterozygous mutation does result in increased EGFR activity. This indicates a dominant negative loss of function mutation. The dominant negative function limits the therapeutic potential of overexpression of *Ptprh* through gene therapy. Any therapeutic involving *Ptprh* would have to be an edit of the mutated gene back to wildtype. This could perhaps be a target for emerging CRISPR based gene editing therapy.

Likely, the utility of the *Ptprh* clinically is as a biomarker of response to erlotinib treatment. More work must be completed to use *Ptprh* as a biomarker. Due to the lack of a hotspot mutation, traditional sequencing of the tumor of circulating DNA will not be possible. However, we believe that due to the consistent signaling changes associated with *Ptprh* a signaling based approach may be more beneficial to identifying those patient populations with *Ptprh* mutations. We predict that this may be feasible through a gene signature approach or an IHC stain of pEGFR.

This research underscores the importance and direct translatability of understanding the genomic landscape of mouse models of breast cancer. In addition, to its obvious benefit to the research community understanding mouse models have an impact in clinical care as well. Mouse models present a relatively stable tumor and only have selective pressure for extremely important events. Many times, these events are conserved in human tumors as well. We believe that this study is a proof of concept work and many events other events beyond *Ptprh* mutation will be identified with the increase in number and diversity of mouse models profiled.

MATERIALS AND METHODS

ANIMAL STUDIES

All animal husbandry and use was conducted according to local, national and institutional guidelines. The MMTV-PyMT[5] mice were in the FVB background. MMTV-PyMT634 and MMTV-Neu mice were obtained from The Jackson Laboratory. Mice were monitored twice weekly for tumor initiation and growth. At a 2000 mm^3 endpoint, mice were necropsied. For mice with multiple tumors the endpoint was established when the primary tumor was at 2000 mm^3.Tumors and lungs were collected for genomic analysis, hematoxylin and eosin staining for histological subtyping and presence of pulmonary metastases. The number of metastasis was quantified using a single cut through the lung and count of the number of micro-metastases in that plane. Masson's trichrome staining was used to examine tumors for collagen deposition using standard methods.

WESTERN BLOTTING

Western blots in this manuscript were completed under manufacturer's specifications. Blocking was performed for 1 hour by incubation at room temperature with the LiCor blocking reagents. Western blots were imaged using the LiCor system. The following antibodies were used: EGFR (CST D38B1), pEGFR (Invitrogen PA5-37553), Beta-tubulin (CST 2128S), anti-rabbit secondary (Licor 926-32211), anti-mouse secondary (Licor 926-68070)

ERLOTINIB SENSITIVITY ASSAY

Cell lines derived from *Ptprh* mutant and wildtype tumors were seeded at a concentration of 250 cells/mL and subjected to erlotinib treatment for 48 hours with the concentrations stated in the manuscript. Eroltinib was purchased from Cayman Chemical. After treatment with erlotinib or DMSO control, cells were given fresh media to grow for 7 days. Cells were then fixed and stained with crystal violet for counting.

HUMAN DATASET

The TCGA pan-lung cancer[24] and breast cancer datasets were used in this analysis and visualized through the use of cbio portal. Survival data was queried using the kmplot.com dataset[25].

EGFR ACTIVITY PREDICTION

The pan-lung cancer dataset was downloaded and interrogated with single sample GSEA. The activity score for EGFR was downloaded and mean centered between 0 and 1. The samples were then matched with their *PTPRH* mutation data for the same sample and mutations mapped to the gene location using UCSC genome browser. It was then color coded as seen on the figure.

ONCOGENIC SIGNATURE APPLICATION

Predefined oncogenic signatures were applied to the dataset. Briefly, the training data was merged with the full dataset and batch effects removed through the use of COMBAT. These samples were then subjected to binary regression analysis with a predefined gene list and conditions for each individual signature[26–28].

CLUSTERING ANALYSIS

Unsupervised hierarchical clustering was performed with Cluster 3.0 and visualized with Java TreeView. The heatmap and sample legends used in the figure were made using Matlab.

APPENDIX

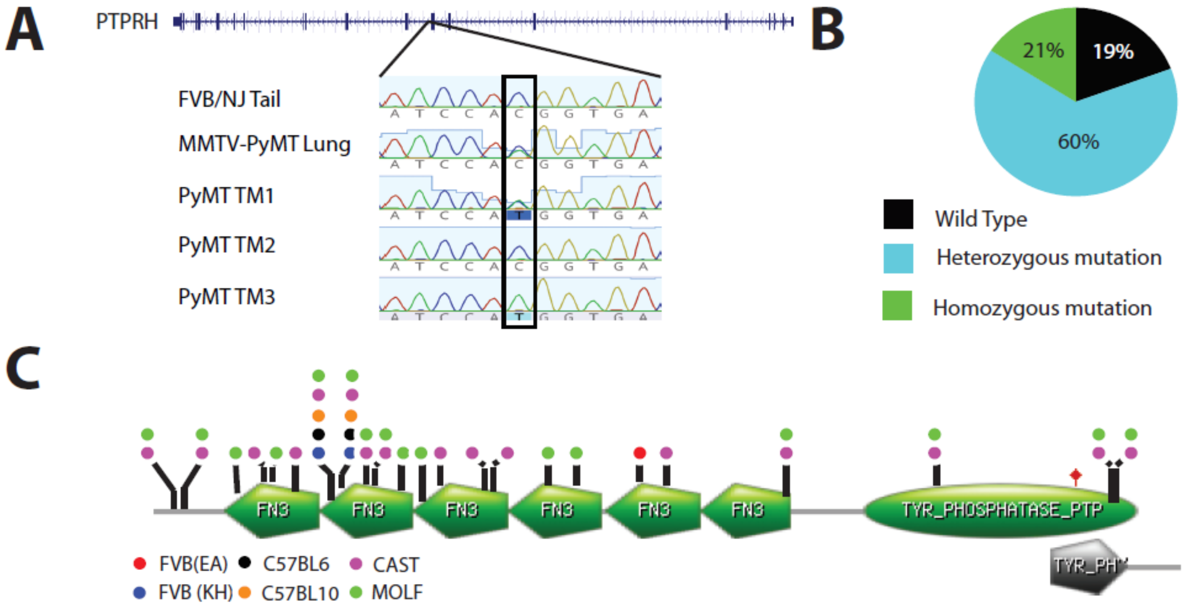

Figure 5.1: *PTPRH* mutations are conserved in MMTV-PyMT and human lung cancer

Phosphotyrosine receptor, *Ptprh* was shown through Sanger sequencing (A) to be heterozygously mutated in 60% and homozygously mutated in 21% (B) of MMTV-PyMT tumors (n=45). Sequencing revealed multiple mouse backgrounds have a variety of mutations clustered in the functional domains of the *Ptprh* proteins

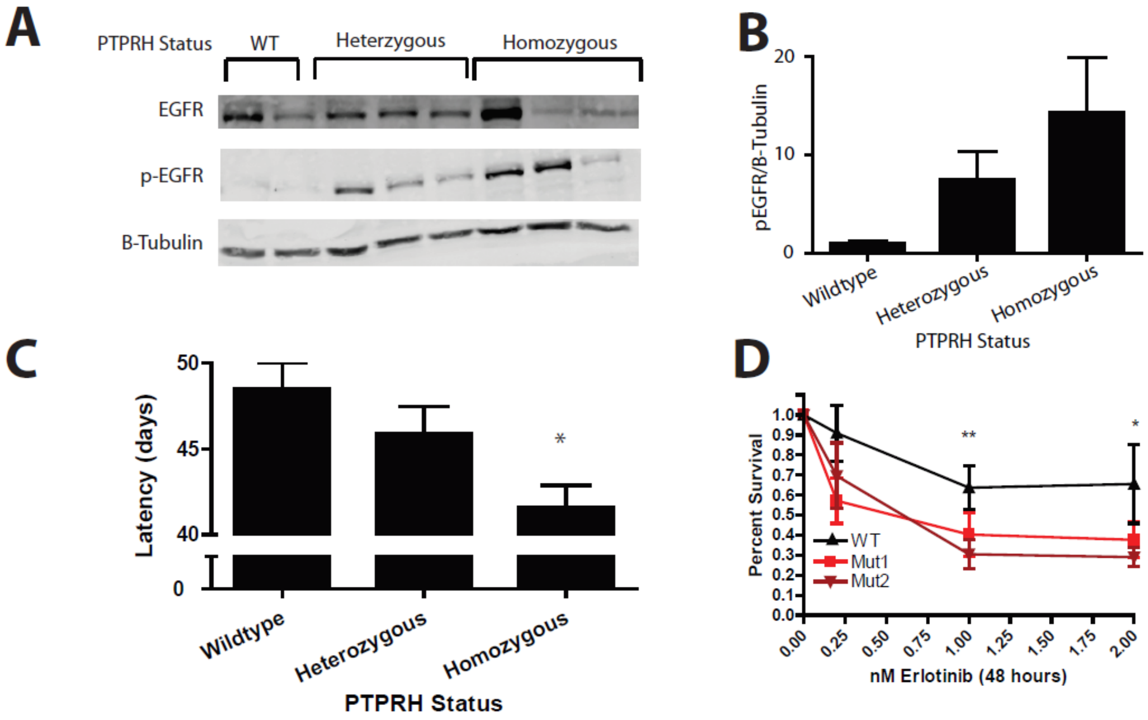

Figure 5.2: *PTPRH* mutation modulates EGFR signaling and treatment response

Increases in pEGFR signaling (A,B) and a decrease in tumor latency (C) (E n=9, *=p<.05) was correlated with mutant *Ptprh* alleles (V483M) within the FVB background. Cell lines derived from *Ptprh* mutant (V483M) PyMT tumors showed an increased response to EGFR targeted therapy including erlotinib (D)(n=3, **=p<.01, *=p<.05)

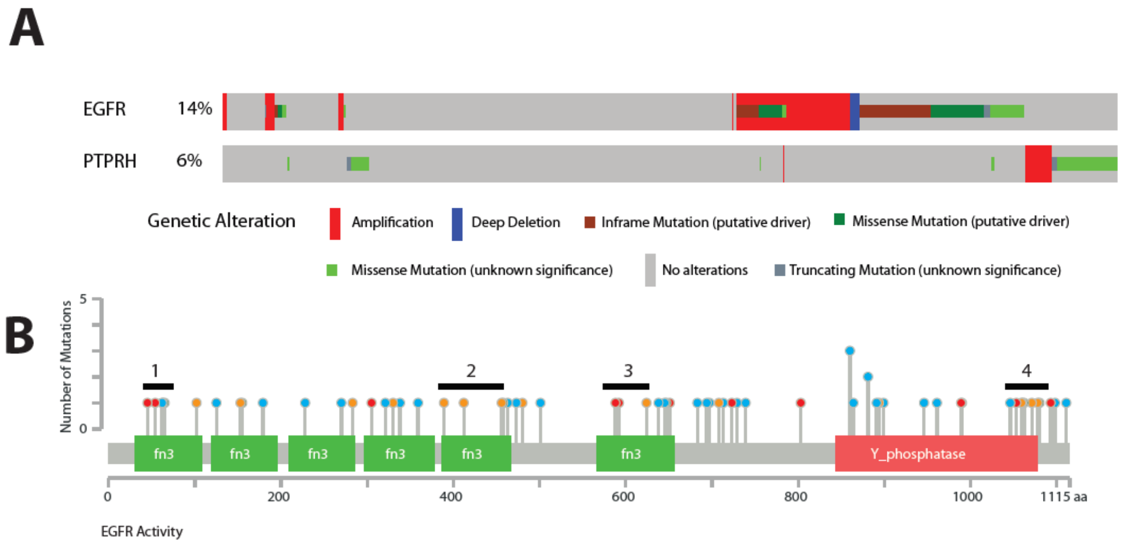

Figure 5.3: *PTPRH* mutations are prevalent in lung cancer

In human lung cancer (TCGA pan-lung cancer) approximately 5% of patients have a mutation in *PTPRH* which are mutually exclusive from EGFR (A). High EGFR activity, as determined by gene set enrichment analysis, is associated with mutations clustered within the structural and functional domains of *PTPRH* as seen by the colors in the lollipop plot (B)

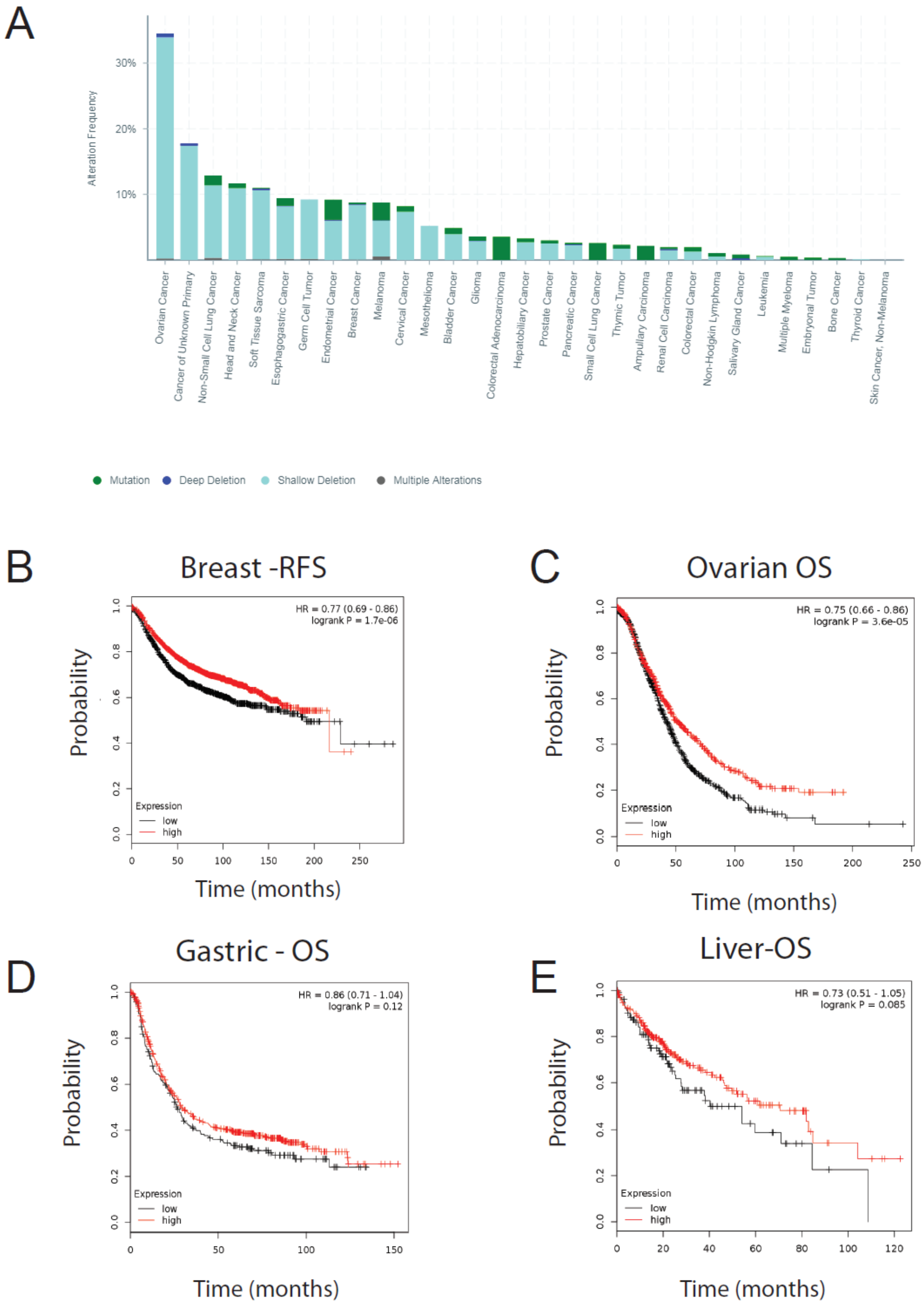

Figure 5.4: Presence and effect of *PTPRH* modulation in other tumor types

PTPRH copy number and mutation was assayed across cancer types (A) and shown to be highly modified in ovarian and lung cancers. Effects in expression show a protective role in Breast (B),

Figure 5.4 (cont'd)

Ovarian (C), Gastric (D), or Liver (E) cancer with respect to overall survival (OS) and relapse free survival (RFS)

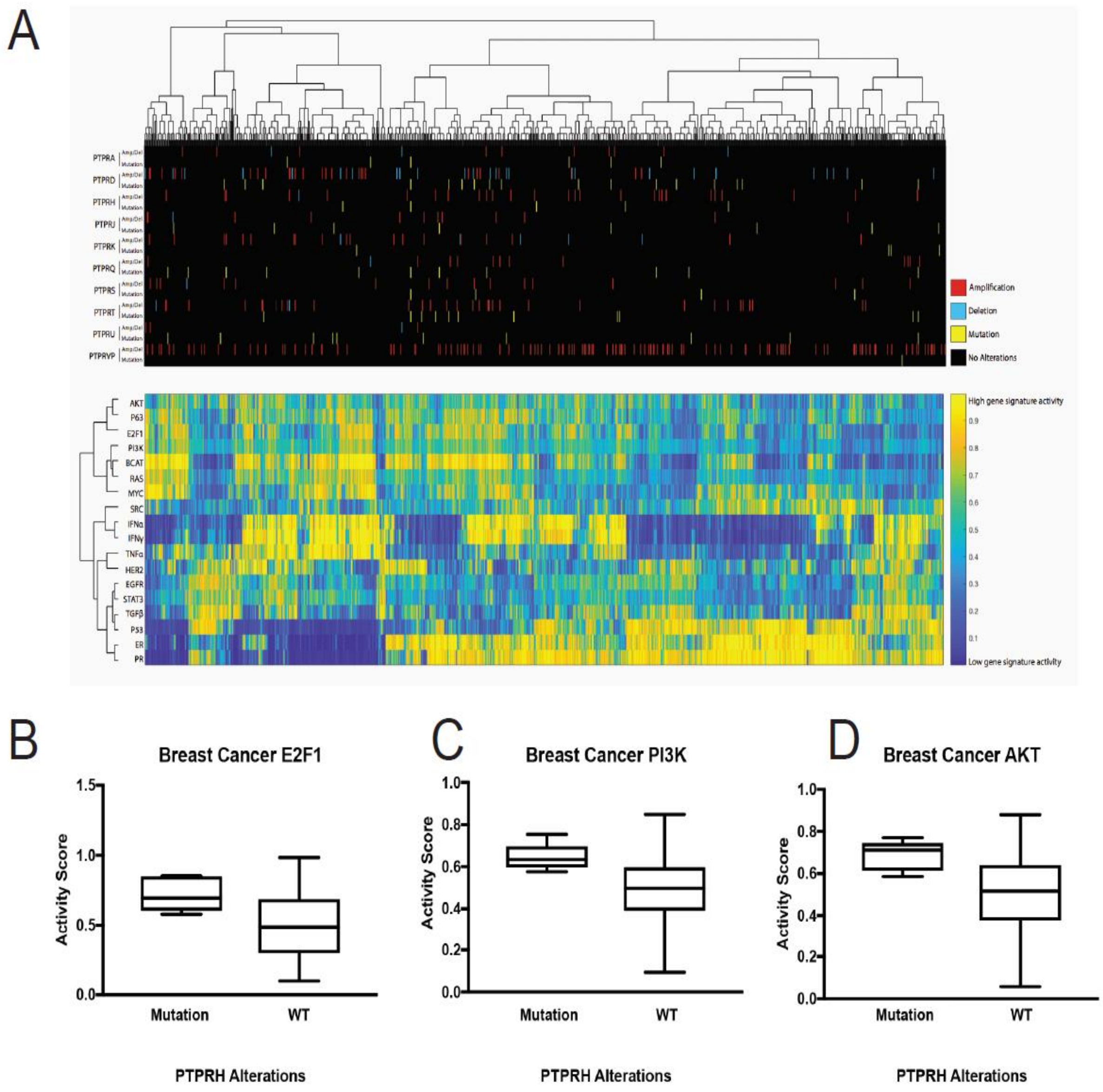

Figure 5.5: *Ptprh* alterations lead to specific Egfr signaling pathway activation

Unsupervised hierarchical clustering identifies unique signaling pathways associated with each ptpr alteration (A). Alterations are categorized into Amplified (Red), Deleted (Blue), Mutation (Yellow), or No Alteration (Black). The pathway activity is presented as a heatmap with high activity being yellow and low activity being blue. Significant differences in E2F1 (B), PI3K (C),

Figure 2.1 (cont'd)

Figure 5.5 (cont'd)

and AKT (D) activities were seen between *Ptprh* modified tumors and non-modified tumors (P<.05).

WORKS CITED

WORKS CITED

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144,** 646–674 (2011).

2. TCGA. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

3. Ciriello, G. *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163,** 506–519 (2015).

4. Lemmon, M. A., Schlessinger, J. & Ferguson, K. M. The EGFR Family: Not So Prototypical Receptor Tyrosine Kinases. *Cold Spring Harb. Perspect. Biol.* **6,** a020768–a020768 (2014).

5. Normanno, N. *et al.* Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene* **366,** 2–16 (2006).

6. Bethune, G., Bethune, D., Ridgway, N. & Xu, Z. Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. *J. Thorac. Dis.* **2,** 48–51 (2010).

7. Xu, H. *et al.* Epidermal growth factor receptor in glioblastoma. *Oncol. Lett.* **14,** 512–516 (2017).

8. Zimmermann, M., Zouhair, A., Azria, D. & Ozsahin, M. The epidermal growth factor receptor (EGFR) in head and neck cancer: its role and treatment implications. *Radiat. Oncol.* **1,** 11 (2006).

9. Downward, J., Parker, P. & Waterfield, M. D. Autophosphorylation sites on the epidermal growth factor receptor. *Nature* **311,** 483–5

10. Yan, X. *et al.* Mesenchymal stem cells from primary breast cancer tissue promote cancer proliferation and enhance mammosphere formation partially via EGF/EGFR/Akt pathway. *Breast Cancer Res. Treat.* **132,** 153–164 (2012).

11. Scaltriti, M. & Baselga, J. The epidermal growth factor receptor pathway: a model for targeted therapy. *Clin. Cancer Res.* **12,** 5268–72 (2006).

12. Manole, S., Richards, E. J. & Meyer, A. S. JNK Pathway Activation Modulates Acquired Resistance to EGFR/HER2–Targeted Therapies. *Cancer Res.* **76,** 5219–5228 (2016).

13. Yamaoka, T., Frey, M. R., Dise, R. S., Bernard, J. K. & Polk, D. B. Specific epidermal growth factor receptor autophosphorylation sites promote mouse colon epithelial cell chemotaxis and restitution. *Am. J. Physiol. Gastrointest. Liver Physiol.* **301,** G368-76 (2011).

14. Vecchione, L., Jacobs, B., Normanno, N., Ciardiello, F. & Tejpar, S. EGFR-targeted therapy. *Exp. Cell Res.* **317,** 2765–2771 (2011).

15. Harbeck, N. *et al.* HER2 Dimerization Inhibitor Pertuzumab - Mode of Action and Clinical Data in Breast Cancer. *Breast Care* **8,** 49–55 (2013).

16. Dowell, J., Minna, J. D. & Kirkpatrick, P. Erlotinib hydrochloride. *Nat. Rev. Drug Discov.* **4,** 13–14 (2005).

17. Sordella, R., Bell, D. W., Haber, D. A. & Settleman, J. Gefitinib-Sensitizing EGFR Mutations in Lung Cancer Activate Anti-Apoptotic Pathways. *Science (80-.).* **305,** 1163–1167 (2004).

18. Minkovsky, N. & Berezov, A. BIBW-2992, a dual receptor tyrosine kinase inhibitor for the treatment of solid tumors. *Curr. Opin. Investig. Drugs* **9,** 1336–46 (2008).

19. Yao, Z. *et al.* A Global Analysis of the Receptor Tyrosine Kinase-Protein Phosphatase Interactome. *Mol. Cell* **65,** 347–360 (2017).

20. Du, Y. & Grandis, J. R. Receptor-type protein tyrosine phosphatases in cancer. *Chin. J. Cancer* **34,** 61–9 (2015).

21. Wälchli, S., Espanel, X. & van Huijsduijnen, R. H. Sap-1/PTPRH activity is regulated by reversible dimerization. *Biochem. Biophys. Res. Commun.* **331,** 497–502 (2005).

22. Politi, K. *et al.* Lung adenocarcinomas induced in mice by mutant EGF receptors found in human lung cancers respond to a tyrosine kinase inhibitor or to down-regulation of the receptors. *Genes Dev.* **20,** 1496–1510 (2006).

23. Lan, L. *et al.* Shp2 signaling suppresses senescence in PyMT-induced mammary gland cancer in mice. *EMBO J.* **34,** 2383–2383 (2015).

24. Campbell, J. D. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48,** 607–616 (2016).

25. Györffy, 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.* **123,** 725–31 (2010).

26. Bild, A. H. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439,** 353–7 (2006).

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

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

CHAPTER 6

FUTURE DIRECTIONS

MULTI-OMIC CLASSIFICATION OF MOUSE MODELS OF BREAST CANCER

The most obvious extension of this project is to perform more sequencing and transcriptomic profiling of mouse models. This study would greatly benefit from the addition of more individuals in the MMTV-Neu and MMTV-PyMT mouse models. The addition of more samples will allow us to detect SNVs, CNVs, and translocations that are present in a much smaller population percentage. Another key will be to expand the pool of mouse models. It will be critical to have a more diverse population of models to profile which have a variety of mechanisms of tumorigenesis. For starters I would like to profile the landscape of other oncogenic drivers and promoters such as the MMTV-Myc or WAP-Myc. It would also be interesting to include tumor suppressive models such as those with alterations to P53 and BRCA as well as carcinogen induced models such as DMBA. It will also be enlightening to capture the diversity of models with the same driving oncogene. The Myc or Neu induced models would be ideal systems to study this in. In both of these systems, there are multiple mouse models with modifications to Myc or Neu that change tumor properties in the model. It would be beneficial to the research community to see how those alterations allow for the selection of different genomic landscapes.

Another key direction to take the profiling study would be to expand the datatypes present in the dataset. An unexplored level of regulation in this work is the epigenetic alterations present in the tumors. I hypothesize that there will be a vast diversity in the epigenomic profile of mouse models. This is due to the fact that the diversity found in this thesis with regards to SNVs and CNVs is not enough to account for the heterogeneity found in mouse models at the transcriptomic level. Epigenetics will also be important to consider in matched tumor and metastatic lesions. Recent human studies have shown that metastatic lesions do not contain unique genomic alterations but instead show differences in their epigenetic profile from their primary tumor. It will be informative to see if this is also true in mouse models since many of these are used to study the process of tumor metastasis.

Some key data pieces have been left unexplored in our current dataset due to the lack of reliable informatic pipelines and validation of findings. However, as the pipelines become more refined it will be important to identify the impact of non-coding variant in the tumors as well as translocations. Recent work has identified key non-coding mutations in lung cancer as well as a pan-cancer approach which has identified a number of novel translocation events. It will be important to continue to profile the mouse tumors to identify if these events are conserved as well as identify novel events in mouse and human tumors.

A unique opportunity that presents itself with this dataset it to understand the cause of the mutational signature present in the mouse models. The mouse models share human signature five with human cancers. Human cancers have been unable to identify an aetology associated with this mutational signature. I propose that a causative mechanism maybe able to be identified within the mouse model due to its relative clean genomic landscape and lack of events. To do this, I propose to take each individual sequenced mouse tumor and compare using a mixture modeling approach to understand the contribution of signature 5 to the mutation profile of the tumor. I will use this as a mutation score and use weighted gene correlation analysis to find gene expression correlates of signature 5. Once the signature is obtained, I will cross compare this to the copy number variants and SNVs both directly and through interactors. From this gene list I will identify those that overlap with repair and instability gene signatures to identify potential mechanisms of signature five. The will then need to be explored through *in vitro* and *in vivo* knockout experiments to confirm the bioinformatic experiments

17Q21.33 AMPLIFICATION

The 17q21.33 amplification event has roles in metastasis and treatment response. Due to the dual role of the amplicon the future directions for this project the future directions for this project are varied and involve everything from biological assays to clinical trials.

Most immediately, the studies that must be completed are the mechanistic studies behind how *Col1a1* and *CHAD* amplification contribute to both early and late stages of metastasis. I hypothesize that the contributions to metastasis are due to increased migration in the presence of high *Col1a1* and *Chad*. The initial experiment to do would be a picrosirius red stain to understand if it is simply more type one collagen that is constructed in the extracellular matrix or if other types of collagen are also upregulate as well to make an abundance of collagen fibrils. The next experiment would be to perform cell adhesion assays in the presence of wild type *Chad* levels as well as an abundance of *Chad*. This will identify if *Chad* is contributing directly to cellular adhesion. I hypothesize that *Chad* binds to *Col1a1* directly to help facilitate movement. To start to investigate this I propose to use IHC staining to identify the location of *Chad* and *Col1a1* both within the cell and within the overall structure of the tumor. This assay will show co-localization. To identify binding of the two proteins, I will use a co-IP approach of the entire tumor and ECM lysate. To understand migration I will complete in vitro migration assays include wound healing and transwell experiments with various combinations of *Col1a1* and *Chad* knockout.

Another area that could be expanded from this experiment it to use a drug screen to block metastasis. *Col1a1* and *Chad* are potential targets for therapeutic intervention to block metastasis from developing. One could envision a therapeutic that either works to block the production of a collagen matrix or blocks the adherence properties of *Chad*. To identify this I propose to use the drug repurposing core to identify compounds which block migration through the use of a high throughput wound healing assay both in the presence of a collagen matrix and without – to identify compounds which inhibit *Col1a1* and *Chad.* Once this compound is identified it will need to be tested in mouse models to identify its ability to block metastasis in an *in vivo* setting.

Next, much more work must be done to confirm the AKT sensitivity present in 17q21.33 patients. To validate the bioinformatic predictions *Phb* KO and overexpression studies must be completed both within a 17q21.33 amplified setting and in a wildtype setting. After the creation of these, AKT targeted therapy will be performed and sensitivity assayed through cell death assays. This will work to confirm the mechanisms proposed by the bioinformatic means. Furthermore, the AKT sensitivity was identified in mostly a high-throughput *in vitro* setting. To confirm this *in vivo* drug targeting studies must be performed both in genetically engineered mouse models and PDX models.

Preliminary work shows that the 17q21.33 amplicon is present both within the metastatic lesion and the primary tumor. Sequencing and or copy number assays need to be performed on both the primary tumor and the metastasis lesions to identify the extent to which the metastatic lesion contains the 17q21.33 amplification event and if it has the propensity to drive metastasis to a certain location. Furthermore, to identify 17q21.33 as a biomarker for therapeutic response it must be seen in a patient setting. Recent studies show that cell free circulating tumor DNA is reflective of the metastatic lesions. It will be important to see if the 17q21.33 amplification event can be detected through a blood draw. If not, FISH techniques must be refined to identify the 17q21.33 amplification event from a tumor biopsy.

PTPRH STUDIES

The *Ptprh* studies present in this work have the most immediate translational impact. However, to move them from the benchside to clinic more work must be completed in both through biochemical assays and with patient samples.

The largest remaining question is the mechanistic impact of *Ptprh* mutation. That is, how exactly does the *Ptprh* mutation disrupt the process of removing the phosphate from active EGFR. There are two hypotheses for this currently. First, the mutation prevents dimerization of *Ptprh* and through that stops it from being active. The other hypothesis is that mutations to *Ptprh* prevent it from binding to phosphorylated EGFR. The most illuminating experiment will be a Co-IP where an anybody will pull down the mutant and wildtype *Ptprh* and identifying binding partners.

Another experiment to perform is to use CRISPR-Cas9 based gene editing to induce mutant *Ptprh* as well as restore wildtype *Ptprh* in a mutant setting. Through these experiments we will be able to identify if the *Ptprh* mutation is acting in a dominant negative or haploinsufficient manner. Once these are created it will also give a controlled setting in which to confirm the erlotinib sensitivity found in this study.

Along with erlotinib sensitivity, it will need to be identified how *Ptprh* mutations effect the response to other first-generation EGFR inhibitors. Furthermore, this study can be expanded to second and third generation drugs. These experiments will be completed using a standard drug sensitivity curve of tumors with mutant *Ptprh* vs tumors with the same genetic background with wildtype *Ptprh*. Drug sensitivity experiments will identify if patients with mutant *Ptprh* will respond to any EGFR targeted therapy or if they will only respond to erlotinib.

A key followup experiment to the *in vitro* data presented in this study it will be important to complete *in vivo* validation of the erlotinib sensitivity. To start with this experiment, I propose to use PDX models with wildtype EGFR. These models will be split into *Ptprh* wild type and mutant and treated with erlotinib or other EGFR targeted therapy that was found to be effective in the above experiments. Once these experiments are completed the transition to the clinic will be relatively rapid due to the already wide use of EGFR targeted therapy in a lung cancer setting.

One potential hurdle to the clinical translation of the finding is identifying a biomarker of *Ptprh* mutation. Due to the lack of a hotspot *Ptprh* mutation traditional genomic sequencing will not be an option for identifying mutations of *Ptprh*. It may however, be cost effect to sequence the RNA present to identify the presence of a point mutation. The most cost-effective options would be to identify phosphorylated EGFR or other downstream targets of EGFR to tell activity. *Ptprh* de-phosphorylates a specific tyrosine residue on EGFR so an IHC approach with an antibody against pospho-tyrosine on EGFR may be an effective way to identify sensitive patients. It also may be possible to develop a gene expression signature to identify key genes which are differentially regulated through in the presence of *Ptprh* mutation. This signature would be developed through the identification of differentially regulated genes and could be assessed at initial biopsy.