GENETIC DRIVERS AND CLONAL HETEROGENEITY OF LETHAL BREAST CANCER

Marni B. Siegel

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Genetics and Molecular Biology in the School of Medicine

Chapel Hill
2017

Approved by:

Charles M. Perou

Carey K. Anders

C. Ryan Miller

Al Baldwin

Ken McCarthy

ABSTRACT

Marni B. Siegel: Genetic Drivers and Clonal Heterogeneity of Metastatic Breast Cancer
(Under the direction Charles M. Perou and Carey K. Anders)

Breast cancer remains the second leading cause of cancer related death in women in the United States. Despite great advances in both early detection and treatment for primary breast cancer, 40,000 women die of breast cancer each year. Metastasis, namely when cancer spreads beyond the original site, is the main cause of breast cancer mortality. A lack of understanding of metastasis continues to thwart prevention and treatment of lethal breast cancer. Genome-wide comparisons of both the genetic composition (DNA) and expression (RNA) of primaries and metastases in multiple patients could help elucidate the underlying mechanisms causing breast cancer metastasis.

In this thesis, next-generation sequencing was performed on a dataset of patients with both primary breast cancers and multiple distant metastases. DNA and RNA sequencing were performed on 16 breast cancer patients with 86 matched tumors (primary + multiple metastases). We confirmed previous work that the primary cancer is extremely diverse with multiple distinct populations of cells. Comparisons of these populations in the original tumor and the distant metastases demonstrates that in some instances, it is likely that a clump of cells containing multiple different genetic populations together leave the breast and seed distant sites. Finally, a novel computational method integrating RNA gene expression, somatic copy number alterations, and somatic mutations identifies drivers of breast cancer in matched primaries, metastases, and in the broader context of breast cancer as a whole. We show that a

majority of the drivers of breast cancer are established in the original cancer and maintained in metastasis.

This work asks clinically impactful questions of the biology of breast cancer metastasis through multiple genomic approaches. The body of knowledge presented here demonstrates that the complex heterogeneity in primary breast cancer is maintained throughout metastasis while also proving that the majority of genetic drivers in metastasis are established in the original breast cancer. Finally, we demonstrate that common mechanisms driving breast cancer are utilized across the previously-described molecular and clinical subgroups of breast cancer, offering novel, tractable therapeutic targets. These findings contribute significantly to our understanding of the genetic diversity and drivers of lethal breast cancer metastasis.

ACKNOWLEDGEMENTS

The work presented here is a reflection of the incredible effort of many talented, driven, and empathetic physicians, scientists, and patients. Dr. Lisa Carey's establishment of the University of North Carolina's Tumor Donation Program is an incredible example of Dr. Carey's dedication and selflessness in a continued effort to cure breast cancer. This program would not be possible without her support, as well as the pathologists, oncologists, scientists, clinical coordinators, and, of course, the patients themselves who selflessly donated their bodies at the end of their lives. In brief, I would like to recognize the physicians: Dr. Leigh Thorne, Dr. Chad Livasy, Dr. Carey Anders, Dr. Lisa Carey; the clinical coordinators: Amy Garrett, Julie Benbow, and Naiam Kianh; and the scientists: Dr. Xiaping He, Dr. Katherine Hoadley, Dr. Elaine Mardis, Dr. Anders, Dr. Carey, and Dr. Chuck Perou. The research presented here utilized these tissue to expand the field's understanding of breast cancer metastasis.

Chapter 2 of this dissertation was originally published in *PLOS Medicine* and is included here as permitted by the PLOS by the Creative Commons Attribution (CC BY) license, with myself, Dr. Katherine Hoadley, and Krishna Kachni as co-first authors (1).

> Hoadley KA*, Siegel MB*, Kanchi KL*, Miller CA, Ding L, Zhao W, He X, Parker JS, Wendl MC, Fulton RS, Demeter RT, Wilson RK, Carey LA, Perou CM†, Mardis ER†. "Tumor Evolution in Two Patients with Basal-like Breast Cancer: A Retrospective Genomics Study of Multiple Metastases." *PLoS Medicine* 13, no. 12 (January 2017): e1002174.

Chapters 2 and 3 utilize DNA and RNA sequencing from the UNC Tumor Donation Program. All DNA and RNA isolation from these tissues was performed by Dr. Xiaping He. The DNA Sequencing of Chapter 2 was performed at the Washington University in St. Louis by Dr. Elaine Mardis and colleagues. The RNA sequencing of Chapter 2 and all DNA and RNA

sequencing of Chapter 3 was isolated, library prepared, and sent to sequencing at the UNC High-Throughput Sequencing Facility by Dr. Xiaping He. This work would not be possible without Dr. He. Chapter 3 also employs a novel method for calling copy number, which was created by Dr. Mengjie Chen and Dr. Grace Silva.

Chapter 4 utilizes data from The Cancer Genome Atlas and METABRIC efforts. I greatly appreciate the commitment of these consortia to making the genomic and clinical information of thousands of tumors publicly available. This work was in collaboration with Dr. Jian Ma at Carnegie Melon University, Dr. Jack Ma at the University of Illinois, and Dr. Grace Silva during her time at the University of North Carolina.

I appreciate and recognize the funding agencies which made this work possible: The Breast Cancer Research Foundation (LAC, CMP); National Institutes of Health (NIH) (LAC, M01RR00046); National Cancer Institute (NCI) P50-CA58223 Breast SPORE Program (LAC; CMP); Susan G. Komen Foundation SAC110006 (LAC); NCI R01-CA195754-01 (CMP); NCI R01-CA148761 (CMP); National Human Genome Research Institute Center Initiated Projects U54HG003079 (ERM); NCI K23-157728 (CKA), and ASCO-BCRF ACRA Award (CKA). I have been supported by the University of North Carolina Medical Scientist Program (MSTP; NIH T32 GM008719), the UNC Cancer Cell Biology Training Program (T32-CA071341-17), and my NCI F30-CA200345 throughout my doctoral work. I am extremely grateful to the late Dr. Eugene Orringer, Dr. Mohanish Deshmukh, Alison Regan, Carol Herion, and Dr. Toni Darville of the UNC MSTP team for their vision and leadership in training me as a future MD/PhD physician scientist.

Finally, I am extremely indebted to my advisors, Dr. Charles Perou and Dr. Carey Anders. Thank you for your investment in me, your advice, and your incredible scientific mentorship.

DEDICATION

To Robin Leigh Siegel, in whose memory shines strength, hard-work, and honesty. Thank you, Mom, for showing me how much we can improve this world through our actions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ER	Estrogen receptor
PR	Progesterone receptor
HER2	Epidermal growth factor receptor 2
HER2E	Her2-enriched molecular subtype
Lum	Luminal molecular subtype
RAP	UNC Rapid Autopsy Program
TCGA	The Cancer Genome Atlas
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
CNA	Copy number alteration
ClaNC	Classification to nearest centroids
SAM	Significance analysis of microarray
DR	DawnRank
CTCs	Circulating tumor cell clusters
ECM	Extracellular matrix
TNBC	Triple negative breast cancer
CNS	Central nervous system
WGS	Whole genome sequencing
WES	Whole exome sequencing
In/dels	Insertions/deletions
VAF	Variant allele frequency
SVs	Structural variants

CHAPTER 1 – INTRODUCTION

Breast cancer is the second leading cause of cancer-related death in women, accounting for 40,000 deaths each year in the United States. Progression to metastasis is the predominant cause of breast cancer mortality. Brain metastases represent a particularly dire consequence of advanced breast cancer with no approved systemic therapeutics and limited survival. Understanding the underlying biology driving the metastatic phenotype (i.e. molecular drivers of seeding, invasion, and growth at a distant site) could provide novel therapeutic targets to prevent and treat metastatic breast cancer.

Breast Cancer Heterogeneity

Breast cancer is not a single disease but rather a collection of diseases having unique morphologies, gene expression profiles, DNA mutation profiles, DNA copy number alterations, widely varying clinical responses, differences in hormone receptor expression, and variations in patterns of metastasis. Systemic treatment of breast cancer begins with identifying hormone receptor positivity based on the estrogen receptor (ER), progesterone receptor (PR), and epidermal growth factor receptor 2 (HER2) expression coming from tumor cells. RNA gene expression studies define four dominant subgroups of breast cancer: Luminal A, Luminal B, HER2-enriched, and Basal-like breast cancer (Perou et al., 2000). Luminal A breast cancers are typically ER positive and have lower proliferation rates than the Luminal B tumors. Luminal B tumors have poorer overall survival and tend to relapse predominantly in the bone. HER2E tumors have increased expression of the HER2 DNA amplicon genes. Finally, the basal-like breast cancers are the most poorly differentiated and typically lack expression of ER, PR, and

HER2. Patients with basal-like tumors represent the greatest clinic need, with a paucity of targeted therapies clinically available and the worst 5-year overall survival.

Breast cancer subtype captures some of the clinical heterogeneity including first site of metastasis. Luminal tumors often first metastasize to the bone, HER2E tumors to the liver, and basal-like tumors to the lung and brain. Furthermore, the timeline for recurrence is dramatically different: basal-likes typically recur within three years following diagnosis but highly unlikely to be past 5 years, while the hormone-positive tumors often may not recur until closer to 10 years after diagnosis. This has been shown to be a result of both treatment differences and the underlying biology.

There is substantial heterogeneity even within the subtypes of breast cancer. Luminal A breast cancers can have highly variable responses to current therapies. Molecular studies of the copy number landscape of luminal breast cancers have further shown 5 subtypes of these tumors: a copy number neutral subgroup which lack *TP53* mutation and have the best prognosis, three intermediate groups, and one highly copy number altered subgroup which harbor *TP53* mutations and have the worst overall prognosis (Ciriello et al., 2013). These two extremes are also reflected in the METABRIC cohort, which defined subgroups of breast cancer based on copy number and gene expression (Curtis et al., 2012). RNA gene expression of these luminal tumors further defined drivers of the tumors with increased proliferation rates, including *MYC* amplification and *RB* loss (Gatza et al., 2014).

Not only is there substantial DNA alteration heterogeneity within subtypes of breast cancer, but there is also differences in the stromal response to these tumors. Immune infiltrate has been shown to have prognostic value in the HER2E and Basal-like breast cancer subtypes, indicating that variability across this subtype is present (Iglesia et al., 2014). Additionally, cancer-cell associated fibroblasts behave differently around basal-like breast cancers in

comparison to luminal breast cancers (Camp et al., 2011). Finally, a distinct host-wound response varies by subtype (Troester et al., 2009), with increased hypoxia and altered metabolic program around basal-like breast cancers (Harrell et al., 2012).

Unfortunately, this heterogeneity extends even within a single patient's tumor. An elegant study of multiregional sequencing of breast cancer demonstrated that primary breast cancers have spatial heterogeneity (Yates et al., 2015). In almost all patients, mutations were observed in one part of the tumor but not another. Thus, there were multiple populations of cancer cells within a single tumor. Heterogeneity by point mutation was shown in 9/12 patients. This is contrast to heterogeneity measured by copy number alteration, which was shown only in 3/12 patients.

Monoclonal vs Polyclonal Seeding of Metastasis

It is currently unknown what portion of the heterogeneity elucidated from Yates *et al*. leaves the original breast cancer and causes distant metastasis. Two possibilities may occur: monoclonal vs polyclonal seeding of metastasis. In the first instance, a single cell may escape the original tumor, representing one distinct population of cells from the original breast cancer, that then seeds a distant site. In contrast, possibly a chunk of the primary breast cancer moves into the circulation and together seeds a distant site. Thus, the distant site of metastasis would reflect the heterogeneity measured in the original breast cancer.

To better understand the process of clonal evolution in metastasis, many groups have studied matched primaries and single sites of metastasis. The seminal work in renal cell carcinoma hypothesized that metastasis is a result of a single clone escaping the primary cancer followed by clonal expansion (Gerlinger et al., 2012). Few mutations from the primary were observed in the distant metastasis. Recently, multi-metastatic sequencing compared prostate

cancer metastases to the matched primary, demonstrating both mechanisms of seeding (Gundem et al., 2015). The authors hypothesize metastasis-to-metastasis seeding in which a chunk of tumor from one metastasis breaks off and seeds another site.

In an ovarian cancer study, multiple primary tumors were compared to later time points following recurrence (Castellarin et al., 2013). Their results demonstrate multiple clones in the primary that are maintained through metastasis, indicating polyclonal seeding of metastasis. This in contrast to AML in which clonal expansion of a therapy-resistant clone was observed following treatment (Ding et al., 2012).

In breast cancer, Nik-Zainal and colleagues published whole genome sequencing of 21 breast cancers and later 560 whole genome sequences of breast cancers (Nik-Zainal et al., 2016). Sequencing of a matched basal-like breast cancer normal, primary, metastasis, and xenograft demonstrated that all of the original mutations were maintained in metastasis to the brain with continued evolution in the brain metastasis (Ding et al., 2010). In addition, the overall copy number structure was extremely similar to the original tumor (Ding et al., 2010). In array CGH comparisons of 23 primary breast cancers and matched metastases, copy number was shown to be highly concordant – 92% for recurrent variants and 73% for non-recurrent variants. 22/23 patients would have similar targeted therapy based on sequencing, further providing evidence of the genetic similarity of primary and metastatic disease (Bertucci et al., 2014).

Whole exome sequencing of matched normal tissue, ductal carcinoma *in situ(DCIS)*, a primary tumor, and a loco-regional lymph node metastasis demonstrate linear progression and monoclonal seeding (Krøigård et al., 2015). Importantly, genetic alterations were stable: if the alteration was observed in the primary, it was also observed in the metastasis. The complete events maintained in the primary and observed in the metastasis argue for a single cell to be the ancestor with a relatively late occurrence of metastasis in this patient.

Single cell sequencing of breast cancer found single clonal expansion from the primary to the liver metastasis as shown in mutations; however, they also showed that a very similar copy number profile was observed in all cells sequenced (Navin et al., 2011). This was done in only one patient and only one site of metastasis. Sequencing of matched brain metastases and primary breast cancers argued that clinically actionable mutations were acquired during spread of disease and not previously observed in the primary tumor (Brastianos et al., 2015).

DNA from cancer cells identified in the blood offer another glimpse into the clonal evolutionary process of breast cancer. In a study of two patients with metastatic breast cancer, whole exome sequencing of both the tumor and cell free DNA were compared (Butler et al., 2015). Both *ESR1* mutations and *PIK3CA* mutations were identified in the metastatic and primary tumors, respectively, but not observed in the others. Authors showed that the cell free DNA (cfDNA) more closely reflects the metastases rather than the primary tumor.

Three elegant *in vivo* study of the actual physiologic process of breast cancer metastasis demonstrate how polyclonal seeding is physically possible. Circulating tumor cell clusters *in vivo* demonstrated that clusters of CTCs have much greater metastatic potential than single CTCs, although both were observed (Aceto et al., 2014). A combined red and green fluorescent transgenic mouse breast cancer was injected into the mammary fat pad of mice and then analyzed lung metastases (Cheung et al., 2016). Cheung and colleagues demonstrate that metastases were between 2 to >1000 cells and all composed of at least red and green tumor cells. They further demonstrate that the extravasation process itself is a bulk tumor process (Cheung et al., 2016). Au and colleagues utilized microfluidic devices that mimic human capillary to study the fluid dynamics and extravasation of tumor cells (Au et al., 2016). Tumor cells were observed to squeeze through as small as 5 um spaces in single file before rounding up once through the passage and continuing their progress.

Timing of Metastatic Drivers

RNA gene expression of a metastasis is 82% identical to the primary breast cancer from which it originated (Harrell et al., 2012; Hoadley et al., 2016), and subtype is generally maintained throughout the metastatic process (Weigelt et al., 2003). This provides evidence that the metastatic potential is likely in the original, primary breast cancer. The underlying biology responsible for successful metastatic seeding and growth are likely present in the primary breast cancer but remain unknown. Understanding genetic features driving metastasis, both in the primary breast cancer and in distant metastasis, could provide prognostic information as well as future, novel therapeutic targets.

Prognostic signatures of metastasis based on genetic features in the primary breast cancer have been developed within our laboratory and independently by others as well. Clinical tools (i.e. PAM50 (Parker et al., 2009), Oncotype Dx (Paik et al., 2004), and MammaPrint (Glas et al., 2006)) stratify patients into high versus low risk of recurrence and are routinely employed in the clinic (Cardoso et al., 2016). In order for these to be prognostic, there must be some amount of metastatic potential in the primary breast cancer. Further research with primary breast cancers and multiple matched sites of metastasis are needed to elucidate the genetics causing these metastases.

Metastasis-Specific Events

Some genetic features specifically enriched in metastasis have been identified through RNA gene expression studies of small cohorts of human breast cancer metastases (Zhang et al., 2009). In human metastases, up-regulation of the hypoxia/VEGF signature (Hu et al., 2009) and down-regulation of extracellular matrix (ECM) signatures are differentially expressed in

metastasis as compared to primary breast cancers, suggesting alteration in the *VEGF* pathway and remodeling of the ECM must occur for successful metastasis. Organ-specific drivers of breast cancer metastasis were identified in *in vivo* mouse models of lung, bone, and brain metastases (Bos et al., 2009; Minn et al., 2005; Sevenich et al., 2014; Valiente et al., 2014; Zhang et al., 2009). In these preclinical studies, overexpression of *SRC* and *COX2* are critical for bone metastasis, *MTDH* is sufficient for lung metastasis, and *neuroserpin* expression is necessary for brain metastasis. The specific DNA alterations that drive these gene expression changes in metastases, and the order in which they occur, remain unknown. Moreover, reproducibility in the human condition has yet to be described.

While some genetic drivers of metastasis are inherent to tumor cells themselves, the tumor microenvironment also plays a vital role in successful tumor cell seeding and survival (Fidler, 2001). Recent literature suggests that some primary breast cancer cells already express proteins essential for the establishment of breast cancer brain metastases (BCBMs), including *serpins* (Valiente et al., 2014), *cathepsin S* (Sevenich et al., 2014), *matrix metalloproteases* (Romagnoli et al., 2014; Wang et al., 2013), and *αB-crystallin* (Malin et al., 2014). Once in the brain, BCBMs up-regulate proteins to enable transport and metabolism of GABA, increasing tumor cell proliferation (Neman et al., 2014). Targeting reactive astrocytes in the tumor microenvironment with drugs decreases brain seeding and growth *in vivo* (Gril et al., 2013), signifying a reliance of BCBMs on the brain microenvironment. Identification of key drivers of breast cancer metastases, both within the tumor and its surrounding microenvironment, will be critical to acquire a comprehensive understanding metastatic biology.

Breast cancer brain metastases have an extremely poor survival with median survival from CNS recurrence at 4.9 months. Within the triple negative breast cancer classification, 46% of patients with metastases will develop brain recurrence and subsequently have a median

survival of 13.3 months following CNS recurrence (Lin et al., 2008). We must put our resources towards understanding the biology of this lethal form of breast cancer in order to develop a cure and save lives.

Genetic Drivers of Breast Cancer

Large-scale sequencing efforts have afforded the opportunity to identify recurrent mutations and copy number alterations. Uncovering the incredible genetic diversity within breast cancer, The Cancer Genome Atlas (TCGA) demonstrated very few recurrent mutations in the most aggressive form of breast cancer, basal-like breast cancer, other than *TP53* (Cancer Genome Atlas, 2012). A recent study of 560 breast cancers with whole genome sequencing identified very few recurrent drivers other than those previously described (Nik-Zainal et al., 2016). Interestingly, in ER+ breast cancer, the mutation burden is often much lower; however, there are more recurrently mutated genes in ER+ positive breast cancer including *PIK3CA, GATA3,* and *FOXA1*. With the decreasing cost of high-throughput sequencing, the ability to integrate multiple platforms of genetic data provides a unique opportunity to better identify genetic drivers. Computational predictions of the impact of a mutation on the cancer cell development, growth, and metastatic potential typically incorporate both the location of the mutation on the protein and the number of mutations in a dataset (Dees et al., 2012b; Lawrence et al., 2013).

Breast cancer is a highly copy number altered disease; however, the large spans of genomic space altered makes identifying the actual driver(s) difficult. Several computational approaches have been previously described to narrow the candidate drivers.

Gatza and colleagues integrated a small interfering RNA screen, with gene expression-based pathway signatures, and copy number data (Gatza et al., 2014). By comparing the most

proliferative ER+ luminal A breast cancers to those with lower proliferation scores, *MYC* and *RB* were identified as the most likely candidates driving this proliferative phenotype (Gatza et al., 2014). Silva and colleagues took a different approach by comparing cross-species conserved regions of basal-like breast cancer (Silva et al., 2015). Comparing known mouse models that faithfully recapitulate human breast cancer and integrating copy number analyses, RNA gene expression, and DawnRank driver analysis, they identified that *NCSTN* and *IKBKE* are critical drivers amplified at the 1q amplicon in basal-like breast cancer (Silva et al., 2015). Finally, integration of known protein-protein interaction networks, RNA gene expression, and DNA alterations allows for ranked candidate drivers (Hou and Ma, 2014).

Tumor Microenvironment in Primary and Metastatic Breast Cancer

Breast cancers develop in a milieu of cell types including epithelial fibroblasts, immune cells, and organ-specific cell types at the final sites of metastasis. Previous publications both by our group and others have demonstrated faithful measurement of tumor-infiltrating immune cells from both microarray data and RNA sequencing (Bindea et al., 2013; Iglesia et al., 2014). Increased immune infiltrate in basal-like and HER2-enriched breast cancer are known to be positively prognostic (Iglesia et al., 2014). Immune infiltrate also predicts response to immunotherapy in melanoma (Daud et al., 2016).

In addition to bulk tumor-infiltrated immune cell measurements from gene expression data, novel computational methods can rebuild both the adaptive T cell receptor (Nazarov et al., 2015), the B cell receptor repertoires (Mose et al., 2016), and predict neoantigens (Kardos et al., 2016). By integrating both DNA sequencing mutation calls with RNA sequencing, expression of the adaptive immune receptors and neoantigens can be computationally determined. The bioinformatics capacity to have further insight into the interaction of the immune and tumor

interface provide new opportunities in cancer research.

The role of the immune system and metastasis is not well understood. Recent research provided evidence of a down-regulation of macrophages, T, B, and NK cells in the 'metastasis' relative to the 'parental' cell lines *in vivo*, hypothesizing that metastases achieve an immune-escape mechanism mediated by Wnt signaling (Malladi et al., 2016); however, these studies were performed only in an immune-comprised mouse model with one human patient-derived xenograft from breast cancer. Therefore, further rigorous research in both immune-competent mouse models and human tissues are needed to understand the interaction of metastasis and the immune system.

Not only is it currently unknown what type and level of immune infiltrate exists in metastasis, but we also do not know if it varies in different organ sites. Certainly, immune surveillance in normal lung, liver, and brain vary widely, with the brain typically thought of as an immune privileged organ. As we move into an era of immune modulatory agents, it will be critical to better understand the role of the immune system in metastasis.

Research Introduction

Research elucidating the underlying mechanisms of metastasis is a great clinical need. Through this thesis, we sought to address three critical questions: (1) is breast cancer metastasis a monoclonal or polyclonal event (Chapters 2 and 3); (2) when are the genetic drivers of metastasis established (Chapters 2 and 3); and (3) what are the genetic drivers of metastasis (Chapter 3). We then explore the heterogeneity of genetic drivers across breast cancer in Chapter 4. Through this research, we hope to contribute to the field's continued effort to find therapeutic targets to prevent and treat breast cancer metastasis.

CHAPTER 2 – TUMOR EVOLUTION IN TWO PATIENTS WITH BASAL-LIKE BREAST CANCER: A RETROSPECTIVE GENOMICS STUDY OF MULTIPLE METASTASES[1]

Preface

This work was previously published in *PLOS Medicine* as a co-first-authorship effort among Dr. Katherine Hoadley, Krishna Kanchi, and myself. I aided in the analysis of the expression of mutations in the RNA, identifying the timing with which DNA mutations were established, and interpreted the clonality studies performed by Chris Miller. In addition, I completed the figures, supplemental material, the writing of the manuscript, and all revisions. RNA sequencing was performed at UNC by Dr. Xiaping He and initially analyzed by Dr. Joel Parker and Dr. Hoadley. The DNA sequencing, mapping, validation, and structural variation calls were performed at the Washington University of St. Louis McDonnell Genome Institute by Krishna Kachni, Chris Miller, Li Ding, Ryan Demeter, Robert Fulton, and Michael Wendl. Chris Miller led the clonality analyses. Dr. Lisa Carey, Dr. Chuck Perou, and Dr. Elaine Mardis conceived, funded, and oversaw the project.

Introduction

Breast cancer patients who die from their disease typically succumb to a metastatic rather than primary tumor. Metastasis is a complex process likely involving many potentially

[1] This chapter previously appeared as an article in *PLOS Medicine.* The original citation is as follows: Hoadley KA*, Siegel MB*, Kanchi KL*, Miller CA, Ding L, Zhao W, He X, Parker JS, Wendl MC, Fulton RS, Demeter RT, Wilson RK, Carey LA, Perou CM†, Mardis ER†. “Tumor Evolution in Two Patients with Basal-like Breast Cancer: A Retrospective Genomics Study of Multiple Metastases.” *PLoS Medicine* 13, no. 12 (January 2017): e1002174.

distinct mechanistic steps. Biologically similar tumors vary in their ability to seed distant metastatic sites. Indeed, different molecular intrinsic subtypes of breast cancer as determined by the PAM50 subtype classifier vary markedly in their preferred sites for metastasis (Harrell et al., 2012; Sihto et al., 2011; Smid et al., 2008). The luminal subtypes often metastasize to the bone, HER2-enriched tumors to the lung and liver, and basal-like and claudin-low tumors to the brain, lung, and liver (Harrell et al., 2012; Sihto et al., 2011). The metastatic process is often described as a slow and continuous process of tumor evolution and acquisition of traits such as increased genomic instability, motility, and the epithelial-to-mesenchymal transition. Recent work in renal, prostate, ovarian, and lung cancer has identified significant amounts of intratumor variability in the primary tumor, as well as identifying new driver mutations that arose in metastases (de Bruin et al., 2014; Gerlinger et al., 2012; Gundem et al., 2015; Schwarz et al., 2015; Zhang et al., 2014). In several breast cancer analyses of targeted gene panel, there was considerable concordance of mutations observed between primary tumors and matched metastases (Brastianos et al., 2015; Cummings et al., 2014; Meric-Bernstam et al., 2014; Moelans et al., 2014). This finding, combined with our increasing understanding that a particular intrinsic subtype predicts the future site(s) of metastasis, suggests that in breast cancer at least some of the metastatic potential already exists within the primary tumor (Harrell et al., 2012; Meric-Bernstam et al., 2014; Sihto et al., 2011; Smid et al., 2008). To examine this further, we studied the genomic relationship between the primary tumors and multiple matched metastases of two patients with triple-negative breast cancer (TNBC), with both cases also of the basal-like breast cancer intrinsic subtype.

A common means of studying intratumor heterogeneity is to sample multiple parts of the same tumor and then perform genetic or genomic assays on these different regions. A more extreme approach to intratumor heterogeneity is to study a primary tumor and its associated

metastases to determine the extent to which the metastatic tumor genome was derived from the primary tumor cells as opposed to being an independent tumor (de Bruin et al., 2014; Gundem et al., 2015; McCreery et al., 2015; Shain et al., 2015; Zhang et al., 2014). Whether metastases can develop from the primary tumor or require continued evolution and gain of additional mutations in order to metastasize remains unknown in basal-like breast cancers, and addressing this issue may have important implications for therapy. In order to study the genomic evolution of basal-like breast cancer, we performed DNA whole genome and mRNA sequencing on two patients with matched primary tumors and multiple distant metastases.

Methods

Patient Consent and Tissue Processing

Tumor tissue was obtained from metastatic breast cancer patients who consented to a rapid autopsy at the University of North Carolina prior to death. Patient consent for the autopsy was obtained in accordance with the UNC Office for Human Research Ethics (OHRE) and criteria established by the US Department of HHS, but was not IRB regulated. There was no prospective analysis plan for this study. Primary, metastatic, and adjacent normal tissues were taken within 6 h of death for all metastatic sites identified prior to death and at time of autopsy. Tissues were frozen in liquid nitrogen, and RNA and DNA were isolated from each tissue using Qiagen RNAeasy and DNAeasy kits, respectively, according to manufacturer protocol (Qiagen, Valencia, California).

Sequencing Methods

RNA was isolated with RNeasy Mini Kit (Qiagen), and sequencing libraries were prepared with Illumina TruSeq RNA Sample Prep Kit (CAT #RS-122-2001) with the polyA select protocol, except for the A7-Brain, which was first prepared using the Epicentre's Ribo-Zero rRNA Removal kit (Cat #RZH11042) (Zhao et al., 2014). RNA-seq was mapped with MapSplice (Wang et al., 2010) and quantified with RSEM (Li and Dewey, 2011). Upper-quantile normalized counts, log2 transformed, were combined with the Cancer Genome Atlas (TCGA) breast RNA-seq data (Ciriello et al., 2015). Samples were median centered and clustered using the human breast cancer intrinsic gene set list (Parker et al., 2009), in Cluster 3.0 (Hoon et al., 2004) and visualized with Java TreeView v. 1.1.6r4 (Saldanha, 2004).

A previously described procedure was followed for library construction and sequencing (Mardis et al., 2009). Briefly, DNA was sheared (Covaris), end repaired (Lucigen), polyadenylated (Lucigen), and ligated to adapters (Illumina) for paired-end data generation.

DNA sequencing was performed on the Illumina Genome Analyzer II and generated between 114 and 260 Gbp of sequence data for each tissue studied and haploid coverage ranging from 29.24 to 72.17.

Somatic Alteration Detection Pipeline

Reads were aligned to human reference build 36 (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.3/special_requests/assembly_variants/; BWA 0.5.5, http://sourceforge.net/projects/bio-bwa/), merged into a single binary alignment map (BAM) file, with duplicate reads removed using Picard 1.07 (http://broadinstitute.github.io/picard/) by the established pipeline, as previously reported (Govindan et al., 2012). To determine somatic variants, we utilized samtools (Li et al., 2009) followed by SomaticSniper using a somatic score ≥40 and mapping quality ≥40 (Larson et al., 2012, 2014). Additional screening against dbSNP was used to remove probable germline variants (Ley et al., 2008; Sherry et al., 2001). Indels were identified with Pindel (Ye et al., 2009) and GATK (McKenna et al., 2010). All variants were further annotated as previously described (Ley et al., 2008; Mardis et al., 2009) using VarScan2 (Koboldt et al., 2012) (parameters: *--min-coverage* = 30, *--min-var-freq* = 0.08, *--normal-purity* = 1, *--p-value* = 0.10, *--somatic-p-value* = 0.001, *--validation* = 1) to classify mutations as reference, germline, somatic, or resulting from loss of heterozygosity (LOH). A Bayesian classifier was applied to retain the somatic variants with a binomial log-likelihood of at least 3 (parameters: *--llr-cutoff* = 3, *--tumor-purity* = 0.95). False positives, as determined by strand specificity, consistent positions near the ends of reads, and poorly mapped qualities, were removed.

Mutations were assigned to four tiers: (1) coding, (2) conserved or regulatory, (3) unique noncoding, and (4) repetitive noncoding regions.

Structural variants (SVs) were called with BreakDancer (Chen et al., 2009) and filtered using TIGRA_SV (Chen et al., 2014). Somatic copy number alterations were detected using CopyCat v1.6.9 (https://github.com/chrisamiller/copycat), with 10,000 bp windows and default parameters.

Experimental Validation of Mutations

Genotypes from Illumina Human OmniExpress BeadChip SNP arrays were used to compare and confirm the heterozygous SNPs detected in the analyzed WGS data.

Putative indels of 1-2bp were converted to BED format and provided as target intervals for the GATK IndelRealigner (DePristo et al., 2011; McKenna et al., 2010). The primary, metastases, and matched normal breast tissue for each patient were then realigned to these BED files independently. To validate the original predictions, we developed a matching algorithm that attempts to match Varscan validation calls with the original indel predictions, as described (Govindan et al., 2012). All validated somatic indels were then manually reviewed using Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al., 2013).

Indels of 3–100 bp were assembled using TIGRA (Chen et al., 2014) and validated as previously described (Govindan et al., 2012). Variants that passed the strict validation were manually reviewed.

Custom sequence capture validation was performed with Roche NimbleGen arrays for 97.3% of the Tier 1–3 somatic alterations and 68.6% of the SVs. Whole genome amplified DNA was prepared for Illumina sequencing according to the manufacturer's protocol (Illumina, San Diego, California). DNA was fragmented with the Covaris S2 DNA Sonicator (Covaris, Woburn, Massachusetts), adapter-ligated, SPRI-bead cleaned, and PCR amplified. One µg of the 300–500 bp fragments was hybridized to the NimbleGen HD2 probe set according to the

manufacturer's protocol (Nimblegen, Madison, Wisconsin). Following hybridization, the library was PCR amplified for 16 cycles and quantified with the KAPA SYBR FAST qPCR Kit (KAPA Biosystems, Woburn, Massachusetts) such that 180,000 clusters were sequenced per lane of the Illumina GAIIx.

Reads were mapped to the NCBI Build 36 reference WUGSC Variant, a subset of the NCBI36 sequences from Ensembl Release 46 (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.3/special_requests/assembly_variants/).

The validation sequence was aligned with BWA v0.5.9, and duplicate reads were marked using Picard (v1.29). Updated versions of BWA and Picard were used for increased alignment speed and variant detection efficiency. The RefCov package was used to evaluate the coverage of target sequences (http://gmt.genome.wustl.edu/packages/refcov/).

Capture validation reads and mates were mapped to both the assembled SV contigs and the reference with CrossMatch (version 1.080721). The threshold for an acceptable alignment is ≤1 mismatch at either end, ≤1% substitutions, 1% indels and a CrossMatch score ≥ 50. An SV-supporting read is required to span the breakpoint on the SV contig, align to 10 bases flanking on each side of the breakpoint, and have no alignment to the reference above the minimum alignment criteria. The somatic status of each SV was determined using Fisher's exact test between the matched tumor and normal sample. All validated calls were manually reviewed.

UNCeqR (Wilkerson et al., 2014) was run on validation mode: the algorithm accepts as input a set of predetermined mutations, such as a list of mutations generated from WGS/WES, and then looks within the RNA-seq data for expression evidence of the variants. Tier 1 mutations were input into UNCeqR along with the RNA-seq BAM files aligned with MapSplice (Wang et al., 2010). UNCeqR then calculated the number of reads of the reference and variant

alleles at each position interrogated. Mutations with less than 5 reads were considered as 0. RNA variant allele fraction (VAF) was calculated as variant allele reads/total reads.

Clonality Analyses

The clonal structure of each tumor was inferred with SciClone (version 1.0.7) (Miller et al., 2014), with parameters minDepth = 75, copyNumberMargins = 0.25, and maximumClusters = 20. Single nucleotide variants (SNVs) in copy number altered regions or with evidence of complete or partial LOH were reviewed and excluded. Phylogeny was inferred using the clonevol R package (https://github.com/hdng/clonevol) with default parameters.

Results

Case Histories

Patient A1 was a 65-y-old white woman who presented with stage IV TNBC and synchronous metastases to the bone of the vertebral column (spinal), lung, adrenal gland, liver, and lymph nodes. She was treated with radiation therapy to the breast, whole brain, and C3/T2 of the spine, had one cycle of palliative paclitaxel without response, and died of disease 2-mo post-diagnosis. Patient A7 was a 60-y-old African-American woman diagnosed with a 5-cm stage IIIA TNBC. A pretreatment primary tumor biopsy was collected as a part of an existing tissue collection protocol (LCCC 9819, NCT01000883), and she subsequently received neoadjuvant doxorubicin plus cyclophosphamide followed by paclitaxel. She underwent mastectomy with T2N2 residual disease, followed by adjuvant radiation therapy to the chest wall (SCV fossa and axillary nodes). Patient A7 remained without evidence of disease recurrence for 17 months before presenting with metastases to the brain, kidney, liver, lung, and ribs. She received single-agent capecitabine for 4 months, with an initial minimal response and then progression both systemically and in the central nervous system (CNS), followed by a single cycle of carboplatin that was discontinued because of poor tolerability and evidence of rapid progression. Patient A7 died of disease 8 months after her metastatic progression. For both patients, fresh frozen tissue was collected at autopsy from primary tumor, distant metastases, and adjacent normal (nonmalignant breast) tissue, except for the primary tumor specimen that was obtained before neoadjuvant treatment was initiated in patient A7 (Figure 2.1).

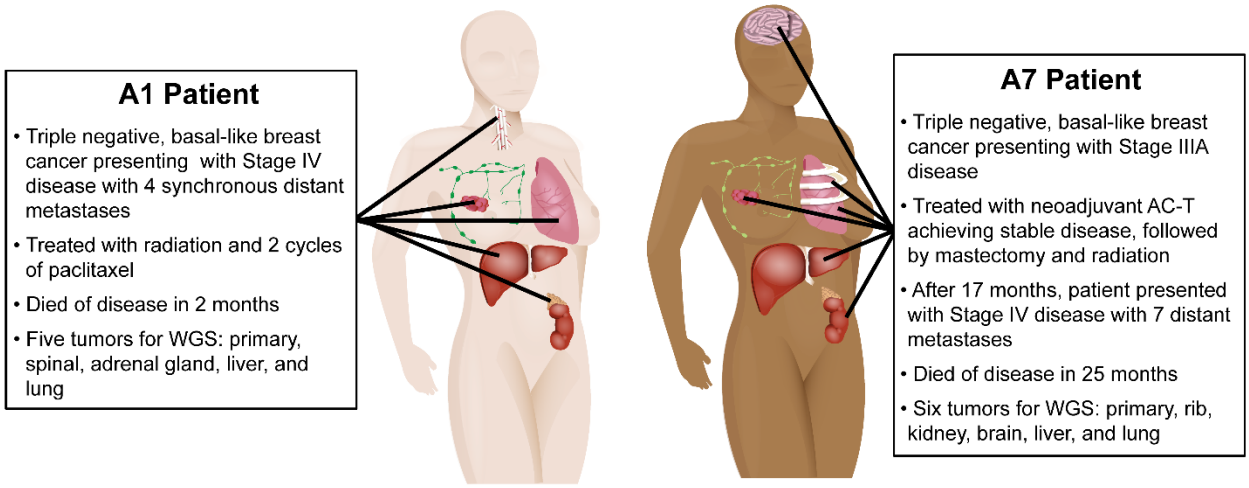

Figure 2.1. Clinical history and distribution of metastases from patients A1 and A7.

Whole Genome Sequencing Coverage and Mutation

For the matched normal tissues, primary tumor (pre-treatment biopsy for A7), and distant metastases from patients A1 and A7, we performed DNA whole genome paired-end sequencing. For A7, we derived 138.38, 118.76, 260.93, 128.69, 204.34, 201.66, and 156.82 Gbp of sequencing data from normal tissue, primary tumor, liver, lung, rib, kidney, and brain metastases, respectively, with corresponding haploid coverages ranging from 33.17X to 70.19X. For A1, we generated 265.53, 134.07, 115.85, 210.45, 114.31, and 131.03 Gbp of data from normal tissue, primary tumor, liver, lung, adrenal, and spinal cord metastases, respectively, with haploid coverage ranging from 30X to 72.16X.

Candidate somatic changes were predicted using multiple algorithms. Confirmatory testing of heterozygous mutations with genotype arrays confirmed bi-allelic detection of 80.47% to 89.63% in all samples. Candidate mutations were further validated with capture probes corresponding to all putative somatic SNVs and small insertions/deletions (indels) that overlap with coding exons, splice sites, and RNA genes (Tier 1), a number of high-confidence SNVs and indels in noncoding conserved or regulatory regions (Tier 2), and nonrepetitive regions of the human genome (Tier 3). In addition, we included predicted somatic SVs for validation. We obtained 40X haploid reference coverage for 87.48% to 94.02% of the targeted sites. For A1, 73 Tier 1 point mutations, 1 Tier 1 indel, and 53 somatic SVs were confirmed across the primary tumor and metastases. For A7, there were 150 Tier 1 point mutations, 47 indels, and 40 SVs confirmed in the primary tumor and five metastatic samples.

Genomic Relatedness of Primary Tumors and Metastases

Common gene expression patterns throughout metastasis. In order to study the degree of relatedness between a primary tumor and its metastases, we performed mRNA-seq gene expression analyses followed by hierarchical clustering analysis using a breast cancer "intrinsic" gene list (Parker et al., 2009) including data from the 11 specimens studied here and 1,100 breast tumors from the Cancer Genome Atlas (TCGA) (Ciriello et al., 2015). Regardless of physical or temporal distance between the primary and its metastases, all tumors from these two patients clustered tightly together by patient (Figure 2.2). By gene expression analysis using the PAM50 intrinsic breast cancer subtype predictor (Parker et al., 2009), the primary tumors and metastases all maintained a basal-like subtype phenotype and clustered with the basal-like samples from TCGA (Figure 2.2B); previous research has demonstrated a high correlation among primaries and matched metastases by microarray gene expression (Bertucci et al., 2016; Harrell et al., 2012).

In patient A1, in whom the primary tumor and distant metastases were found synchronously and who had limited exposure to chemotherapy and radiation prior to death, the gene expression hierarchical cluster node correlation for the primary and the four metastases was 0.77 (Figure 2.2C). In patient A7, who received neoadjuvant chemotherapy and radiation and had a 17-mo interval separating the discovery of the primary tumor and distant metastases, the node correlation for the six samples was 0.79 (Figure 2.2C). This demonstrates that subtype was maintained throughout metastasis in these two patients and that, as we and others have shown (Bertucci et al., 2016; Harrell et al., 2012), distant metastases are typically much more similar to their original primary than they are to other primary tumors or metastases from other patients.

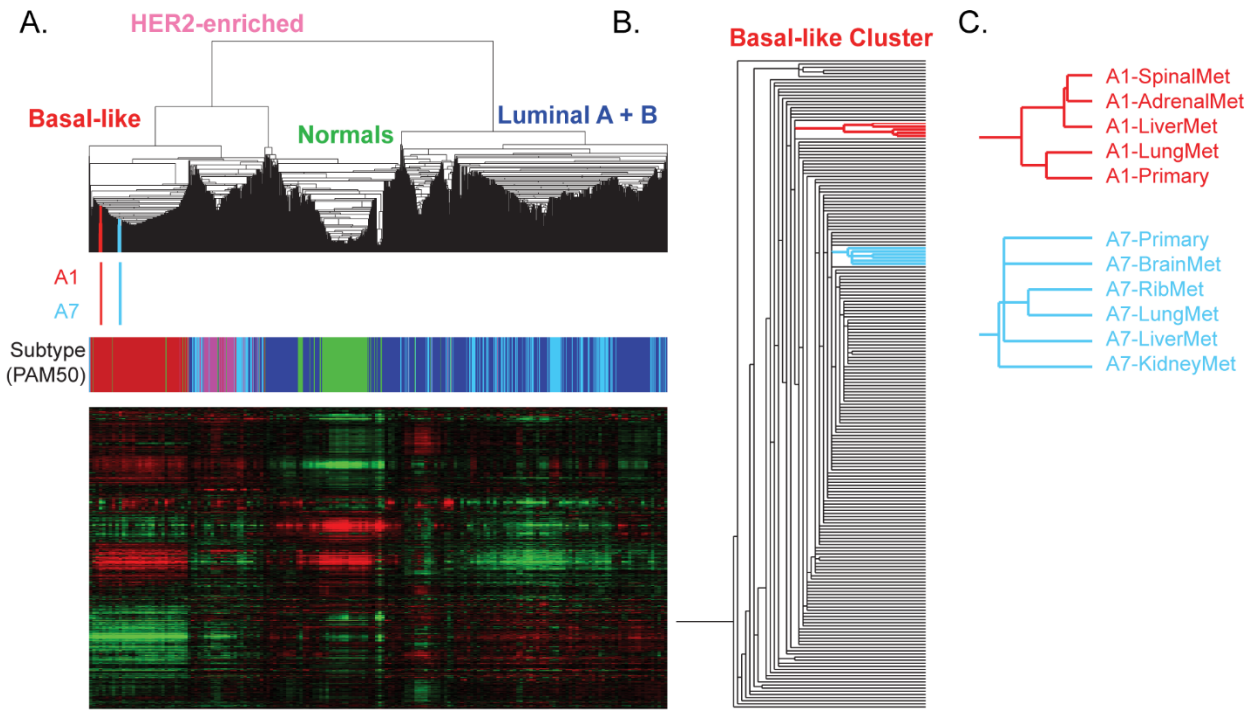

Figure 2.2. Molecular relatedness of matched primary and metastases. (A) Hierarchical clustering of patient A1 and A7's tumors with 1,100 TCGA Primary samples and 98 normal breast samples analyzed using a breast cancer intrinsic gene list. The color bars under the dendrogram indicate (i) where A1 (red) and A7 (blue) specimens are clustered and (ii) the PAM50 subtype of each sample (basal-like, red; HER2-enriched, pink; luminal A, dark blue; luminal B, light blue; and normal-like, green). (B) The position of A1 (red) and the position of A7 (blue) within the basal-like cluster are highlighted. (C) The relationship of the primary and metastases for each patient based upon gene expression patterns.

Functional mutations are maintained and enriched during metastasis. We next studied DNA-based data from each primary tumor and its multiple distant metastases. In patient A1, 54 genes were mutated with a VAF greater than 0.5% in the primary tumor (13 non-silent mutations were in the Catalogue of Somatic Mutations in Cancer [COSMIC] (Forbes et al., 2015)) (S4 Table). Almost every Tier 1 mutation present in the A1 primary tumor was identified in one or more of the metastases (52/54), and in many cases the VAF was enriched in the metastasis (median: 5-fold enrichment, average: 8.8-fold, range: 1- to 38-fold; Figure 2.3A). Eleven mutated genes were shared among the primary and all matched metastases: *TARBP1, FCRL1, XIRP1, TRMT1, PANX3, MYSM1, PHLDB3, TBC1D25, LOC284288, MDS2,* and *TP53*. The adrenal metastasis and spinal metastasis contained the most unique SNVs, with seven and nine, respectively. The liver metastasis and lung metastasis did not have any private mutations at a VAF > 1%, although the lung metastasis did share two mutations with the adrenal metastasis that were not observed in the primary.

In patient A7, 75 Tier 1 genes were mutated with a VAF ≥ 0.4% in the primary tumor (14 of these non-silent mutations were in COSMIC) (Figure 2.3B). The VAF in all of the metastases had a median enrichment of 1.4-fold, closer to the primary tumor than in patient A1. All of the mutations identified in the primary tumor were detected in at least one metastasis, and 65 mutations, including mutations in *RUNX1T1, ADGRB2, KMT2C, RP1, TP53,* and *AKT3,* were shared across the primary and all matched metastases. There were 75 mutations identified in one or more of the metastases that were not observed in the primary tumor (8 of these nonsilent mutations also were in COSMIC). The majority of these metastasis-specific mutations (54/75) were present in two or more metastases. Of the 21 mutations private to a single metastasis, the liver and kidney metastases had the most, with 7 and 8 private mutations, respectively. The rib metastasis contained no unique mutations.

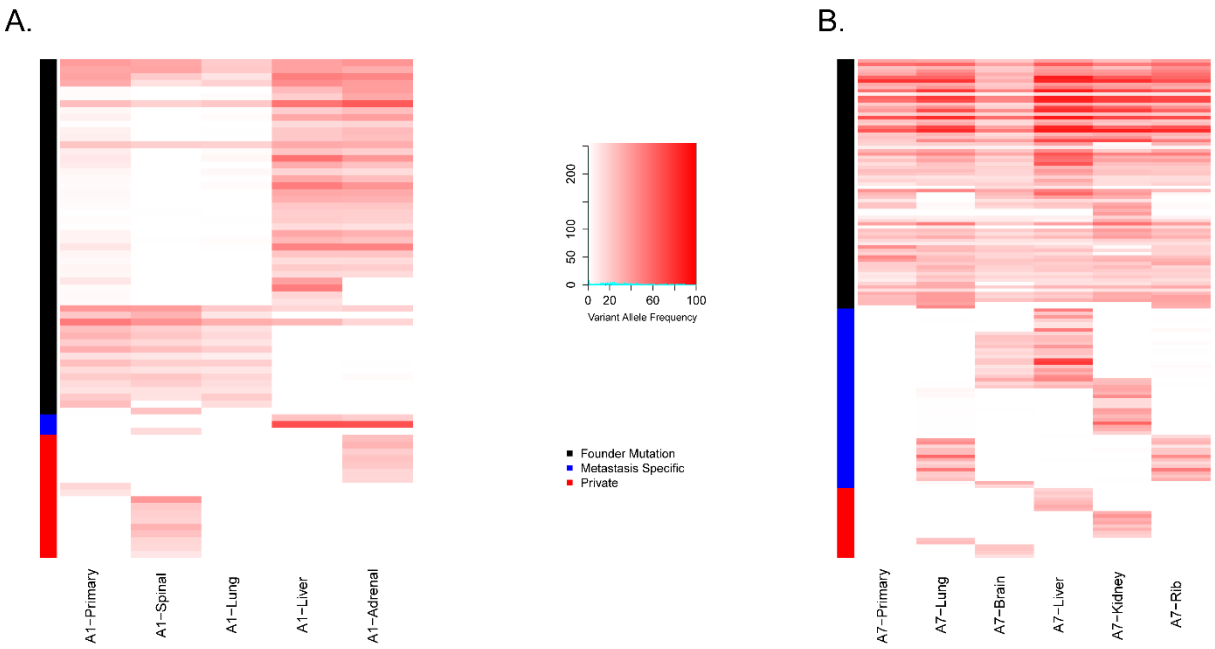

Figure 2.3. **Heat map of the DNA variant allele frequency of Tier 1 mutations in patients A1 and A7.** The vertical bar to the left of each heat map designates genes shared with the primary and metastases (black), genes mutated in metastases but not in the primary (blue), and genes private to a single individual metastasis (red) in (A) Patient A1 and (B) Patient A7.

TP53 as a common driver of metastasis. *TP53* alterations are frequently observed in basal-like breast cancers (Cancer Genome Atlas, 2012). *TP53* was the only shared somatic mutated gene between the two patients and was present in every tumor specimen sequenced. Close examination of patient A1 data identified an 11 bp deletion in *TP53* that was common to all samples (Figure 2.4). In patient A7, the *TP53* missense mutation H168R had a greater than 68% VAF in all tumors except the brain metastasis (31%). While this exact mutation was not observed in the TCGA breast cohort, a missense mutation was identified at the same position in one case (H168P) (Cerami et al., 2012; Gao et al., 2013), supporting the likelihood that alteration of *TP53* is a founding event critical for the development of basal-like breast cancer (Shah et al., 2012) and subsequent metastasis.

Mutations established early tend to be expressed and enriched in metastasis. We examined the mRNA expression data for evidence of expression of the somatic point mutations in primary tumors and metastases. Interestingly, mutations shared between the primary and metastatic tumors were more likely to be expressed (Figure 2.5, black dots) and were expressed at higher levels than mutations unique to metastasis (Figure 2.5, blue dots). In patient A1, 21/52 (40%) of the mutations established in the primary were expressed in the metastases (Figure 2.5A, black dots). In patient A7, 47/75 (63%) of the mutations established in the primary were expressed both in the primary and metastases (Figure 2.5B, black dots).

Fewer mutations were detected only in the metastases, and those mutated transcripts had lower RNA expression than mutations shared with the primary (Figure 2.5, blue dots). In patient A1, 2/3 mutations shared among more than one metastasis but not in the primary tumor were expressed (Figure 2.5A, blue dots), while only 4/18 private mutations (detected only in one tumor) were expressed (Figure 2.5A, red dots). In patient A7, 23/54 (43%) of the mutations that

were shared across the metastases but not with the primary tumor were expressed, and 8/21 (38%) of the private mutations were expressed (Figure 2.5B).

Interestingly, many of the expressed metastasis-specific mutations occur in genes that are involved in DNA damage responses, RNA processing, and degradation of the extracellular matrix (ECM). In patient A1, metastasis-specific mutations included *FANCF* and *SMC6* (DNA double-stranded break repair), *DDX6* (promotes mRNA degradation), and *HYAL3* (degrades hyaluronan in the ECM) (Rebhan et al., 1997). In patient A7, *AQR*, *DOCK6,* and *HLTF* were shared across metastases and expressed. Metastasis-specific mutations in patient A7 included *CASC3* (the core of the exon junction complex), *TIMP3* (degrades ECM), and *LAMA5* (part of the ECM) (Rebhan et al., 1997). These could represent convergent evolutionary paths to the resistance of DNA damaging agents and promotion of cell mobility and survival.

Figure 2.4. TP53 Deletion in A1. Genome view of the 11 bp deletion of TP53 in Patient A1 at chr17:7,579,474 to chr17:7,579,485, present in the primary tumor and all of the metastases.

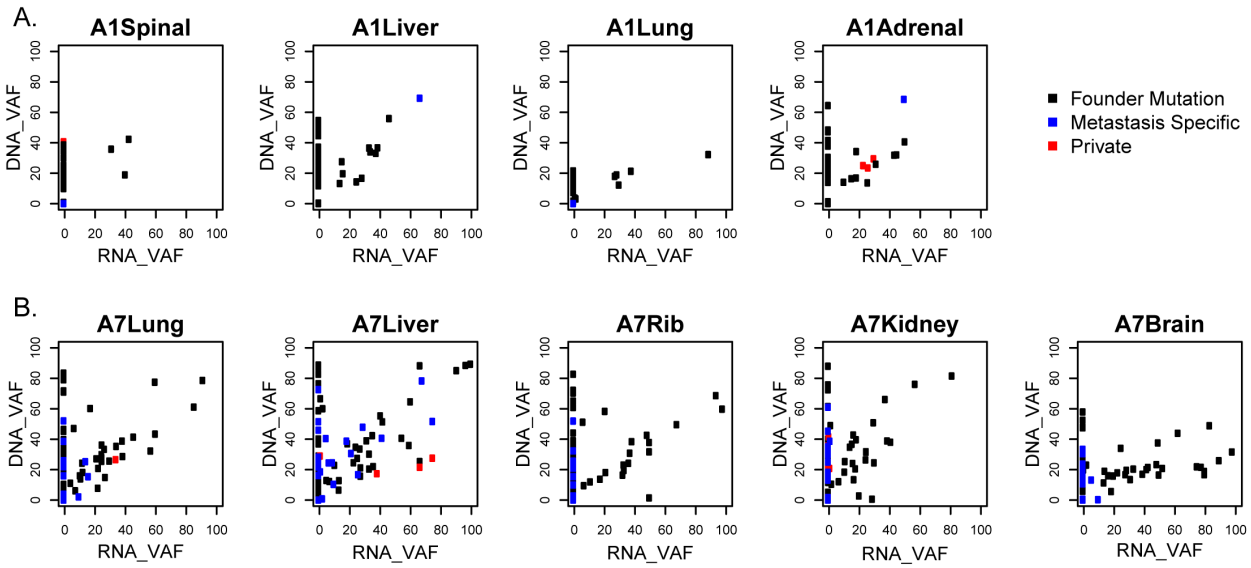

Figure 2.5. Gene expression of variant alleles. Variant allele fractions (VAFs) of each point mutation were determined from mRNA-sequencing data and compared to those from combined whole genome sequencing (WGS) and validation sequencing data. Gene variants shared in the primary and metastases (shared mutations, black), metastases but not primaries (metastases specific, blue), or only in one metastasis (private, red) in patients A1 (A) and A7 (B) are shown.

Structural variations tend to be established early in metastasis. To further explore the development of larger genomic alterations during metastasis, Circos plots were generated to illustrate the combined Tier 1 somatic mutations, DNA copy number alterations, and SVs for each sequenced tumor (patient A1: Figure 2.6; patient A7: Figure 2.7). These illustrate that, overall, SVs were mostly established in the primary tumor and maintained through the different metastatic processes.

In patient A1, all 8 of the SVs in the primary tumor were shared with the metastases (Figure 2.6), including one that was specifically shared with the adrenal and liver metastases. The metastases had few additional interchromosomal SVs, and these were shared, except in the spinal metastasis. Interestingly, the spinal metastasis evolved to have many more rearrangements between chromosomes 2 and either 3, 8, 12, or 16.

In patient A7, the brain and kidney metastases shared most interchromosomal SVs with the primary (Figure 2.7). The rib and liver metastases had three private SV alterations each (of a total of six and eight alterations, respectively), while the lung metastasis showed many more private interchromosomal SVs than the other metastatic samples.

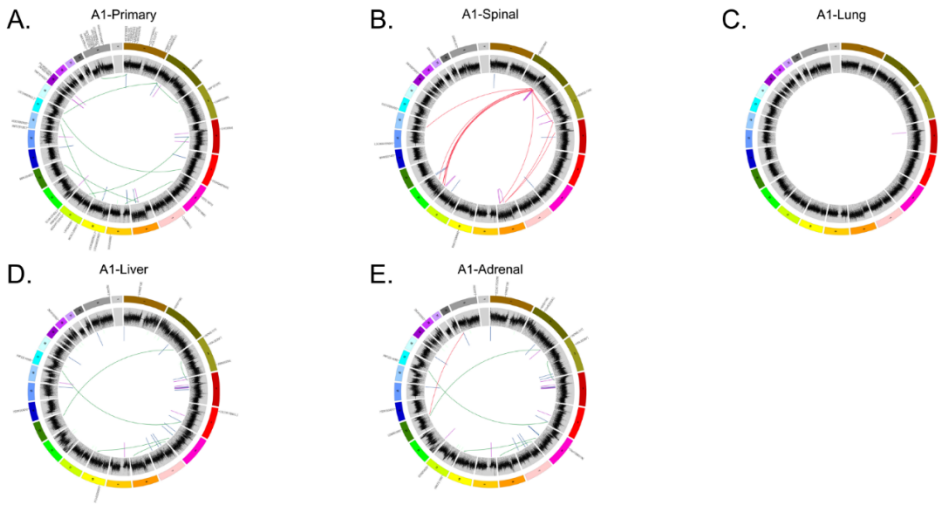

Figure 2.6. DNA alterations of matched primary and metastases of patient A1. (A–F): Circos plot displays mutations, copy number, and structural rearrangements in the (A) primary, (B) spinal, (C) lung, (D) liver, and (E) adrenal metastases. Translocations with significant read coverage include shared (green) and private (red) interchromosomal and shared (purple) and private (blue) intrachromosomal translocations.

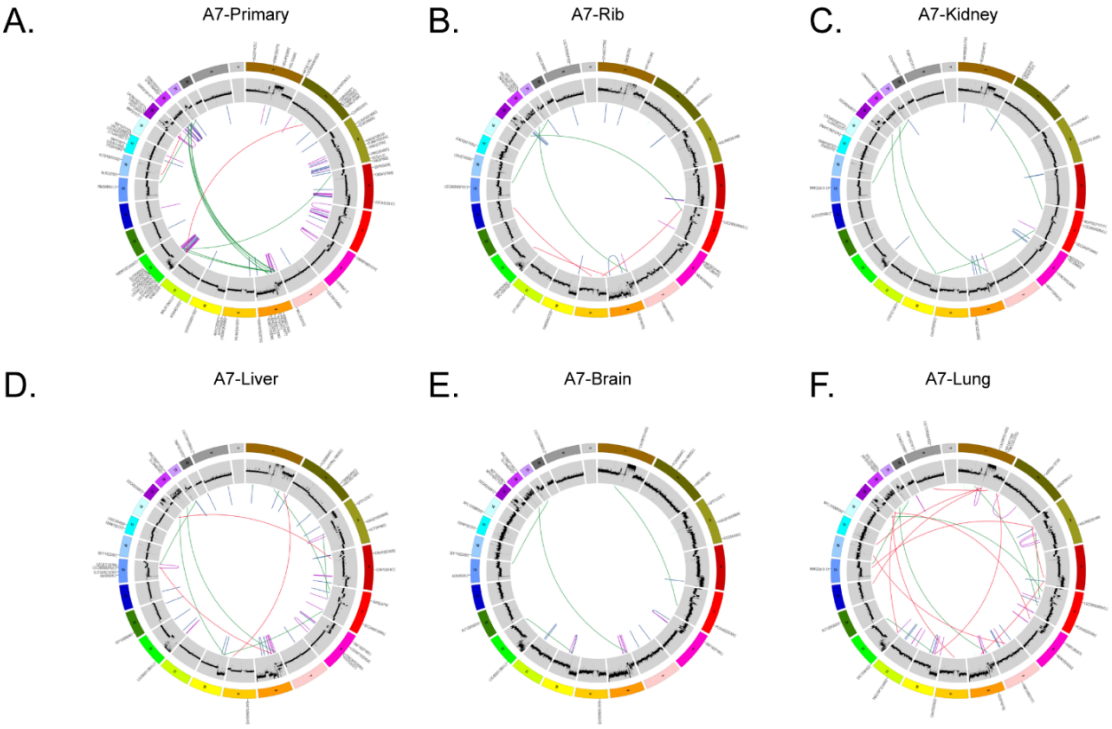

Figure 2.7. Circos plots of matched primary and metastases of patient A7. Circos plots displaying mutations, copy number landscape, and structural rearrangements (order starting from outside) in the (A) primary, (B) rib, (C) kidney, (D) liver, (E) brain, and (F) lung metastases. Translocations with significant read coverage include shared (green) and private (red) interchromosomal and shared (purple) and private (blue) intrachromosomal translocations.

FBXW7-INPP4B fusion in patient A7. To confirm SVs, we created a modified genome that represented the possible new alignments in RNA space. Realigning A7 data to this map demonstrated expression of an *FBXW7-RNF150* fusion gene observed in all A7 samples, indicating early fusion of this gene in the development of this patient's breast cancer (Figure 2.8). Interestingly, deletion of the last ten exons of *FBXW7* was previously reported as a founding event in a basal-like breast cancer (Ding et al., 2010). The 5′ end of the fusion in patient A7 began at exon 3 or 4 of *FBXW7*, which likely inactivated *FBXW7*. The 3′ end of the fusion occurred just before *RNF150*, resulting in deletion of *INPP4B*. There was decreased RNA expression of *INPP4B* in this patient, further supporting the deletion of *INPP4B* by the *FBXW7-RNF150* fusion gene event. *INPP4B* has important implications in breast cancer that include DNA repair defects (Ip et al., 2015), increased genomic instability (Weigman et al., 2012), and inhibition of the PI3K pathway (Gewinner et al., 2009).

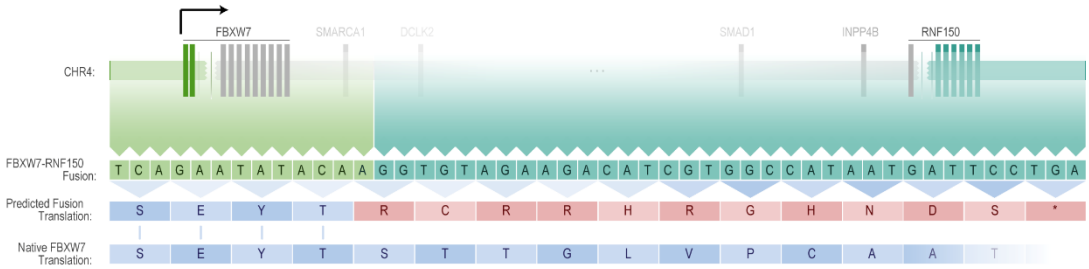

Figure 2.8. FBXW7 fusion. Representative illustration of FBXW7 fusion and INPP4B deletion in all tumors from A7.

Multiclonal Evolution of Metastasis in Two Patients with TNBC

To understand the Darwinian evolution occurring in the primary tumor and throughout metastasis (Campbell et al., 2008), we established the subclonal relationships and phylogenetic trees for patient A1 (Figure 2.9, S6–S8 Figs) and patient A7 (Fig 5, S9–S10 Figs).

Subclonality analysis using SciClone of the A1 patient samples demonstrates that the primary tumor predominantly contained clones 1, 3, 5, and 8, with very low allele fractions of minor clones 2, 4, and 7 (Figure 2.9,Figure 2.10A). Clone 1, established in the primary tumor, seeded all other metastases. Of the other major clones in the primary, clones 3 and 5 seeded the lung metastasis, while clone 3 additionally seeded the spinal metastasis. This metastasis then continued to evolve, developing private clone 9 (Figure 2.10A). These clones (3 and 5) were mutually exclusive with minor clone 2, which was found in the primary tumor, lung, liver, and adrenal metastases (Figure 2.11). Two of the minor clones in the primary tumor (clones 2 and 4) became the dominant clones in the liver and adrenal metastases, with additional private subclonal evolution in the adrenal metastasis (clone 6). Interestingly, clone 7 was established in the primary tumor and also metastasized to the liver, but not to the adrenal, metastasis (Figure 2.10B). Using ClonEvol, there were two potential models for clone 7 development that we were not able to fully resolve; either it evolved (1) from clone 4 (Figure 2.11) or (2) independently from clone 2 (Figure 2.12). This result demonstrates that the multiclonal metastatic potential residing in the primary tumor is maintained through metastasis.

Importantly, patient A1 presented at stage IV and only received two doses of radiation and one cycle of single-agent taxane before death. Thus, her primary-to-metastatic disease likely is representative of the natural course of basal-like breast cancer rather than representing selection from the evolutionary pressure imposed by therapy.

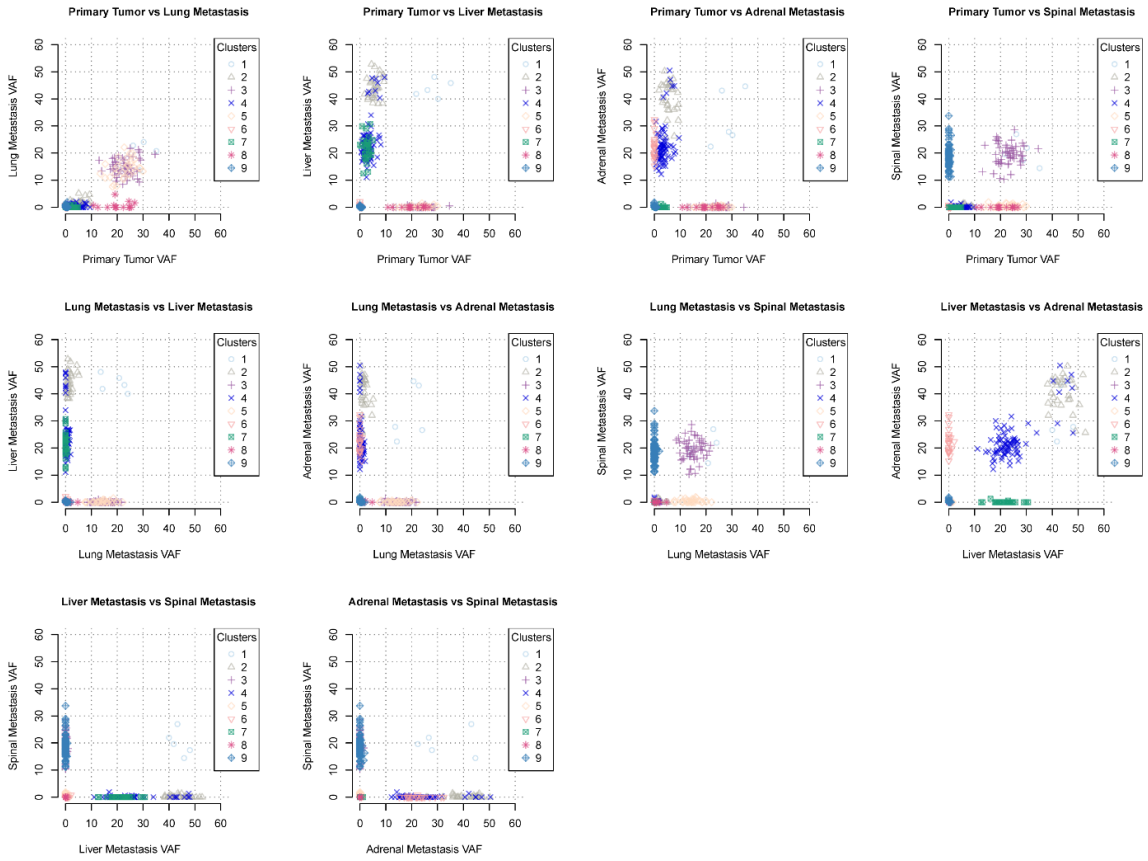

Figure 2.9. SciClone analysis of A1. SciClone analysis of variant allele frequencies in copy number neutral regions of each tumor using Bayesian beta mixture modeling and multi-dimensional clustering of tumors from patient A1. Multiple clones are shared in the primary and metastases, with Clone 1 in the primary and all matched metastases; Clone 2: primary, adrenal, and liver; Clone 3: primary, adrenal, and liver; Clone 4: primary, lung, and spine; Clone 5: primary, adrenal, and liver; Clone 6: primary and lung; Clone 7: adrenal; Clone 8: primary and liver; Clone 9: primary; and Clone 10: spinal.

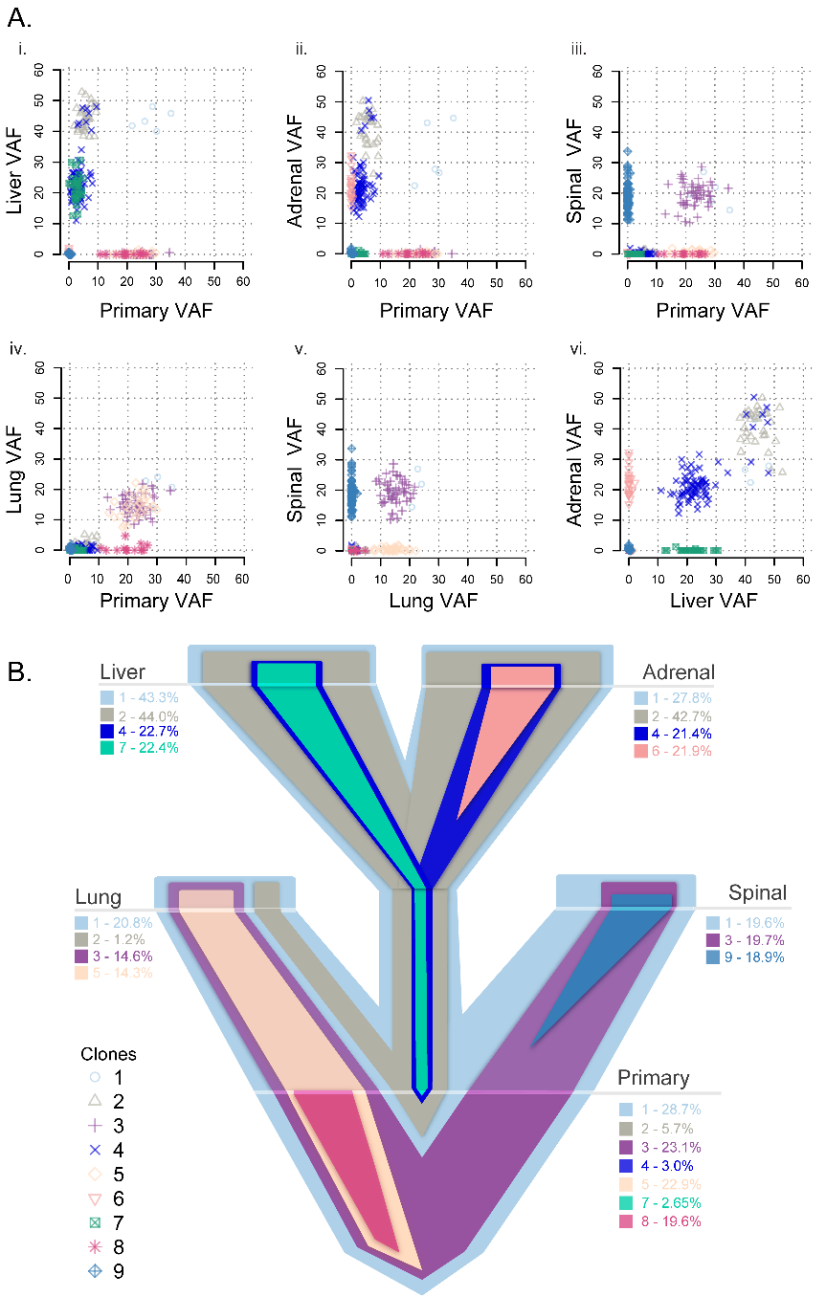

Figure 2.10. Clonality analysis of each tumor from patient A1. VAFs among the primary and matched metastases in patient A1 (A) and a representative evolutionary tree (B) colored by subclone based on the clonality plots in panel A, with the width of the branch indicating the approximate percentage of that clone within the tumor. Clone 1 is established in the primary tumor and seeded all distant metastases. Clones 2 and 4 from the primary tumor seeded the liver and the adrenal gland, with clone 7 concurrently seeding the liver from the primary tumor. Clones 3 and 5 from the primary tumor seeded the lung, with clone 3 also seeding the spine. Private clones include clone 6, specific to the adrenal metastasis; clone 8, specific to the primary tumor; and clone 9, specific to the spinal metastasis.

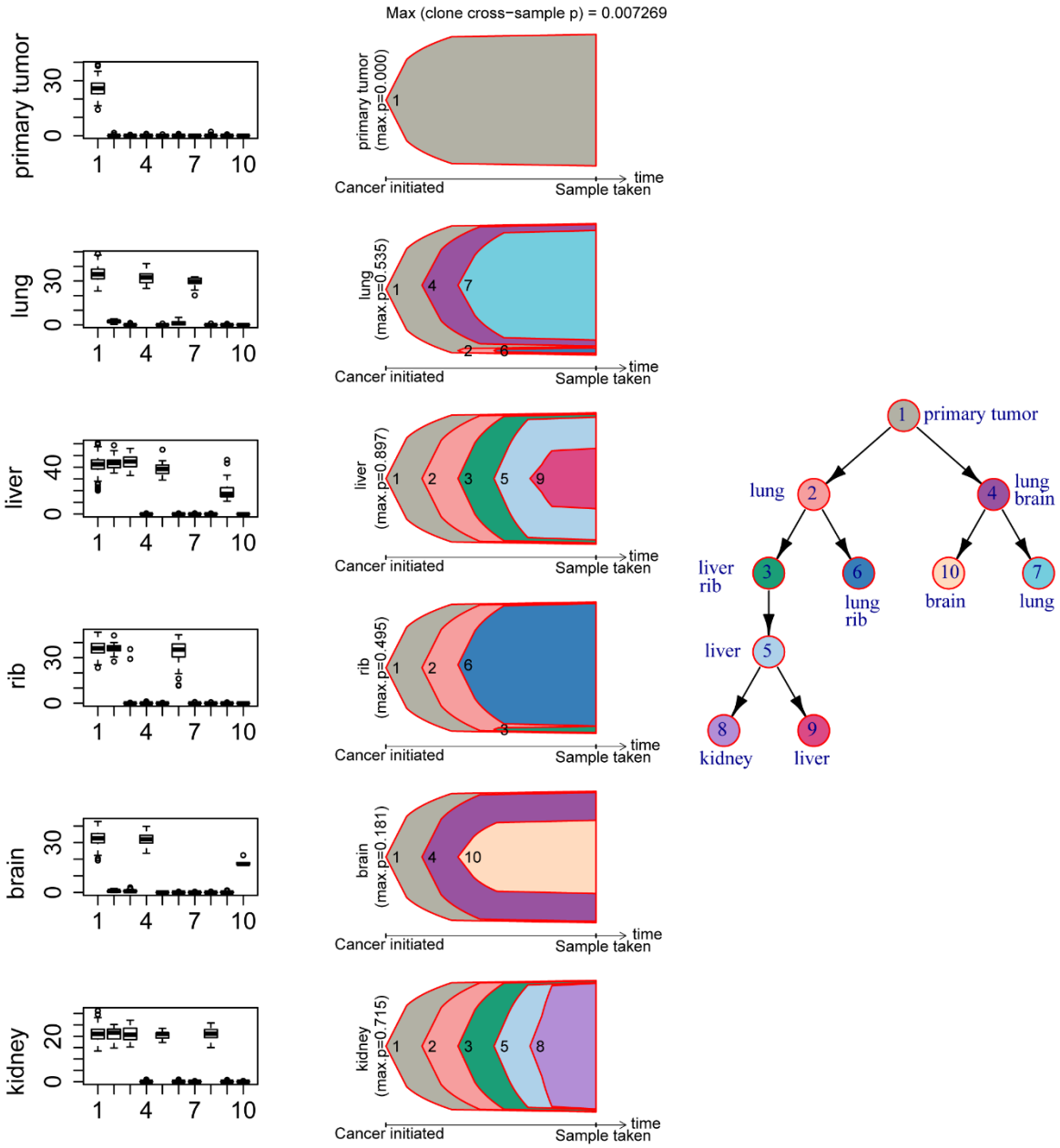

Figure 2.11. ClonEvol analysis of A1. ClonEvol demonstrates that Clones 1 and 2 are founding clones that seed the distant metastases at different percentages. Clone 2 and Clone 3 are exclusive of one another, leading to separate lineages. The proportion of each clone is demonstrated by the width of the nested shapes.

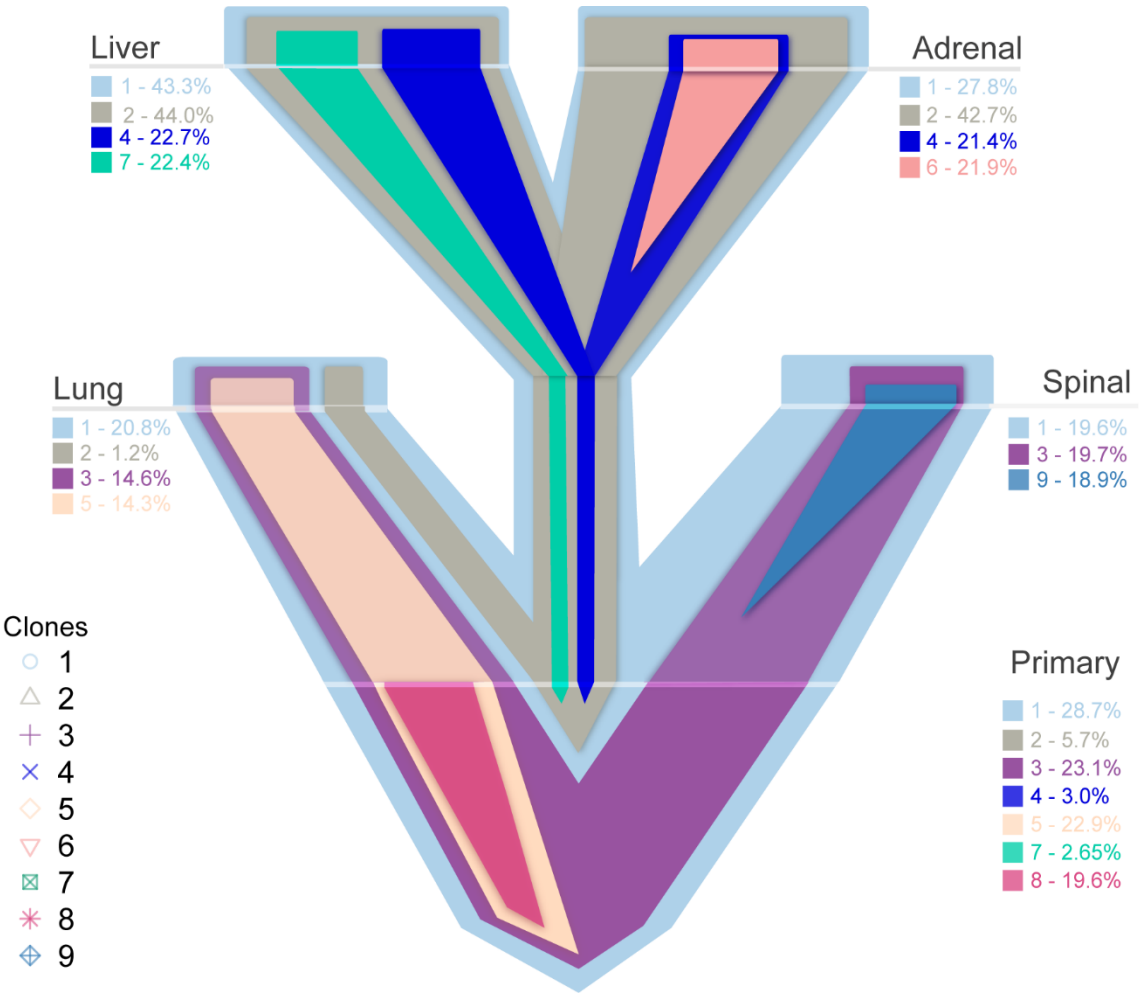

Figure 2.12. Representative evolutionary tree of an alternative model of A1. ClonEvol predicted two possible evolutionary lineages of clones in patient A1. The first model is in Fig 5B. The alternative model demonstrating that Clone 7 is independent of Clone 4 is presented.

In patient A7, the subclonal structure was determined by SciClone (Figure 2.13), and a single model of evolution was suggested by ClonEvol (Figure 2.14). The primary tumor consisted of one main clone (Figure 2.15), seeding all other sites of metastasis at the highest VAF observed. The main clone then diverged to two lineages, giving rise to clone 2 predominantly in the liver, kidney, and rib and clone 4 predominantly in the lung and brain (Figure 2.15B). Clone 4 is present in the lung and brain metastases at an almost equivalent VAF to the founding clone 1. Clones 2 and 6 in the rib are also present at an almost equivalent VAF to clone 1; clones 2 and 6 are seen at a low VAF in the lung. These clonal data paint a complex picture with two possible explanations: either the split of clone 1 into clones 2 and 6 and clone 4 occurred prior to metastatic spread (Solution A, Figure 2.15B) or these clones cross seeded from the rib metastasis to the lung metastasis (Solution B, Figure 2.15B). Clone 2 further evolved to clones 3 and 5 in the liver and kidney metastases. We favor the first hypothesis, namely that clone 2 in the rib, liver, and kidney metastases is at a VAF equivalent to the founding clone, indicating that the evolution of this clone occurred before metastatic seeding. All metastases aside from the rib metastasis also contained private subclones.

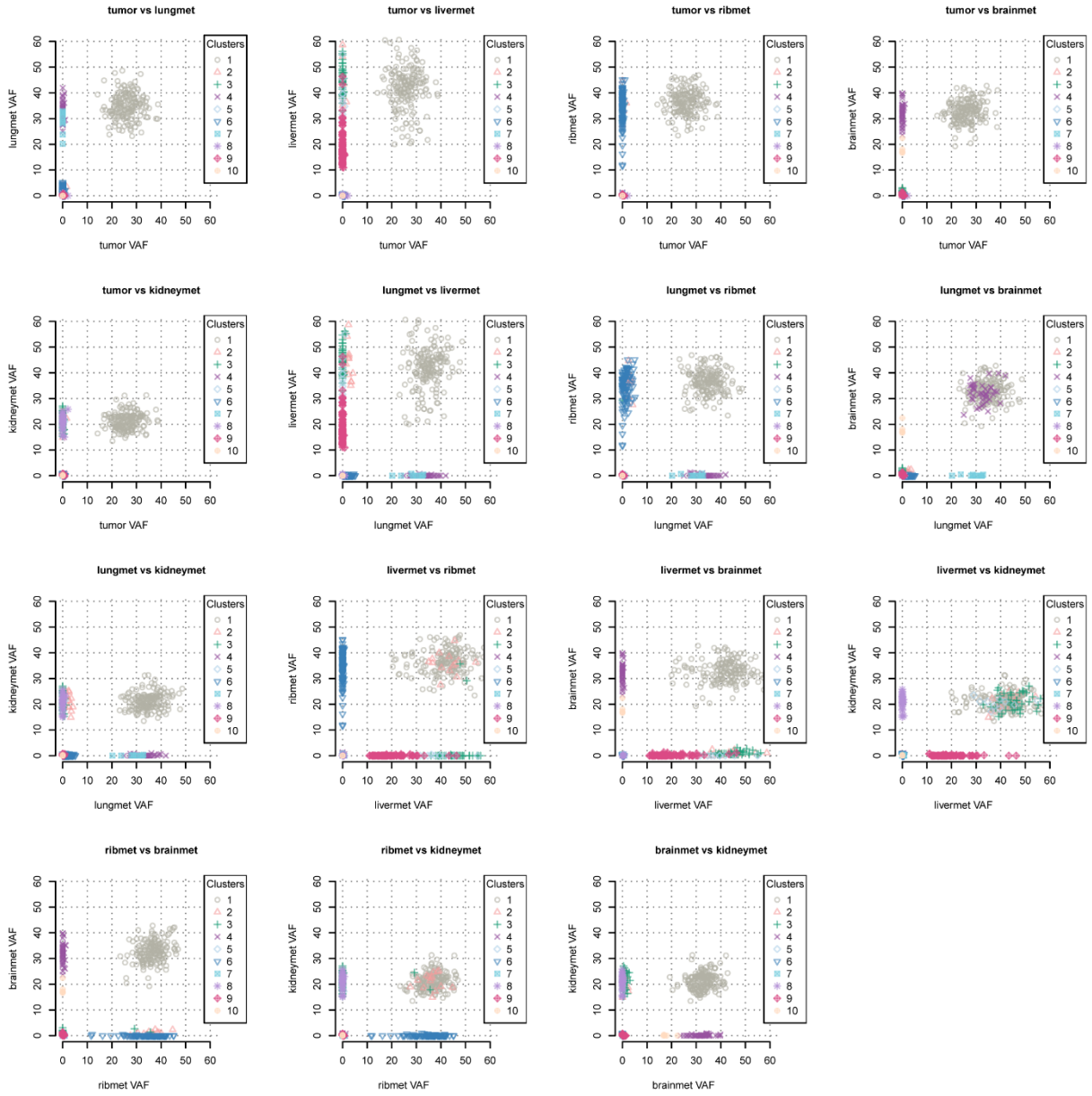

Figure 2.13. SciClone analysis of only copy number neutral regions demonstrates multiclonal seeding of metastases. The Lung metastasis contains both branches of the clonal tree, predominantly containing Clone 4 but with a small fraction of Clone 2. In contrast, the rib metastasis contains predominantly Clone 2 with a small minority of Clone 3. Private clones are seen in all metastases.

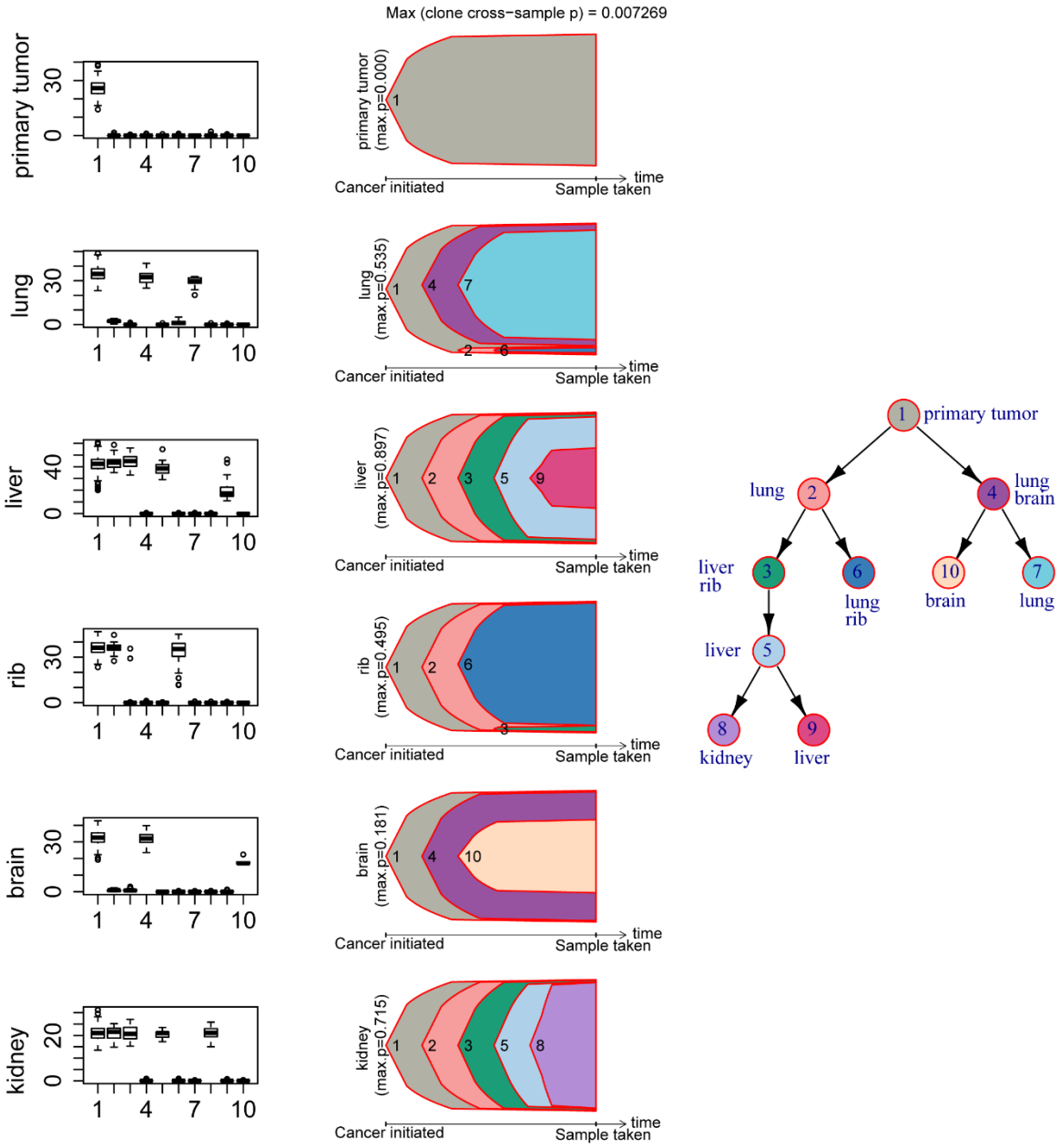

Figure 2.14. Clonevol analysis of A7. ClonEvol of the copy number neutral mutations from SciClone analysis demonstrates one founding clone leading to a branched pattern of Clones 2 and 4. Private clones are present in all metastases.

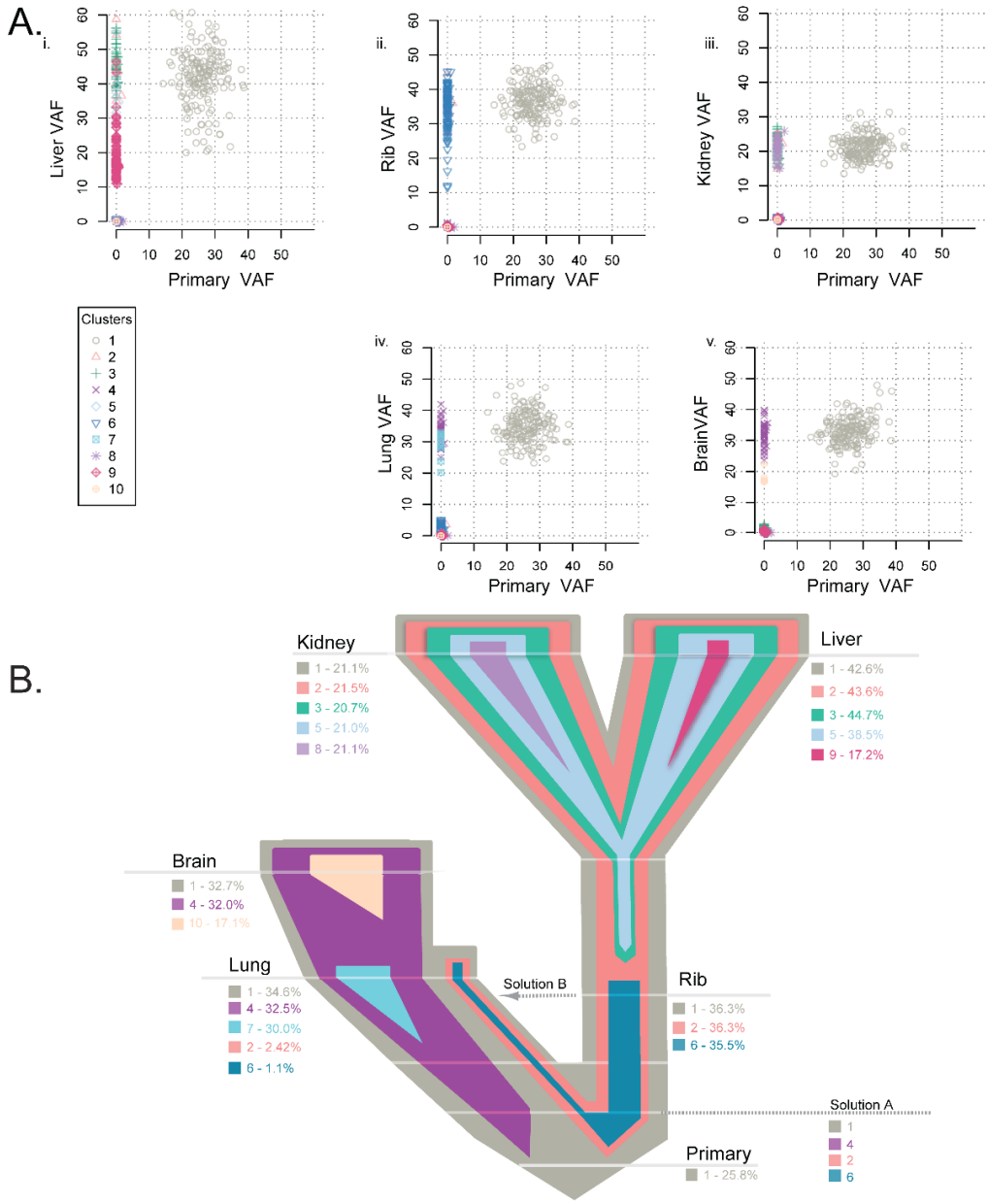

Figure 2.15. Clonality analysis of each tumor from patient A7. Clonality shared among the primary tumor and matched metastases in patient A7 (A) and the representative evolutionary tree (B) colored by subclone identity based on the clonality plots in panel A, with the width representative of the percentage of the clone within that tumor. Clone 1 was established in the primary tumor and maintained through metastatic spread in every tumor. Clone 2 was present in the liver, kidney, and rib and at a low frequency in the lung, while clones 3 and 5 were additionally shared by the liver and kidney metastases. Clone 6 was present in the rib and a low frequency in the lung metastases. Brain and lung metastases shared clone 4. Each tumor had a private clone not shared with any other tumor: clone 7 specific to the lung, clone 8 specific to the kidney, clone 9 specific to the liver, and clone 10 private to the brain.

Discussion

Whole genome sequencing and mRNA sequencing of two TNBC/basal-like breast cancer patients with primary tumors and multiple matched metastases demonstrated significant genetic similarity between the primary breast cancers and their matched metastases. Patient A1 demonstrated significant intratumoral heterogeneity established in the primary tumor and multiclonal seeding of metastasis. Interestingly, patient A7 possibly contained a more homogenous primary breast cancer that then led to diverse, heterogeneous metastases. Even though there is continued evolution, the acquisition of mutations private to a single metastasis likely had limited impact on the metastatic potential, as these mutations were rarely expressed or were expressed at low levels. In contrast to earlier findings in renal cell carcinoma of monoclonal metastasis seeding (Gerlinger et al., 2012), basal-like breast cancer metastases can be the result of multiclonal seeding of cells established in the primary. The results presented here are inconsistent with a single cell of a primary breast cancer seeding a distant metastasis (Navin et al., 2011). Herein, we describe an example of multiple subclones that resided within a primary tumor followed by multiclonal seeding of all distant metastases as well as a common disruption of *TP53*.

In both patients, relatively few mutations occurred once the tumor cells left the primary site, and of those that did alter protein coding sequences, the mutations were not highly expressed at the RNA level in general. The high correlation of gene expression among primaries and matched metastases illustrates that subtype is typically maintained throughout metastasis (Harrell et al., 2012), and that specific intrinsic subtypes have an inherent tendency to metastasize to specific organs (Harrell et al., 2012; Smid et al., 2008). Taken together, these results suggest that the metastatic potential was present within the primary tumor of these two basal-like breast cancer patients. Here, we uncover a genetic explanation for the close

correlation of gene expression in metastases and matched primaries—namely that, in the two cases examined, the samples from a given individual were much more genetically similar than they were dissimilar, both on the DNA and RNA levels.

While the majority of genetic alterations present in metastases were shared with the matched primary cancer in these two patients examined, we also identified a significant amount of intratumoral heterogeneity, evident because multiple subclones were detected within each metastasis. Patient A1 demonstrates that more than one subclone from the primary seeded each metastasis, and the intratumoral heterogeneity in the primary tumor setting was mostly reflected in each metastasis. In patient A7, the lung metastasis exhibited diverse intratumoral heterogeneity, with two small subclones (2 and 6) found at high frequency in three of the other metastases. There are two possible explanations for the complex clonal patterns seen in patient A7: either the two dominant clones (clones 2 and 4) were established in the primary and were not sampled in the piece of the primary tumor that was actually sequenced, or clones 2 and 6 in the rib cross seeded into the lung metastasis. While one metastasis seeding another metastasis has been previously demonstrated in prostate cancer (Gundem et al., 2015), we also recognize that the A7 primary breast cancer likely had spatial heterogeneity that was not fully captured by our sequencing (Yates et al., 2015). In fact, the A7 primary breast cancer piece sequenced was a skin punch biopsy taken from a 5 cm primary breast cancer, rather than a tumor resection. Hence, samples from multiple portions of this tumor were not sequenced. Of the two possibilities, the most parsimonious explanation for the observations relevant to patient A7 is that multiclonal seeding of the metastases did occur and that our limited sample did not permit detection of clones 2 and 6. Hence, only subsequent deep sequencing of additional portions of the A7 tumor would resolve the issue of monoclonal versus multiclonal seeding from the primary. Unfortunately, no additional specimens exist for this patient. Regardless of this, in

patient A7 multiple multiclonal seeding events were discovered, such as the rib metastasis seeding the kidney and liver.

The genetic heterogeneity in both of the primary tumors and the resulting metastases may explain why many metastatic TNBC patients fail to have a durable treatment response and instead progress within a few years (Anders and Carey, 2008). In particular, heterogeneity provides for a wealth of individual genotypes, thus yielding a genetic diversity from which chemotherapy resistance may arise. Treatment has been shown to select for therapy-resistant clones in primary breast cancer (Juric et al., 2015; Li et al., 2013; Miller et al., 2016), and therapy can select for subclones in the metastatic setting.

While our studies provided evidence of multiclonal seeding of metastasis in these two patients, both with basal-like breast cancer, our results may or may not apply to a larger cohort of patients with basal-like breast cancers, to other subtypes of breast cancers, or to other cancer types. Even within the poor-prognosis basal-like subtype, patients often receive many more lines of therapy and have more favorable responses to their therapies for a longer duration than the two patients presented here. Furthermore, patients with luminal and HER2-enriched breast cancer have comparatively more opportunities to benefit from targeted therapies such as tamoxifen, aromatase inhibitors, and/or HER2 agonists such as trastuzumab, lapatinib, or pertuzumab. Since neither patient A1 nor A7 was treated with targeted therapies, there were different selective pressures in the metastatic setting compared to current standard of care for ER+ and HER2+ patients.

The basal-like subtype is a highly aggressive cancer that often metastasizes to the lung and brain within 5 y of diagnosis. This is in contrast to luminal A breast cancers, which are typically more indolent, are less likely to progress to stage IV, and typically metastasize first to the bone (Haque et al., 2012). The difference in these patterns of relapse and the timing with

which they occur suggest fundamental differences in disease progression between the subtypes (Ellis et al., 2012) within the context of drastically different treatment strategies. Continued analyses of larger datasets representing each of the subtypes and patients with varying clinical histories will be necessary to identify consistently altered genes to define early versus late drivers, metastasis-site specific alterations, and differences among the mechanism of metastasis across various subtypes of breast cancer.

CHAPTER 3 – THE EVOLUTION OF LETHAL BREAST CANCER METASTASIS: MULTICLONAL SEEDING DRIVEN BY TP53 AND COPY NUMBER ALTERATIONS

Preface

This work is currently under review and is a first author manuscript. The UNC Tumor Donation Program was started by Dr. Lisa Carey, with tissues collected by Niamh Kieran, Julie Benbow, and Amy Garret. Autopsies were performed by Vincent Moylan and Claudia Brady. Tissue quality control was performed by pathologists. Chad Livasy and Leigh Thorne. DNA and RNA isolation, library preparation, and sequencing was performed mostly by Dr. Xiaping He with the latter DNA samples done by me. Sequencing data was mapped by Alan Hoyle and Joel Parker. Copy number was evaluated by Mengjie Chen. Droplet PCR of *ESR1* mutations was performed by Sunil Kumar and Gaorav Gupta. I performed all scientific investigation of both the RNA and DNA sequencing analyses, with significant oversight and mentorship from Katherine Hoadley, Joel Parker, Elaine Mardis, Lisa Carey, Carey Anders, and Charles Perou. I designed the figures, supplemental data, and written text of this manuscript.

Introduction

Breast cancer is the second leading cause of cancer related death in women and is typically caused by metastasis. Breast cancer is a heterogeneous disease comprised of multiple "intrinsic" expression-based subtypes (Perou et al., 2000), wherein the subtype predicts future sites of recurrence and survival (Harrell et al., 2012; Smid et al., 2008). While this evidence strongly supports the hypothesis that the primary tumor contains information about metastatic potential, the factors responsible for this metastatic potential are still not well understood.

It also remains unknown whether metastasis is the result of a single cell from the primary tumor circulating in the blood to seed and survive at distant sites (i.e. monoclonal seeding), or instead results from a collection of multiple cells of the primary that seed together and survive at distant sites (i.e. multiclonal seeding). Additionally, it is unclear when the ability to metastasize is acquired: by the original primary cells, over time during some dormancy period that follows treatment, or with adaptation at the final site of metastasis. Understanding the heterogeneity of metastatic sites, for example whether they correspond genetically to one or multiple clones from the primary tumor, could more accurately inform treatment decisions.

Several recent studies in other tumor types have demonstrated both single clones leaving the primary to seed distant metastases (Gerlinger et al., 2012) and multiclonal seeding (Gundem et al., 2015; Maddipati and Stanger, 2015). In small cohorts of breast cancer patients, previous breast cancer studies have also demonstrated both monoclonal (Ding et al., 2010; Krøigård et al., 2015) and multiclonal (Murtaza et al., 2015) seeding of single, matched metastatic sites. Multiregional sequencing of breast cancer has demonstrated that significant heterogeneity existed within 8/12 primary tumors (Yates et al., 2015). Brastianos and colleagues demonstrated continued acquisition of new driver mutations in the context of brain metastases (Brastianos et al., 2015). Using two patients and whole genome sequencing of multiple matched metastases and primaries, we previously reported multiclonal seeding of triple negative, basal-like breast cancers (Hoadley et al., 2016). These studies were, however, limited by studying small cohorts of patients and typically only one or two matched metastatic sites per patient.

Additionally, these studies defined “genetic drivers” as genes previously shown in large scale sequencing projects to be significantly mutated above the background rate that is expected by chance (Cancer Genome Atlas, 2012; Ciriello et al., 2015; Dees et al., 2012a; Lawrence et al., 2013). The actual biological or functional impact of these alterations in individual patients therefore was not measured. Computational approaches incorporating gene expression from RNAseq data and known protein interaction networks could help to predict the

functional impact of individual DNA-based somatic alterations (Hou and Ma, 2014). Previous work has demonstrated the power of integrating gene expression and DNA alterations to define unique driver sets beyond mutational background (Silva et al., 2015). By employing the DawnRank method to determine the functional impact of mutations and copy number alterations, we can empirically define *drivers* on an individual tumor basis as well as the timing with which they occur during the development of breast cancer metastasis.

Here, we present the underlying evolutionary processes of breast cancer metastasis in a large cohort of primaries with matched multiple metastases per patient. Utilizing a Rapid Autopsy Program established at the University of North Carolina at Chapel Hill, we have collected matched primary and metastatic breast cancers from 16 individuals and performed RNA-sequencing (RNAseq) and DNA whole exome sequencing on the primary, 67 matched metastases (2-7 per patient) and a matched normal tissue comparator for each patient. We examine the clonal evolution of metastasis within each patient, copy number and mutational spectrum of the metastatic process in a subtype-specific manner, and apply a novel computational approach that integrates RNA and DNA sequencing data to identify genomic drivers. These results demonstrate the genetic diversity of the metastatic process and highlight the potential of using the primary tumor data as a means of targeting metastases.

Methods

Patient consent and tissue processing

Tumor tissue was obtained from metastatic breast cancer patients who consented to Rapid Autopsy at the University of North Carolina prior to death. Primary, metastatic, and normal tissue were taken within 6 hours of death for all metastatic sites, both known and found, at time of autopsy. Tissues were frozen in the -80C freezer, and RNA and DNA were isolated from each tissue using Qiagen RNAeasy and DNAeasy kits, respectively, according to the manufacturer protocols (Valencia, CA). Primary breast cancer tissues taken at diagnosis were also acquired as available. Archived tissues in formalin-fixed paraffin-embedded (FFPE) tissues had total RNA isolated with Roche High Pure RNA paraffin kit Cat #03270289001 and DNA isolated with the Maxwell 16 FFPE Tissue LEV DNA Purification Kit (San Diego, CA). Quality of RNA was checked with the Agilent BioAnalyzer RNA 6000 Nano Kit (Santa Clara, CA).

DNA Whole Exome Sequencing

DNA was prepared for sequencing with the Agilent SureSelect XT library protocol (Santa Clara, CA). Fresh-frozen tumors were processed according to manufacturer's protocol 3ug input, while FFPE tumors were processed with the low-volume input according to manufacturer's protocol for 200 ng input. DNA libraries were captured and amplified with Agilent SureSelect Human All Exon v5 or v6 (Santa Clara, CA) according to the manufacturer's protocol. Quality of both DNA libraries and DNA exome capture quality and concentration were quantified with Agilent ScreenTape DNA 1000 and High Sensitivity D1000 respectively (Santa Clara, CA).

2x100 bp paired-end sequence data was generated from the Illumina HiSeq 2500 for each tumor or normal sample with 3 samples per lane. Illumina reads were mapped to the NCBI Build 36 reference sequence with BWA (Li and Durbin, 2009), realigned with ABRA (Mose et al., 2014), processed by biobambam2 (Tischler and Leonard, 2014), and called as somatic variants with STRELKA (Saunders et al., 2012).

We used minor allele frequency of highly variable SNPs in the general population for sample identity. All samples had an expected 87-100% identity with tumors from the same patient.

Copy number was called with SynthEx (*Silva GO* et al., manuscript in preparation). Briefly, the ratio of on-target and off-target exome reads of tumor were compared to a normal selected from the dataset by highest degree of similarity in library size and fold enrichment. Segment level ratios were calculated and log2 transformed. Copy number levels greater than 0.25 were considered as gains, and less than -0.32 as losses.

RNA Sequencing

Fresh-frozen (FF) RNA was prepared for sequencing with Illumina TruSeq polyA Select protocol. If libraries failed the protocol, they were then prepared with Illumina TruSeq RiboZero Gold protocol according to the manufacturer's protocol. FFPE RNA was prepared with Illumina TruSeq FFPE RiboZero Gold protocol according to the manufacturer's protocol. RNA libraries were sequenced as 2x50 base-paired end read with two samples per lane on an Illumina HiSeq 2500 sequencers. Reads were aligned with MapSplice (Wang et al., 2010), genes values were quantitated with RSEM (Li and Dewey, 2011), and counts were upper quartile normalized and log2 transformed for analysis.

Because of bias in FFPE and Total RNASeq data as compared to mRNAseq data, a normalization vector was calculated. Previously published matched samples of FFPE, total RNAseq, and mRNA sequenced samples with the same protocol were used to find the mean difference for each gene across each platform (Zhao et al., 2014). This was then applied to total RNASeq runs where a gene by gene adjustment was made in the total RNAseq samples.

Droplet PCR for ESR1 Mutations

Digital droplet PCR for wild-type (*WT*) and four hotspot *ESR1* alleles (D538G, Y537C, Y537S, and Y537N) was performed using the Raindrop Source and Sense instruments (Raindance^{TM} Technologies, Billerica, MA). Primers for a 75bp amplicon that includes these

hotspot mutations were used in conjunction with locked nucleic acid Taqman probes for wild-type (conjugated to TET) or mutant *ESR1* alleles (conjugated to FAM), purchased from Integrated DNA Technologies (IDT, Coralville, IA). The multiplexed genotyping reaction was validated using synthesized 125bp DNA fragments (gBlocks, IDT, Coralville, IA). Details of primer and probe sequences are available upon request. TaqMan Genotyping Master Mix (Applied Biosystems, Foster City, CA) was used for 10-100 ng of Covaris-sheared genomic DNA in a 50 µl reaction volume. After PCR amplification in a thermocycler (C1000 Touch™ Thermal Cycler, Bio-Rad®, Hercules, CA), the emulsion was analyzed on the Raindrop Sense instrument (RainDance™ Technologies, Billerica, MA) to measure the end-point fluorescence signal from each droplet using standard manufacturer's protocols. The fluorescence intensity and duration for each droplet in the FAM and TET channels were analyzed using RainDrop Analyst Software II (RainDance™ Technologies, Billerica, MA). Two-dimensional (FAM and TET intensity) plots were made for each sample and gates were used to define graphical areas with specific fluorescence properties. The number of droplet events specific for *WT* or mutant *ESR1* alleles was used to calculate the mutation frequency.

Computational Analyses

Hierarchical Clustering of Gene Expression. TCGA 1098 primary breast cancers (Ciriello et al., 2015) were merged with tissues from this study and median centered. Correlation centered hierarchical clustering of the median centered dataset with the PAM50 50 genes was performed with Cluster and visualized with Java TreeView.

Computational re-interrogation of somatic mutations in Related Tumors. Low read coverage or low tumor cell purity can cause our rigorous somatic mutation caller to miss mutations (Mose et al., 2014; Wilkerson et al., 2014). Thus, we examined high-confident somatic mutations from a related individual in all tumors from that patient. First, all of the somatic mutations from the tumors within one patient were collapsed into one vector, excluding

any guanine to adenine or cytosine to thymine mutations from FFPE tissues. For each mutation from a single patient, we then counted the mutant and reference alleles at that position from the original BAM file of each tumor from that patient. Variant Allele Fractions (VAF, alternate counts/total read counts) were recalculated from the new calls. All mutations from the dataset were interrogated in the normal sequence for all tumors in this dataset to account for false positives. Mutations with variant allele frequencies greater than 20% in at least two normal tissues from unrelated patients were excluded from future analyses.

DawnRank. We generated a binary matrix of 0 indicating no alteration and 1 indicating any alteration (mutation or copy number) for genes in the published DawnRank network (Hou and Ma, 2014). We combined TCGA log2 transformed normalized RNASeq data with RAP RNASeq data, median centered the data for each gene, and further transformed scores to the absolute value. DawnRank was then run for each individual tumor with a mu = 3. DawnRank scores were saved, and the top 5% of scores within each tumor were considered to be candidate drivers. These candidate drivers were then filtered by non-silent mutations and copy number laterations such that if an alteration was present, it was then identified as a driver.

RNA Interrogation of DNA Mutations. Using the 'union' list of mutations for each patient, UNCeqR (Wilkerson et al., 2014) was employed to count the number of mutated reads from the RNA BAMs at each position within a patient. Mutations with read counts from non-normalized RNA counts less than 5 reads in the RNA were considered to be 0. Mutations within each tumor were only considered if at least 5 reads of that gene were detected with RNASeq. Any genes in which the RNA gene expression of the gene was less than 5 were removed from the total number of DNA mutations in that tumor. UNCeqR was additionally run on the *de novo* mutation identification with default parameters.

Subclonal analysis. SciClone (Miller et al., 2014) was applied to all related tumors for each patient using the mutation calls following computational re-interrogation. The final clone for each patient was excluded due to wide scatter across all samples. SciClone was then rerun, and

clusters were tested for significance using SigClust (Huang et al., 2015). Clusters that overlapped, with the same pattern, and non-significant p values were collapsed into one clone. The mean VAF of the mutations comprising each clone was then calculated per tumor. Circles were then drawn with the radius of the circle proportionate to the mean VAF.

R Version. All statistical analyses were performed using R v.3.3.0 in RStudio (RStudio Team, 2015).

Results

Patient Characteristics

To explore the genetic evolution and drivers of breast cancer metastasis, we performed DNA whole exome sequencing and RNA sequencing for gene expression on 16 primary invasive breast cancers and 67 matched metastases (Figure 3.1, Figure 3.2). Our cohort had a median age at diagnosis of breast cancer of 45.5 years old, a median time to relapse of 14.5 months, and overall survival of 36.5 months (Table 3.1). These patients all received at least 1 chemotherapeutic agent, and all but one patient received radiation, predominantly to the breast and/or brain (Table 3.2).

We examined the clinical features and intrinsic molecular subtype of each of the primary tumors and their matched metastases. We applied the PAM50 subtype predictor (Parker et al., 2009) to determine the intrinsic molecular subtype (Appendix 3.1). Breast tumors from 4 patients were positive for estrogen receptor (ER) expression, but negative for HER2 amplification (ER+/HER2-) at diagnosis. All four of these patients are luminal or second closest to the luminal centroid (due to normal contamination). Primary breast tumors from 3 patients were clinically HER2-positive: 1 of the HER2-enriched subtype, 1 of the luminal subtype, and 1 of the basal-like subtype. Breast tumors from 9 patients were triple negative (negative for ER, progesterone receptor (PR), and HER2), with 6 patients classified as the basal-like subtype and 3 patients second closest to the basal centroid but called as normal-like due to normal tissue contamination (Appendix 3.1).

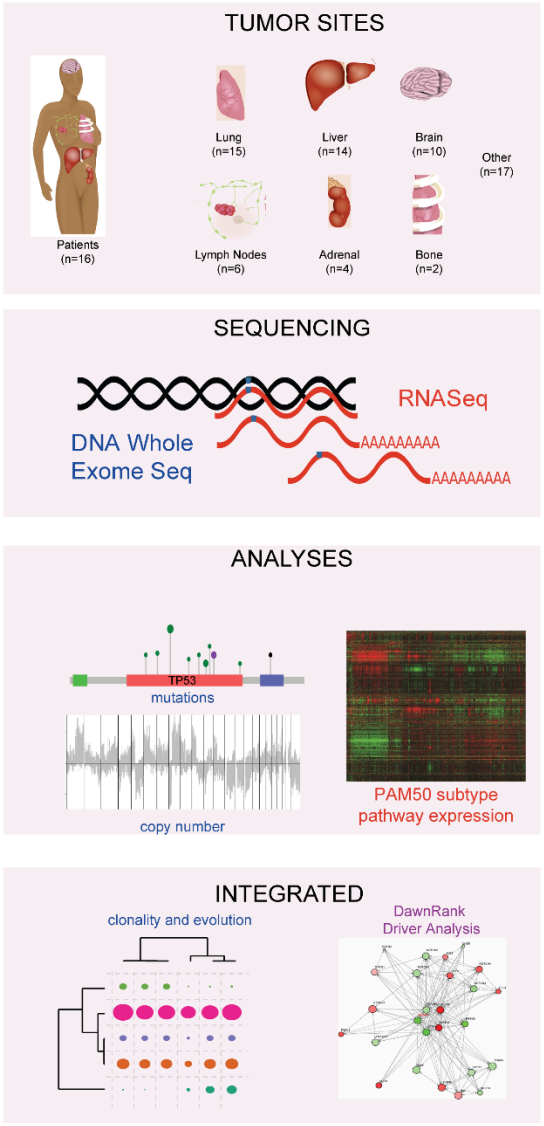

Figure 3.1. **Experimental design.** (A) 16 patients with primary breast cancers and matched metastases. (B) DNA Whole exome sequencing and RNA sequencing was performed on all tumors. (C) Gene expression, mutations and copy number alterations for all tumors were determined for each patient, and each tumor specimen. Subclonality analysis was performed with SciClone to define clones, then SigClust to perform posterior significance testing on the subclones, and hierarchical clustering to depict relationship of subclones and tumors in each patient. DawnRank driver analysis evaluated the network impact of copy number alterations and mutations to identify individual drivers in each tumor specimen.

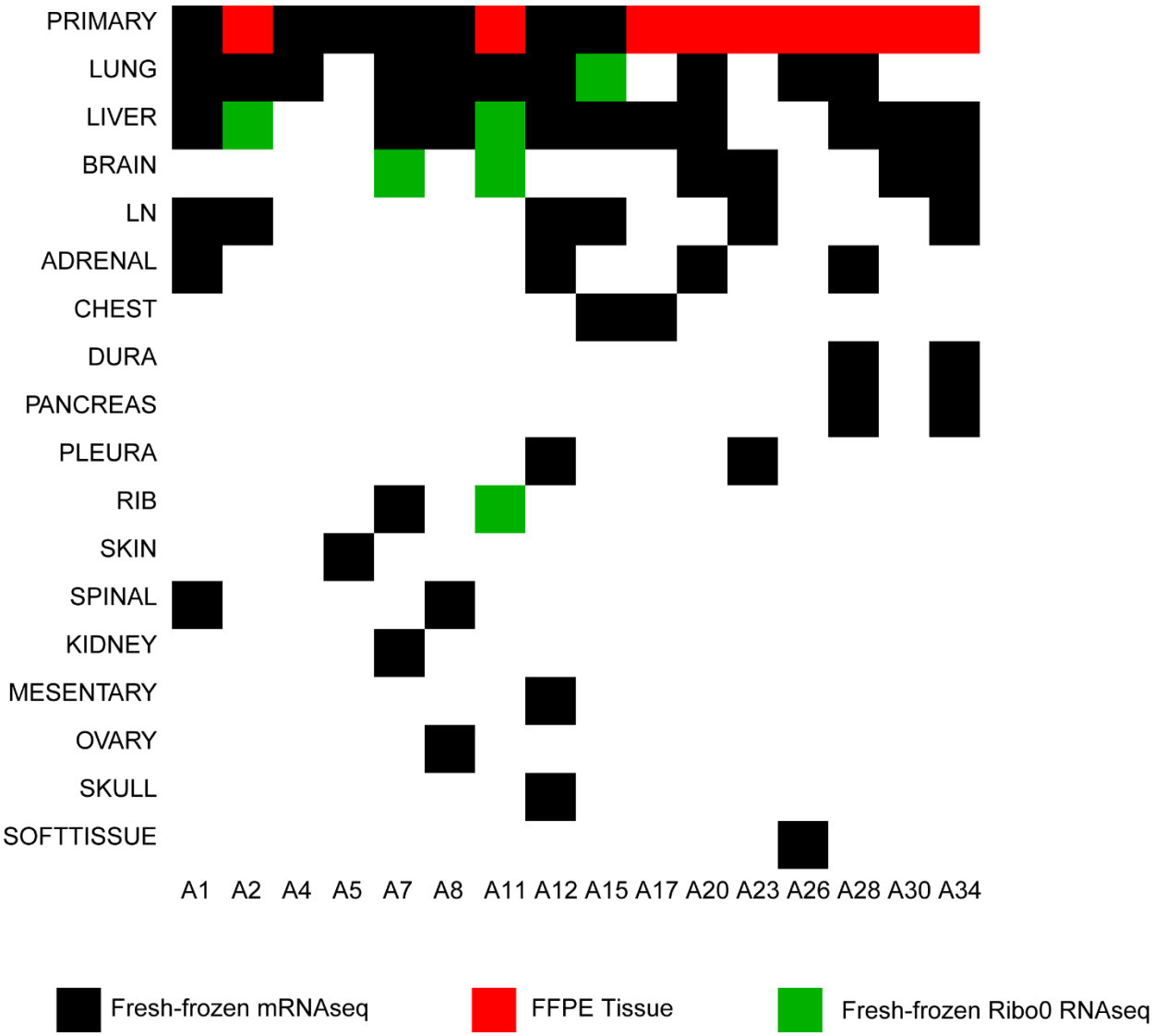

Figure 3.2. **Distribution of tumor specimens for each patient.** Diagrammatic view of the tumors from the 16 patients in the UNC Rapid Autopsy Program (RAP) that were sequenced with both RNA and DNA whole exome sequencing by site of disease (black = sequenced).

Table 3.1. Clinical History for each patient.

Patient	Race	ER Status (0 = negative; 1 = positive)	PR Status (0 = negative; 1 = positive)	HER2 Status (0 = negative; 1 = positive)	Age at Diagnosis	Stage at Diagnosis	Time to relapse (months)	Overall Survival (months)
A1	Caucasian	0	0	0	64	T4N2M1	0	1.5
A2	African American	1	0	0	57	T3N1M1	0	12
A4	Caucasian	1	1	1	42	T4N2M1	0	22
A5	African American	0	0	0	65	T4N0M0	23	26
A7	African American	0	0	0	57	T2N2M0	17	24
A8	Caucasian	1	1	1	45	T1N1M1	0	48
A11	Caucasian	0	0	0	46	T2N0M0	35	56
A12	Caucasian	1	1	0	64	T3N2MX	9	61
A15	Caucasian	0	0	0	59	T4N0M0	8	12
A17	Caucasian	0	0	0	74	T2N3M0	63	72
A20	Caucasian	0	0	0	63	T2N2M0	22	38
A23	Caucasian	0	0	0	49	T4N2M0	17	37
A26	Caucasian	0	1	1	66	T4N0M0	12	14
A28	African American	1	1	1	38	T1N1M0	91	121
A30	Caucasian	0	0	0	53	T2N0M0	12	36
A34	Caucasian	1	1	0	30	T2N1M0	36	73

Table 3.2. Therapeutic interventions received for each patient.

Pt	**Chemotherapy**	**Estrogen-directed therapy**	**Her2 directed therapy**	**Other biologics**
A1	taxol			
A2	doxorubicin/cytoxin	letrozole, alendronate		
A4	doxorubicin/cytoxin, paclitaxel, gemcitabine		trastuzumab, navelbine	
A5	docetaxel, 5-fluorouracil, epirubicin, cyclophosphamide, capecitabine			pamidronate
A7	doxorubicin/cytoxin, paclitaxel, capectiabine, carboplatin			
A8	doxorubicin/cytoxin, paclitaxel, capecitabine	letrozole, fulvestrant	trastuzumab lapatinib	
A11	doxorubicin/cytoxin, paclitaxel, gemcitabine, carboplatin			Ispinesib
A12	doxorubicin/cytoxin, paclitaxel, capecitabine, vinorelbine, gemcitabine, carboplatin, irinotican	tamoxifen, letrozole, exemestane, fulvestrant		bevacizumab
A15	doxorubicin/cytoxin, paclitaxel, carboplatin, capecitabine, bevacizumab		lapatinib	cetuximab
A17	fluorouracil/epirubicin/ cyclophosphamide, paclitaxel	tamoxifen		
A20	doxorubicin/cytoxin, paclitaxel, gemcitabine, carboplatin, capecitabine, vinorelbine			bevacizumab, denosumab
A23	doxorubicin/cytoxin, paclitaxel, carboplatin, capecitabine, gemcitabine			bevacizumab, anti-death receptor 5
A26	capecitabine, doxorubicin/cytoxin, paclitaxel	tamoxifen, letrozole	trastuzumab	
A28	doxorubicin/cytoxin, paclitaxel, gemcitabine, vinorelbine	tamoxifen, letrozole, luprolide, anastrazole, exemestane	trastuzumab, TDM1	samarium, denosumab
A30	gemcitabine, doxorubicin, paclitaxel			denosumab
A34	doxorubicin/cytoxin, paclitaxel, capecitabine, eribulin, carboplatin, gemcitabine	luprolide, tamoxifen, letrozole, goserelin, exemestane		bevacizumab, denosumab, everolimus

As reported previously (Harrell et al., 2012), gene expression of tumors from an individual patient are highly correlated with one another, regardless of spatial and temporal distance from the primary tumor and/or exposure to different therapies (Figure 3.3). This result was recapitulated in our sample set: of 16 patients, 4 had all specimens from the same patient clustered immediately together, and 14 patients had all tumors contained within the same subtype-defining dendrogram branch (Figure 3.3). Two patients, A2 and A4, had primaries of the luminal subtype with mixed HER2-enriched and luminal metastases. All of the basal-like primaries had metastases of the basal-like subtype.

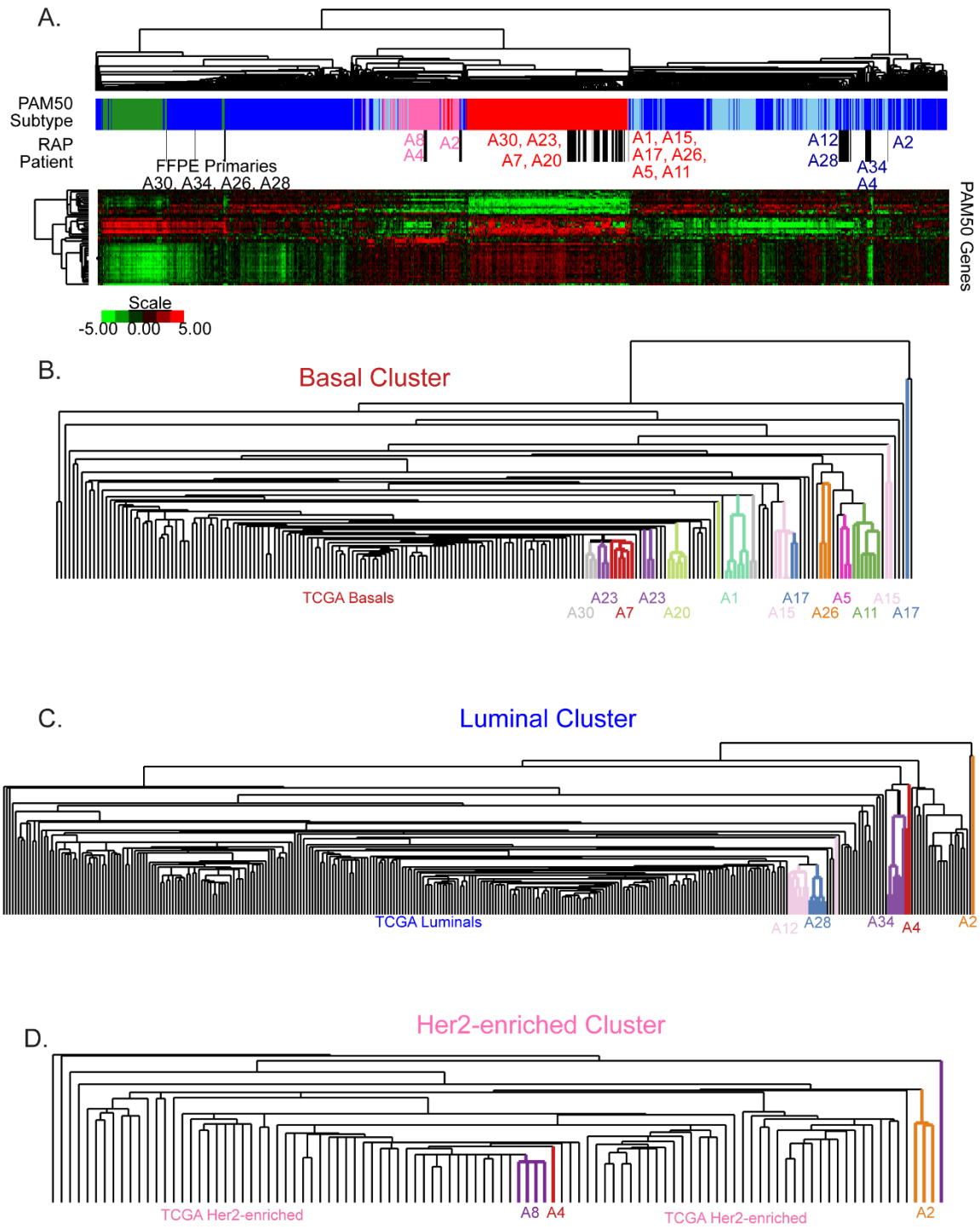

Figure 3.3. **Hierarchical clustering of 1098 TCGA Primary breast cancers with the RAP primaries and metastases.** A. Supervised hierarchical clustering using the PAM50 gene set with TCGA and RAP tumors. PAM50 subtype represented and positioning of RAP tumors shown in the second row of the color bar**.** Zoomed in view of the dendrogram of each subtype showing the location of tumors from each RAP patient for A. basal-like, B. luminal, and C. HER2-enriched sample associated clusters.

Computational Re-Interrogations of Mutations in Related Tumors Identifies

Previous work from our group demonstrated that low frequency clones present at 1-5% in the primary tumor are enriched to >40% in the related metastases (Hoadley et al., 2016). Other groups have also identified an increased sensitivity and specificity of utilizing genomic alignments from multiple related tumors in the whole exome space to identify low frequency mutations (Josephidou et al., 2015). Based on these results, we investigated whether mutations called with high confidence in one tumor from an individual were present in other tumors from that same patient.

To first control for false positive calls, all germline variants in the population were removed using dbSNP (Landrum et al., 2014) as well as mutations with ≥ 20% variant allele frequency (VAF) in at least two normal tissues from unrelated patients. All high-quality somatic mutations across tumors within each patient were computationally re-interrogated from the original DNA binary alignment map (BAM) files in each tumor from that patient. Quantification of the VAF was calculated as the read depth of the variant allele/total counts at that position. For example, in Patient A20, 304 mutations identified with high confidence across all 6 specimens from that patient were computationally re-interrogated in each tumor from this patient (Figure 3.4). For all tumors, a median of 58 mutations were additionally identified per tumor (Figure 3.5). A median of 28.6% of mutations would have been missed if not for the computational re-interrogation method.

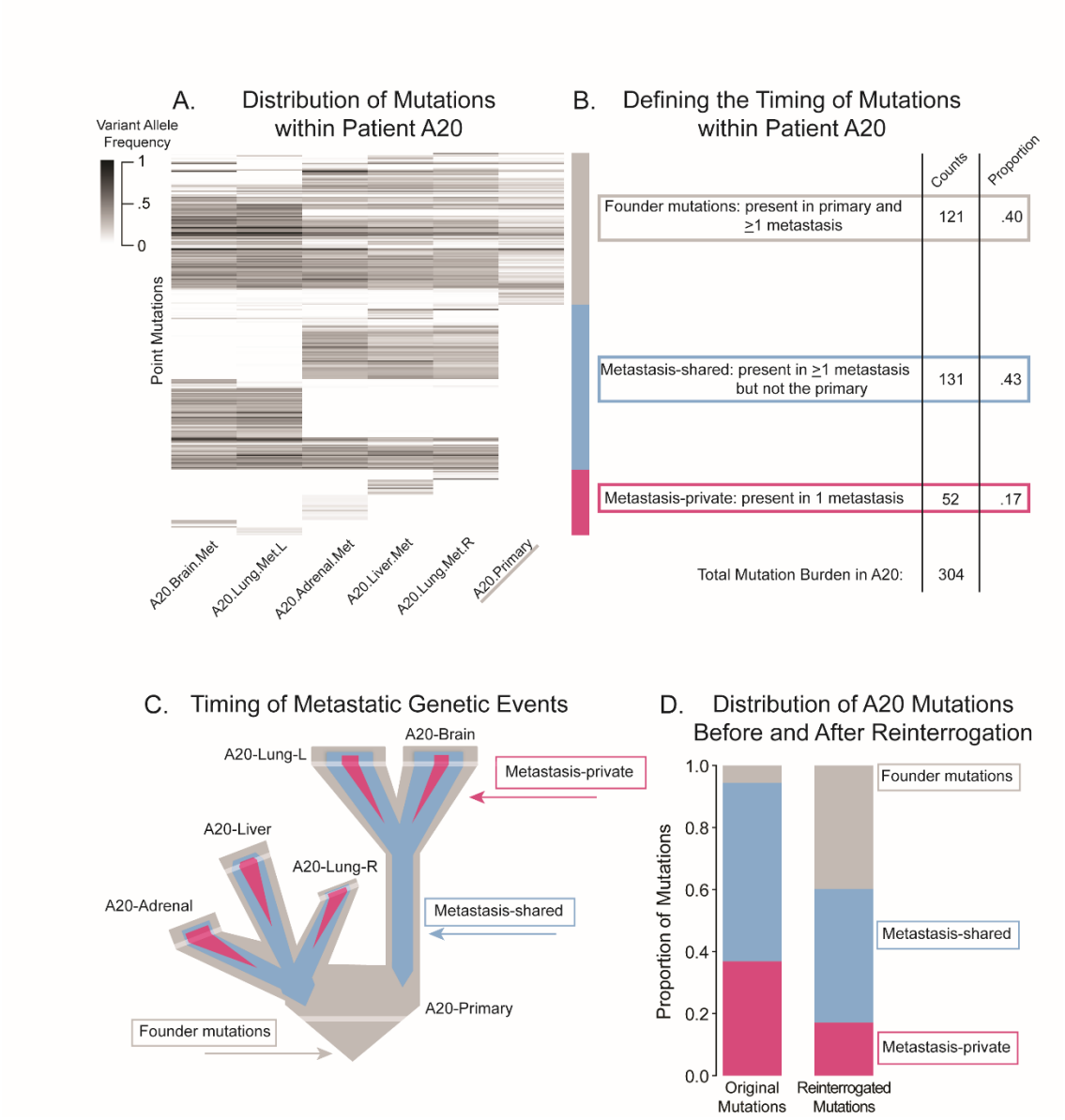

Figure 3.4. Timing with which somatic alterations are acquired. (A) A heatmap of the variant allele frequencies for Patient A20 for all somatic mutations identified following computational re-interrogation. (B) Each genetic alteration is categorized as a founder alteration if present in the primary and at least 1 metastasis (gray); metastasis-shared if present in ≥ 2 metastases and not the primary (blue); or metastasis-private if present in only 1 tumor from that patient (pink). Total counts for each category and relative proportions within that patient are then calculated. (C) Representative drawing of when during the development of metastasis each category of mutations could have occurred: founder mutations established in the original breast cancer and maintained throughout metastasis (gray); metastasis-shared mutations occurring after metastasis but along shared branches of the tree (blue); or metastasis-private mutations, acquired at the final site of metastasis (pink). (D) Quantification of the proportion of mutations within each category described in (C) before and after computational re-interrogation.

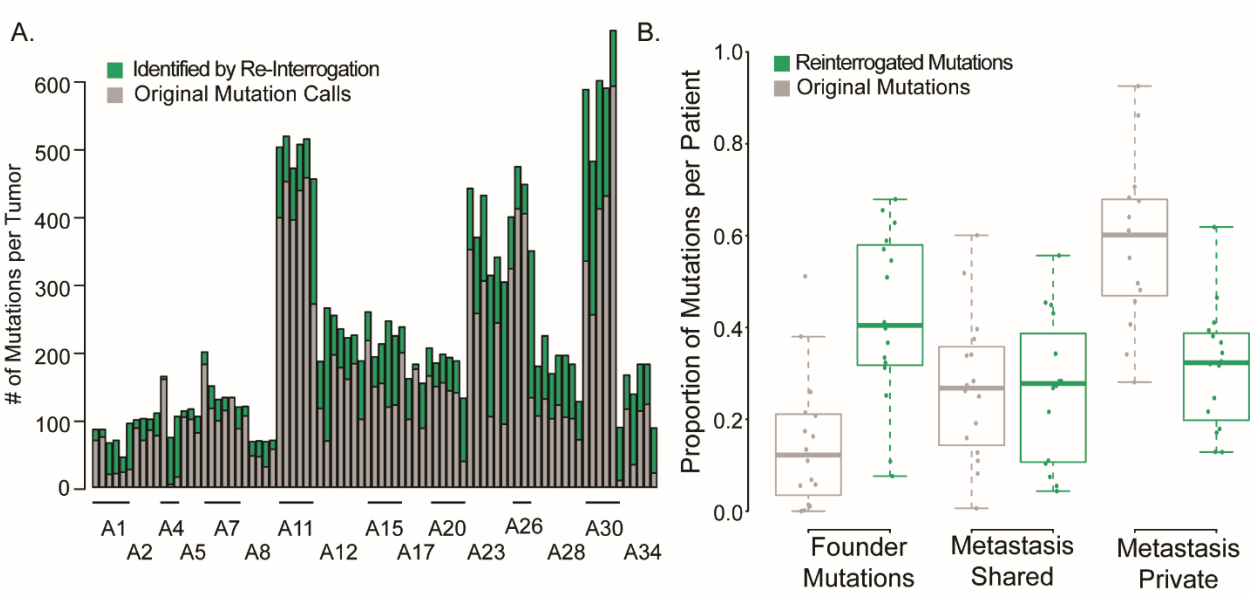

Figure 3.5. Computational re-interrogation of high quality mutation calls related tumors. (A) Mutation load per tumor before (gray) and after (green) computational re-interrogation. (B) Proportion of mutations within each patient were categorized as founder, metastasis-shared, or metastasis-private before re-interrogation (gray boxes) and following re-interrogation (green boxes).

It is critical to understand when during the development of metastasis genetic drivers are acquired. Therefore, we wanted to determine whether this computational re-interrogation altered our main conclusions of when during the metastatic process somatic mutations are acquired. We categorized mutations on when they possibly occurred in the metastatic process: (1) in the primary setting and thus shared between the primary and all metastases (founder, gray); (2) during metastatic spread, thus shared with at least 2 metastases but not measured in the primary (metastasis-shared, blue); (3) or at the final site of metastasis and thus not shared with any other tumor in the patient (metastasis-private, pink) (Figure 3.4A). The total number of mutations per category was counted, and the proportion of mutations in each category was calculated with the denominator including all mutations observed in the metastases (Figure 3.4B).

In Patient A20, there are clear metastatic specific clones as well as common founder mutations. Additionally, private mutations are measured in every tumor from this patient. This can be represented by an evolutionary tree rooted in the primary with branches representing shared genetic events in the metastases but not shared with other branches (Figure 3.4C). Within Patient A20, computational re-interrogation of mutations altered the distribution: 37% of the mutations originally classified as private to 17% with re-interrogation and 5% originally classified as founder mutations shifting to 40% of mutations with re-interrogation (Figure 3.4D).

Computational re-interrogation across the entire cohort significantly altered the categorization of mutations: with the original mutation calls, 60% of mutations per patient were considered as private and 12% as founders (Figure 3.5B, gray boxes) compared to 32% of mutations per patient considered as private and 40% as founders following re-interrogation (Figure 3.5B, green boxes).

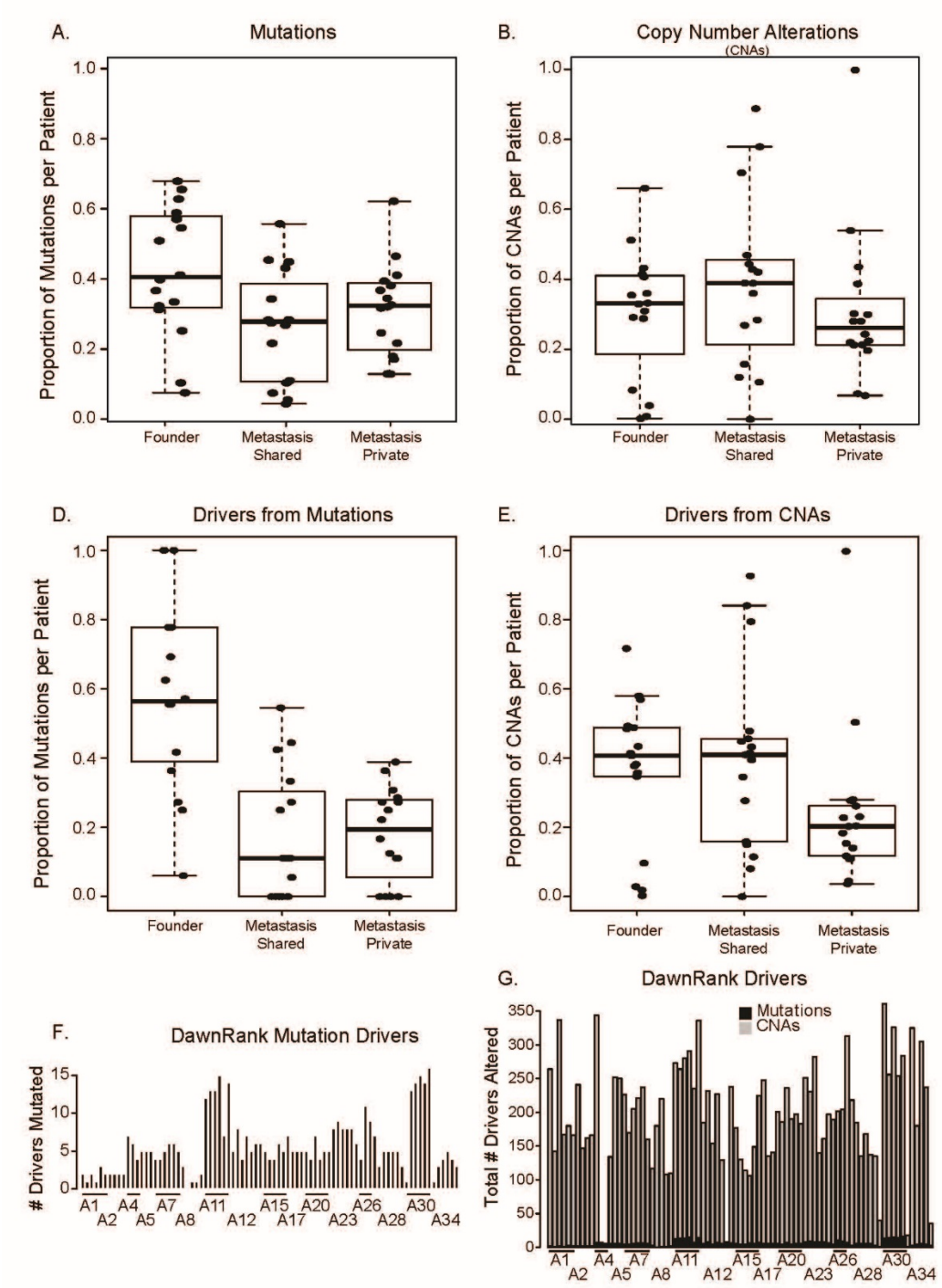

Figure 3.6. Timing of genetic alterations and driver acquisition in metastasis. Categorization of DNA alterations into founder alterations (established in the primary and observed in at least 1 metastasis), metastasis-shared (shared between at least 2 metastases but not the primary), or in only one metastasis (private) for (A) mutations and (B) copy number alterations (CNAs). The analyses in A and B were repeated for DawnRank drivers in (D) mutations and (E) CNAs. (F) Total number of DawnRank driver genes altered by mutation. (G) Total DawnRank driver counts for each tumor including both CNA (gray) and mutation (black).

The shift in the majority of mutations being considered as 'private' to being measured in the original primary tumor has significant clinical implications: if the majority of somatic mutations are already established in the original breast cancer, these could be potentially targeted to prevent future metastatic seeding with effective therapies. We demonstrate a critical need for re-interrogation of mutations in related samples. In matched tumor sets, ignoring low frequency mutations could alter the conclusions of a study.

Evolutionary Progression of Genetic Alterations in Breast Cancer Metastasis

To understand when during the metastatic process somatic mutations and copy number alterations (CNAs) occurred, we categorized all somatic alterations within each patient into the previously described categories of founder, metastasis-shared, or metastasis-private. Of these categories, the majority of mutations and CNAs in the metastases were shared with at least one other tumor (mutations: 40% founder; 28% metastasis-shared; 32% private; and for CNAs: 37% founder; 40% metastases-shared; 25% metastasis-private; Figure 3.6A-B). Each tumor had a median of 185 genes mutated and 8706 genes copy number altered. Only 3 non-synonymous mutated genes were common across the dataset: *TP53* (13/16 patients), *MT-ATP6* (10/13), and *TTN* (9/13); in contrast, large portions of the genome were commonly amplified or deleted in most patients across the dataset.

Many mutations and copy number alterations are likely passenger alterations without functional biologic consequences. We therefore used a novel computational tool called DawnRank (Hou and Ma, 2014) that integrates DNA alterations, protein-protein interaction network, and the expression of these networks via RNA gene expression data for each individual tumor. By evaluating the perturbation of the network through RNA gene expression data, DNA alterations can be scored and identified as "genetic drivers" on an individual patient level (Hou and Ma, 2014). DawnRank network analysis was applied to each tumor, and genes with DawnRank network scores in the top 5% of all genes (of 8710 total genes) were then

examined for DNA alteration via somatic mutation and CNA. Genes with copy number alteration and/or mutation within this top 5% were considered as “genetic drivers”.

Using this methodology, genetic drivers were even more likely to be “founder” events that were established in the primary breast cancer and maintained throughout metastasis when compared to the original mutation spectrum (Figure 3.6D; mutations: median of genetic drivers in founders was 56%; metastasis-shared was 11%; metastasis-private was 18%). Genetic drivers as a result of CNA were also more likely to be “founder” events than the original proportions (Figure 3.6E; median of genetic drivers in founders was 41%; metastasis-shared was 41%; metastasis-private was 21%). CNAs again comprised the numerically dominant somatic mechanism behind driver genes, with each tumor having on average, 6 mutation-based driver alterations (Figure 3.6F; Figure 3.6G, black) as compared to 189 CNA-based driver alterations (Figure 3.6G, gray).

TP53 Drives Breast Cancer Metastasis

We next examined common DawnRank drivers and the timing with which these alterations were established during the progression of metastasis. All 16 patients in our cohort harbored *TP53* alterations identified by DawnRank as drivers, with 14/16 present in the primary and all metastases. Interestingly, tumors from 13/16 patients' primary tumors had a *TP53* mutation that was not only in the primary, but in every metastasis from that patient (Figure 3.7A); tumors from the additional 3 patients had copy number loss of *TP53*, also identified by DawnRank as drivers. *TP53* mutations were diverse across the protein and altered protein function regardless of subtype: Patient A12's luminal tumors had a 45 base deletion between exons 4/5 incorporating the splice site, patient A8's HER2-enriched tumors had a premature stop codon introduced at Arg306*, and tumors from 9/10 of the basal-like patients had either nonsense or deleterious missense mutations (Bouaoun et al., 2016).

RNASeq validation of the presence of these *TP53* mutations utilized UNCeqR (Wilkerson et al., 2014) in two ways: re-interrogating known mutations in the RNA BAM file as well as *de novo* discovery of mutations with combined DNA and RNA BAM files. Interestingly, re-interrogation UNCeqR identified 3/16 mutations while the *de novo* caller identified an additional 4/16 *TP53* mutations. These 7 mutations comprise the missense and non-synonymous *TP53* mutations. 6 additional mutations in *TP53* were not observed in the RNA: the 45 bp deletion in A12, 2 frame shift deletions, 2 splice site deletions, and an in-frame deletion previously validated with whole genome sequencing (Hoadley et al., 2016).

DawnRank driver identification of *TP53* in every patient in our dataset coupled with RNASeq validation of the expression of most of these mutations provides conclusive evidence of *TP53* disruption as a critical, early event in the formation of aggressive breast cancer. *TP53* is the only founding driver disrupted by mutation in our metastatic breast cancer patients.

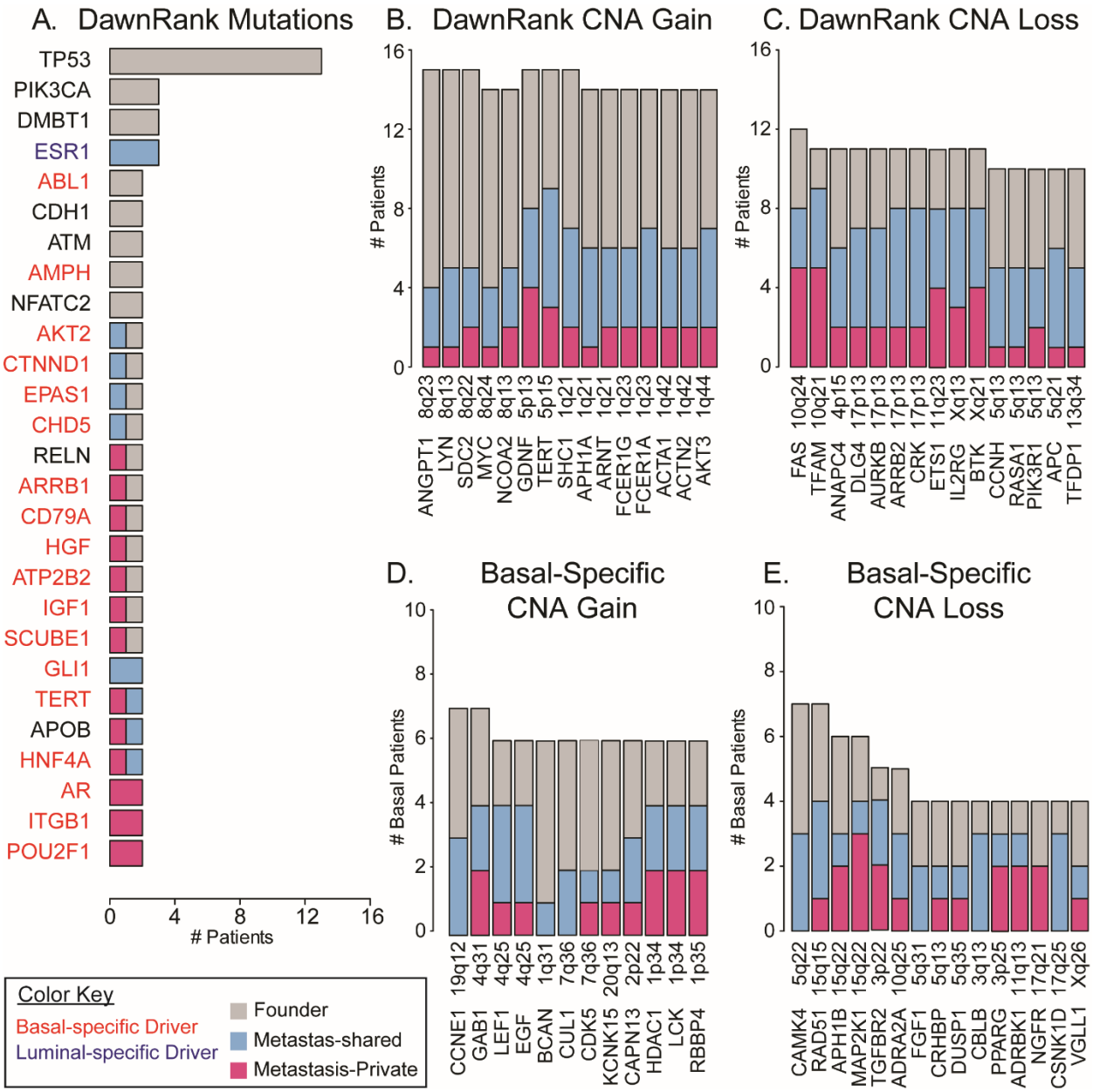

Figure 3.7. Timing and frequency of predicted drivers in primary and metastatic breast cancers. (A) DawnRank drivers from somatic mutations in at least 2 patients in the cohort (blue gene = only luminal patients; red gene = only basal patients). DawnRank copy number (B) amplfications and (C) deletions in 14/16 patients. The most frequent copy number drivers seen exclusively in basal-like patients for (D) gains and (E) losses are presented. Each driver is annotated with chromosomal cytoband location and characterized per patient as a founder alteration (gray), metastasis-shared (blue), or metastasis-private (pink) as described in Figure 2.

Cohort-Wide and Subtype-Specific Genetic Drivers of Breast Cancer Metastasis

Beyond *TP53*, genetic drivers caused by mutation were observed in only 3/16 patients: *ESR1*, *PIK3CA*, and *DMBT1* (Figure 3.7A). All other mutation drivers were identified in only 1 or 2 patients in the dataset, and many were specifically observed in the basal-like patients (Figure 3.7A, red font).

In contrast to the low frequency of common mutational drivers in our dataset, many copy number amplifications and deletions were consistently identified as drivers in almost all patients. Previously identified common regions of amplification in breast cancer (8q, 5p, and 1q) included the DawnRank hits *ANGPT1*, *LYN*, *SDC2*, *SHC1*, *GDNF*, and *TERT* identified as drivers in 15/16 patients, with 6/10 of these events showing amplification in the primary that was maintained in metastases in those patients (Figure 3.7B, gray). Common copy number losses included *FAS*, a critical member of the apoptosis cascade, *PIK3R1,* the repressive subunit of PIK3CA, and *AURKB*, a central inhibitor of the cell cycle pathway Figure 3.7C).

In an analysis restricted to the basal-like subset of patients (n=10), we collectively identified common copy number amplifications of genes involved in cell cycle genes, specifically the G1/S transition including *CCNE1, CUL1, CDK5* and chromatin associated-proteins *RBBP4* and *HDAC1* (Figure 3.7D). *BCAN* gain specifically in the basal-like patients has not been previously described in breast cancer but has been shown to be highly overexpressed in aggressive gliomas via STAT3 signaling (Natesh et al., 2015). Interestingly, concurrent basal-specific copy number loss of non-canonical STAT signaling and brain-specific genes include *ADRBK1, ADRA2A,* and *DUSP1* (Figure 3.7E). Basal-like copy number loss of the DNA damage cascade regulator *RAD51* was also called as a common basal-specific driver (Figure 3.7E).

Resistance to Aromatase Inhibitor Therapy via ESR1 Mutations is Subtype Dependent

DawnRank driver analysis identified *ESR1* mutations specifically in the metastatic samples only in 3 ER-positive, luminal patients (Figure 3.7A). *ESR1* mutations in the binding pocket of the estrogen receptor have been previously described as effectors of resistance mechanisms to estrogen suppression by aromatase inhibitors (AIs) (Li et al., 2013; Miller et al., 2016). Upon re-examining the medical histories of the patient's in this dataset, 6 patients had ER-positive breast cancer and all had received both a nonsteroidal aromatase inhibitor (letrozole) and a steroidal aromatase inhibitor (exemestane). Three of the 6 ER-positive patients exhibited *ESR1* mutations in the metastases but not the primary, and all were called as drivers by DawnRank. Interestingly, the 3 ER-positive patients who had received AIs but did not develop *ESR1* mutations were not of the luminal molecular subtype: A26 = basal-like; A8 = HER2-enriched; A2 = mixed luminal/HER2-enriched.

Confirmatory testing of *ESR1* mutations in these 3 patients' tumors was performed via two orthogonal approaches: expression of the mutant version in the RNA via UNCeqR and confirmation of DNA mutations with the highly sensitive Droplet PCR system RainDrop. In Patient A34, a T to A mutation at chr6:152419922 was called as a somatic mutation in the lymph node metastasis (A34-LN-Met; Figure 3.8A, gray), two liver metastases (data not shown), and the pancreatic metastasis upon re-interrogation (A34-Pancreatic-Met; Figure 3.8A). This variant was confirmed in the RNASeq BAM file for all metastases from this patient (Figure 3.8B). Fluorescence measurement of wild-type (y-axis) versus mutant (x-axis) *ESR1* confirmed mutant *ESR1* in both the A34-LN-Met (Figure 3.8C) and A34-Pancreatic-Met (Figure 3.8D) at VAFs extremely comparable to those identified in the DNA. Droplet PCR validation across all 3 patients demonstrated a sensitivity down to 0.4% VAF in the DNA when using the re-interrogation method.

In addition to Patient A34, Patient A12's and Patient A28's metastases also exhibited *ESR1* mutations that cause constitutive activation of *ESR1* in the presence of AI therapy (Li et

al., 2013, Miller et al., 2016). In A12, 2 of the 5 metastases contained a p.Tyr537Ser mutation in *ESR1,* which was not observed in the primary. Interestingly, Patient A28 had 1 metastasis with the p.Tyr537Asn mutation while the other 3 metastases from this patient exhibited a different p.Ser463Pro mutation.

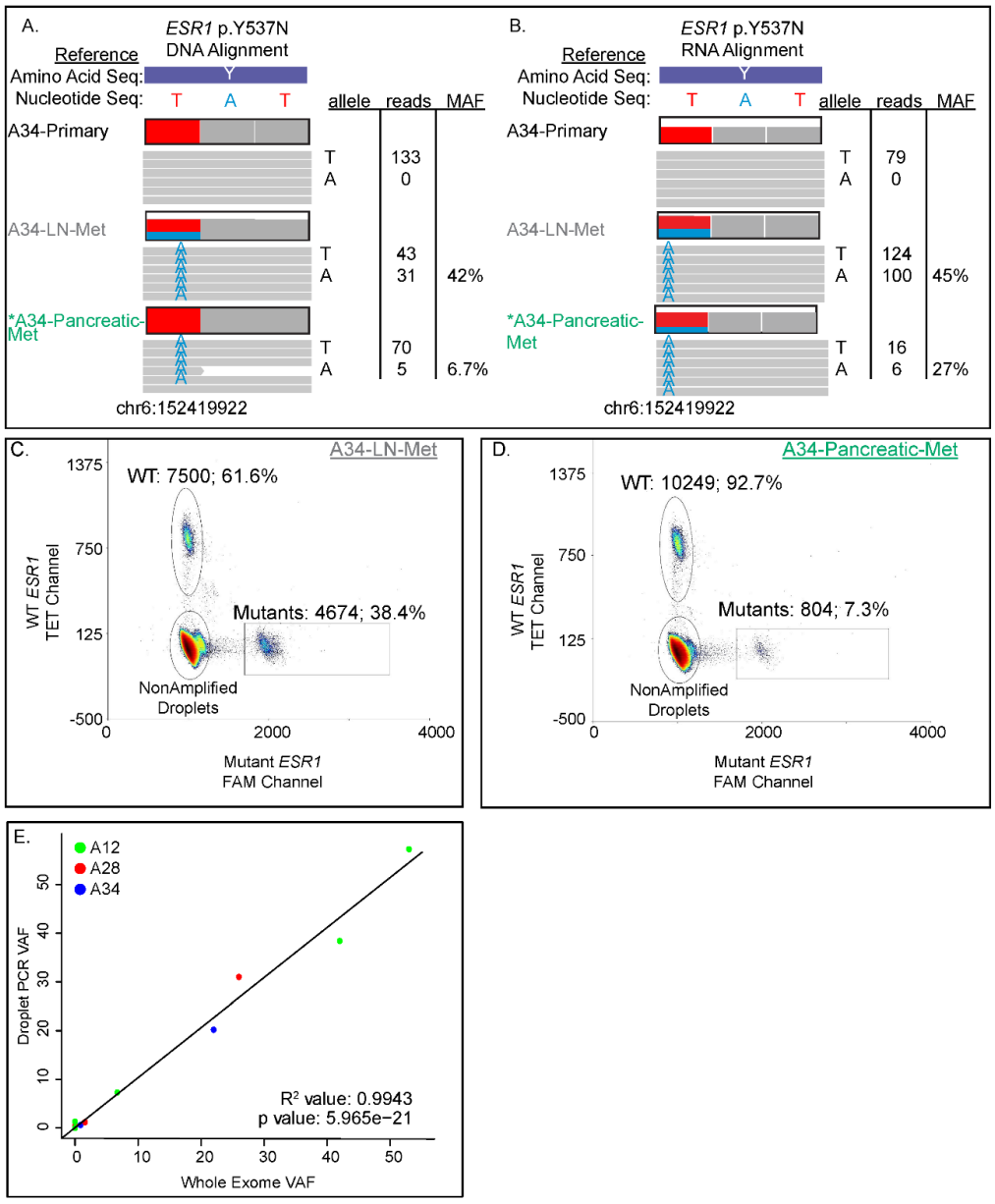

Figure 3.8. Resistance to aromatase inhibitor therapy via ESR1 mutations. (A) DNA sequencing alignment of Patient A34's ESR1 mutation: wild-type in the primary, chr6-152419922-A-T mutation originally discovered in the lymph node metastasis (A34-LN-Met), and discovered in the pancreatic metastasis (A34-Pancreatic-Met) following re-interrogation. (B) RNA Sequencing alignment at the same genomic location, confirming the re-interrogated mutation in A34-Pancreatic-Met. Confirmatory testing with Droplet PCR fluorescence quantified wild-type (y axis) versus mutant (x axis) ESR1 in (C) A34-LN-Met and (D) A34-Pancreatic-Met. (D) Comparison of the variant allele frequency measured from Droplet PCR (y-axis) versus whole exome sequencing (x axis) for three luminal patients who received aromatase inhibitor therapy: Patient A12 (green dots), A28 (red dots), and A34 (blue dots). R2 and p value are reported for the Spearman correlation of the two methods.

Multiclonal Seeding of Metastasis is Present in ER+ and ER- Patients

The subclonal heterogeneity of primary breast cancer was elegantly demonstrated in recent publications (Miller et al., 2016; Yates et al., 2015). Whole genome sequencing of two triple-negative, basal-like patients in this dataset also demonstrated multiclonal seeding of metastasis (Hoadley et al., 2016). To evaluate the clonal evolution of metastasis, we performed subclonality analysis with SciClone (Miller et al., 2014). Posterior significance testing of SciClone clusters with SigClust was applied to clusters, wherein the radius of the point plotted demonstrates the mean variant allele fraction (VAF) of the mutations in a given cluster.

Of the 16 patients examined, 13 patients had multiple clones in the primary that collectively seeded each distant metastasis (Figure 3.9, Figure 3.10, Figure 3.11A). Patients A8, A17, and A30 had only a single clone detected in the primary (Figure 3.11A). Within all patients, the metastases were multiclonal, meaning each metastasis had at least two clones present. Most patients also had metastasis-shared clones, indicating further evolution after metastasis occurred. This could be a result of either one metastasis seeding other metastases, clones that arose in a separate part of the primary that were not sequenced, or clones present in the primary that simply were below our level of detection with NGS-based methods.

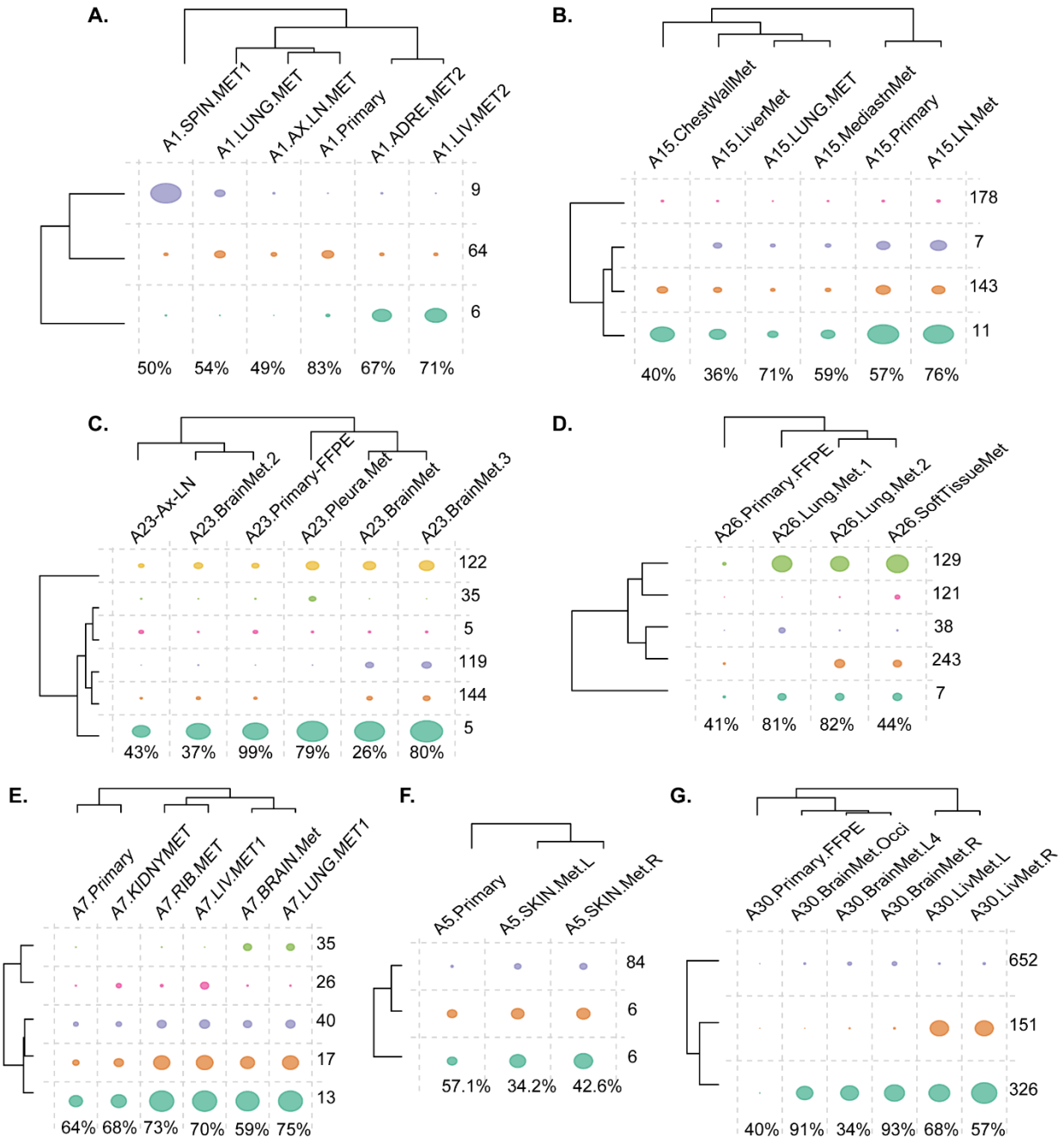

Figure 3.9. Clonality plots for each basal-like patient. Clones as determined by SciClone and posteriorly tested with SigClust are plotted with the radii of the circle proportional to the mean variant allele frequency of the mutations in that clone per each tumor. Total number of mutations per clone are demonstrated in the last column, and tumor purity is reported at the bottom of each plot for A. A1; B. A15; C. A23; D. A26; E. A7; F. A5; G. A30.

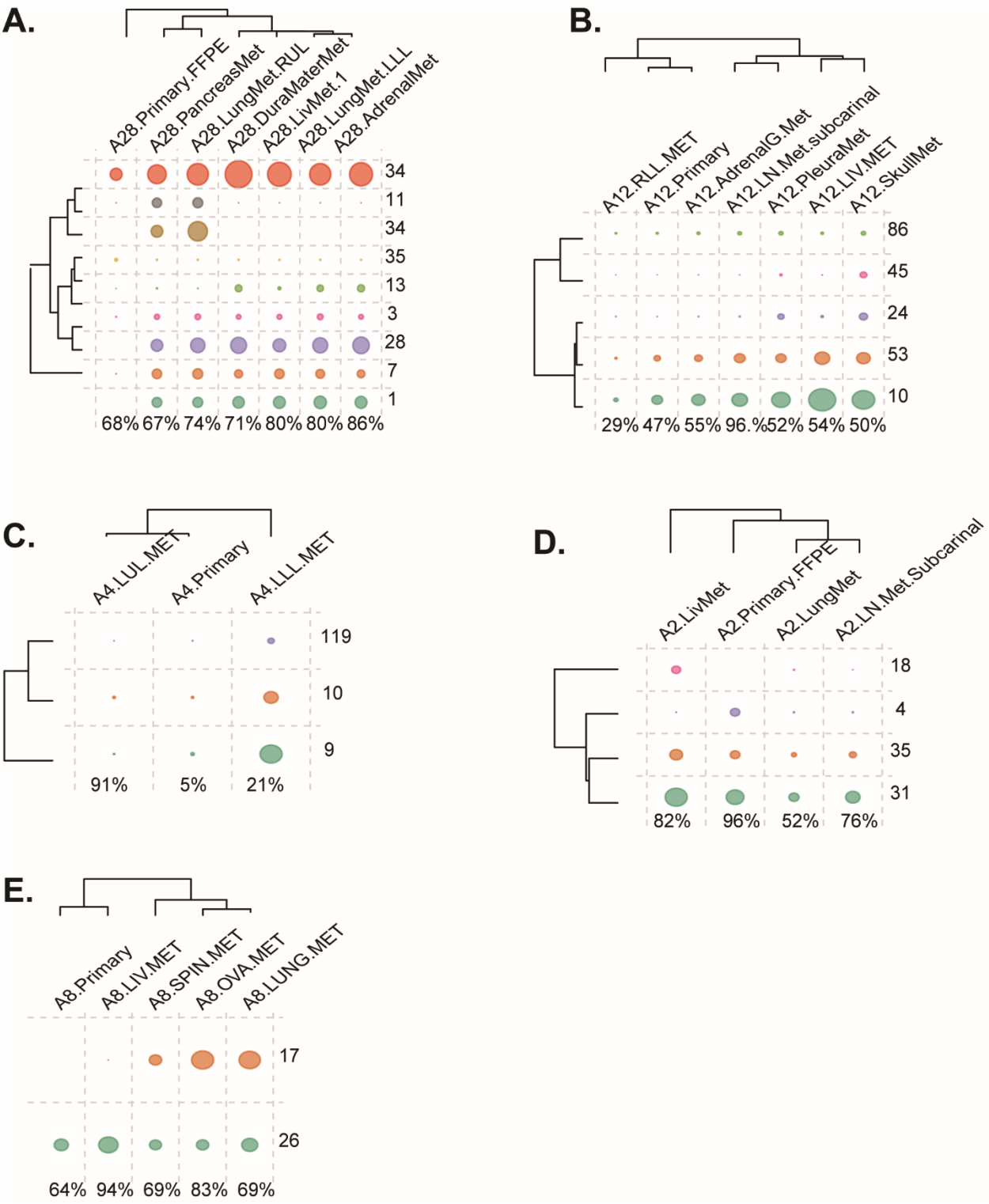

Figure 3.10. Clonality plots for luminal and HER2-enriched patients. Clones as deter- mined by SciClone and posteriorly tested with SigClust are plotted with the radii of the circle proportional to the mean variant allele frequency of the mutations in that clone per each tumor. Total number of mutations per clone are demonstrated in the last column, and tumor purity is reported at the bottom of each plot for A. A28; B. A12; C. A4; D. A2; E. A8.

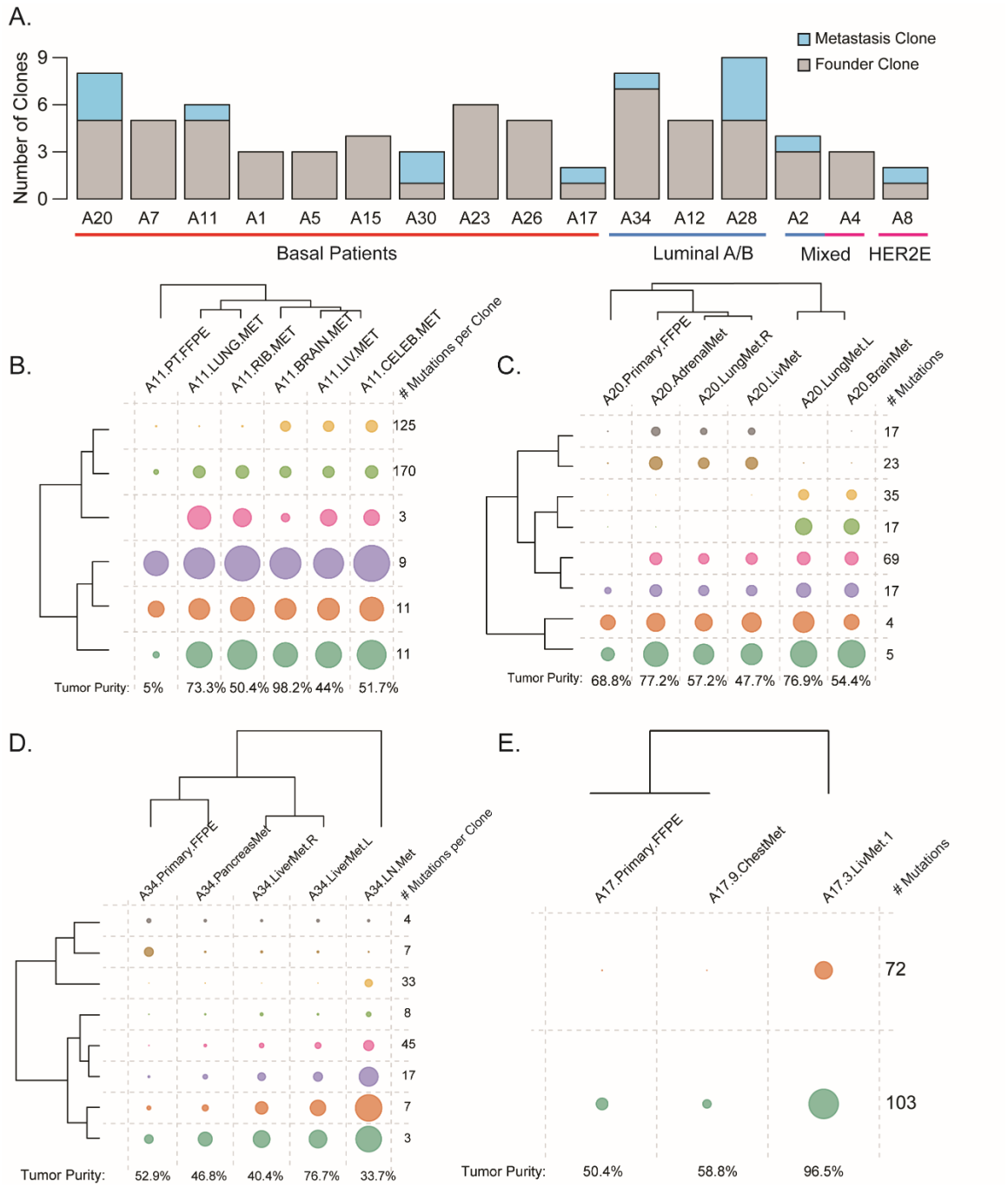

Figure 3.11. Metastatic seeding patterns. (A) Clones present in the primary and metastases (founder) as compared to a clone shared by at least 2 metastases in a given patient (metastatic clone), arranged according to molecular subtype. (B-E) Each subclone detected in a patient is represented as a separate color along the x axis for each primary and metastasis down the y axis. The radius of each circle is proportionate to the mean variant allele frequency of that clone in each tumor. Private mutations were excluded in clonality analysis. Tumor purity estimates are reported on the bottom row, and the total number of mutations per clone are in the right-most column. Multi-clonal seeding patterns are observed in basal-like patients (B) A11, (C) A20, and luminal patient (D) A34. Monoclonal patterns are identified in 3 patients including basal-like patient (E) A17.

Clonality plots for basal-like patients are presented in Figure 3.9 and for the luminal and HER2-enriched patients in Figure 3.10, with four interesting cases presented in Figure 3.11B-E. With each row indicating a distinct clone and each column representing a different tumor, the radius of the circle plotted is proportionate to the mean VAF of the mutations in the clone in each tumor. Multiclonal seeding was observed in basal-like patients A11 and A20 (Figure 3.11B-C), and in luminal patient A34 (Figure 3.11D). In Patient A11, the primary contained four founding clones present in all specimens (green, purple, orange, and teal), and a subclone (gold) that seeded the brain, liver, and cerebellar metastases. The pink subclone is enriched in the lung and rib metastases as compared to the brain, liver, and cerebellar metastases. A similar complex subclonal pattern was seen in patient A20, with two subclones predominantly present in the left lung metastasis and brain metastasis (gold and green) and two separate subclones in the liver, adrenal, and right lung metastases (brown and gray). The primary of A20 also contained 3 subclones that seeded all metastases (purple, organe, and teal). Patient A34, mentioned earlier with *ESR1* mutations, had an equally complex subclonal pattern, with 5 subclones present in the primary and in every metastasis at different variant allele fractions identified in individual metastases (Figure 3.11D).

Patient A17 demonstrates monoclonal seeding, with the dominant clone in the primary (teal) further evolving separately in the metastases (Figure 3.11E). This pattern was observed in 3/16 patients, confirming previous reports of monoclonal seeding in breast cancer (Krøigård et al., 2015). Private mutations were observed in almost all tumors across the dataset (but are not displayed), indicating continued evolution after metastasis, even within each primary tumor.

Discussion

The molecular mechanisms driving the metastatic process are critical to understand in order to better prevent and treat existing metastases. Utilizing the UNC Rapid Autopsy Program and next generation sequencing of multiple tumors from 16 breast cancer patients, we demonstrated that metastasis is largely a result of multiclonal seeding of breast cancer metastases in the majority of cases examined. Moreover, our data illustrates that the majority of genetic drivers were established in the primary breast cancer and maintained throughout the metastatic process; this was observed in both luminal and basal-like breast cancers. We also demonstrate that *TP53* is the only mutational driver common across all subtypes of breast cancer metastasis, and that the majority of drivers were predominantly altered by virtue of somatic copy number alterations. Finally, we provide evidence that computational re-interrogation of high quality somatic mutations is a requirement when studying related tumors, as previously classified private mutations are often, in-fact, shared across tumors often with lower coverage.

Previous work using two patients identified multiclonal seeding as a mechanism in breast cancer metastasis with the majority of functional mutations established in the primary and maintained throughout metastasis (Hoadley et al., 2016). Here, we build upon this very small study and demonstrate, at least for basal-like tumors, that multiclonal seeding is a common mechanism of metastasis. In patients where multiclonal seeding occurred, the metastasis formation must have occurred via a large mass of cells breaking off from the primary (i.e. large enough to contain 2, often 3 subclones), which then travels to the distant site and seeds this site. This has significant clinical implications including that if the metastasis seed is a clump of cells with distinct subclonal populations, then successful therapy to prevent metastasis that are focused on inhibiting individual cell migration/motility may have no effect upon tumors that use multiclonal seeding.

Historically, point mutations or small intragenic in/dels have been regarded as the driving force behind oncogenesis. One of the novel aspects of this work was to utilize a more functional-based assessment of genetic drivers, DawnRank, which integrates prior knowledge of protein interaction networks with mutations, copy number alteration, and RNA expression data to refine our ability to identify drivers for each individual patient. By contrast, DNA-only based methods can only identify drivers based upon correlation to previous datasets from which population-based enrichments of specific genes were determined. Our novel, functional-based genetic driver approach demonstrated that the majority of drivers were the result of copy number alteration, which also was suggested from a DNA-only approach (Ciriello Nat Gen 2014). Finally, our results confirm that copy number alterations are established early in the development of breast cancer and maintained throughout the evolution of breast cancer metastasis (Krøigård et al., 2015). This is contrast to earlier literature in breast cancer metastasis demonstrating that a majority of drivers are private and acquired at the final site of metastasis (Brastianos et al., 2015).

We discovered that computational re-interrogation of high quality mutations in one tumor are often at lower coverage in the original primary tumor. Furthermore, when considering DawnRank computationally predicted functional drivers, the vast majority of drivers are indeed established in the primary. In genomic studies of matched tumors from a single patient, it is critical to re-examine the sequencing files to fully characterize the timing with which mutations are acquired. Clinically, if most of the drivers of breast cancer metastasis are indeed established in early development of breast cancer, more effective therapies could possibly prevent or treat existing metastases.

Strikingly, *TP53* mutations were seen repeatedly in both basal-like and luminal breast cancers, with the mutation always established in the primary and maintained in every metastasis from that patient. Beyond *TP53,* no other driver mutations were present in more than 3/16 patients in this dataset. Driver analysis identified *ESR1* mutations in patients with luminal

subtype breast cancers who received aromatase inhibitors (AIs) in the binding pocket of *ESR1*, consistent with previous reports demonstrating the mechanism of resistance to AIs (Li et al., 2013; Miller et al., 2016). Interestingly, patients who received aromatase inhibitors for non-luminal yet clinically ER positive tumors did not demonstrate *ESR1* resistance mutations in the metastatic setting. This molecular diversity of ER-positive tumors (Ciriello et al., 2013; Gatza et al., 2014) may explain differential response of many patients' metastases to aromatase inhibitor therapy.

Our study had a number of limitations, most notably including the sample size of 16 patients; this inhibited our ability to identify recurrent somatic mutations common to the metastatic setting, although our sample size was large enough to identify the importance of *TP53* and *ESR1*. A larger sample size will also be needed to identify site-specific (i.e. lung or brain) differences and adaptations. In addition, with only 2 HER2-enriched patients in our analysis, additional patients in this subtype are necessary to confirm clonality of metastasis and understand resistance mechanisms that develop in HER2-positive breast cancer. Finally, many of the primary breast cancers in this dataset were treated with neo-adjuvant (preoperative) therapy prior to mastectomy. Future studies comparing matched therapy-naïve, post-neo-adjuvant therapy, axillary lymph nodes, liquid biopsies, and distant metastases will be needed to understand the earlier steps of clonal evolution.

In summary, this study validates and further expands upon the compelling evidence of multiclonal seeding across multiple subtypes of breast cancer, especially for TNBC/Basal-like tumors. Additionally, we demonstrate that most genetic drivers arise from copy number alterations. The mechanism to generate genetic diversity is largely unknown; however, the consistency across our cohort and previous literature suggests that *TP53* dysfunction is an early and critical event in the development of aggressive breast cancer. Despite the high degree of heterogeneity in primary breast cancer (Yates et al., 2015) maintained through metastasis via multiclonal seeding, these results also show that the majority of genetic drivers are established

in the primary breast cancer and maintained throughout metastasis. This gives hope that therapeutic targeting of the founding events that drive the metastatic phenotype might prevent metastatic spread or inhibit the progression in the advanced setting.

APPENDIX 3.1: PAM50 Analysis of A16 Tumors

Sample Name	Tissue Type	RNASeq Method	Basal	Her2	LumA	LumB	Normal	Call
A11.PT.FFPE	FFPE	Ribo0	0.60	0.07	-0.54	0.02	-0.08	Basal
A17.Primary.FFPE	FFPE	Ribo0	0.24	-0.08	-0.02	-0.23	0.32	Normal
A2.Primary.FFPE	FFPE	Ribo0	-0.11	-0.17	-0.01	0.27	-0.21	LumB
A20.Primary.FFPE	FFPE	Ribo0	0.63	-0.38	-0.23	-0.33	0.34	Basal
A23.Primary.FFPE	FFPE	Ribo0	0.81	0.01	-0.67	-0.09	-0.08	Basal
A23.LN.Met.FFPE	FFPE	Ribo0	0.79	0.09	-0.74	0.05	-0.19	Basal
A26.Primary.FFPE	FFPE	Ribo0	0.17	-0.54	0.22	-0.25	0.45	Normal
A28.PT.FFPE	FFPE	Ribo0	-0.02	-0.62	0.52	-0.45	0.66	Normal
A30.Primary.FFPE	FFPE	Ribo0	-0.07	-0.55	0.59	-0.47	0.75	Normal
A34.Primary.FFPE	FFPE	Ribo0	-0.25	-0.53	0.67	-0.37	0.68	Normal
A1.LUNG.MET	FF	polyA	0.52	-0.10	-0.24	-0.51	0.39	Basal
A1.ADRE.MET2	FF	polyA	0.58	-0.17	-0.31	-0.37	0.29	Basal
A1.AX.LN.MET	FF	polyA	0.58	-0.05	-0.35	-0.36	0.25	Basal
A1.PRIMT.2	FF	polyA	0.56	-0.09	-0.29	-0.43	0.31	Basal
A1.LIV.MET2	FF	polyA	0.62	-0.12	-0.37	-0.35	0.25	Basal
A1.SPIN.MET1	FF	polyA	0.69	-0.13	-0.47	-0.33	0.21	Basal
A11.BRAIN.MET	FF	polyA	0.57	0.03	-0.48	-0.19	0.00	Basal
A11.LUNG.MET	FF	polyA	0.56	0.08	-0.52	-0.10	-0.07	Basal
A12.LIV.MET	FF	polyA	-0.39	0.10	0.10	0.30	-0.15	LumB
A12.SkullMet	FF	polyA	-0.17	-0.20	0.08	0.23	-0.13	LumB
A12.AdrenalG.Met	FF	polyA	-0.39	-0.24	0.49	0.03	0.27	LumA
A12.PleuraMet	FF	polyA	-0.36	-0.08	0.24	0.18	0.07	LumA
A12.LN.Met.subcarinal	FF	polyA	-0.13	0.04	-0.02	0.11	-0.06	LumB
A12.RLL.MET	FF	polyA	-0.28	-0.22	0.28	0.02	0.22	LumA
A12.PRIMT020076B	FF	polyA	-0.64	-0.05	0.53	0.10	0.03	LumA
A15.LiverMet	FF	polyA	0.54	-0.11	-0.28	-0.43	0.33	Basal
A15.ChestWallMet	FF	polyA	0.41	-0.04	-0.15	-0.44	0.40	Basal
A15.PRIMT070427B	FF	polyA	0.61	-0.12	-0.44	-0.27	0.14	Basal

A15.LN.MET	FF	polyA	0.55	-0.14	-0.35	-0.30	0.19	Basal
A15.MediastnMet	FF	polyA	0.42	-0.06	-0.14	-0.45	0.41	Basal
A17.3.LivMet.1	FF	polyA	0.55	0.07	-0.40	-0.33	0.25	Basal
A17.9.ChestMet	FF	polyA	0.53	0.09	-0.41	-0.36	0.28	Basal
A2.LungMet	FF	polyA	-0.27	0.21	0.08	0.04	-0.01	Her2
A2.LN.Met.Subcarinal	FF	polyA	-0.25	0.12	0.13	-0.02	0.06	LumA
A20.BrainMet	FF	polyA	0.67	-0.26	-0.33	-0.45	0.33	Basal
A20.AdrenalMet	FF	polyA	0.69	-0.21	-0.43	-0.36	0.25	Basal
A20.LungMet.L	FF	polyA	0.67	-0.22	-0.32	-0.48	0.39	Basal
A20.LivMet	FF	polyA	0.67	-0.08	-0.53	-0.26	0.16	Basal
A20.LungMet.R	FF	polyA	0.68	-0.18	-0.41	-0.41	0.31	Basal
A23.BrainMet.3	FF	polyA	0.67	-0.09	-0.38	-0.44	0.27	Basal
A23.PleuraMet	FF	polyA	0.62	0.00	-0.41	-0.31	0.23	Basal
A23.BrainMet.2	FF	polyA	0.67	-0.08	-0.40	-0.38	0.24	Basal
A23.BrainMet	FF	polyA	0.63	-0.11	-0.35	-0.46	0.28	Basal
A26.Lung.Met.2	FF	polyA	0.57	-0.06	-0.47	-0.17	0.04	Basal
A26.Lung.Met.1	FF	polyA	0.39	-0.06	-0.46	-0.09	-0.02	Basal
A26.SoftTissueMet	FF	polyA	0.39	0.11	-0.50	-0.08	-0.17	Basal
A28.LungMet.LLL	FF	polyA	-0.42	0.10	0.29	0.25	-0.10	LumA
A28.LivMet.1	FF	polyA	-0.54	-0.10	0.49	0.20	0.04	LumA
A28.LungMet.RUL	FF	polyA	-0.49	-0.15	0.48	0.14	0.07	LumA
A28.AdrenalMet	FF	polyA	-0.26	-0.01	0.15	0.22	-0.06	LumB
A28.PancreasMet	FF	polyA	-0.49	-0.16	0.48	0.11	0.12	LumA
A28.DuraMaterMet	FF	polyA	-0.19	0.01	0.11	0.20	-0.05	LumB
A30.LivMet.R	FF	polyA	0.69	0.12	-0.46	-0.33	0.17	Basal
A30.LivMet.L	FF	polyA	0.75	0.11	-0.53	-0.33	0.11	Basal
A30.BrainMet.L4	FF	polyA	0.69	-0.09	-0.32	-0.54	0.29	Basal
A30.BrainMet.Occi	FF	polyA	0.70	-0.11	-0.43	-0.32	0.15	Basal
A30.BrainMet.R	FF	polyA	0.76	-0.09	-0.42	-0.42	0.18	Basal
A34.PancreasMet	FF	polyA	-0.42	-0.28	0.61	-0.14	0.41	LumA
A34.LiverMet.L	FF	polyA	-0.65	-0.11	0.61	0.14	0.13	LumA

A34.LN.Met	FF	polyA	-0.56	-0.11	0.51	0.18	0.06	LumA
A34.LiverMet.R	FF	polyA	-0.57	-0.19	0.63	-0.02	0.22	LumA
A4.LLL.MET	FF	polyA	0.11	0.42	-0.26	-0.09	-0.10	Her2
A4.PRIMT020306B	FF	polyA	-0.18	-0.39	0.50	-0.31	0.44	LumA
A4.LUL.MET	FF	polyA	-0.21	-0.05	0.46	-0.46	0.50	Normal
A5.SKIN.Met.L	FF	polyA	0.69	-0.10	-0.49	-0.22	0.13	Basal
A5.SKIN.Met.R	FF	polyA	0.61	-0.13	-0.40	-0.29	0.24	Basal
A5.PRIMT030065B	FF	polyA	0.71	-0.17	-0.49	-0.19	0.10	Basal
A7.RIB.MET	FF	polyA	0.75	-0.06	-0.41	-0.41	0.20	Basal
A7.LUNG.MET1	FF	polyA	0.70	-0.15	-0.33	-0.47	0.35	Basal
A7.LIV.MET1	FF	polyA	0.74	0.01	-0.54	-0.29	0.15	Basal
A7.PRIMT020552B	FF	polyA	0.78	-0.28	-0.36	-0.54	0.43	Basal
A7.KIDNYMET	FF	polyA	0.56	-0.06	-0.23	-0.46	0.37	Basal
A8.LUNG.MET	FF	polyA	-0.16	0.50	-0.25	0.16	-0.29	Her2
A8.030222BSPIMET	FF	polyA	-0.39	0.40	0.08	0.22	-0.18	Her2
A8.LIV.MET	FF	polyA	0.00	0.49	-0.34	0.26	-0.32	Her2
A8.OVA.MET	FF	polyA	-0.20	0.19	0.09	-0.13	0.16	Her2
A8.SPIN.MET	FF	polyA	-0.12	0.11	0.13	-0.27	0.25	Normal
A11.RIB.MET	FF	Ribo0	0.58	0.08	-0.50	-0.26	0.07	Basal
A11.CELEB.MET	FF	Ribo0	0.56	0.03	-0.44	-0.27	0.12	Basal
A11.LIV.MET	FF	Ribo0	0.60	0.03	-0.51	-0.22	0.03	Basal
A15.LUNG.MET	FF	Ribo0	0.40	-0.25	-0.03	-0.65	0.54	Normal
A2.LivMet	FF	Ribo0	0.10	0.00	0.01	-0.35	0.23	Normal
A7.BRAIN.Met	FF	Ribo0	0.68	-0.15	-0.26	-0.57	0.42	Basal

CHAPTER 4 – INTEGRATED MUTATIONS AND COPY NUMBER COMPUTATIONAL DRIVER CLASSIFICATION IDENTIFIES NOVEL AND KNOWN SUBTYPES OF BREAST CANCER

Preface

This manuscript is a shared first co-authorship between Jack Hou and myself. Grace Silva curated the TCGA and METABRIC copy number data. Jack Hou created the DawnRank method with Jian Ma, identified the subgroups, and performed the ClaNC classification. The classification was edited by myself and Dr. Charles Perou. I performed all subsequent analyses. Dr. Perou conceived this project.

Introduction

Breast cancer remains the second leading cause of cancer related death in women each year. Breast cancer is a heterogeneous disease with distinct molecular and clinical subgroups (Perou et al., 2000); however, even patients within the same molecular subgroup can have highly variable first sites of metastasis and differential response to targeted therapeutics(Carey et al., 2006; Ciriello et al., 2013). Identification of the underlying drivers of breast cancer could help identify the molecular cause of this heterogeneity while concurrently providing novel therapeutic targets.

Large efforts to identify the genetic underpinnings causing breast cancer have led to unprecedented amounts of both DNA and RNA genomic data (Cancer Genome Atlas, 2012; Ciriello et al., 2015; Curtis et al., 2012). Genomic instability is a hallmark of cancer, leading to many mutations and copy number alterations per tumor; however, the significance of these alterations often is not well understood. Previous efforts to identify drivers rely heavily on mutation data, defining drives as those genes with mutations above the background rate of

mutation (Dees et al., 2012b; Forbes et al., 2015; Lawrence et al., 2013). Copy number alteration (CNA), however, is known to be an early, common, and critical factor in the development of breast cancer (Cancer Genome Atlas, 2012; Hoadley et al., 2016; Krøigård et al., 2015). Additionally, CNAs alter large numbers of genes in the same area of the genome, making it difficult to identify the actual driver in an area of alteration. Genetic driver analyses incorporating both mutation and CNAs provide a novel method of defining breast cancer drivers.

Integration of RNA gene expression network analyses can accurately reflect oncogenic pathway alteration. DawnRank (Hou and Ma, 2014) allows for the integration of RNA gene expression, DNA mutations, and DNA copy number data. The proportion of drivers from CNA as compared to mutation is not well known. Additionally, it is not well known if drivers on an individual tumor level are consistent within and across subtypes or private to a tumor. A better understanding of the biology driving breast cancer needs to be explored with the hopes of identifying novel, tractable therapeutic targets.

In this analysis, we have applied the novel computational method DawnRank (Hou and Ma, 2014), which predicts potential driver genes in individual samples, to the Cancer Genome Atlas breast cancer freeze set (Ciriello et al., 2015) followed by ConsensusClusterPlus (Wilkerson and Hayes, 2010) to identify novel subgroups. A ClaNC classifier (Dabney, 2006) was then identified, and the classifier applied to METABRIC (Curtis et al., 2012; Pereira et al., 2016) for validation. Finally, we characterize the molecular and clinical phenotypes of each subgroup with additional analyses of publicly available clinical data, RNA gene expression signatures, protein expression, and survival outcomes.

Methods

Patient sample selection

We selected tumors with gene expression, mutation, and copy number data from The Cancer Genome Atlas (TCGA) breast cancer dataset (n = 871) and the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset (n = 1,992). We randomly selected 500 samples from each dataset, keeping relative distribution of PAM50 subtypes consistent between the two datasets. The composition of samples is 19.3% Basal, 10.9% Her2, 39.5% Luminal A, and 30.3% Luminal B. Normal-like breast cancers are not included in this analysis.

DawnRank score calculation

To assess a mutation or copy number alteration's impact on the differential gene expression of downstream genes in the network, DawnRank (Hou and Ma, 2014) was applied. Briefly, gene networks were built from both curated and non-curated human gene interactions obtained from the MEMo paper (Ciriello et al., 2012) and KEGG analyses (Kanehisa et al., 2012). DawnRank's default dynamic damping factor parameter μ was 3, and the default Condorcet penalty parameter δ 0.85. Gene expression data was first converted into a Z-score. DawnRank scores were next calculated according to the previously published method, such that the rank reflects their driver potential in a given sample. A non-parametric score based on the rank-order of DawnRank genes in each sample was used due to the uncertainty that individual DawnRank scores followed a distribution (QQ correlation 0.569).

For each patient, segmented CNAs were converted into a discrete copy number gene matrix, with significant segment means greater than .1 assigned to 1 and means less than -.1 assigned to -1. Using the hg19 gene annotation, genes that were completely encompassed within a segment based on genomic location were assigned that segment's discrete copy number value. For mutations, a mutation in known tumor suppressors were assigned -1 while mutations in known oncogenes are assigned 1 (Schroeder et al., 2014). If no mutation or copy

number alteration is present, the gene has a score of 0. DawnRank scores were converted into a rank-based percentage and then multiplied by both the mutation and copy number alteration matrices. Thus, DawnRank scores ranged from -100 (copy number loss/tumor suppressor mutation) to 100 (copy number gain/oncogene mutation). Only tumors with at least 5 non-zero DawnRank genes were further considered.

Cohort level DawnRank scores were also calculated using a modified Condorcet voting scheme to assess the population-level driver potential of a given gene, as previously described (Hou and Ma, 2014). A p-value was calculated by fitting a normal distribution over the cohort-level scores. Only genes with a cohort-wide p-value < 0.05 were considered for subtype classification.

Alteration based subtype classification using consensus clustering

TCGA DawnRank scores were clustered using ConsensusClusterPlus (Wilkerson and Hayes, 2010), testing k =2 to k = 10 with 1,000 iterations and 80% sampling. Sample distances were calculated using the Pearson distance over 1,000 iterations. k = 5 was selected as the maximum number of groups with the minimal number of misclassifications.

Validation classifier

To define a robust classifier, we employed ClaNC (Dabney, 2006). ClaNC ranks features based on *t*-statistics and then employs a custom Linear Discriminant Analysis to define centroids for each group. TCGA DawnRank drivers were tested with ClaNC beginning with 10 features and increasing by 5 features with the default parameters. Performance of the ClaNC classifier was defined by comparing misclassification rate of TCGA tumors with the ClaNC classifier to ConsensusClusterPlus group identity.

Statistical analyses

Gene Expression Signatures. 420 previously published signatures were curated from multiple sources (Bindea et al., 2013; Fan et al., 2011; Gatza et al., 2010; Hoadley et al., 2007; Hu et al.,

2009; Iglesia et al., 2014). For each signature, the mean of the genes comprising that signature was calculated for each tumor including TCGA normal breast samples, the 1098 freeze lobular dataset(Ciriello et al., 2015), and METABRIC tumors.

Subtype-defining expressed features. To define subtype-specific features, significance analysis of microarray (Tusher et al., 2001) was applied in two ways: first, multiSam (an ANOVA permutated 100 times) compared variation across all subgroups; secondly, each subgroup was compared to all others in a two-class, unpaired parametric t-test again permutated 100 times. Features were considered significant if both the false discovery rate from the multiSam and at least one of the two-class comparisons were 0.

Survival analyses. Overall survival data was calculated up to 10 years and plotted using a Kaplan-Meier survival curve using the *survival* package in R. Patient samples with greater than 10 year survival are censored at the 10-year mark. Log-ranked likelihood test was used to compare significance among subgroups.

R Version. All statistical analyses were performed in R v.3.3.2 .

Results

Identification of driver-based subtypes

In order to define the genetic drivers of breast cancer through an integrated analysis of gene expression, copy number, and mutation, we first applied DawnRank to the TCGA breast cancer dataset (Figure 4.1). Genes with cohort-wide DawnRank p values <= .05 were considered for clustering to define driver-subtypes. 65 copy number altered genes and 38 mutated genes were significant across the cohort (Figure 4.2A).

ConsensusClusterPlus was applied to the tumors with at least 5 features identified utilizing 1000 iterations and 80% resampling of genes and samples. Varying the number of groups from k = 2 to k = 10, we identified five as the ideal number of clusters by observing the maximum cophenetic correlation (Figure 4.2A, color bar). We compared the clusters after 25 different runs of ConsensusClusterPlus and observed consistent clustering results with a pairwise Rand Index of 0.97. Silhouette widths were calculated as the distance to the centroid, and samples were ordered accordingly.

To define a robust predictor of driver subtype, we evaluated the ClaNC classifier using these 113 features. A t-test of each feature in one subgroup compared across the subgroups was calculated to obtain subtype-defining features (Figure 4.2B). The only mutations in the first 50 features were *PIK3CA, TP53, CDKN1B,* and *CDH1*. This highlights copy number alterations as a critical mechanism driving breast cancer biology. Using the ConsensusClusterPlus subgroups to compare, we increased the number of features in the classifier by 5 and tested the performance of the ClaNC classifier (Figure 4.3). Misclassification rates varied by subtype (Figure 4.3), with the CN0 subtype (solid black line) having the highest misclassification rate. 73 features were selected to build the centroid for each subtype, with 61 copy number alterations and 12 mutations.

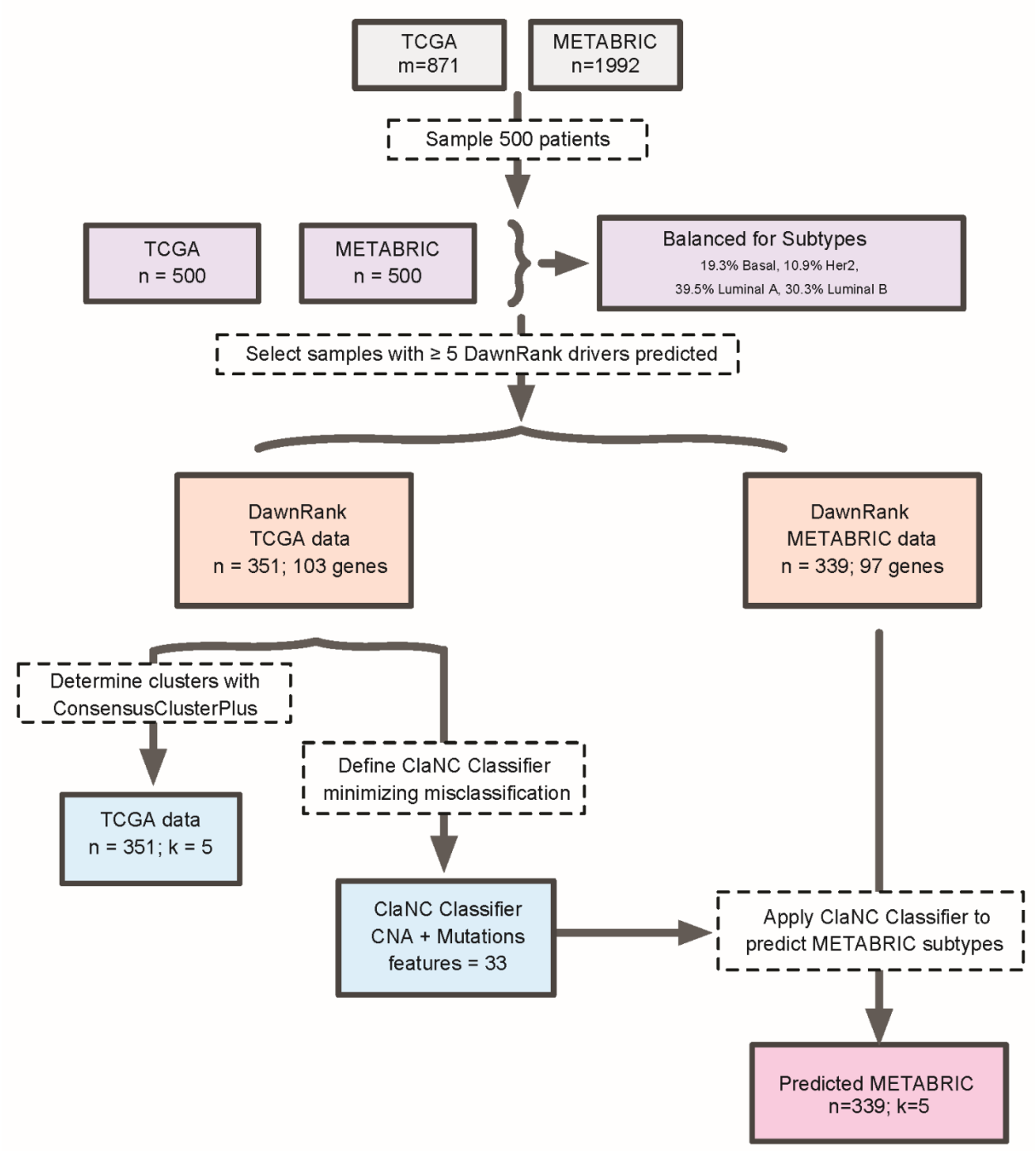

Figure 4.1. Overview of Method. A schematic diagram detailing the selection of tumor samples from TCGA and METABRIC, calculation of DawnRank scores, clustering to define subgroups, building the classifier using ClaNC, and finally applying this classifier to the validation dataset with METABRIC tumors.

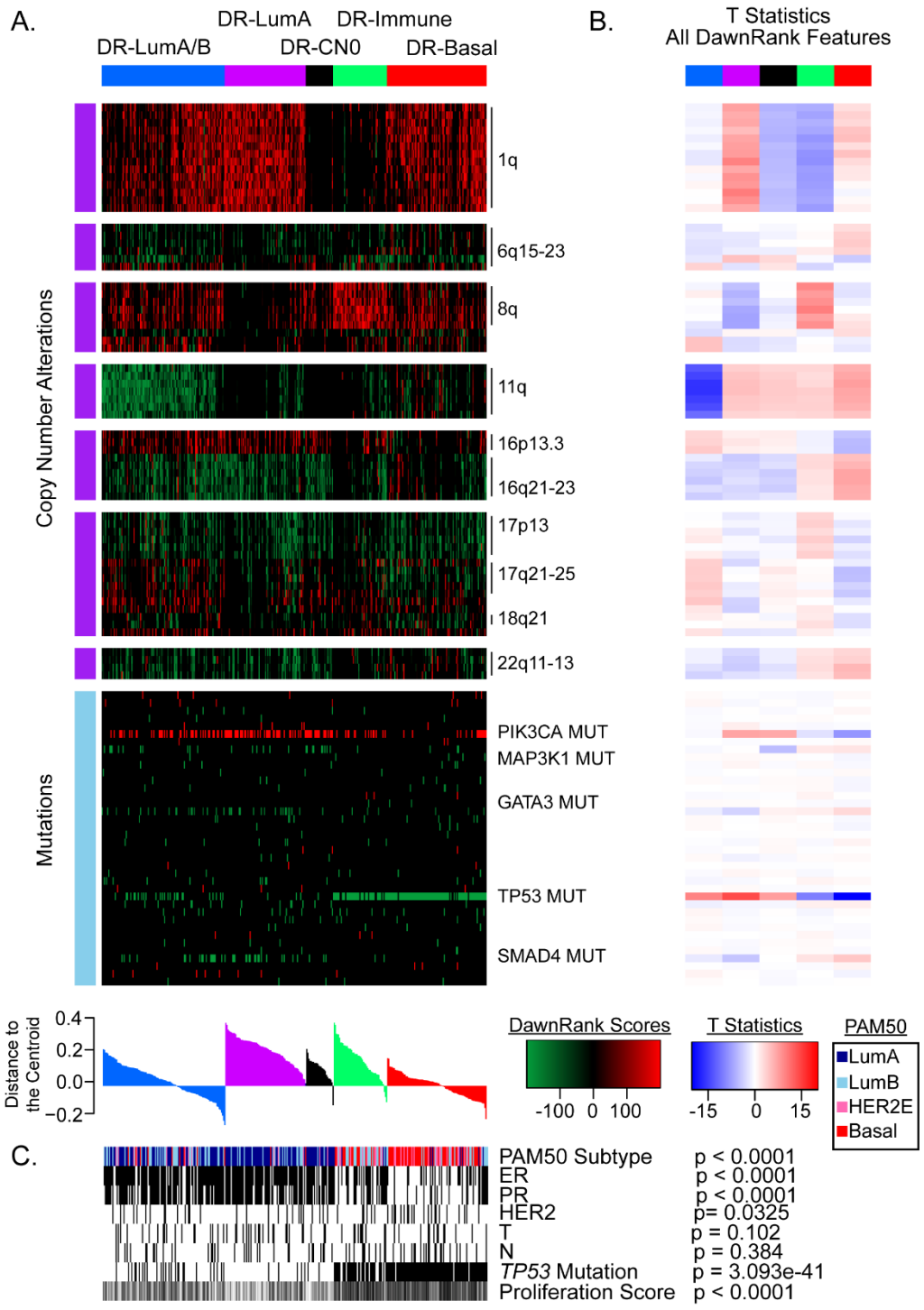

Figure 4.2. DawnRank Subtype Identification. (A) DawnRank scores were calculated (red = high score, oncogene; green = high score, tumor suppressor), and clusters identified using ConsensusClusterPlus. Tumors are ordered by the distance to the centroid. Copy number alterations and mutations are ordered by chromosome and position in the genome. (B) T statistics are plotted for each group, calculated by comparing one group to all other tumors (red = maximum association; blue = negative association). (C) Molecular and clinical characteristics of each tumor with Chi square test p values reported (black = positive or mutated; white = negative or wild-type).

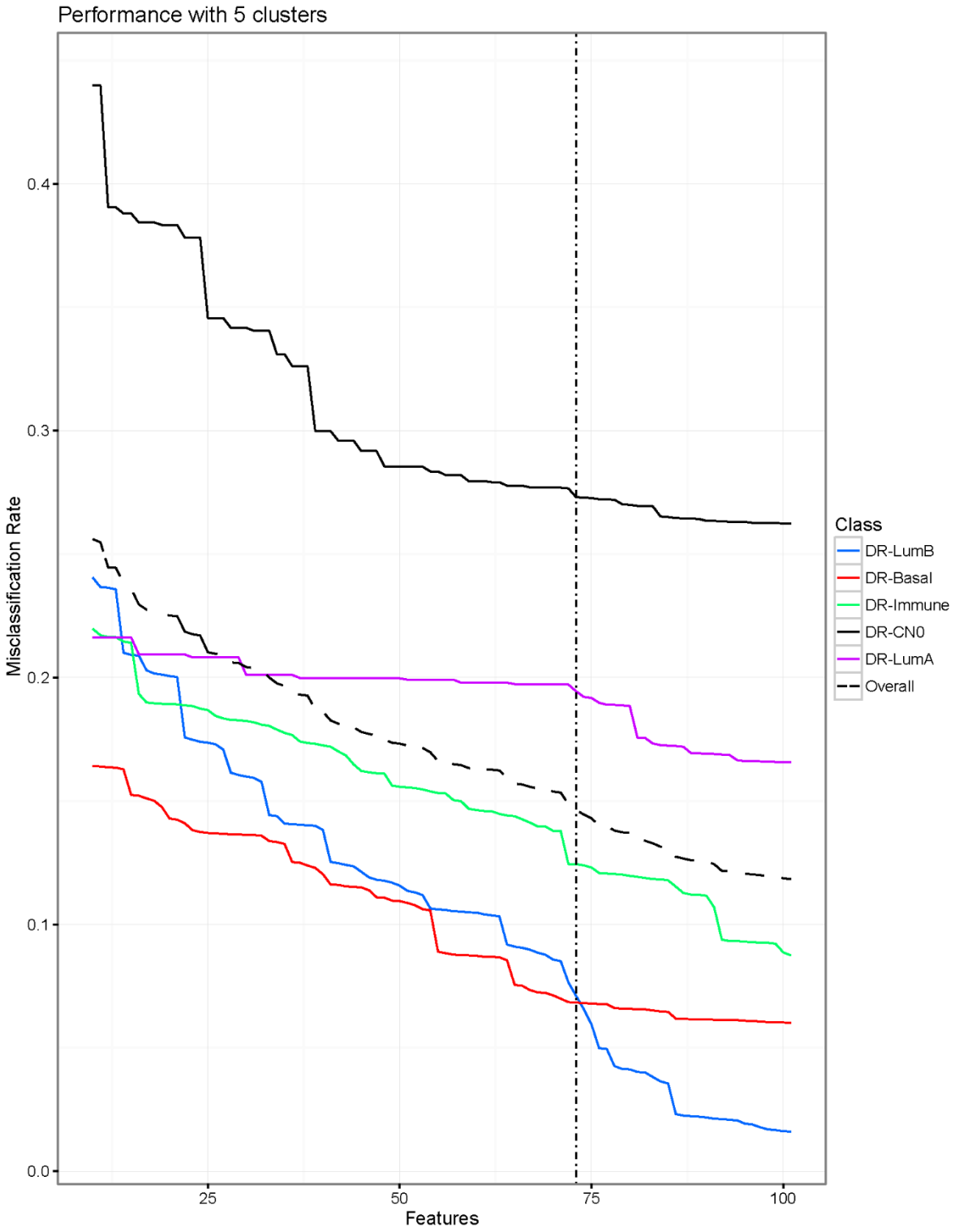

Figure 4.3. Misclassification Rate of ClaNC. Increasing features by 5, misclassification rate was calculated by comparing ClaNC classifier to the original ConsensusClusterPlus subgroup identification (blue = DR-LumA/B; red = DR-Basal, green = DR-Immune; black solid = DR-DR-CN0; purple = DR-LumA; black dashed = overall rate). Vertical line indicates the chosen number of features in the final classifier.

Subtype-defining drivers

To define subtype-specific drivers, each driver (68 CNAs and 38 mutations) was tested by t statistic with one class against all others (Figure 4.4B). DR-LumA/B is defined by chr11q loss including *BIRC3*, *ATM*, *CBL,* and the T cell receptor family genes *CD3E/D/G*. Examination of the *BIRC3* DawnRank network demonstrates a distinct up-regulation of *PAK1*, a known oncogene that activates MAPK and MET signaling (Figure 4.5) (Shrestha et al., 2012). This network may be the cause of the increased proliferation rate of DR-LumA/B compared to other ER-positive tumors.

Interestingly, the Immune subgroup lacks 1q amplification but has distinct gain of 8q amplification. *ERBB2* is located at 8q, but other drivers defined here include *IKBKB, LYN, COPS5, NCOA2,* and *SDC2. LYN* is a known oncogene that can mediate anti-estrogen resistance in ER-positive breast cancer (Schwarz et al., 2014). In contrast, the DR-Basal subgroup has 1q amplification and a lack of focal 16p13.3 amplification and 8q amplification.

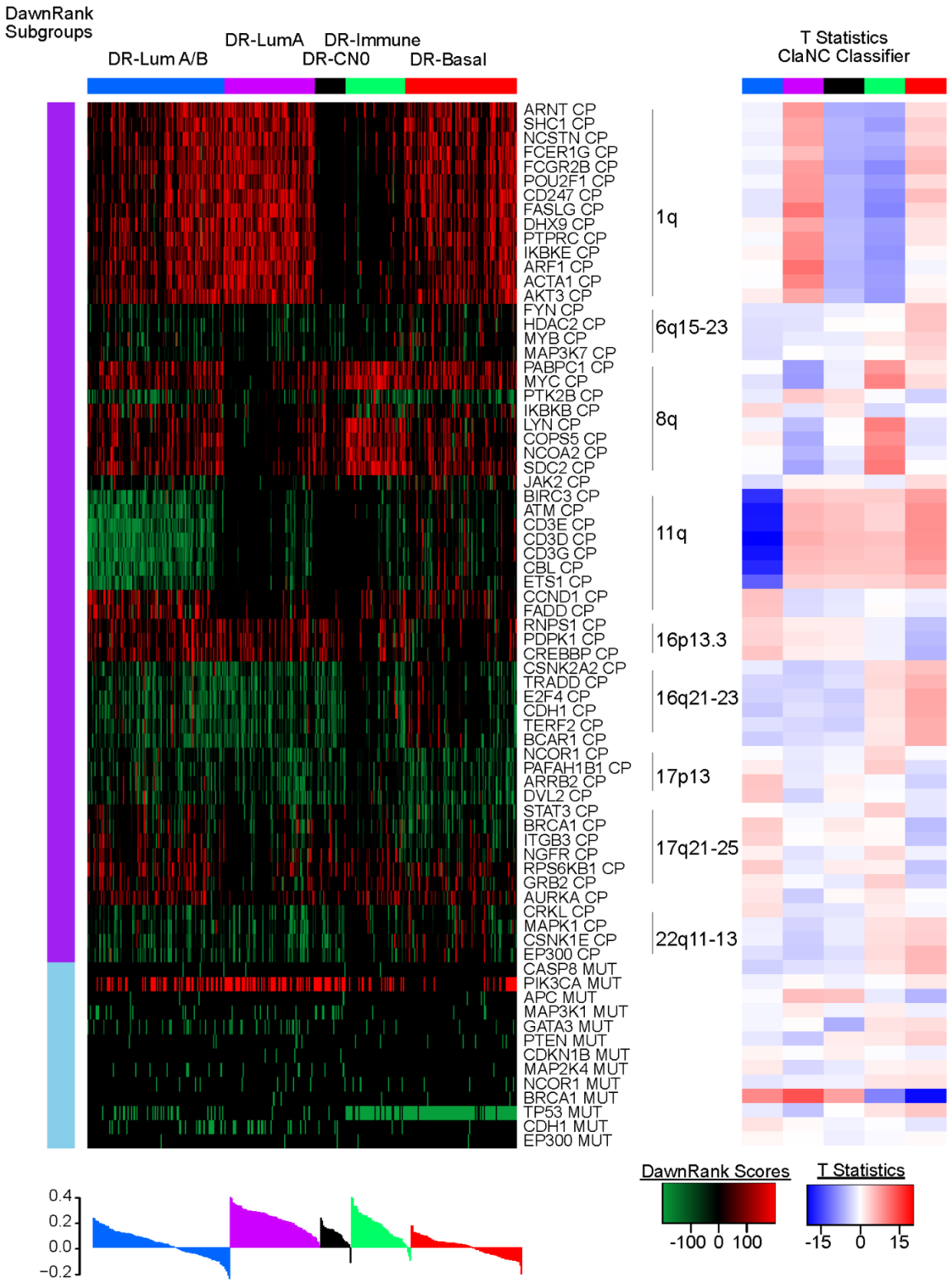

Figure 4.4. DawnRank scores and statistics of ClaNC features. DawnRank scores of the 75 features selected in the ClaNC classifier, with tumors ordered by subgroup and the distance to the centroid and features ordered first by copy number alteration or mutation and secondly by genomic location. T statistic is reported as the t-estimate for each subgroup compared to all other tumors.

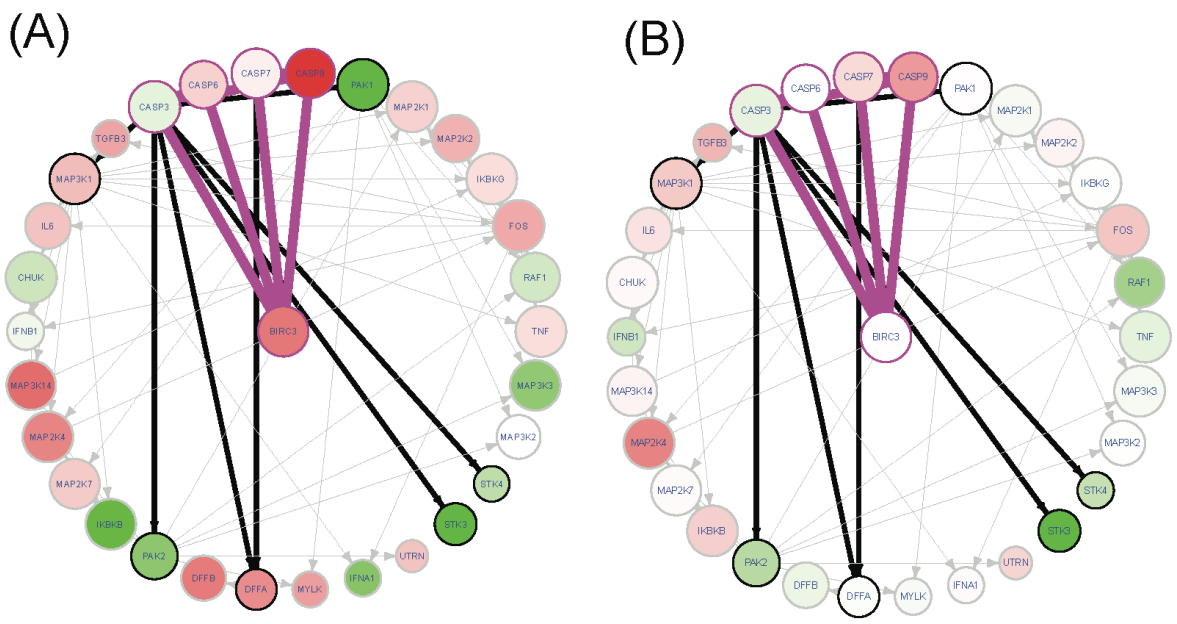

Figure 4.5. BIRC3 network distinctly altered in DR-LumA/B tumors. (A) A network view in the DR-LumA/B subtype detailing the gene interactions between BIRC3 and nearby genes in the network up to two levels downstream. Red nodes represent down-regulation and green nodes represent overexpression. The intensity of the node represents the magnitude of gene expression. Edge thickness and color represent the distance between these gene in question and BIRC3. Magenta edges represent 1 degree of separation from BIRC3, black represents 2, and gray represents 3. (B) A network view of BIRC3 in non-Luminal subtypes.

Clinical and Molecular Heterogeneity within the Driver Subgroups

To classify the driver-subtypes in context of previously defined clinical predictors and molecular taxonomy, we examined the correlation with PAM50 subtype, known clinical predictors, and IntClust. Predominantly tumors of the PAM50 Luminal B subtype comprise the first subgroup, defined by ER and PR positivity as well as an increased proliferation rate (Figure 4.2C). Two luminal A subtypes were identified, one with a distinct lack of 1q amplification thus called copy number neutral (CN0). This subset of luminal A breast cancers have been previously reported both in a subset analysis of TCGA luminals (Ciriello et al., 2013) and defined in the METABRIC IntClust taxonomy(Curtis et al., 2012).

Interestingly, two subgroups are comprised of a mixture of basal-like, HER2-enriched, and Luminal B tumors. These tumors have significantly increased proliferation rates, higher rates of *TP53* mutation, and lack ER and PR expression (Figure 4.2C). This is the first classifier, to our knowledge, to categorize PAM50 basal-like and HER2E tumors into two subgroups.

Clinically, tumor stage (T) and nodal status (N) do not correlate with the driver subtype classification (Figure 4.2C). This demonstrates the added knowledge of these subtypes above known clinical characteristics.

Protein and Pathway Expression Varies by Driver Subgroup

To identify pathways differentially expressed, we performed parametric t-tests and ANOVA testing of each class versus all others to identify protein and gene expression differences among the subgroups. Utilizing previously published gene signatures, the mean gene signature scores for each tumor were compared. Interestingly, the luminal progenitor signatures were highly expressed in the DR-Basal subgroup with concurrent down-regulation of mature luminal and estrogen signaling gene expression signatures (Figure 4.6A). Not surprisingly, those subgroups dominated by ER-negative tumors had lower expression of estrogen markers; however, even the PAM50-Luminal B tumors that were grouped into the DR-Basal subgroup

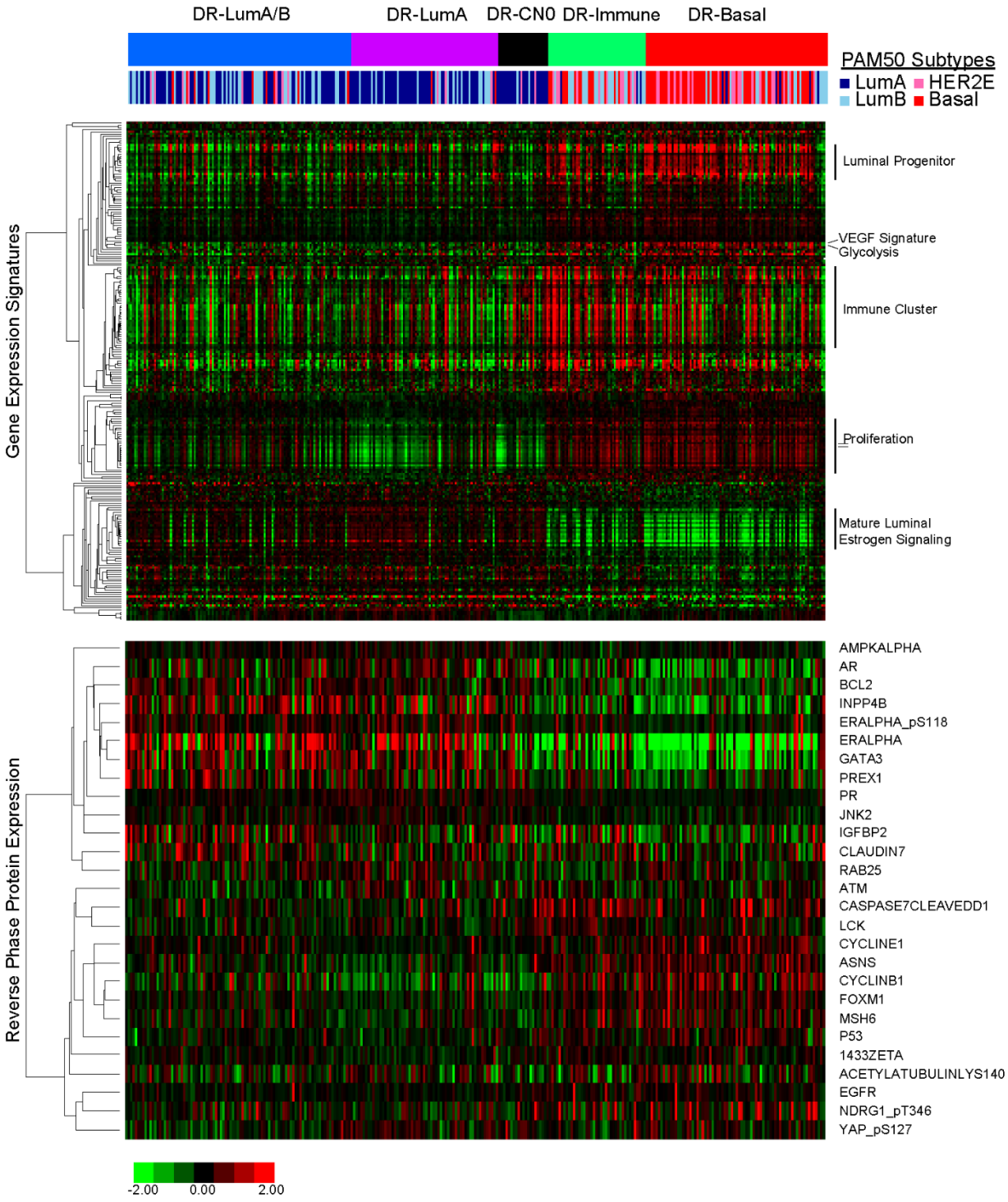

Figure 4.6. RNA Pathway signatures and protein alterations. Signatures and protein expression significant across the cohort by a parametric ANOVA and also significant in at least one subgroup compared to all others (FDR = 0) were median centered across the cohort and clustered. Tumors were ordered by the distance to the centroid within each subgroup.

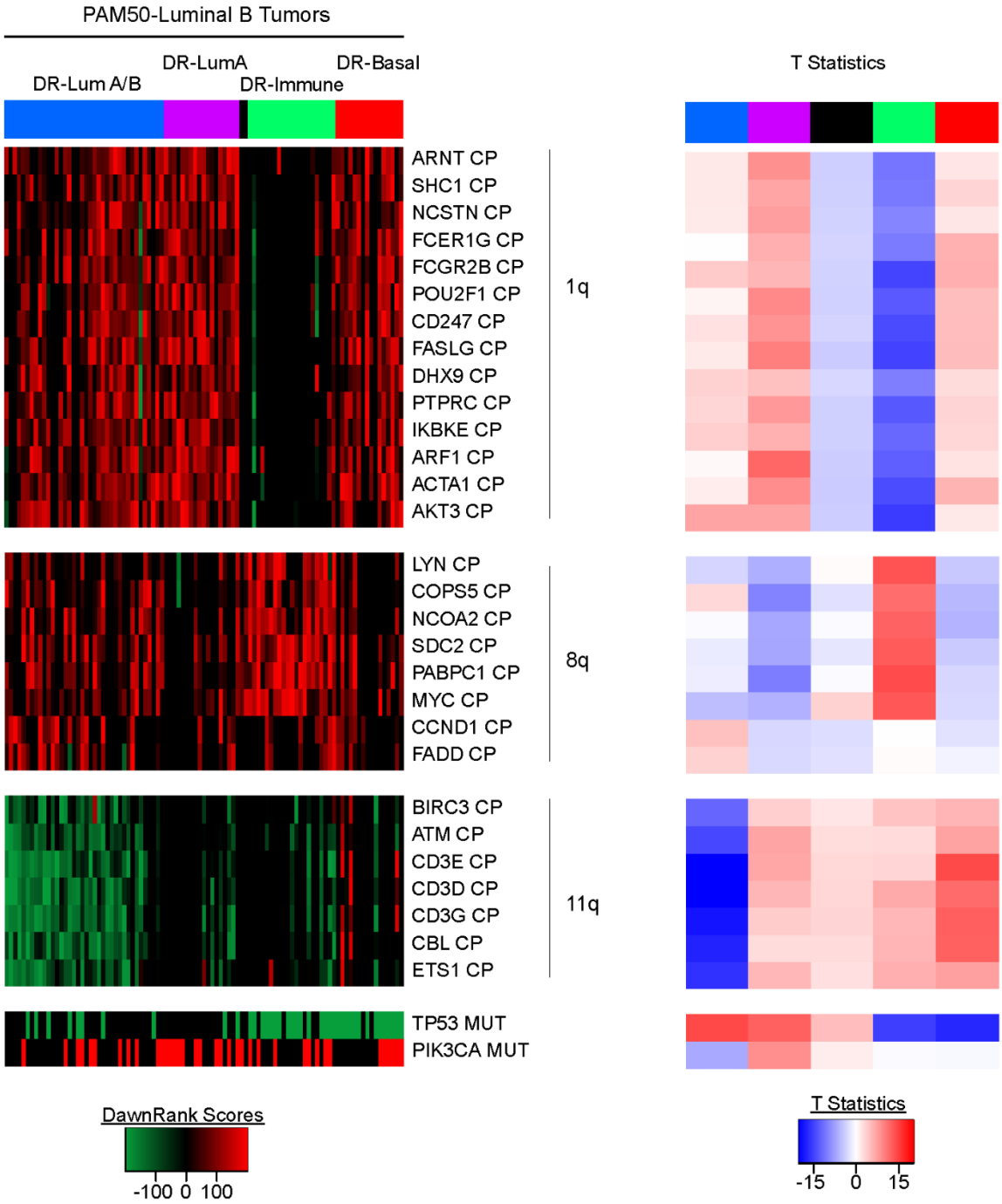

Figure 4.7. PAM50-Luminal B tumors reclassified into DawnRank subgroups. (A) DawnRank scores for the n=98 PAM50-Luminal B tumors demonstrate DR-subtype defining features. (B) T statistics are reported as each subtype compared to all other tumors.

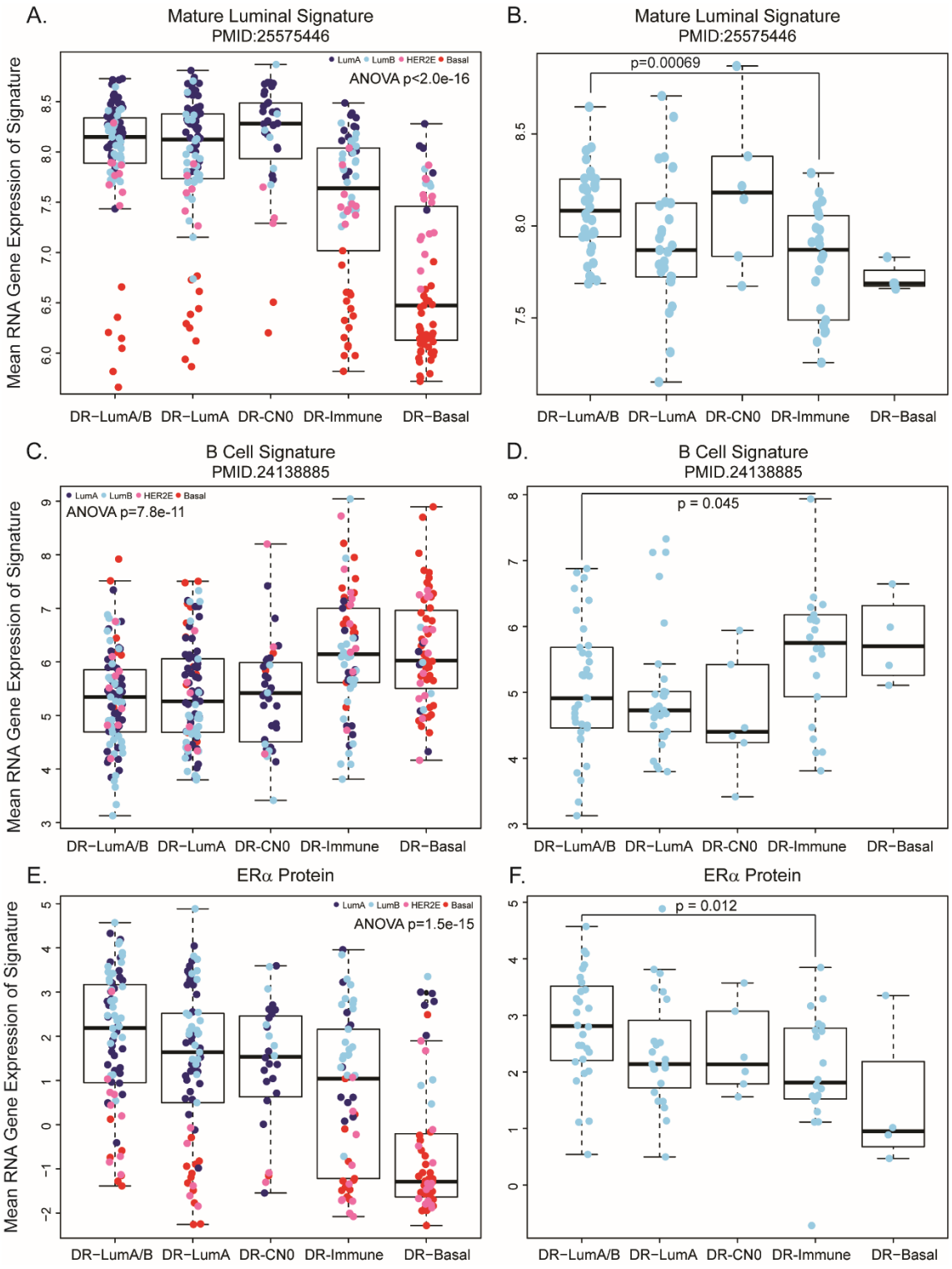

Figure 4.8. Gene expression signature differences. Boxplots demonstrating variability of gene sigantures and protein expression both across all TCGA samples as well as those specifically characterized by the PAM50 classifier as Luminal B for the Mature Luminal signature (A,B), B cell signature (C,D), and ERα protein expression (E,F).

have lower expression of the mature luminal and estrogen signaling pathways. This indicates an alteration in estrogen signaling in these tumors concurrent with an increased proliferation.

The other group of signatures significantly expressed are the immune signatures. Significant up-regulation of the immune gene signatures define the Immune subgroup. Again, PAM50-Luminal B tumors that were classified into the Immune DawnRank subtype have significantly higher immune infiltrate than those in the DR-LumA/B subgroup (Figure 4.8A).

In addition to the gene expression signatures, we also analyzed publicly-available reverse phase protein arrays (RPPA) data from the Cancer Genome Atlas to investigate protein expression differences. Known estrogen signaling proteins including GATA3, INPP4B, and AR are overexpressed in the more luminal subtypes, DR-LumA/B, DR-Luminal A, and CN0 (Figure 4.6B). This confirms protein expression of the gene expression measured in Figure 4.6A: distinct down-regulation of estrogen signals in ER-positive tumors are classified by our DawnRank classifier.

DawnRank Subtypes Confer Improved Survival Differences Beyond Current Clinical and Molecular Predictors

Utilizing gene expression, recently published mutation data (Pereira et al., 2016), and copy number data from the METABRIC dataset (n = 339 patients), we calculated the DawnRank scores for each tumor and applied the ClaNC classifier. METABRIC confirms the association of the DawnRank subtypes with the PAM50 classifier. We further confirmed the gene expression signature associations with subtype as analyzed above, validating the increased immune expression in the Immune subgroup and loss of differentiation in ER-positive tumors of the DR-Basal subtype Figure 4.10. To first examine the differences in survival prediction, we performed Kaplan-Meier plots and survival analysis of the DawnRank subtypes and PAM50 classifier.

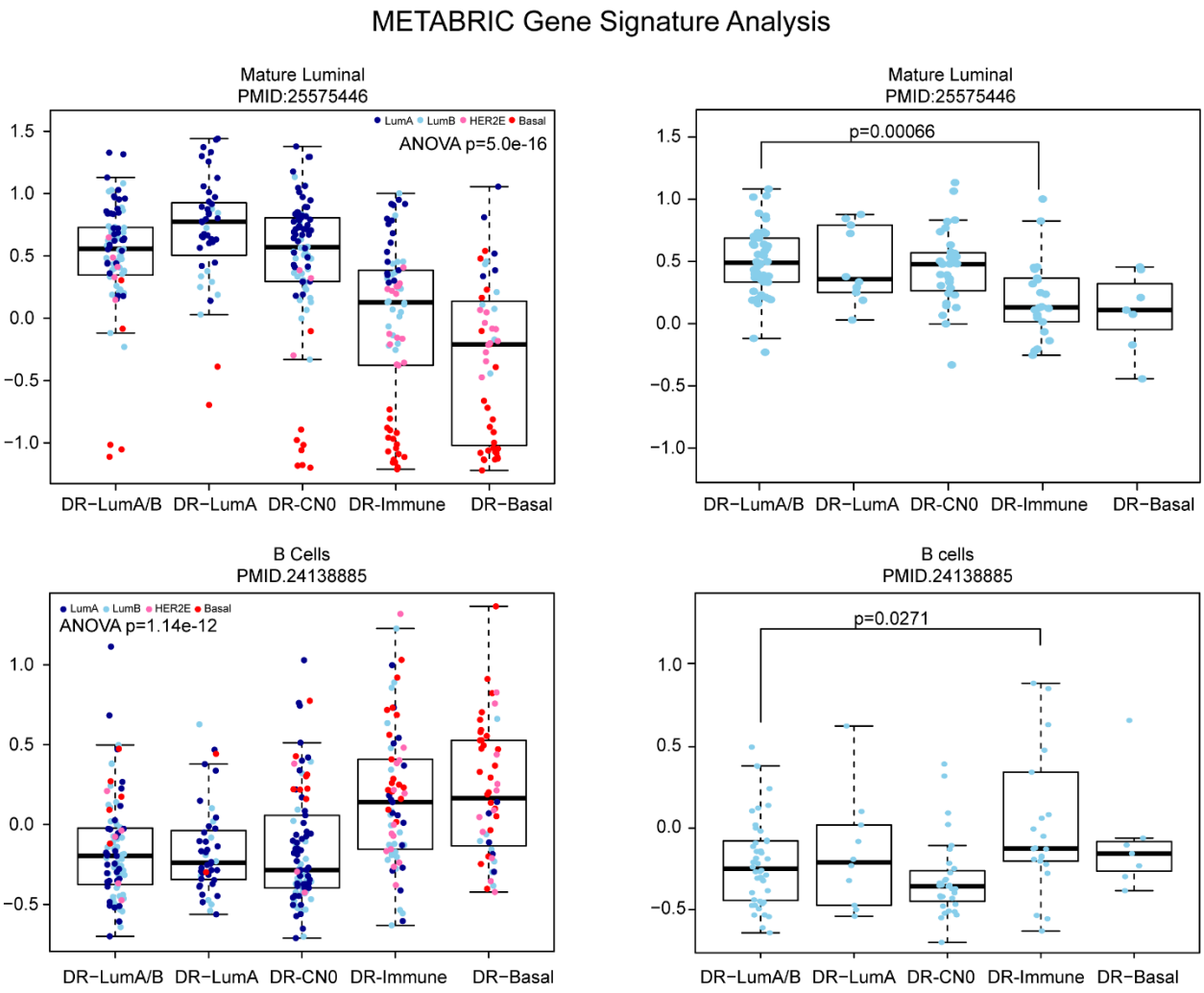

Figure 4.9. Validation of expression differences with METABRIC. Gene signature scores for the Mature Luminal signature (A,B), and B cell signature (C,D) both across all of TCGA and only in the PAM50 Luminal B subtype.

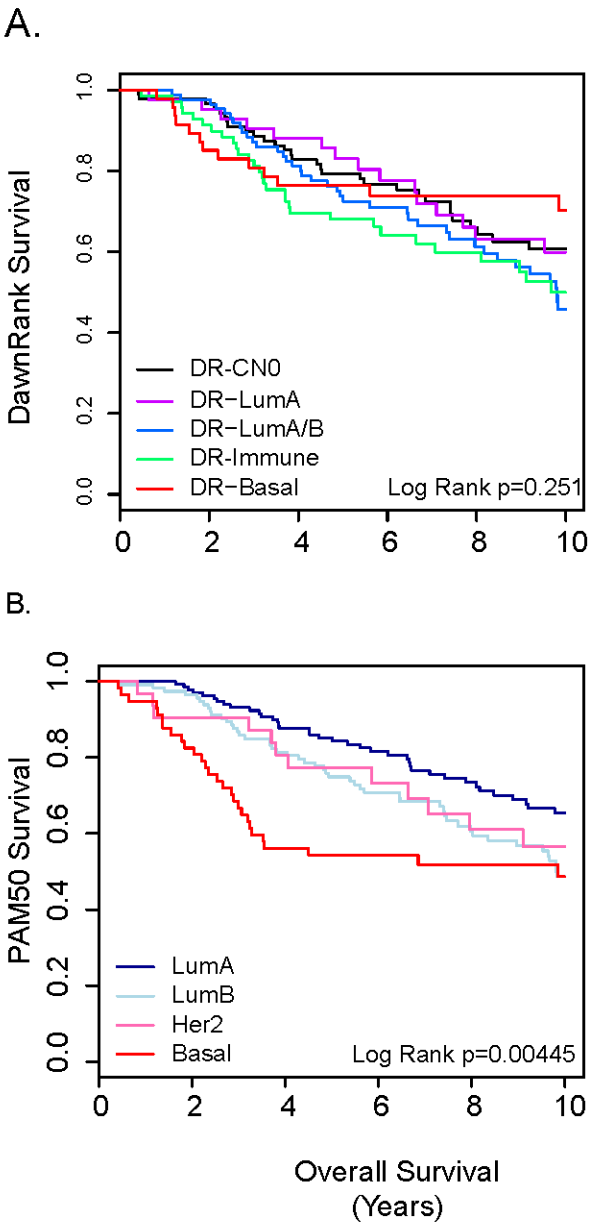

Figure 4.10. Association of survival by subtype. Kaplan-Meier plots and log rank tests of the overall survival up to 10 years from the METABRIC dataset of three classifiers: (A) DawnRank, (B) PAM50, and (C) IntClust classifiers.

While DawnRank is not significant by log likelihood test, PAM50 demonstrate significant separation of subtypes (Figure 4.9). We next tested univariate and multivariate Cox proportional hazard comparing survival and clinical or molecular variables. When incorporating DawnRank subtype in addition to the clinical variable, DawnRank subtype contributed survival outcome information in addition to ER status, PAM50 subtype, IntClust classification, *TP53* mutation status, nodal status, ERBB2 status, and tumor size. Interestingly, ERBB2 status and IntClust subtype were not significantly predictive of survival alone.

We then tested DawnRank subtype in a multivariate Cox proportional hazard test, incorporating ER status, Stage, nodal status, PAM50-subtype, and DawnRank (Table 4.1). Nodal Status had the most significant prediction power, followed by DR-LumA/B and Immune subtypes. PAM50-Basal and DR-LumA were also significant. Interestingly, ER status was not significant.

In this study, we have demonstrated the ability to use an indirect classifier of empirical driver analyses to generate robust subgroups associated with both clinically relevant features as well as clinical outcome.

Table 4.1. Cox proportional hazard test of DawnRank subgroups with other known molecular and clinical features.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	Significance
ER status (positive)	-0.768	0.464	0.457	-1.680	0.093	.
Stage1	-0.334	0.716	0.357	-0.936	0.349	
Stage2	-0.265	0.767	0.257	-1.029	0.303	
Stage3	0.301	1.352	0.415	0.727	0.467	
PAM50-LumB	0.493	1.637	0.266	1.851	0.064	.
PAM50-Her2	0.463	1.589	0.462	1.002	0.316	
PAM50-Basal	1.008	2.740	0.516	1.953	0.051	.
Nodal status	0.558	1.746	0.169	3.308	0.001	***
DR-LumA	0.968	2.632	0.476	2.031	0.042	*
DR-LumA/B	1.208	3.346	0.436	2.770	0.006	**
DR-Immune	0.995	2.705	0.387	2.573	0.010	*
DR-CN0	0.694	2.001	0.438	1.583	0.113	

Significance. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

Discussion

We present a novel classification of breast cancer by calculating heuristic driver scores from and integration of gene expression, copy number, and mutation data. Utilizing both The Cancer Genome Atlas (TCGA) as a test set and the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) as the validation set, we demonstrate five robust driver-subtypes. Our subtypes include a pure Luminal A subtype, a copy number neutral subtype, a pure Luminal B subtype, and two mixed subtypes: one with an increase in immune infiltration and the other demonstrating a de-differentiation phenotype.

Known hotspots of copy number alteration in breast cancer, including 1q amplification, 8q amplification, 11q loss, and 16q loss, demonstrate subtype-specific differences. Chromosome 11 loss is specific to the DR-LumA/B subtype including *BIRC3* and *CBL* loss. *BIRC3* network analysis demonstrates loss of *BIRC3* and concurrent up-regulation of *PAK1*, a known oncogene downstream of *BIRC3*. A second interesting result is the loss of *CBL*, an E3 ubiquitin protein ligase which recognizes known oncogenes including *FGFR2*, *KIT*, and *PDGFRA*. *CBL* loss has not been previously described in the context of Luminal breast cancer. Targeting of FGFR family members with dovitinb has been showing to be effective in a small cohort of breast cancer patients in Phase 2 trial (André et al., 2013). *CBL* loss could be a second marker for FGFR sensitivity in patients who lack FGFR amplification but still may be dependent on this pathway.

Integrating gene expression to evaluate the impact of a genomic alteration, both mutation and copy number alterations, allows for novel subgroup identification. We demonstrate that these subgroups have survival differences beyond known clinical and molecular markers. There is information to be gained by utilizing a dynamic, integrated driver analysis including identification of novel therapeutic targets such as *PAK1* in DR-LumA/B tumors. Improving our understanding of the molecular drivers of underpinning different subtypes of breast cancer are necessary to develop more targeted therapies.

Future *in vitro* and *in vivo* confirmation will be needed to confirm our findings. In addition, we are limited by known, curated pathways used to evaluate the networks. Assessment of these drivers through both therapeutic selection (comparing pre-treatment and post-treatment samples) and the selection of these drivers through the metastatic process are needed. DawnRank network-based analysis on both metastases and clinical trial samples with available gene expression, mutation, and copy number data are needed to understand the shift in drivers during these selection processes.

The heterogeneity of breast cancer has long been described and understood from a clinical, histopathologic, and molecular lens. Through a novel computational framework, we were able to capture this heterogeneity and assess novel molecular drivers for each breast cancer subtype. Future functional studies confirming the role of these drivers in a subtype-specific manner are needed in order to lead to novel therapeutic development. Incorporation of mutations, copy number alterations, and gene expression confirm the importance of evaluating not only mutations but also copy number variations in understanding the underlying biology driving breast cancer.

CHAPTER 5 – DISCUSSION AND CONCLUSION

Breast cancer metastasis is still a devastating diagnosis with limited treatment options, especially for women with hormone receptor negative breast cancer. A better understanding of the process of metastasis, the timing with which the metastatic potential is established, and the common drivers in metastasis are needed to develop better therapeutic interventions. By comparing matched primary breast cancers and multiple metastatic sites, we described the clonal process of evolution in 16 patients, defined computationally-determined drivers, and identified the timing with which these drivers were acquired. Through these research projects, we attempted to clarify three questions: (1) is metastasis a monoclonal or polyclonal event; (2) when are metastatic drivers acquired; (3) what are common mechanisms of metastasis.

Polyclonal Seeding in Breast Cancer Metastasis

We determined that polyclonal seeding can occur in both luminal and basal-like breast cancers. We first performed whole-genome sequencing of two patients with triple-negative, basal-like breast cancer with a primary, 4 and 5 matched metastases, and a matched normal tissue to define the germline genotype. We then defined clones by SciClone (Miller et al., 2014), demonstrating that multiple clones are present in the primary and leave to metastasize. Additionally, these metastases are made up of more than one clone in every instance. This is in contrast to previous research mostly from other cancer types where monoclonal seeding appears to predominate. We then expand these findings in Chapter 2 in 16 patients, demonstrating both monoclonal seeding and polyclonal seeding. Both basal-like and luminal breast cancers demonstrate each model of seeding.

Previous literature suggests a single cell escapes the primary and then diversifies during metastasis. A seminal paper using DNA sequencing of a matched tumor and metastasis was in renal cell carcinoma, demonstrating branched evolution with a single cell of origin seeding distant metastasis (Gerlinger et al., 2012). A study of basal-like breast cancer, the matched metastasis, and a xenograft show high percentage of shared genetics across all 3 tumors (Krøigård et al., 2015). Single cell sequencing of one breast cancer with one matched liver metastasis suggest a single cell seeded the distant site (Navin et al., 2011). This is corroborated by a large panel of matched primary and brain metastases sequenced, showing continued evolution and acquisition of resistance mechanisms in the brain metastasis specifically (Brastianos et al., 2015). A larger study in prostate cancer suggests metastasis can seed other metastases (Gundem et al., 2015). Other studies in non-small cell lung cancer (Govindan et al., 2012), colorectal cancer, and ovarian cancer (Castellarin et al., 2013; Schwarz et al., 2015) all shed light on the cancer evolution through metastasis.

Recent *in vivo* evidence, however, sheds light on how polyclonal seeding might be possible. First in a genetically engineered pancreatic cancer mouse model, metastasis was shown to be a result from at least two distinct populations (Maddipati and Stanger, 2015). Recent evidence using a breast cancer genetically engineered mouse model further demonstrates not only polyclonal seeding of lung metastases but also that tumors cells self-seed the contralateral fat pad of the mouse (Cheung et al., 2016). Furthermore, recent investigation demonstrates the fluid dynamics and video imaging of how exactly polyclonal seeding could occur (Au et al., 2016).

Why would polyclonal seeding occur in metastasis? What is the evolutionary advantage? In order for a breast cancer to break off, survive through circulation, successfully land in a distant organ, and survive all while escaping immune surveillance, some level of genetic diversity and adaptation is needed. Others have observed clumps of circulating tumor cells in cancer patients (Aceto et al., 2014), further suggesting that multiple cells are needed in order for

metastasis to be successful. Potentially, this genetic diversity would be better shared across multiple tumor cells rather than the entire genetic burden existing in one cell. Furthermore, cross-talk between multiple types of cells could be beneficial: studies in our lab have demonstrated that most basal-like cancers show a mixed population of tumor cells containing both claudin-low (stem-cell like population) and basal-like cells, when we isolate only 1 population of cells, the cells can repopulate both populations such that both populations are in the final culture (Prat et al., 2010). In addition, Zhang et al (Zhang et al., 2015) showed there is growth factor cross talk between these two populations, such that one makes the ligand the other the receptor; thus both populations would be needed in order to keep a tumor going. This suggests both tumor cell plasticity as well as a need for both populations to exist for the cancer cells to continue to grow.

Similarity of Primary and Metastatic Breast Cancer

When we looked at the RNA expression profiles of our 86 tumors compared to over 1000 breast cancers from TCGA, the metastases were more similar to the matching primary tumor than other breast cancers. This confirms previous findings from our group with a smaller number of metastases and primaries (Harrell et al., 2012). Additionally, the primary breast cancer carries significant prognostic information including future site of first metastasis and overall survival. All of these conclusions provide evidence for much of the metastatic phenotype residing within the original primary breast cancer. Identifying these genetic features could provide therapeutic targets in the neo-adjuvant and adjuvant settings to ultimately prevent metastatic spread if these critical factors are identified.

Recent publications have identified that a majority of the ‘genetic drivers’ are private to distant brain metastasis, not established in the original primary (Brastianos et al., 2015). While investigating our DNA sequencing data, we observed that mutations called in one or two tumors were present at very low coverage in the other tumors from that patient, especially in the

primary breast cancer. We computationally re-interrogated these mutations in two ways: first, we took the union set of mutations from one patient and counted the mutant allelic reads in each tumor. Second, we took all sequencing runs from a single patient, collapsed them into one "tumor" and *de novo* called mutations. We confirmed all of the mutations that had been identified from each tumor individually compared to the normal, only missing some insertions/deletions (which are notoriously difficult to identify). In the whole genome sequencing paper, we identify 2-3% clones in Patient A1 in the primary breast cancer specific to the liver and adrenal metastasis, proving that the original breast cancer contains multiple clones that together seed distant metastasis. These were only identified through computational re-interrogation. In Chapter 2, we formally examine this re-interrogation and demonstrate a 30% increase in 'founder' mutations across the dataset that would have been otherwise missed.

It is absolutely critical for evolutionary metastatic studies to perform computational re-interrogation. In order to understand the timing with which drives are acquired, we must first accurately identify *when* in that patient's cancer the genetic alteration occurred. Improper conclusions will be drawn if a depth coverage is required upon re-interrogation. Multi-regional sequencing of primary breast cancers demonstrated that primary breast cancers can have >10 clones in them at times. Thus, a clone that seeds a metastasis and has 20-40% variant allele frequency in the metastasis could be as low as 1-2% in the primary when performing bulk sequencing of the primary. If we require a variant frequency cutoff of 5%, these would be missed. Re-analysis of publicly available metastatic datasets will be needed in the future to fully appreciate how different the conclusions could be in an independent dataset.

Timing of Copy Number and Point Mutation Alterations in Cancer Development

Recent literature described the timing of acquired genetic alteration. The authors showed that mutations are acquired in a linear function of time such that the older the tumors, the more mutations that tumor would have. This fits with other publications that demonstrate

pediatric cancers have relatively simple genetics – maybe one or two mutations – compared to melanoma or lung cancer, which tend to have the highest mutation burden caused by DNA mutation inducing origins (i.e. smoking and UV light). In contrast, the authors demonstrate that copy number events occur in large smatterings wherein the genome is significantly disrupted in one point in time, followed by large changes, and then relative stability. This difference in the process of acquiring genetic alterations fits with our findings: there seems to be a relative steady increase in the number of mutations as these clonal populations grow in the primary, metastasize, and seed distant sites. In contrast, when copy number alterations occur, it seems as though huge amounts of the genome are altered all at once. There is relatively little 'private' copy number alteration in the genomes of our metastatic patients.

Copy number alterations as a mechanism driving breast cancer progression is incredibly important. Copy number changes alter a large number of genes effectively in comparison to mutations. Thus, the cancer has a mechanism for generating large genetic diversity quickly. In our metastatic breast cancer patients as well as in primary breast cancer, copy number was the dominant mechanism for causing drivers. In trying to identify common mechanisms of metastasis, only one mutation was shared among our two patients with whole genome sequencing data and among 13/16 patients with whole exome sequencing (i.e. *TP53*). In contrast, 15/16 patients had common copy number altered regions, and these alterations were almost always established in the primary and maintained throughout metastasis.

The only common mutation across both our two patients with whole genome sequencing and in the whole exome dataset was *TP53. TP53* is the most highly mutated gene in cancer and is known to be negatively prognostic in breast cancer. Interestingly, the basal-like breast cancer patients in our dataset were more likely to have missense mutations, and when these missense mutations occurred, they were also expressed in the RNA. In contrast, the luminal metastatic patients often had complete frame-shift, insertions/deletions, or early stop-codon mutations. These alterations were not expressed in the RNA, as they likely produced nonsensical RNA

transcripts which would be degraded by nonsense mediated decay. Finally, these alterations were all established in the original primary and carried in every single metastasis in every patient in which the *TP53* alteration was present. This provides definitive evidence that *TP53* disruption is a critical event to generate breast cancer metastasis regardless of subtype.

Aggressive breast cancer is a heavily copy-number altered disease. Triple negative breast cancers of the basal-like molecular subtype have the worst 5-year overall survival and the largest burden of copy number alterations. Studying only the ER+, luminal breast cancers, the poorest prognostic breast cancers again have *TP53* alteration and significant copy number destabilization. Potentially, *TP53* disruption is a critical event to generate genomic destabilization and ensuing copy number alteration. This also fits with the previous hypothesis that copy number alterations occur in bursts. Finally, this is consistent with one previous study which demonstrate that *TP53* disruption and copy number alteration are the only occurrences shared in pre-invasive ductal carcinoma *in situ*, invasive breast cancer, and a matched lymph node.

Heterogeneity of Primary Breast Cancer Genetic Drivers

Previous work in our group as well as many others have demonstrated the large amount of variation in primary breast cancer. Even within the PAM50 molecular subtypes or clinical subtypes of hormone receptor positive versus hormone receptor negative disease, there are large variations in clinical response and survival. In Chapter 4, we strove to identify variation of genetic drivers in primary breast cancer with the similar computational strategy as applied to our metastatic tumors.

Defining 5 distinct ‘driver’ subtypes of breast cancer, we observed that these divisions were not based on estrogen receptor positivity. In contrast, we had two subgroups with mixed HER2 positive, ER positive, and ER negative breast cancers. When stratifying by ER positivity,

there was still significant survival differences across our ‘driver’ subgroups. This demonstrates that there are common mechanisms driving breast cancer across the different subgroups.

Interestingly, our most mixed subgroup of cancer had the worst overall survival but an elevated immune infiltrate. Immune infiltrate typically predicts an improved prognosis in both HER2-enriched and basal-like molecular subtypes but not the luminal subtypes. These immune infiltrated PAM50 Luminal B tumors that end up in the mixed driver subtype have a poor prognosis. They also demonstrate a loss of estrogen regulation and loss of the mature luminal phenotype. Whether this is due to the increased immune infiltration thus decreasing the differentiation or an actual biologic down-regulation of estrogen receptor is unknown; however, previous research in the fields of both oncology and rheumatology have demonstrated estrogen’s immune repressive affects. The possibility that estrogen may mediate the ineffectiveness of immune infiltrate has very interesting therapeutic implications. In metastasis, often estrogen receptor-positive primaries lose estrogen signaling and become more dedifferentiated. Potentially, in the metastatic setting, immune modulatory-therapies could be harnessed with more power than in the adjuvant setting when ER-positive tumors are highly estrogen dependent. Further investigation is needed to test this hypothesis directly in an *in vivo* and *in vitro* setting.

Clinical Implications of Our Research Findings

Our findings of the relationship between primaries and metastases has significant clinical implications. First, if polyclonal seeding is indeed common in triple negative breast cancer metastases, then therapeutics targeting multiple clones will be needed to effectively eradicate these subclones. This supports previous evidence of why single targeted therapies sometimes do not work in breast cancer due to resistance mechanisms: a primary breast cancer as multiple subclonal populations residing within the tumor that together mediate metastasis. Thus, an

understanding of the heterogeneity existing in a primary breast cancer is critical for effective halting of metastatic progression.

While dual targeted therapy is necessary to block metastatic progression, it is also encouraging that much of the metastatic potential resides within the original tumor. We demonstrate that a majority of the copy number alterations occur in both the primary tumor and the matched metastases. In addition, when filtering the mutations (which are likely acquired as a linear progression of time) to those that significantly alter the gene expression network and thus are called 'drivers', a significant majority of them are established in the primary breast cancer. Therefore, it is possible that therapies targeting this original population that has metastatic potential could be delivered effectively in the adjuvant setting thus preventing metastatic spread. Alternatively, if metastasis does occur prior to detection, potentially effective targeting of these 'founder' mutations that are present both in the primary breast cancer and in the metastasis could prevent and treat future sites of metastasis.

Future research targeting copy number alterations is desperately needed. It has long been known that a majority of driver alterations are a result of copy number alteration. 1q amplification, 5q amplification, 8p loss, and 8q amplification are recurrent copy number alterations across all of breast cancer and are not subtype specific. In addition, triple negative breast cancer have subtype specific copy number alterations known to drive the tumor phenotype and are conserved across species. In our study as well as previous research have demonstrated that copy number alteration is an extremely early event, present in pre-invasive ductal carcinoma *in situ*, and are maintained not only in primary breast cancer but also distant metastasis. Therefore, it should not be ignored as a therapeutic target. Clearly, copy number alteration is a fundamental mechanism of oncogenic activity in breast cancer. Our research in Chapters 1 and 2 support that copy number alteration is a shared mechanism across metastasis. Furthermore, in Chapter 3, we demonstrate that copy number alteration causes more drivers that can separate subtypes of breast cancer than mutation. Finally, it is generally

accepted in the field that copy number alteration is a much more effective mechanism of genomic alteration than mutation, as whole arm amplifications can amplify a host of oncogenic factors at once. Further research in targeting these copy number alterations is desperately needed, especially in hormone receptor negative breast cancer.

Conclusions

In summary, our research has demonstrated that both polyclonal and monoclonal seeding can occur and both are common mechanisms of metastasis across both hormone receptor positive and negative breast cancers, that most genetic drivers are established in the original breast cancer, and that common mechanisms exist across the molecular subtypes of breast cancer. We hope to continue investigating the >50 women's metastases that have graciously donated their bodies to medical research at the end of their lives to continue to enhance our understanding of clonality, evolution, and genetic drivers of metastasis. We furthermore envision a day in which the proper combinatorial therapies that target *TP53* as well as copy number amplifications can effectively prevent and treat metastatic progression in breast cancer. We firmly believe that continue research in metastatic breast cancer is needed to fully understand the molecular mechanisms of therapeutic resistance, site-specific metastasis, and therapeutic targets in the advanced setting. Only then can we begin to develop the proper therapeutic approaches to ultimately help our patients live longer, healthier lives.

REFERENCES

Aceto, N., Bardia, A., Miyamoto, D.T., Donaldson, M.C., Wittner, B.S., Spencer, J.A., Yu, M., Pely, A., Engstrom, A., Zhu, H., et al. (2014). Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. Cell *158*, 1110–1122.

Anders, C., and Carey, L.A. (2008). Understanding and treating triple-negative breast cancer. Oncol. Williston Park *22*, 1233–1239; discussion 1239–1240, 1243.

André, F., Bachelot, T., Campone, M., Dalenc, F., Perez-Garcia, J.M., Hurvitz, S.A., Turner, N., Rugo, H., Smith, J.W., Deudon, S., et al. (2013). Targeting FGFR with dovitinib (TKI258): preclinical and clinical data in breast cancer. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *19*, 3693–3702.

Au, S.H., Storey, B.D., Moore, J.C., Tang, Q., Chen, Y.-L., Javaid, S., Sarioglu, A.F., Sullivan, R., Madden, M.W., O'Keefe, R., et al. (2016). Clusters of circulating tumor cells traverse capillary-sized vessels. Proc. Natl. Acad. Sci. U. S. A. *113*, 4947–4952.

Bertucci, F., Finetti, P., Guille, A., Adélaïde, J., Garnier, S., Carbuccia, N., Monneur, A., Charafe-Jauffret, E., Goncalves, A., Viens, P., et al. (2014). Comparative genomic analysis of primary tumors and metastases in breast cancer. Oncotarget.

Bertucci, F., Finetti, P., Guille, A., Adélaïde, J., Garnier, S., Carbuccia, N., Monneur, A., Charafe-Jauffret, E., Goncalves, A., Viens, P., et al. (2016). Comparative genomic analysis of primary tumors and metastases in breast cancer. Oncotarget.

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A.C., Angell, H., Fredriksen, T., Lafontaine, L., Berger, A., et al. (2013). Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity *39*, 782–795.

Bos, P.D., Zhang, X.H., Nadal, C., Shu, W., Gomis, R.R., Nguyen, D.X., Minn, A.J., van de Vijver, M.J., Gerald, W.L., Foekens, J.A., et al. (2009). Genes that mediate breast cancer metastasis to the brain. Nature *459*, 1005–1009.

Bouaoun, L., Sonkin, D., Ardin, M., Hollstein, M., Zavadil, J., and Olivier, M. (2016). TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. Hum. Mutat. *37*, 865–876.

Brastianos, P.K., Carter, S.L., Santagata, S., Cahill, D.P., Taylor-Weiner, A., Jones, R.T., Van Allen, E.M., Lawrence, M.S., Horowitz, P.M., Cibulskis, K., et al. (2015). Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets. Cancer Discov. *5*, 1164–1177.

de Bruin, E.C., McGranahan, N., Mitter, R., Salm, M., Wedge, D.C., Yates, L., Jamal-Hanjani, M., Shafi, S., Murugaesu, N., Rowan, A.J., et al. (2014). Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science *346*, 251–256.

Butler, T.M., Johnson-Camacho, K., Peto, M., Wang, N.J., Macey, T.A., Korkola, J.E., Koppie, T.M., Corless, C.L., Gray, J.W., and Spellman, P.T. (2015). Exome Sequencing of Cell-Free

DNA from Metastatic Cancer Patients Identifies Clinically Actionable Mutations Distinct from Primary Disease. PloS One *10*, e0136407.

Camp, J.T., Elloumi, F., Roman-Perez, E., Rein, J., Stewart, D.A., Harrell, J.C., Perou, C.M., and Troester, M.A. (2011). Interactions with fibroblasts are distinct in Basal-like and luminal breast cancers. Mol. Cancer Res. MCR *9*, 3–13.

Campbell, P.J., Pleasance, E.D., Stephens, P.J., Dicks, E., Rance, R., Goodhead, I., Follows, G.A., Green, A.R., Futreal, P.A., and Stratton, M.R. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc. Natl. Acad. Sci. U. S. A. *105*, 13081–13086.

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.

Cardoso, F., van't Veer, L.J., Bogaerts, J., Slaets, L., Viale, G., Delaloge, S., Pierga, J.-Y., Brain, E., Causeret, S., DeLorenzi, M., et al. (2016). 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N. Engl. J. Med. *375*, 717–729.

Carey, L.A., Perou, C.M., Livasy, C.A., Dressler, L.G., Cowan, D., Conway, K., Karaca, G., Troester, M.A., Tse, C.K., Edmiston, S., et al. (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA *295*, 2492–2502.

Castellarin, M., Milne, K., Zeng, T., Tse, K., Mayo, M., Zhao, Y., Webb, J.R., Watson, P.H., Nelson, B.H., and Holt, R.A. (2013). Clonal evolution of high-grade serous ovarian carcinoma from primary to recurrent disease. J. Pathol. *229*, 515–524.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods *6*, 677–681.

Chen, K., Chen, L., Fan, X., Wallis, J., Ding, L., and Weinstock, G. (2014). TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. Genome Res. *24*, 310–317.

Cheung, K.J., Padmanaban, V., Silvestri, V., Schipper, K., Cohen, J.D., Fairchild, A.N., Gorin, M.A., Verdone, J.E., Pienta, K.J., Bader, J.S., et al. (2016). Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters. Proc. Natl. Acad. Sci. U. S. A. *113*, E854–E863.

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. *22*, 398–406.

Ciriello, G., Sinha, R., Hoadley, K.A., Jacobsen, A.S., Reva, B., Perou, C.M., Sander, C., and Schultz, N. (2013). The molecular diversity of Luminal A breast tumors. Breast Cancer Res. Treat. *141*, 409–420.

Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., Rhie, S.K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., et al. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell *163*, 506–519.

Cummings, M.C., Simpson, P.T., Reid, L.E., Jayanthan, J., Skerman, J., Song, S., McCart Reed, A.E., Kutasovic, J.R., Morey, A.L., Marquart, L., et al. (2014). Metastatic progression of breast cancer: insights from 50 years of autopsies. J. Pathol. *232*, 23–31.

Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352.

Dabney, A.R. (2006). ClaNC: point-and-click software for classifying microarrays to nearest centroids. Bioinforma. Oxf. Engl. *22*, 122–123.

Daud, A.I., Loo, K., Pauli, M.L., Sanchez-Rodriguez, R., Sandoval, P.M., Taravati, K., Tsai, K., Nosrati, A., Nardo, L., Alvarado, M.D., et al. (2016). Tumor immune profiling predicts response to anti–PD-1 therapy in human melanoma. J. Clin. Invest. *126*, 3447–3452.

Dees, E.C., Cohen, R.B., von Mehren, M., Stinchcombe, T.E., Liu, H., Venkatakrishnan, K., Manfredi, M., Fingert, H., Burris, H.A., 3rd, and Infante, J.R. (2012a). Phase I study of aurora A kinase inhibitor MLN8237 in advanced solid tumors: safety, pharmacokinetics, pharmacodynamics, and bioavailability of two oral formulations. Clin Cancer Res *18*, 4775–4784.

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R., et al. (2012b). MuSiC: identifying mutational significance in cancer genomes. Genome Res. *22*, 1589–1598.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Ding, L., Ellis, M.J., Li, S., Larson, D.E., Chen, K., Wallis, J.W., Harris, C.C., McLellan, M.D., Fulton, R.S., Fulton, L.L., et al. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature *464*, 999–1005.

Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506–510.

Ellis, M.J., Ding, L., Shen, D., Luo, J., Suman, V.J., Wallis, J.W., Van Tine, B.A., Hoog, J., Goiffon, R.J., Goldstein, T.C., et al. (2012). Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature *486*, 353–360.

Fan, C., Prat, A., Parker, J.S., Liu, Y., Carey, L.A., Troester, M.A., and Perou, C.M. (2011). Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. BMC Med. Genomics *4*, 3.

Fidler, I.J. (2001). Seed and soil revisited: contribution of the organ microenvironment to cancer metastasis. Surg Oncol Clin N Am *10*, 257–269, vii – viiii.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. *43*, D805–D811.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1.

Gatza, M.L., Lucas, J.E., Barry, W.T., Kim, J.W., Wang, Q., Crawford, M.D., Datto, M.B., Kelley, M., Mathey-Prevot, B., Potti, A., et al. (2010). A pathway-based classification of human breast cancer. Proc. Natl. Acad. Sci. U. S. A. *107*, 6994–6999.

Gatza, M.L., Silva, G.O., Parker, J.S., Fan, C., and Perou, C.M. (2014). An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. Nat. Genet. *46*, 1051–1059.

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. *366*, 883–892.

Gewinner, C., Wang, Z.C., Richardson, A., Teruya-Feldstein, J., Etemadmoghadam, D., Bowtell, D., Barretina, J., Lin, W.M., Rameh, L., Salmena, L., et al. (2009). Evidence that inositol polyphosphate 4-phosphatase type II is a tumor suppressor that inhibits PI3K signaling. Cancer Cell *16*, 115–125.

Glas, A.M., Floore, A., Delahaye, L.J.M.J., Witteveen, A.T., Pover, R.C.F., Bakx, N., Lahti-Domenici, J.S.T., Bruinsma, T.J., Warmoes, M.O., Bernards, R., et al. (2006). Converting a breast cancer microarray signature into a high-throughput diagnostic test. BMC Genomics *7*, 278.

Govindan, R., Ding, L., Griffith, M., Subramanian, J., Dees, N.D., Kanchi, K.L., Maher, C.A., Fulton, R., Fulton, L., Wallis, J., et al. (2012). Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell *150*, 1121–1134.

Gril, B., Palmieri, D., Qian, Y., Anwar, T., Liewehr, D.J., Steinberg, S.M., Andreu, Z., Masana, D., Fernandez, P., Steeg, P.S., et al. (2013). Pazopanib inhibits the activation of PDGFRbeta-expressing astrocytes in the brain metastatic microenvironment of breast cancer cells. Am J Pathol *182*, 2368–2379.

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L.B., Tubio, J.M.C., Papaemmanuil, E., Brewer, D.S., Kallio, H.M.L., Högnäs, G., Annala, M., et al. (2015). The evolutionary history of lethal metastatic prostate cancer. Nature *520*, 353–357.

Haque, R., Ahmed, S.A., Inzhakova, G., Shi, J., Avila, C., Polikoff, J., Bernstein, L., Enger, S.M., and Press, M.F. (2012). Impact of Breast Cancer Subtypes and Treatment on Survival: An Analysis Spanning Two Decades. Cancer Epidemiol. Biomarkers Prev. *21*, 1848–1855.

Harrell, J.C., Prat, A., Parker, J.S., Fan, C., He, X., Carey, L., Anders, C., Ewend, M., and Perou, C.M. (2012). Genomic analysis identifies unique signatures predictive of brain, lung, and liver relapse. Breast Cancer Res Treat *132*, 523–535.

Hoadley, K.A., Weigman, V.J., Fan, C., Sawyer, L.R., He, X., Troester, M.A., Sartor, C.I., Rieger-House, T., Bernard, P.S., Carey, L.A., et al. (2007). EGFR associated expression profiles vary with breast tumor subtype. BMC Genomics *8*, 258.

Hoadley, K.A., Siegel, M.B., Kanchi, K.L., Miller, C.A., Ding, L., Zhao, W., He, X., Parker, J.S., Wendl, M.C., Fulton, R.S., et al. (2016). Tumor evolution in two patients with basal-like breast cancer a retrospective genomics study of multiple metastases. PLOS Med.

Hoon, M.J.L. de, Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. Bioinformatics *20*, 1453–1454.

Hou, J.P., and Ma, J. (2014). DawnRank: discovering personalized driver genes in cancer. Genome Med. *6*, 56.

Hu, Z., Fan, C., Livasy, C., He, X., Oh, D.S., Ewend, M.G., Carey, L.A., Subramanian, S., West, R., Ikpatt, F., et al. (2009). A compact VEGF signature associated with distant metastases and poor outcomes. BMC Med. *7*, 9.

Huang, H., Liu, Y., Yuan, M., and Marron, J.S. (2015). Statistical Significance of Clustering using Soft Thresholding. J. Comput. Graph. Stat. Jt. Publ. Am. Stat. Assoc. Inst. Math. Stat. Interface Found. N. Am. *24*, 975–993.

Iglesia, M.D., Vincent, B.G., Parker, J.S., Hoadley, K.A., Carey, L.A., Perou, C.M., and Serody, J.S. (2014). Prognostic B-cell signatures using mRNA-seq in patients with subtype-specific breast and ovarian cancer. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *20*, 3818–3829.

Ip, L.R.H., Poulogiannis, G., Viciano, F.C., Sasaki, J., Kofuji, S., Spanswick, V.J., Hochhauser, D., Hartley, J.A., Sasaki, T., and Gewinner, C.A. (2015). Loss of INPP4B causes a DNA repair defect through loss of BRCA1, ATM and ATR and can be targeted with PARP inhibitor treatment. Oncotarget *6*, 10548–10562.

Josephidou, M., Lynch, A.G., and Tavaré, S. (2015). multiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples. Nucleic Acids Res. *43*, e61.

Juric, D., Castel, P., Griffith, M., Griffith, O.L., Won, H.H., Ellis, H., Ebbesen, S.H., Ainscough, B.J., Ramu, A., Iyer, G., et al. (2015). Convergent loss of PTEN leads to clinical resistance to a PI(3)Kα inhibitor. Nature *518*, 240–244.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. *40*, D109–D114.

Kardos, J., Chai, S., Mose, L.E., Selitsky, S.R., Krishnan, B., Saito, R., Iglesia, M.D., Milowsky, M.I., Parker, J.S., Kim, W.Y., et al. (2016). Claudin-low bladder tumors are immune infiltrated and actively immune suppressed. JCI Insight *1*.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576.

Krøigård, A.B., Larsen, M.J., Lænkholm, A.-V., Knoop, A.S., Jensen, J.D., Bak, M., Mollenhauer, J., Kruse, T.A., and Thomassen, M. (2015). Clonal expansion and linear genome evolution through breast cancer progression from pre-invasive stages to asynchronous metastasis. Oncotarget *6*, 5634–5649.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–D985.

Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., Mardis, E.R., Wilson, R.K., and Ding, L. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinforma. Oxf. Engl. *28*, 311–317.

Larson, D.E., Abbott, T.E., and Wilson, R.K. (2014). Using SomaticSniper to Detect Somatic Single Nucleotide Variants. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al *15*, 15.5.1–15.5.8.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature *456*, 66–72.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. *25*, 2078–2079.

Li, S., Shen, D., Shao, J., Crowder, R., Liu, W., Prat, A., He, X., Liu, S., Hoog, J., Lu, C., et al. (2013). Endocrine-Therapy-Resistant ESR1 Variants Revealed by Genomic Characterization of Breast-Cancer-Derived Xenografts. Cell Rep. *4*.

Lin, N.U., Claus, E., Sohl, J., Razzak, A.R., Arnaout, A., and Winer, E.P. (2008). Sites of distant recurrence and clinical outcomes in patients with metastatic triple-negative breast cancer: high incidence of central nervous system metastases. Cancer *113*, 2638–2645.

Maddipati, R., and Stanger, B.Z. (2015). Pancreatic Cancer Metastases Harbor Evidence of Polyclonality. Cancer Discov. *5*, 1086–1097.

Malin, D., Strekalova, E., Petrovic, V., Deal, A.M., Al Ahmad, A., Adamo, B., Miller, C.R., Ugolkov, A., Livasy, C., Fritchie, K., et al. (2014). alphaB-Crystallin: A Novel Regulator of Breast Cancer Metastasis to the Brain. Clin Cancer Res *20*, 56–67.

Malladi, S., Macalinao, D.G., Jin, X., He, L., Basnet, H., Zou, Y., de Stanchina, E., and Massagué, J. (2016). Metastatic Latency and Immune Evasion through Autocrine Inhibition of WNT. Cell *165*, 45–60.

Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., Koboldt, D.C., Fulton, R.S., Delehaunty, K.D., McGrath, S.D., et al. (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. N. Engl. J. Med. *361*, 1058–1066.

McCreery, M.Q., Halliwill, K.D., Chin, D., Delrosario, R., Hirst, G., Vuong, P., Jen, K.-Y., Hewinson, J., Adams, D.J., and Balmain, A. (2015). Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers. Nat. Med. *21*, 1514–1520.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Meric-Bernstam, F., Frampton, G.M., Ferrer-Lozano, J., Yelensky, R., Pérez-Fidalgo, J.A., Wang, Y., Palmer, G.A., Ross, J.S., Miller, V.A., Su, X., et al. (2014). Concordance of genomic alterations between primary and recurrent breast cancer. Mol. Cancer Ther. *13*, 1382–1389.

Miller, C.A., White, B.S., Dees, N.D., Griffith, M., Welch, J.S., Griffith, O.L., Vij, R., Tomasson, M.H., Graubert, T.A., Walter, M.J., et al. (2014). SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS Comput. Biol. *10*, e1003665.

Miller, C.A., Gindin, Y., Lu, C., Griffith, O.L., Griffith, M., Shen, D., Hoog, J., Li, T., Larson, D.E., Watson, M., et al. (2016). Aromatase inhibition remodels the clonal architecture of estrogen-receptor-positive breast cancers. Nat. Commun. *7*, 12498.

Minn, A.J., Gupta, G.P., Siegel, P.M., Bos, P.D., Shu, W., Giri, D.D., Viale, A., Olshen, A.B., Gerald, W.L., and Massague, J. (2005). Genes that mediate breast cancer metastasis to lung. Nature *436*, 518–524.

Moelans, C.B., van der Groep, P., Hoefnagel, L.D.C., van de Vijver, M.J., Wesseling, P., Wesseling, J., van der Wall, E., and van Diest, P.J. (2014). Genomic evolution from primary breast carcinoma to distant metastasis: Few copy number changes of breast cancer related genes. Cancer Lett. *344*, 138–146.

Mose, L.E., Wilkerson, M.D., Hayes, D.N., Perou, C.M., and Parker, J.S. (2014). ABRA: improved coding indel detection via assembly-based realignment. Bioinforma. Oxf. Engl. *30*, 2813–2815.

Mose, L.E., Selitsky, S.R., Bixby, L.M., Marron, D.L., Iglesia, M.D., Serody, J.S., Perou, C.M., Vincent, B.G., and Parker, J.S. (2016). Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with V'DJer. Bioinforma. Oxf. Engl.

Murtaza, M., Dawson, S.-J., Pogrebniak, K., Rueda, O.M., Provenzano, E., Grant, J., Chin, S.-F., Tsui, D.W.Y., Marass, F., Gale, D., et al. (2015). Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. Nat. Commun. *6*, 8760.

Natesh, K., Bhosale, D., Desai, A., Chandrika, G., Pujari, R., Jagtap, J., Chugh, A., Ranade, D., and Shastry, P. (2015). Oncostatin-M differentially regulates mesenchymal and proneural signature genes in gliomas via STAT3 signaling. Neoplasia N. Y. N *17*, 225–237.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Nazarov, V.I., Pogorelyy, M.V., Komech, E.A., Zvyagin, I.V., Bolotin, D.A., Shugay, M., Chudakov, D.M., Lebedev, Y.B., and Mamedov, I.Z. (2015). tcR: an R package for T cell receptor repertoire advanced data analysis. BMC Bioinformatics *16*, 175.

Neman, J., Termini, J., Wilczynski, S., Vaidehi, N., Choy, C., Kowolik, C.M., Li, H., Hambrecht, A.C., Roberts, E., and Jandial, R. (2014). Human breast cancer metastases to the brain display GABAergic properties in the neural niche. Proc Natl Acad Sci U A *111*, 984–989.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47–54.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T., et al. (2004). A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. N. Engl. J. Med. *351*, 2817–2826.

Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *27*, 1160–1167.

Pereira, B., Chin, S.-F., Rueda, O.M., Vollan, H.-K.M., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nat. Commun. *7*, 11479.

Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752.

Prat, A., Parker, J.S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J.I., He, X., and Perou, C.M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. BCR *12*, R68.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1997). GeneCards: integrating information about genes, proteins and diseases. Trends Genet. TIG *13*, 163.

Romagnoli, M., Mineva, N.D., Polmear, M., Conrad, C., Srinivasan, S., Loussouarn, D., Barille-Nion, S., Georgakoudi, I., Dagg, A., McDermott, E.W., et al. (2014). ADAM8 expression in invasive breast cancer promotes tumor dissemination and metastasis. EMBO Mol Med *6*, 278–294.

RStudio Team (2015). RStudio: Integrated Development Environment for R (Boston, MA: RStudio, Inc.).

Saldanha, A.J. (2004). Java Treeview—extensible visualization of microarray data. Bioinformatics *20*, 3246–3248.

Saunders, C.T., Wong, W.S.W., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinforma. Oxf. Engl. *28*, 1811–1817.

Schroeder, M.P., Rubio-Perez, C., Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2014). OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. Bioinforma. Oxf. Engl. *30*, i549–i555.

Schwarz, L.J., Fox, E.M., Balko, J.M., Garrett, J.T., Kuba, M.G., Estrada, M.V., González-Angulo, A.M., Mills, G.B., Red-Brewer, M., Mayer, I.A., et al. (2014). LYN-activating mutations mediate antiestrogen resistance in estrogen receptor-positive breast cancer. J. Clin. Invest. *124*, 5490–5502.

Schwarz, R.F., Ng, C.K.Y., Cooke, S.L., Newman, S., Temple, J., Piskorz, A.M., Gale, D., Sayal, K., Murtaza, M., Baldwin, P.J., et al. (2015). Spatial and Temporal Heterogeneity in High-Grade Serous Ovarian Cancer: A Phylogenetic Analysis. PLOS Med *12*, e1001789.

Sevenich, L., Bowman, R.L., Mason, S.D., Quail, D.F., Rapaport, F., Elie, B.T., Brogi, E., Brastianos, P.K., Hahn, W.C., Holsinger, L.J., et al. (2014). Analysis of tumour- and stroma-supplied proteolytic networks reveals a brain-metastasis-promoting role for cathepsin S. Nat Cell Biol *16*, 876–888.

Shah, S.P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., et al. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature *486*, 395–399.

Shain, A.H., Yeh, I., Kovalyshyn, I., Sriharan, A., Talevich, E., Gagnon, A., Dummer, R., North, J., Pincus, L., Ruben, B., et al. (2015). The Genetic Evolution of Melanoma from Precursor Lesions. N. Engl. J. Med. *373*, 1926–1936.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Shrestha, Y., Schafer, E.J., Boehm, J.S., Thomas, S.R., He, F., Du, J., Wang, S., Barretina, J., Weir, B.A., Zhao, J.J., et al. (2012). PAK1 is a breast cancer oncogene that coordinately activates MAPK and MET signaling. Oncogene *31*, 3397–3408.

Sihto, H., Lundin, J., Lundin, M., Lehtimäki, T., Ristimäki, A., Holli, K., Sailas, L., Kataja, V., Turpeenniemi-Hujanen, T., Isola, J., et al. (2011). Breast cancer biological subtypes and protein expression predict for the preferential distant metastasis sites: a nationwide cohort study. Breast Cancer Res. BCR *13*, R87.

Silva, G.O., He, X., Parker, J.S., Gatza, M.L., Carey, L.A., Hou, J.P., Moulder, S.L., Marcom, P.K., Ma, J., Rosen, J.M., et al. (2015). Cross-species DNA copy number analyses identifies multiple 1q21-q23 subtype-specific driver genes for breast cancer. Breast Cancer Res. Treat. *152*, 347–356.

Smid, M., Wang, Y., Zhang, Y., Sieuwerts, A.M., Yu, J., Klijn, J.G.M., Foekens, J.A., and Martens, J.W.M. (2008). Subtypes of breast cancer show preferential site of relapse. Cancer Res. *68*, 3108–3114.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

Tischler, G., and Leonard, S. (2014). biobambam: tools for read pair collation based algorithms on BAM files. Source Code Biol. Med. *9*, 13.

Troester, M.A., Lee, M.H., Carter, M., Fan, C., Cowan, D.W., Perez, E.R., Pirone, J.R., Perou, C.M., Jerry, D.J., and Schneider, S.S. (2009). Activation of Host Wound Responses in Breast Cancer Microenvironment. Clin. Cancer Res. *15*, 7020–7028.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U A *98*, 5116–5121.

Valiente, M., Obenauf, A.C., Jin, X., Chen, Q., Zhang, X.H., Lee, D.J., Chaft, J.E., Kris, M.G., Huse, J.T., Brogi, E., et al. (2014). Serpins promote cancer cell survival and vascular co-option in brain metastasis. Cell *156*, 1002–1016.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. *38*, e178.

Wang, L., Cossette, S.M., Rarick, K.R., Gershan, J., Dwinell, M.B., Harder, D.R., and Ramchandran, R. (2013). Astrocytes directly influence tumor cell invasion and metastasis in vivo. PLoS One *8*, e80933.

Weigelt, B., Glas, A.M., Wessels, L.F.A., Witteveen, A.T., Peterse, J.L., and van't Veer, L.J. (2003). Gene expression profiles of primary breast tumors maintained in distant metastases. Proc. Natl. Acad. Sci. U. S. A. *100*, 15901–15905.

Weigman, V.J., Chao, H.-H., Shabalin, A.A., He, X., Parker, J.S., Nordgard, S.H., Grushko, T., Huo, D., Nwachukwu, C., Nobel, A., et al. (2012). Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. Breast Cancer Res. Treat. *133*, 865–880.

Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinforma. Oxf. Engl. *26*, 1572–1573.

Wilkerson, M.D., Cabanski, C.R., Sun, W., Hoadley, K.A., Walter, V., Mose, L.E., Troester, M.A., Hammerman, P.S., Parker, J.S., Perou, C.M., et al. (2014). Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. Nucleic Acids Res. *42*, e107.

Yates, L.R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L.B., Larsimont, D., Davies, H., et al. (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat. Med. *21*, 751–759.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinforma. Oxf. Engl. *25*, 2865–2871.

Zhang, J., Fujimoto, J., Zhang, J., Wedge, D.C., Song, X., Zhang, J., Seth, S., Chow, C.-W., Cao, Y., Gumbs, C., et al. (2014). Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science *346*, 256–259.

Zhang, M., Tsimelzon, A., Chang, C.-H., Fan, C., Wolff, A., Perou, C.M., Hilsenbeck, S.G., and Rosen, J.M. (2015). Intratumoral Heterogeneity in a Trp53-Null Mouse Model of Human Breast Cancer. Cancer Discov. *5*, 520–533.

Zhang, X.H.-F., Wang, Q., Gerald, W., Hudis, C.A., Norton, L., Smid, M., Foekens, J.A., and Massagué, J. (2009). Latent bone metastasis in breast cancer tied to Src-dependent survival signals. Cancer Cell *16*, 67–78.

Zhao, W., He, X., Hoadley, K.A., Parker, J.S., Hayes, D.N., and Perou, C.M. (2014). Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics *15*, 419.