COMPREHENSIVE CHARACTERIZATION OF DNA COPY NUMBER ALTERATIONS IN MOUSE AND HUMAN BREAST TUMORS

Grace O. Silva

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 Bioinformatics and Computational Biology

Chapel Hill
2015

Approved by

Charles M. Perou

J.S. Marron

Joel S. Parker

Ivan Rusyn

Andrew B. Nobel

© 2015
Grace O. Silva
ALL RIGHTS RESERVED

ABSTRACT

Grace O. Silva: Comprehensive Characterization of DNA Copy Number Alterations in Mouse and Human Breast Tumors
(Under the direction of Charles M. Perou)

Breast cancer is a heterogeneous disease as evident through the diversity observed between the molecularly identified "intrinsic subtypes". These intrinsic subtypes are based upon patterns of gene expression, are predictive of relapse-free survival, overall survival, and responsiveness to treatment. Furthermore, these subtypes are in part driven by specific genomic DNA copy number alterations (CNAs), such that the identification of these intrinsic subtype-defining genetic events is of research and clinical value.

To robustly identify breast cancer "driver" genes within frequently occurring DNA CNAs, we implemented multiple integrative strategies using genomic data from various human breast tumors and genetically engineered mouse (GEM) mammary models. One strategy, a cross-species conservation based method, identified "conserved genes" that are the subtype-specific DNA copy number altered genes found in both human breast tumors and GEM mammary tumors. Another strategy, incorporated gene expression signatures of oncogenic pathway activity to identify patterns of oncogenic signaling within each breast cancer subtype that correlated directly with DNA CNAs. In both strategies, additional functional data from genome-wide RNA-mediated interference screens and/or a molecular interaction network analysis were included highlighting multiple Basal-like-specific 1q21-23 amplified genes and also amplified genes unique in highly proliferative luminal breast tumors.

In addition to using CNAs as a base for identifying therapeutic targets, we demonstrated that CNAs play other important roles in the advancement of "personalized medicine". For example, when tumor DNA is used as the source DNA for genotyping, we demonstrate that CNAs should be taken into consideration as they can lead to erroneous classification of germline genotypes. We examined two separate breast cancer cohorts and observed frequent loss of heterozygosity at the *CYP2D6* locus, which is a predictive marker of tamoxifen response. As result, when tumor tissue was used to determine germline *CYP2D6* genotype, we observed departure from Hardy Weinberg equilibrium and misclassification of intermediate metabolizers (of tamoxifen) as either extensive or poor metabolizers.

In summary, my work utilized multiple genomic data types to develop novel methods of analysis and data visualization to identify driver gene(s) within regions of DNA copy number change, which can and should be used to guide personalized treatment decisions.

To my loving parents and their many years of advice – “And this too shall pass”

ACKNOWLEDGEMENTS

I would like to thank:

My advisor, Charles M. Perou

The Perou Lab

The Lineberger Bioinformatics Group

Mouse Phase 1 Unit

My committee members

My Family: Lawrence, Moji, James, Emmanuel, Kristin, Daniel and Evan Silva

Team Dynasty

PREFACE

All scientific research is a team effort, and in that spirit I wish to acknowledge the contributions of others and give details concerning my specific contributions to the work discussed in each Chapter. Chapter 2 represents work that was recently accepted for publication in Breast Cancer Research and Treatment. I was the first author on this paper and was responsible for most aspects of this work. I performed the data analyses of gene expression and DNA copy number alterations, figure creation, and contributed to the writing of the manuscript. I would like to thank the following scientists for their collaborative help on this project:

Conception and design: Joel S. Parker, Charles M. Perou

Collection and assembly of data: Xiaping He, Lisa A. Carey, Jack P. Hou, Stacey L. Moulder, Paul K. Marcom, Jian Ma

Provision of study materials or analytics tools: Victor Weigman, Andrey A. Shabalin, Joel S. Parker, Andrew Cherniack

Analysis and interpretation of data: Jian Ma, Jack P. Hou, Charles M. Perou

Manuscript writing: Michael L. Gatza, Jeffrey M. Rosen, Charles M. Perou

Chapter 3 was previously published in Nature Genetics. I provided the analyses of the copy number data used to identify copy number alterations as a function of pathway activity and contributed to the writing of the manuscript. I would like to thank the following scientists for their collaborative help on this project:

Conception and design: Michael L. Gatza, Joel S. Parker, Charles M. Perou

Conception and design: Michael L. Gatza, Joel S. Parker, Charles M. Perou

Development of methodology: Michael L. Gatza, Joel S. Parker, Chris Fan, Charles M. Perou

Acquisition of data: Michael L. Gatza

Analysis of data: Michael L. Gatza, Chris Fan

Manuscript writing: Michael L. Gatza, Charles M. Perou

Chapter 4 was previously published in the Journal of National Cancer Institute. My role included data analyses of copy number alterations and loss of heterozygosity frequency for the gene of interest. I contributed to the writing of the manuscript and figure creation. I would like to thank the following scientists for their collaborative help on this project:

Acquisition of data: Roman Yelensky, Mark J. Ratain

Analysis and interpretation of data: Matthew P. Goetz, James X. Sun, Roman Yelensky, Charles M. Perou

Manuscript writing: Matthew P. Goetz, James X. Sun, Vera J. Suman

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

ABCSG	Austrian Breast and Colorectal Cancer Study Group
aCGH	array-based Comparative Genomic Hybridization
ATAC	arimidex, tamoxifen, alone or in combination
CBS	Circular Binary Segmentation
CNA	Copy Number Alteration
ER	oestrogen receptor
FFPE	formalin-fixed paraffin-embedded
FM	Foundation Medicine
GARP	Gene Active Ranking Profile
GEM	genetically engineered mouse
H&E	hematoxylin and eosin
HER2	human epidermal growth factor receptor
HWE	Hardy Weinberg Equilibrium
LOH	loss of heterozygosity
LUMA	Luminal A
LUMB	Luminal B
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
NCCTG	North Central Cancer Treatment Group
PAM50	Subtype classification algorithm consisting of 50 genes
PR	Progesterone Receptor
RNAi	RNA-mediated interference
SAM	Significance Analysis of Microarrays

SNP	single-nucleotide polymorphism
SWITCHdna	SupWald Identification of Copy cHanges in DNA
TCGA	The Cancer Genome Atlas
TNBC	Triple Negative Breast Cancer

CHAPTER 1

INTRODUCTION

Breast cancer is a heterogeneous disease that is characterized by distinct histological forms, genetic alterations, and patient outcomes. In addition, breast cancer is the most common cancer in women living in the Unities States and is also common in women living in less developed countries across the world. In 2012, the World Health Organization reported that an estimated 1.67 million new cases of breast cancer occurred among women worldwide [1]. In 2014, the estimated new cases of breast cancer in the United States was 14% of all cancer cases, whereas the estimated deaths was 6.8% of all cancer deaths [2]. Interestingly, the NCI SEER program demonstrated that between 2002-2011 the death rates have falling yearly at an average of 1.9% [2]. This decrease in death rates attests to the improvements in screening, and patient treatment options including the advancement of hormone and HER2 targeted therapies.

Genomic studies of breast cancer have highlighted the molecular differences observed between, and even within, tumors. Genome-wide gene expression pattern analyses identified the molecular subtypes existing within breast cancer. These “intrinsic subtypes” demonstrate many genetic differences and also varying frequency of clinical features such as differences in incidence, survival (both relapse-free and overall), and responsiveness to therapies. Breast tumors are categorized into therapeutic groups using clinical-pathological markers based on the estrogen receptor (ER), progesterone receptor (PR), and the human epidermal growth factor

receptor (HER2; also known as ERBB2). The presence of ER, PR and HER2 proteins dictates the administration of specific targeted drug therapies. For example, patients with amplified and/or over-expressed HER2 are treated with trastuzumab, a monoclonal antibody targeting HER2 [3], while all ER and/or PR positive (+) patients are treated with endocrine therapy (tamoxifen or aromatase inhibitors) [4]. However, there are subgroups of HER2+ patients with worsen response to trastuzumab [5, 6] and also ER+ patients that failed to response to endocrine therapies [7, 8]. In addition, there are the poor prognoses "triple-negative (-)" breast cancer (TNBC) patients that lack the expression of ER, PR and HER2, and therefore are not candidates to receive standard endocrine or trastuzumab therapy options. This heterogeneity in responsiveness, even within more homogenous clinical groups like ER+/HER2- patients, suggested that additional genetic diversity exists that must be responsible for this behavior. Therefore, the focus of this thesis was to identify some of this genetic heterogeneity by using the DNA copy number landscape as an arena of potential causation.

DNA Copy Number Alterations as Drivers of Carcinogenesis

Numerous somatic mutations occur in all cancer cells and some are known to have important clinical implications [9]. These types of genetic mutations included base substitutions, small insertions and deletions, translocation, inversions and copy number alterations. Copy number alterations (CNAs) are biologically relevant due to gene changes that may affect gene expression levels, function, and/or sequence [10, 11]. Specifically, CNAs are imbalances that lead to an altered diploid status and result in regions of gains (amplifications) or losses (deletions) of genetic information. CNAs range in size, varying in the number of base pairs

altered and can also encompass just a handful of bases or even millions of bases encompassing hundreds of genes [10, 12].

Somatic CNAs, which are separate from germline copy number variations, are common in cancer and specific CNAs are associated with numerous cancer types [12, 13]. Frequently occurring CNAs are important in understanding the cellular defects that promote cancer and the identification of potential therapeutic strategies [11, 14]. A shining example of this was the copy number gain and therapeutic targeting of HER2 in breast cancer, which has now literally saved thousands of lives. The current and future challenge is in identifying alterations that promote cancer growth from passenger mutations in the many existing CNAs where we do not have an obvious candidate like HER2 or MYC [15]. Specifically, copy number losses could lead to deletion of tumor suppressor genes, whereas copy number gains could lead to amplification of oncogenes [10, 11]; thus the identification of these regions and their potential driver genes are of great value.

The development of microarray technology provided an exciting new resource for estimating CNAs across the genome. In our work, the two array-based technologies used to infer CNAs include oligonucleotide array-based comparative genomic hybridization (aCGH) and single nucleotide polymorphism (SNP) array platforms. Both technologies were able to infer relative copy number values at specific known locations across the genome by using the ratio from a simultaneous analysis of tumor and reference DNAs [16, 17]. However, there are caveats about using relative copy number values, as the actual DNA content of cancer cells (ploidy) may be unknown or difficult to determine, and also cancer cells can be heterogeneous comprising of subclones with and without specific CNAs [18–20]. In addition, tumor tissue can be contaminated with an unknown amount of normal cells, thus complicating estimates of copy

number change. Today, even more computationally consuming and expensive resources are currently in development that infers CNAs through single-cell array-based technologies and next-generation sequencing data. However, despite all these caveats, microarray-based (and now sequencing-based) assessments of genome-wide DNA copy number levels are feasible, robust, and yield a large amount of clinically useful data.

One benefit of current microarray based technologies includes the ability to specify unique regions of the genome with higher coverage or target an even distribution genome-wide [11, 17, 21]. Specifically, aCGH analysis can use bacterial artificial chromosomes, cDNAs, or long oligonucleotides to target specific regions of the genome [11, 21]. Genomic DNA from tumor and normal/reference are uniquely fluorescently labeled and hybridized onto the arrays [11, 17, 21]. Relative copy number values are inferred by measuring the fluorescence intensity ratio, between tumor and reference [11, 17, 21]. In addition, aCGH can detect whole chromosome aneuplodies and submicroscopic deletions and duplications [11, 21]. However, because aCGH arrays co-hybridize the tumor with the reference DNA, aCGH arrays may miss copy-neutral alterations and regions of LOH, if the LOH is a copy neutral event.

SNP arrays use short-base-pair sequences (i.e. regions of single nucleotide polymorphisms) to target genomic regions [16, 22]. Unlike aCGH arrays, SNP arrays do not require the reference/normal sample to be hybridized along with the target of interest [16, 22]. However, we ran SNP arrays on matching normal samples to create relative copy number ratios from the two SNP arrays. Relative copy number is inferred at each SNP by comparing the ratio of combined intensity signal at both alleles from a target sample against the reference [16, 22]. Additionally, SNP arrays provided genotyping information by also determining the minor-allele frequency, the relative proportion of one allele with respect to total intensity signal [16, 22].

Therefore, a benefit of SNP arrays is the ability to identify copy-neutral loss of heterozygosity events across the genome [16, 22]. However, a limitation of SNP arrays is the coverage, being restricted by the location of SNPs throughout the genome.

To ultimately identify regions of the genome where CNAs have occurred, two main additional analyses are performed on data produced from aCGH or SNP array platforms. The first stage is the normalization and the second stage is segmentation. Normalization involved probe-specific adjustments to remove intensity biases due to binding affinity differences and other artifacts [11]. Additional quality control pre-processing steps are also needed according to each individual platform-specific protocol. Segmentation incorporated statistical methods to identify the exact regions of the genome, on a patient level basis, where the relative copy number value is greater or less than the observed diploid value coming from the control/normal DNA sample. Segmentation is implemented under the observation that, within an individual sample, adjacent positions in the genome are likely to have the same underlying copy number [22, 23]. Therefore, segmentation groups the genome into regions that share the same DNA copy number, and identifies the region/location of DNA copy number changes. The two main segmentation methods selected for this research was the circular binary segmentation (CBS) tool by Olshen *et al.* 2004, and the sup-Wald identification of copy changes in DNA (SWITCHdna) tool by Weigman *et al.* 2011 [23, 24]. Both segmentation tools incorporated maximum likelihood statistics to test for significant breakpoints separating neighboring regions (i.e. change-points) and to apply their test procedures iteratively until no more changes are detected [23, 24].

Molecular Intrinsic Subtypes of Breast Cancer

Genomic studies using microarray technology have identified “intrinsic” human breast

cancer subtypes through differential gene expression, resulting in the PAM50 [25] and Claudin-low predictors [26]. These gene expression defined subtypes include the Basal-like (typically triple-negative), Claudin-low, HER2-enriched, Luminal A, Luminal B and a subgroup of normal-like samples (some tumors and all true normals). Furthermore, these breast cancer subtypes demonstrated prognostic value as patients with Basal-like, HER2-enriched and Claudin-low tumors exhibited a worse prognosis compared to patients with Luminal (in particular Luminal A) tumors [26]. Additionally, there are specific gene expression features and frequently copy number altered regions of the genome observed for each intrinsic subtype, thus highlighting the heterogeneity of breast cancers. The genomic features of the intrinsic subtypes will be described further below, with an emphasis on the DNA copy number landscape, which has been a constant focus of my thesis research.

Basal-like Subtype. The Basal-like subtype represents roughly 10-20% of all breast tumors [27]. The Basal-like subtype disproportionally affects African American women, is very prevalent among younger African American women (27%), and also demonstrates higher mortality rates [28, 29]. The majority of Basal-like tumors lack expression of the hormone receptors (ER and PR) or the amplification and/or over-expression of HER2, and therefore are commonly referred to as "triple-negative" (ER-/PR-/HER2-) [27]. Lacking the standard targets of drug therapy, these tumors often only have chemotherapy options and demonstrat high risk of recurrence and disease progression. Therefore, not surprisingly, Basal-like patients also demonstrate a worse overall and relapse free survival in the absence of systemic therapies [26, 30, 31].

These tumors were initially referred to as Basal-like due to the unique expression of basal epithelial genes including cytokertain 5, 6, and 17 [31, 32]. Additionally, these tumors are highly

proliferative, demonstrating high expression of many cell cycle regulated genes and other markers of proliferations (i.e. proliferation index and the immunohistochemical marker Ki-67) [27, 33]. The high proliferative nature of Basal-like tumors is partially due to the lack of RB1 protein function, a key regulator of the cell cycle [34]. Other important features of the Basal-like subtype include high occurrence of *TP53* and *BRCA1* mutations which, in combinations with RB1 loss, leads to high aneuploidy and high level of genomic instability [24, 27, 35]. Regions of previously observed high frequent CNAs in the Basal-like subtype include copy number losses at 4q, 5q and copy number gains at 1q, 6p, 8q and 10p (Figure 1.1a) [24, 36].

Claudin-Low Subtype. The Claudin-low subtype is the newest identified subtype, occurring in 5-7% of all breast cancers [26]. Claudin-low tumors also demonstrate poor prognosis representing an intermediate time of disease-free survival; better than the HER2-enriched and Basal-like subtypes but worse than Luminal A and normal-like patients [26, 37]. The immunohistochemical definitions of the Claudin-low subtype are consistent with the "triple-negative" characteristics of Basal-like tumors. However, Claudin-low tumors lack the hallmark basal expression features, such as high expression of basal keratins and proliferation genes.

The subtype was named "Claudin-low" based on the unique lack of expression of the claudin family of genes. Specifically, the expression of the epithelial cell-cell adhesion genes claudin 3, claudin 4, and claudin 7 are all significantly lower in these breast tumors [26, 27, 28]. In addition, Claudin-low tumors lack the E-cadherin protein, an epithelial cell interaction protein involved in tight junctions [26, 27, 28]. Recently, Claudin-low tumors have been classified as having an intense immune cell infiltrate, with high expression of immune system response genes including *CD4*, *CD79b* and *CD14* [26, 38]. Other characteristics of Claudin-low tumors include low expression of luminal cell surface markers, and enrichment of breast stem cell and epithelial-

mesenchymal transition features [26, 38]. Regions of previously observed high frequent CNAs include copy number gains at 7p, 8q, and copy number losses at 8p (Figure 1.1c) [37].

HER2-enriched Subtype. The HER2-enriched subtype is characterized by high expression of several genes on the *ERBB2* (commonly known as *HER2*) amplicon at 17q22.24, which is the region that includes *HER2* and *GRB7* [31]. *HER2* encodes for a cell surface protein with tyrosine kinase receptor activity, that activates signal transduction pathways involved in multiple cellular functions ranging from cell division, migration, adhesion, differentiation and apoptosis [39]. The subtype was named based on the observation that most samples within this subtype were clinically HER2 positive (HER2+) [38]. Clinically HER2+ tumors represent 15-20% of all tumors, however, not all clinically HER2+ tumors fall within the HER2-enriched subtype, and not all HER2-enriched subtype tumors are HER2 amplified. The clinically HER2+ tumors that fall within the HER2-enriched subtype versus the Luminal subtypes are separated predominantly by ER status [40]. Specifically, 30-40% of HER2-enriched tumors are ER positive (ER+), meanwhile the majority lack the expression of ER (i.e. ER negative, ER-) [27].

Along with the Basal-like subtype tumors, HER2-enriched patients also demonstrate a significantly shorter overall and relapse free survival prognosis [31], which is likely due to deregulation of the ERBB signaling network [39]. However, amplification and/or overexpression of HER2 are associated with benefiting from trastuzumab treatment (an antibody targeting HER2) [41], with this targeting agent showing large improvements in patient outcomes. In addition, HER2-enriched tumors show high genomic instability with frequent copy number gains at 1q, 8q, 17q, and copy number losses at 8p (Figure 1.1b) [36, 40].

Luminal Breast Cancers. Luminal breast tumors are characterized by high gene and protein expression of the luminal signature, which includes the Estrogen Receptor (*ESR1)*, the

Progesterone Receptor (*PGR*), *FOXA1* and *BCL2* [30, 40]. From an immunohistochemical stand point, Luminal tumors are generally ER+ and/or PR+ [29]. Additional profiling studies highlighted two distinct subtypes within Luminal tumors, separating into Luminal A and Luminal B [30]. In a given population, Luminal A tumors tend to occur more often then Luminal B tumors (roughly 2/3 versus 1/3 frequency). The Luminal B subtype is distinctly separate from the Luminal A tumors by exhibiting lower expression of the luminal signature and higher expression of proliferation-related genes including *MKI67*, *BIRC5*, *CCNB1* and *MBL2* [33, 42]. Highly proliferative Luminal tumors have worst prognosis and poor response to standard therapies [43]. Consequently, Luminal A tumors demonstrate a more favorable survival outcome, living longer before developing metastatic disease [30].

Luminal tumors are also heterogeneous within the mutation and copy number landscapes. *PIK3CA* and *TP53* genes are frequently mutated within both Luminal subtypes. However, Luminal A tumors demonstrate the largest set of recurrently mutated genes including *GATA3*, *CDH1*, *MAP3K1* and *FOXA1* [40]. Regions of previously observed high frequent CNAs within Luminal A tumors include copy number gains at 16p (Figure 1.1d), whereas, Luminal B tumors demonstrate copy number gains at 17q (Figure 1.1e) [36, 44]. Interestingly, both Luminal subtypes share a highly frequent copy number loss at 16q [44] (Figure 1.1d and Figure 1.1e).

Conservation Based Approach Using Genetically Engineered Mouse Models

In science, the word "conservation" can have many meanings depending on the context. For an artist, conservation might refer to the restoration, protection and/or care of cultural heritage. For a mathematician, conservation might refer to a quantity that does not change over time. Consequentially, the word "conservation" may lead to a variety of hypotheses within

science. However, the underline concept of conservation is the same throughout science, and that is the preservation of some important feature(s).

Cancer progression is an evolutionary process driven by somatic cell factors [45]. One such factor in tumor progression is the acquisition of mutations in oncogenes and tumor suppressor genes [45], which can occur in the form of single nucleotide variants (SNV), or DNA CNAs (gains or losses of pieces of chromosomal DNA). We took an evolutionary approach to define "conservation", and refer to conservation as the selection of shared breast tumor phenotypes in individuals from distinctly separate species (i.e. mice and humans). Understanding cancer development not only at the level of cells and tissue, but also cross-species can be an important feature in defining key carcinogenic events [46, 47].

Numerous genetically engineered mouse (GEM) mammary models are available that were created to mimic specific genetic properties observed in human breast cancers [48–50]. Incorporating GEM models provide an ability to study the interaction of different cell types and other physiological functions that are not represented in tissue culture cell analyses. However, certain caveats are included in murine to human comparative studies such as species-specific pathway differences and physiology differences [47]. However, the increase in available GEM mammary models and the advancement of gene expression profiling technologies provides an unique ability to group GEM mammary models together that share distinct gene expression patterns with human tumors. As a result, large combined human and murine gene expression-based studies have identified gene expression features within mouse mammary model groups that are also observed in a specific human breast cancer subtype [49, 50]. However, on the gene expression level, there is not a single mouse mammary model that completely covers the complexity of any individual breast cancer subtype [49, 50].

On a DNA copy number level, mouse mammary models are particularly interesting as these models were designed to represent specific genetic alterations found in human breast cancers. Popular targets of mouse mammary modeling include overexpression of *HER2*/*ERBB2* or *MYC,* and inactivation of *BRCA1*, *RB* or *TP53* [51–53]. Interestingly, we hypothesize that if a selected genetic alteration occurs frequently in a human breast cancer subtype, and a corresponding mouse mammary model is made with the same frequent genetic alteration, and if both then shares the same defining gene-expression features, this would be suggestive of causation of that subtype and a "conserved" genetic event. As a result, conserved alterations likely represent regions of importance in tumor progression, and the identification of these regions would be helpful to the scientific community.

Overlapping of CNAs across human and murine genomes created unique computational challenges. The first challenge was due to the inability to directly compare intensity values cross-species from different copy number array technologies [11, 54]. To circumvent this challenge, individual significance tests are applied independently within a species/platform, and then only normalized values compared. Another challenge occurred due to variability in copy number segment sizes, and the differences in genetic annotation cross-species [23, 54]. To address this challenge, we initially increased the complexity from segments to gene level, with the observation that genes are the underlining target of an alteration and therefore are a more unambiguous cross-species. Next, given that blocks of synteny between human and murine genome may span multiple copy number segments, and given that a copy number altered segment many encompass numerous genes, we used the list of overlapping syntenic regions and the genes within them as the link. Using this overlapping gene list provided a straightforward

way to remap the murine genome in human order, with the caveat that mouse to human orthologues represent roughly 85% of all human genes [55].

Research Introduction

In order to address the key topics discussed above, my thesis work highlighted the computational analyses and bioinformatics tools necessary to characterize CNAs in breast cancer. The studies in Chapter 2 and Chapter 3 utilized copy number and gene expression data to define drivers of CNAs in highly recurrent regions of copy number alterations. In addition, the studies in Chapter 2 and Chapter 3 also utilized functional data from RNAi screens to identify CNAs harboring genes essential for proliferation; thus representing novel integrated computational analyses utilizing multiple genomic data types, both quantitative and functional.

In Chapter 2, we characterized over 600 human tumors and 70 mouse mammary models of breast cancer to provide the largest human-to-mouse mammary dataset to date, which identified key human-to-mouse conserved copy number features. In addition, we demonstrated the usefulness of a new tool that can identify shared copy number features cross-species (SWITCHdna). Furthermore, we provided a resource for human to murine (and vice versa) cancer-related projects that can guide model selection during preclinical study designs. In Chapter 3, gene-expression based pathway activity was integrated with DNA copy number analyses to identify the impact of copy number regions with altered signaling pathway in tumorigenesis. Specifically, patterns of oncogenic signaling essential for cell proliferation were correlated with CNAs observed in highly proliferative luminal breast tumors.

In Chapter 4, the focus turns to the examination of another role CNAs play in the clinical setting, specifically when tumor DNA is used as the source of genomic DNA to make treatment

decisions. In this chapter, CNAs in tumors resulted in inaccurate genotype calls at a gene that is a marker of tamoxifen response. Throughout all this work, we highlighted the importance of understanding CNAs in both the laboratory and clinical setting and laid the foundation for how to integrate copy number data with other genomics data to determine subtype-specific drivers of tumorigenesis in breast cancer.

FIGURES

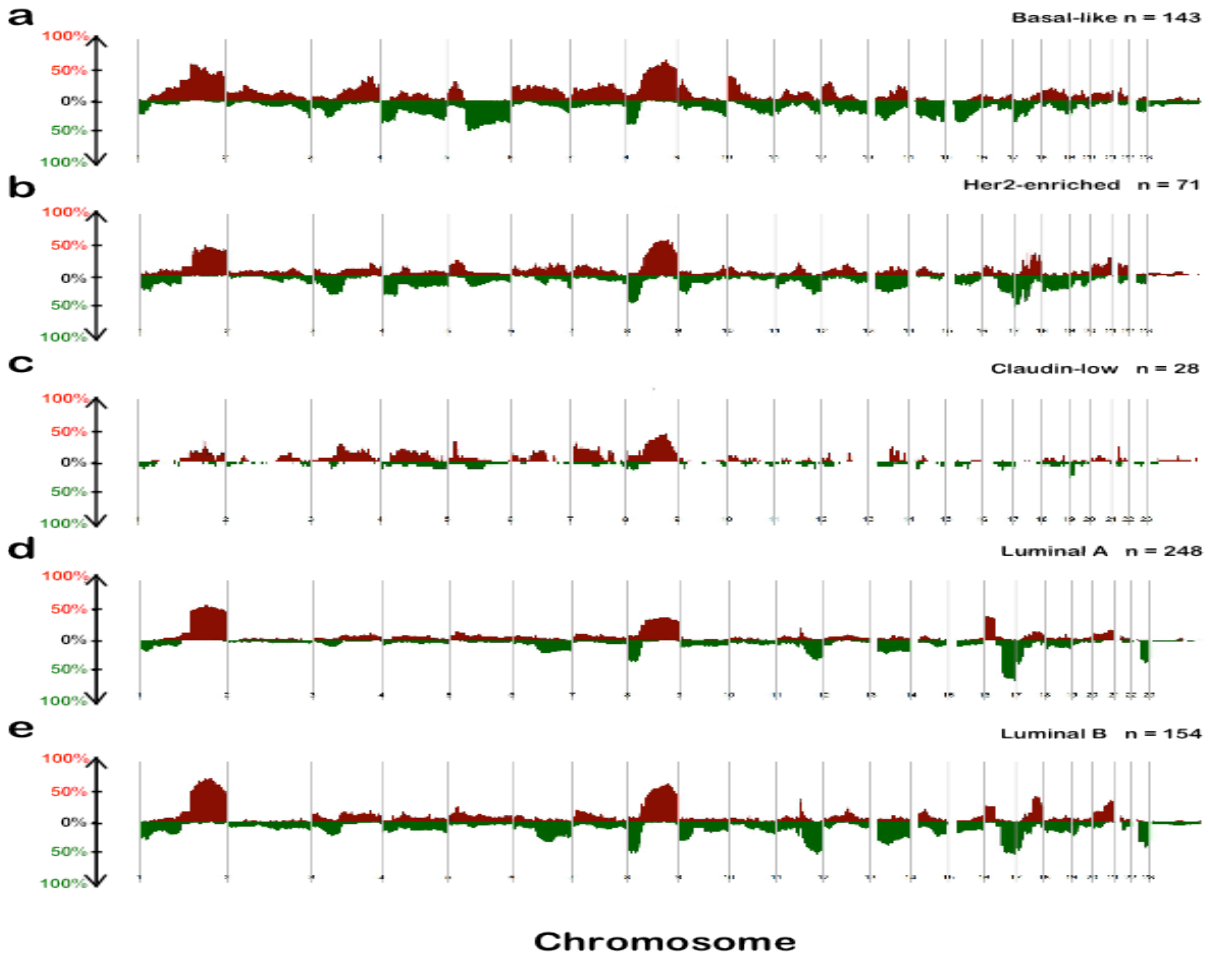

Figure 1.1 Copy number frequency plots from SWITCHplus. Segments of copy number gains are plotted above the x-axis in red and segments of copy number loss plotted below the x-axis in green. The frequency of alterations in each subtype is indicated on the y-axis from 0-100%. **a** Basal-like, **b** Her2-enriched, **c** Claudin-low, **d** Luminal A **e** Luminal B copy number landscapes.

REFERENCES

1. International Agency for Research on Cancer (IARC) and World Health Organization (WHO). GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence worldwide in 2012. http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx, 2014.

2. Cancer of the Breast - SEER Stat Fact Sheets [http://seer.cancer.gov/statfacts/html/breast.html]

3. Slamon, D. & Eiermann, W. Adjuvant trastuzumab in HER2-positive breast cancer. *New England Journal of Medicine* **365**, 1273–1283 (2011).

4. Mauri, D., Pavlidis, N., Polyzos, N. P. & Ioannidis, J. P. a Survival with aromatase inhibitors and inactivators versus standard hormonal therapy in advanced breast cancer: meta-analysis. *Journal of the National Cancer Institute* **98**, 1285–91 (2006).

5. Scaltriti, M. *et al.* Expression of p95HER2, a truncated form of the HER2 receptor, and response to anti-HER2 therapies in breast cancer. *Journal of the National Cancer Institute* **99**, 628–38 (2007).

6. Esteva, F. J. *et al.* PTEN, PIK3CA, p-AKT, and p-p70S6K status: association with trastuzumab response and survival in patients with HER2-positive metastatic breast cancer. *The American Journal of Pathology* **177**, 1647–56 (2010).

7. Berry, D. & Cronin, K. Effect of screening and adjuvant therapy on mortality from breast cancer. *New England Journal of Medicine* **353**, 1784–92(2005)

8. Berry, D. & Cirrincione, C. Estrogen-receptor status and outcomes of modern chemotherapy for patients with node-positive breast cancer. *Jama* **295**, 1658–1667 (2006).

9. Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353–60 (2012).

10. Albertson, D., Collins, C., McCormick, F. & Gray, J. Chromosome aberrations in solid tumors. *Nature Genetics* **34**, 369–376 (2003).

11. Pinkel, D. & Albertson, D. G. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics* **37 Suppl**, S11–7 (2005).

12. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature Genetics* **45**, 1134–1140 (2013).

13. Hoadley, K. a. *et al.* Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell* **158**, 929–944 (2014).

14. Incorvati, J. a, Shah, S., Mu, Y. & Lu, J. Targeted therapy for HER2 positive breast cancer. *Journal of Hematology & Oncology* **6**, 38 (2013).

15. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–8 (2007).

16. Zhao, X., Li, C., Paez, J., Chin, K. & Jänne, P. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Research* **10372**, 3060–3071 (2004).

17. Pinkel, D., Segraves, R., Sudar, D. & Clark, S. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20**, 207–211 (1998).

18. Storchova, Z. & Pellman, D. From polyploidy to aneuploidy, genome instability and cancer. *Nature Reviews. Molecular Cell Biology* **5**, 45–54 (2004).

19. Navin, N., Krasnitz, A., Rodgers, L. & Cook, K. Inferring tumor progression from genomic heterogeneity. *Genome Research* **20**, 68–80 (2010).doi:10.1101/gr.099622.109.

20. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–4 (2011).

21. Ylstra, B., Van den Ijssel, P., Carvalho, B., Brakenhoff, R. H. & Meijer, G. a BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Research* **34**, 445–50 (2006).

22. Zhang, N. DNA Copy Number Profiling in Normal and Tumor Genomes. *Frontiers in Computational and Systems Biology* **1**, (2010).

23. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–72 (2004).

24. Weigman, V. J. *et al.* Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. *Breast Cancer Res Treat.* 1–16 (2011)

25. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology* **27**, 1160–7 (2009).

26. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research* **12**, R68 (2010).

27. Perou, C. Molecular stratification of triple-negative breast cancers. *The Oncologist* **15**, 39–48 (2011).
28. Millikan, R., Newman, B. & Tse, C. Epidemiology of basal-like breast cancer. *Breast Cancer Res Treat.* **109**, 123–139 (2008).

29. Carey, L., Perou, C. & Livasy, C. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *Jama* **295**, (2006).

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

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

32. Perou, C., Sorlie, T. & Eisen, M. Molecular portraits of human breast tumours. *Nature* **533**, 747–752 (2000).

33. Cheang, M. C. U. *et al.* Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *Journal of the National Cancer Institute* **101**, 736–50 (2009).

34. Herschkowitz, J. I., He, X., Fan, C. & Perou, C. M. The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas. *Breast Cancer Research* **10**, R75 (2008).

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

36. Bergamaschi, A. & Kim, Y. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes, Chromosomes and Cancer* **1040**, 1033–1040 (2006).

37. Sabatier, R. *et al.* Claudin-low breast cancers: clinical, pathological, molecular and prognostic characterization. *Molecular Cancer* **13**, 228 (2014).

38. Prat, A. & Perou, C. M. Deconstructing the molecular portraits of breast cancer. *Molecular Oncology* **5**, 5–23 (2011).

39. Yarden, Y. & Sliwkowski, M. Untangling the ErbB signalling network. *Nature Reviews Molecular Cell Biology* **2**, (2001).

40. Cancer, T. & Atlas, G. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

41. Yeon, C. & Pegram, M. Anti-erbB-2 antibody trastuzumab in the treatment of HER2-amplified breast cancer. *Investigational New Drugs* **23**,391–409 (2005).

42. Creighton, C. The molecular profile of luminal B breast cancer. *Biologics: Targets & Therapy* **6**,289–297 (2012).

43. Perreard, L. *et al.* Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. *Breast Cancer Research* **8**, R23 (2006).

44. Ades, F. *et al.* Luminal B breast cancer: molecular characterization, clinical management, and future perspectives. *Journal of Clinical Oncology* **32**, 2794–803 (2014).

45. Merlo, L. M. F., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nature Reviews. Cancer* **6**, 924–35 (2006).

46. Casás-Selves, M. & DeGregori, J. How cancer shapes evolution and how evolution shapes cancer. *Evolution: Education and Outreach* **4**, 624–634 (2011).

47. Richmond, A. & Su, Y. Mouse xenograft models vs GEM models for human cancer therapeutics. *Disease Models & Mechanisms* **1**, 78–82 (2008).

48. Fantozzi, A. & Christofori, G. Mouse models of breast cancer metastasis. *Breast Cancer Research* **8**, 212 (2006).

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

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

51. Hennighausen, L. Mouse models for breast cancer. *Oncogene* **19**, 966–967 (2000).

52. Dankort, D. & Maslikowski, B. Grb2 and Shc adapter proteins play distinct roles in Neu (ErbB-2)-induced mammary tumorigenesis: implications for human breast cancer. *Molecular and Cellular Biology* **21**, 1540–1551 (2001).

53. Evers, B. & Jonkers, J. Mouse models of BRCA1 and BRCA2 deficiency: past lessons, current understanding and future prospects. *Oncogene* **25**, 5885–97 (2006).

54. Mosén-Ansorena, D., Aransay, A. M. & Rodríguez-Ezpeleta, N. Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data. *BMC Bioinformatics* **13**, 192 (2012).

55. Eyre, T. A, Wright, M. W., Lush, M. J. & Bruford, E. A HCOP: a searchable database of human orthology predictions. *Briefings in Bioinformatics* **8**, 2–5 (2007).

CHAPTER 2

CROSS-SPECIES DNA COPY NUMBER ANALYSES IDENTIFIES MULTIPLE 1q21–q23 SUBTYPE-SPECIFIC DRIVERS OF BREAST CANCER

A large number of DNA copy number alterations (CNAs) exist in human breast cancers, and thus characterizing the most frequent CNAs is key to advancing therapeutics because it is likely that these regions contain breast tumor 'drivers' (i.e. cancer causal genes). This study aims to characterize the genomic landscape of breast cancer CNAs and identify potential subtype-specific drivers using a large set of human breast tumors and genetically engineered mouse (GEM) mammary tumors. Using a novel method called SWITCHplus, we identified subtype-specific DNA CNAs occurring at a 15% or greater frequency, which excluded many well-known breast cancer related drivers such as amplification of *ERBB2*, and deletions of *TP53* and *RB1*. A comparison of CNAs between mouse and human breast tumors identified regions with shared subtype-specific CNAs. Additional criteria that included gene expression-to-copy number correlation, a DawnRank network analysis, and RNA interference functional studies highlighted candidate driver genes that fulfilled these multiple criteria. Numerous regions of shared CNAs were observed between human breast tumors and GEM mammary tumor models that shared similar gene expression features. Specifically, we identified chromosome 1q21-23 as a Basal-like subtype enriched region with multiple potential driver genes including *PI4KB*, *SHC1*, and *NCSTN*. This step-wise computational approach based on a cross-species comparison is

applicable to any tumor type for which sufficient human and model system DNA copy number data exists, and in this instance, highlights that a single region of amplification may in fact harbor multiple driver genes.

INTRODUCTION

Breast cancer is a heterogeneous disease that is characterized by distinct histological forms, genetic alterations, and patient outcomes [1–6]. Consistent with these observations, differential gene expression can distinguish molecular subtypes that separates breast cancer into distinct groups including Basal-like, Claudin-low, HER2-enriched, Luminal A, and Luminal B subtypes [2–4, 7–9]. These so called "intrinsic subtypes" are predictive of relapse-free survival, overall survival and responsiveness to treatment [7–11]. Previous work highlighted numerous somatic mutations [12] and DNA CNAs [13] that are linked to specific intrinsic subtypes, suggesting that these genetic events may be causative of these subtypes. Beyond a few well-known drivers, the identification of genetic drivers present in many of these recurrent regions of DNA copy number change remains to be determined. Specifically, numerous copy number alterations are located on chromosome 1 and occur at high frequency among various cancer types including breast and liver [12, 14]. In breast cancer, copy number loss frequently occurs at 1p while copy number gains are frequent at 1q [13]. Furthermore, copy number gains at 1q often encompass the majority of the 1q arm, which include hundreds of genes.

To identify additional genetic drivers of breast cancer in common regions of amplification, we have taken a cross-species conservation approach based upon the hypothesis that important etiological events in breast tumors will occur both in human breast cancers and mouse mammary tumor models. Through combined DNA copy number analyses of human breast tumors and multiple genetically engineered mouse (GEM) mammary tumor models, we identified 662 copy number alteration (CNA) regions conserved between these two species. Our ultimate selection strategy also incorporated gene expression data, an RNAi screen, and a network analysis to focus the list to the most likely driver genes within CNAs. Furthermore,

using published functional studies, we provide new insights on the potential implications of Basal-like tumor-specific chromosome 1 drivers, some of which are therapeutically targetable.

MATERIALS AND METHODS

Breast cancer tumor datasets. For these comparative studies, two human datasets and one mouse dataset were used that contained both gene expression and DNA copy number data (Table 2.1). The two human datasets were: (1) tumors collected at the University of North Carolina at Chapel Hill and the Oslo University Hospital, Radiumhospitalet, Norway ("UNC", n=159, GSE52173), and (2) The Cancer Genome Atlas (TCGA) Project dataset [12] ("TCGA", n=485). The third dataset contained tumors from numerous mouse mammary tumor models including GEM mammary models with inactivation of *TP53*, *BRCA1*, *BRG1*, and over-expression of c*MYC*, *HER2/ERBB2/Neu*, *PyMT* and *WNT1* ("mouse", n = 73, GSE52173). The publically available level 3 segmented copy number data for the TCGA dataset was downloaded through the TCGA data portal and the published PAM50 subtype calls were used [12]. Demographic and clinical characteristics of the UNC tumors, and an extended methods section are provided in the online version of the paper.

Cross-species assessment of subtype-specific changes in genomic DNA copy number. To identify subtype-specific CNAs from segmentation data generated by the various copy number array platforms, we produced an add-on script to the SWITCHdna method of DNA copy number change point detection [13]. We created an R suite of functions called *SWITCHplus*, which can identify segments of the genome with copy number changes specific for a user determined set of tumors, thus providing a supervised method for analyzing copy number data. *SWITCHplus* is provided as a source script in R and available for download at:

https://genome.unc.edu/SWITCHplus/. Note, that we did not perform multiple hypothesis testing corrections as we chose alternative biologically based filtering criteria (Figure 2.1) based upon cross-species conservation.

Computational analysis of candidate driver genes within conserved CNAs. In order to identify putative driver alterations within regions of copy number gains or losses, we began with all the conserved CNAs with a subtype segment frequency of 15% or greater. To distinguish putative drivers from passengers, three further criteria were used. We first identified genes within a CNA that demonstrate concordance between the DNA and RNA expression. The second criterion filtered for conserved CNAs that contained genes with a breast cell line RNAi-associated phenotype as published in the Solimini *et al.* 2012 RNAi screen on Human Mammary Epithelial Cells [15]. The third criterion was to identify top ranking genes when scored using DawnRank [16]. By combining all these features together, we further decrease the false positive genes by filtering out genes without functional implications (Table 2.2).

RESULTS

Subtype-specific breast cancer copy number landscapes

In order to identify both known and novel genetic drivers of breast cancer on the DNA copy number level, we developed a multi-step and multi-platform computational strategy (Figure 2.1). This strategy is predicated on using a "cross-species" comparative genomics approach where we searched for spontaneous copy number events across two different species (human and mouse). For this study, we created a new murine genomic resource of 73 mammary tumors profiled by both gene expression and DNA copy number microarray data (GSE52173); this new

resource complements our human data set that contains 644 human breast tumors that have both gene expression and DNA copy number data (GSE52173 and https://tcga-data.nci.nih.gov/tcga).

We began by using gene expression data to identify subtypes, separately for human tumor samples and GEM mammary models. For clarity, we refer to the classification of mouse tumors as "groups" to distinguish them from human classes that are termed "subtypes". Using the PAM50 [8] algorithm and the Claudin-low predictor [9] we assigned each of the human tumor samples within the dataset to a specific intrinsic breast cancer subtype (Table 2.1). However, since there is no established expression-based classifier for mouse mammary tumors, we performed a supervised hierarchical cluster analysis of the murine mRNA expression data using the Herschkowitz *et al.* 2007 intrinsic mouse list of 866 genes. SigClust [17] analysis was used to identify 7 significant mouse groups (Figure 2.2), which were given an unique group name based on the majority mouse model contributor in that group (i.e. Myc, Neu/PyMT, Wnt1, C3Tag, Mixed, p53null-Basal, and p53null-Luminal). The "Mixed" mouse group lacked a single dominant mouse model contributor, however, this group was comprised of mouse tumors that all demonstrate the previously described Claudin-low gene expression features [18, 19], and hence forth this mouse group is referred to as "ClaudinLow".

To identify subtype-specific, and mouse group-specific regions of DNA copy number gains and/or losses we developed a new bioinformatics visualization tool called *SWITCHplus*. Applying this tool to the mouse dataset identified group-specific DNA copy number changes for each of the seven expression-defined groups (Figure 2.3). These results suggest that most mouse groups are characterized by numerous DNA copy number changes, many of which are specific to a given model/group (source data available in the online version of the paper). However, by comparing the copy number landscape between mouse groups, we also identified CNAs that

were present in multiple models (Figure 2.3), which can be considered common CNAs of murine mammary oncogenesis. Therefore, these data support the notion that common spontaneous events may occur within different GEM mammary models irrespective of the initiating genetic event (i.e. transgene). Consistent with previous work, we identified multiple GEM mammary p53null groups based on gene expression patterns [18, 19]. Interestingly, these p53null groups demonstrated not only differences in mRNA expression patterns, but also exhibited differences in the DNA copy number landscapes (Figures 2.3c and 2.3d). Additionally, we noticed that the p53null-Luminal, p53null-Basal and C3-Tag groups contained more group-specific CNAs than any of the other mouse groups (source data available in the online version of the paper); this observation is likely due to the loss of TP53 in these three groups. On average, each mouse group exhibited nearly twice the number of group-specific copy number gains versus losses.

We next analyzed the human DNA copy number landscape in the combined UNC/TCGA breast cancer dataset (Figure 2.4). Our results, not surprisingly, were consistent with previous publications [6, 12, 13]. For example, our analyses confirmed previously identified breast cancer copy number gains of 8q that is common and present irrespective of breast cancer subtype, as well as a number of subtype-specific CNAs. For instance, we again identified Basal-like-specific DNA copy number losses at 4q, 5q and gains of 10p; Luminal A-specific copy number gains at 16p; Luminal B-specific copy number gains at 17q; and a Luminal-associated (encompassing both Luminal A and Luminal B) copy number loss at 16q (Figure 2.4 and source data available online) [6, 12, 13, 20, 21]. The HER2-enriched subtype contained few subtype-specific CNAs, noting that the HER2/ERBB2 amplicon was not a HER2-enriched subtype specific copy number gain event as it also occurred in many Luminal tumors. Additionally, the Basal-like subtype contained the highest number of subtype-specific CNAs (source data available online). In

contrast to what was observed in the mouse groups, human tumors on average demonstrated more frequent subtype-specific regions of copy number loss compared to copy number gains (source data available online).

Comparisons of copy number landscapes of mouse and human breast tumors

The extent to which mouse models of breast cancer recapitulate human phenotypes has been examined at the gene expression level [18–20], as well as on the copy number level, albeit only in a much smaller subset of these data [20]. We examined sub-chromosomal events and compared human subtype-specific copy number landscape plots to mouse group-specific landscape plots and identified shared cross-species CNAs events (after re-ordering the mouse chromosomal landscape into human chromosome order). We first selected for "conserved regions", which were DNA segments/regions that were altered at high frequency ($\geq$ 15%) and in the same direction (i.e. amplified or lost) in both human and mouse copy number landscapes. Applying this selection criterion reduced the search space for potential subtype-specific drivers more than 2-fold, leaving a total of 662 conserved regions when all mouse groups and human subtypes were considered (Figure 2.5 and source data available online).

In comparison amongst subtypes, the Claudin-low subtype had the fewest number of conserved regions (and the fewest CNAs overall) (source data available online). Conversely, the Basal-like subtype contained the most conserved CNAs; however, this may be due to the fact that the Basal-like subtype also contained the most subtype-specific CNAs (source data available online). Consistent with a previous publication [20], shared Basal-like-specific and murine p53null-Basal-specific regions of DNA copy number loss was observed spanning human 4q31–q35.2 and encompassing *INPP4B*, and also spanning 14q22.1-23.1 (source data available online).

By comparing shared sub-chromosomal CNAs between the human Basal-like subtype and all mouse groups, we noted that the C3-Tag mouse group contained the most human Basal-like-specific copy number amplified regions, while the p53null-Basal mouse group contained the most human Basal-like-specific copy number loss regions (source data available online). Both of these mouse models were previously shown to have the Basal-like tumor gene expression phenotypes [18, 20], therefore, for this study, we largely focused on copy number commonalities between human Basal-like tumors and these two mouse groups.

Identification of Basal-like tumor chromosome 1 amplification driver genes

Across all breast tumors, amplification of human chromosome 1q was the most frequent copy number altered event (not depicted). However, as can be seen in Figures 2.4 and 2.5, the "shape" of the chromosome 1 amplification varies by subtype, with the subtype-enriched amplification regions being identified within this largest of human chromosome arms. Among the 662 conserved regions identified across the genome, chromosome 1 harbored 18% of all conserved CNAs (source data available online). Focusing on chromosome 1, we determined that chromosome 1q harbored more than twice the number of conserved segments when compared to the 1p arm (source data available online). Of particular note, a number of 1q amplified regions that were identified as human Basal-like-specific were also altered in the mouse C3-Tag and/or p53null groups (Figure 2.6 and Appendix 1); thus our results indicate that this region of human chromosome 1q21-23 is being repeatedly selected for in both mouse and human Basal-like breast cancers.

In order to identify the driver(s) present on chromosome 1, we next applied our filtering criteria outlined in Figure 2.1. Of the 120 chromosome 1 conserved CNAs, 79 contained at least

one gene that showed DNA to RNA concordance (Appendix 2); 25 CNAs contained at least one RNAi identified essential gene (Table 2.3), and 20 CNAs contained genes showing DNA to RNA concordance *and* a RNAi identified essential gene (Table 2.4). Interestingly, all 20 CNAs were copy number gained segments, even among the 1p CNAs (Table 2.4).

To further study the biology of the conserved chromosome 1 genes, we performed a cohort based DawnRank [16] analysis using genes from human chromosome 1. DawnRank uses gene-gene interaction networks to measure the impact of genomic alterations on the differential gene expression of downstream genes in the network. Then, DawnRank scores (as previously described [16]) the level of perturbation on the gene interaction network caused by the alteration (either amplification or deletion) of the gene of interest. We selected human chromosome 1 gene blocks with shared synteny with the mouse genome for the DawnRank analysis. There were 7 such gene blocks, totaling 1509 genes (source data available online). Using the chromosome 1 syntenic regions, we identified 44 chromosome 1 genes that represented the top 5% DawnRank scores using DNA copy number changes as the input "mutation" features along with the gene expression for each human tumor sample (Appendix 3). The 44 DawnRank genes mapped to 9 copy number gained segments, which also harbored genes with DNA to RNA concordance, or an RNAi identified essential gene (Table 2.4). Within the 9 CNAs, encompassing a total of 182 potential genes, only 3 genes met all four filtering criteria of 1) subtype-specific CNA, 2) DNA to RNA concordance, 3) a RNAi "GO" gene, and 4) a DawnRank hit: these genes were phosphatidylinositol 4-kinase (*PI4KB*), src homology 2 domain-containing (*SHC1*), and nicastrin (*NCSTN*) (Figure 2.6 and Table 2.4).

The three chromosome 1 potential driving genes span 1q21-q23 and are altered with an average segment subtype frequency of 47% (Table 2.4). Interestingly, *PI4KB* and *SHC1* span

1q21, falling less than the average Basal-like subtype segment length apart (Figure 2.6), thus suggesting that on chromosome 1q21-23 multiple target genes lie within a single amplicon. Furthermore, *SHC1* is in a subtype-specific high frequency altered segment among Basal-like tumors only (Figure 2.6 and Appendix 1), while *NCSTN* and *PI4KB* CNAs appeared across multiple subtypes, passing the significance threshold in the Basal-like and Luminal A subtypes (Appendix 2). However, *NCSTN* and *PI4KB* also passed the significance threshold for the p53null-Luminal, p53null-Basal, and C3-Tag mouse groups (Appendix 1), the last two of which are models linked to human Basal-like disease as determined in previous gene expression comparative studies [18, 19].

Notch pathway features in 1q21-23 amplified Basal-like breast cancers

Numerous studies have implicated the Notch signaling pathway in Basal-like breast and/or Triple-Negative Breast Cancers [22, 23]. Importantly, numerous studies on the functional role of *NCSTN* have already been performed [24–26]. To evaluate the effect of 1q21-23/*NCSTN* amplification, we first examined the DawnRank network space around *NCSTN* and noted that when *NCSTN* was amplified *NOTCH1*, *NOTCH2*, and *NOTCH3* were also more highly expressed (Figure 2.7). In addition, *NCSTN* is one of three components of the gamma-secretase complex, a protein complex that cleaves and activates Notch receptors. Two other gamma-secretase complex members, namely *APH1A* and *PSEN2*, were also both altered within the network (Figure 2.7), and were also higher in *NCSTN* amplified samples versus not amplified (Figure 2.8a). Also, *APH1A* and *PSEN2* are physically located on human chromosome 1q21.2 and 1q42, and are often co-amplified along with *NCSTN* (although *PSEN2* is not within a Basal-like-specific CNA). Thus, three components of the gamma-secretase complex are often co-

amplified together, and more highly expressed, and the *NCSTN*/Notch Network is perturbed in these *NCSTN* amplified tumors. Following up on these network findings, *NCSTN* amplification was also correlated with higher *NOTCH1* and *NOTCH3* mRNA levels (Supplemental Figure 4b), with this feature showing an even greater difference when examined just amongst Basal-like breast cancers (Figure 2.8c). As expected from previous work, Basal-like tumors as a whole exhibited significantly lower *LFNG* expression (i.e. a negative regulator of Notch signaling) along with significantly higher expression of *NOTCH1, NCSTN, APH1A*, *MYC,* and *HEY2* mRNAs (Figure 2.9), the latter two of which are thought to be a targets of activated Notch-pathway.

DISCUSSION

In breast cancer, there are many copy number gains and losses, a few of which like amplification of *ERBB2*, are of known clinical and biological significance. Over the years, many of these CNAs have been studied and candidate genes identified [12, 13, 27–30], but there are still many regions for which the genetic drivers remain unknown. The simultaneous analysis of DNA copy number change in both human and mouse tumors, and their corresponding gene expression patterns, provides for a biologically meaningful way to identify important regions of CNAs. The basic hypothesis being that a CNA found to spontaneously occur in two different mammalian species breast cancers is being repeatedly selected for and must therefore contain an important tumor causing gene(s).

Although many studies have identified frequent CNAs within groups of human breast tumors [13, 21], most do not functionally narrow down the candidate genes within a specific segment. In addition to the mere presence of a highly frequent CNAs being identified across

species, we took a biologically based approach to refine the list of genes within a given segment into a subset of candidate driver genes. These analyses prompted the development of a new a bioinformatics tool (*SWITCHplus*) to identify and highlight subtype-specific DNA copy number events using a visual display in a user-friendly format. Using this tool and a systematic data-mining schema that includes identifying regions that show: 1) shared DNA CNAs cross-species, 2) concordance between mRNA expression and relative DNA copy number value, 3) functional effects in a genome-wide RNAi screen, and 4) functional effects in a network analysis (i.e. DawnRank), we identified a limited number of CNAs that harbored potential breast cancer driver genes. From these analyses, we identified human chromosome 1q21-23 as a region of amplification consistently present in human and mouse Basal-like tumors, and which contains at least three potential driver genes (Figure 2.6).

The first of these three genes, *PI4KB* encodes for a lipid kinase member of the phosphoinositide signaling pathway. The phosphoinositide signaling system regulates cell migration [31–33], proliferation [31–33], and activation of this signaling pathway is observed in many aggressive tumors [33–35]. Specifically, phosphatidylinositol 4-phosphate is utilized by phosphoinositide kinases, such as PI3KCA, to signal to downstream protein kinase targets including AKT and PDK1 [33, 35, 36]. In the 2012 TCGA publication on breast cancer, it was noted that Basal-like cancers showed high activity of the PIK3CA/AKT pathway, and that these tumors tended to show few *PIK3CA* mutations, but frequent loss of *PTEN* and/or *INPP4B* (negative regulators of the pathway) and amplification of *PIK3CA* and *AKT3* (positive regulators of the pathway) [12]. Here we show yet another positive regulator of the pathway is amplified in Basal-like cancers.

SHC1 encodes for a member of the Shc family of adapter proteins. SHC1 is composed of multiple protein domains that can bind to multiple transmembrane receptors including phosphorylated insulin-like growth factor 1 receptor (IGF1R), and the platelet-derived growth factor receptor (PDGFR), thus potentially activating multiple pathways involved in cell proliferation and differentiation [37, 38]. Specifically, *SHC1* is a key signaling mediator, and can act as a scaffold between an activated receptor and downstream signaling proteins [39]. In addition, growth factor signaling through PDGFR is known to occur in many TNBC [40], and thus *SHC1* amplification may be contributing to these key signaling processes.

NCSTN encodes for a component of the gamma-secretase complex (GSC), which is a multi-protein complex that cleaves a number of transmembrane proteins to typically activate their functions [41, 42]; the GSC targets include Notch 1-4, ErBB4, CD44 and E-cadherin [24, 41, 42]. Importantly, Hu *et al.* 2002 demonstrated, in *Drosophila*, that NCSTN provides structural support and is required for GSC cleavage of Notch receptor [43]. In our data, when Basal-like tumors were examined, those with copy number gains at *NCSTN* showed 1) perturbation/activation of the Notch pathway via the DawnRank network analysis (Figure 2.7), 2) significantly higher expression of *NOTCH1* and *NOTCH3* (Figure 2.8c), and 3) high expression of other markers of the Notch pathway (Figure 2.8d). Further support for Notch-pathway importance comes from previous mouse model experiments where genetic inactivation of a negative regulator of Notch signaling (i.e. lunatic fringe) resulted in Basal-like mammary tumors [22]. Interestingly, Notch activity is also higher in Basal-like breast cancer cell lines compared with luminal breast cancer cell lines [44]. *In vitro*, by RNAi-mediated silencing of *NCSTN* in the TNBC cell-line MDA-MB-231, Filipovi *et al.* 2011 showed reduced transcription of Notch pathway targets, and a reduction in cell motility and invasion [41]. In total, these results

strongly suggest that activation of Notch-pathway signaling is occurring within Basal-like/TNBC tumors, and we now provide additional evidence for a mechanistic explanation for this *in vivo*.

Other investigators using different computational approaches have also identified this region, but identified other genes (i.e. *NIT1* and *PVRL4)* as potential drivers [45]. The observed differences in potential driver genes is mostly likely due to the "filtering criteria", where we focused on species conservation, and they focused on somatic mutation targets. It is clear that a multitude of targets and drivers are present, and that 1q21-23 is a region that is the target of selection as opposed to any single gene being the target of selection. In conclusion, our work here provides an objective analysis path for identifying potential driver genes using a cross-species computational approach, which can be applied to any tumor type for which sufficient mouse and human tumor data exist.

TABLES

Table 2.1 Copy number array sample information of Human and Mouse tumors

Subtype	Number of Samples	Total
Basal-Like	UNC: 54 TCGA: 89	143
Claudin-Low	UNC: 20 TCGA: 8	28
HER2-enriched	UNC: 16 TCGA: 55	71
Luminal A	UNC: 35 TCGA: 213	248
Luminal B	UNC: 34 TCGA: 120	154

Expression SigClust Group	Number of Samples
Wap Myc	10
Neu/PyMT	11
Wnt1	16
C3Tag	8
Mixed	6
p53null-Basal	9
p53null-Luminal	13

Table 2.2 Basal-specific pipeline segments count

Pipeline	Segment Count
Stage 1: Basal-specific segments	1511
Stage 2: Basal-specific segments with frequency at least 15%	1067
Stage 3: Basal-specific conserved segments with frequency at least 15%	429
Stage 4: Basal-specific conserved segments with RNAi screen genes & frequency at least 15%	104
Stage 5: Basal-specific conserved segments showing DNA & RNA concordance in Human samples & frequency at least 15%	341
Stage 6: Basal-specific conserved segments showing DNA & RNA concordance in Mouse Samples & frequency at least 15%	126

Table 2.3 Chromosome 1 subtype specific conserved CNAs with frequency >= 15 and genes found in RNAi screen

Chr	Start	Stop	%	Genes in Segment	CNA	Mouse group-specific CNA	Solimini *et al* 2012 GO/STOP RNAi	Subtype
1	78405135	78535584	0.20	FUBP1	GAIN	C3Tag	FUBP1	Basal
1	113655140	116683656	0.22	RSBN1, BCL2L15, AP4B1, DCLRE1B, HIPK1, OLFML3, SYT6, TSHB, TSPAN2, NGF, CASQ2, NHLH2, SLC22A15	GAIN	C3Tag	DCLRE1B, HIPK1	Basal
1	145792064	150025833	0.50	BOLA1, SV2A, SF3B4, MTMR11, OTUD7B	GAIN	C3Tag	SF3B4	LumA
1	146101240	148205520	0.57	PRKAB2, FMO5, CHD1L, BCL9, ACP6	GAIN	C3Tag, Myc	PRKAB2	Basal
1	149850351	149935164	0.59	HIST2H2BE, HIST2H2AC, HIST2H2AB, BOLA1, SV2A, SF3B4, MTMR11	GAIN	C3Tag	SF3B4	Basal
1	150929687	151773763	0.62	LASS2, ANXA9, FAM63A, PRUNE, BNIPL, CDC42SE1, MLLT11, GABPB2, SEMA6C, TNFAIP8L2, LYSMD1, SCNM1, TMOD4, VPS72, PIP5K1A, PI4KB, RFX5, PSMB4, POGZ, CGN, TUFT1, MIR554, SNX27, CELF3, MRPL9, TDRKH	GAIN	C3Tag	GABPB2, PI4KB, PSMB4, TDRKH	Basal
1	151062957	151321770	0.50	SEMA6C, TNFAIP8L2, LYSMD1, SCNM1, TMOD4, VPS72, PIP5K1A, PI4KB, RFX5	GAIN	C3Tag	PI4KB	LumA
1	151321770	151409843	0.50	PSMB4	GAIN	C3Tag	PSMB4	LumA
1	151659023	151789548	0.50	CELF3, MRPL9, TDRKH, LINGO4	GAIN	ClaudinLow, C3Tag	TDRKH	LumA
1	151880754	152292453	0.50	S100A10, S100A11, TCHH, RPTN, HRNR	GAIN	ClaudinLow, C3Tag	RPTN	LumA
1	152067728	152208144	0.62	TCHH, RPTN, HRNR	GAIN	ClaudinLow, C3Tag	RPTN	Basal
1	152233280	152643406	0.64	FLG2, CRNN, CRCT1, LCE3C, LCE3B, LCE3A	GAIN	C3Tag	LCE3C	Basal
1	152447359	152617731	0.51	CRCT1, LCE3C, LCE3B, LCE3A	GAIN	C3Tag	LCE3C	LumA
1	152661380	153346263	0.64	KPRP, LCE1F, LCE1E, LCE1C, LCE1B, LCE6A, SMCP, IVL, SPRR4, SPRR1A, SPRR3, SPRR1B, SPRR2D, SPRR2B, SPRR2E, SPRR2F, SPRR2G, LELP1, LOR, PGLYRP3, PGLYRP4	GAIN	C3Tag	PGLYRP3	Basal
1	153354347	153576396	0.64	S100A8, S100A7A, S100A6, S100A5, S100A4, S100A3	GAIN	C3Tag	S100A5	Basal
1	153576396	154012535	0.63	S100A16, S100A14, S100A13, C1orf77, SNAPIN, ILF2, NPR1, INTS3, SLC27A3, GATAD2B, DENND4B, SLC39A1, CREB3L4, JTB, JTB, RPS27	GAIN	C3Tag	NPR1	Basal

Table 2.3 Chromosome 1 subtype specific conserved CNAs with frequency >= 15 and genes found in RNAi screen (Continued)

Chr	Start	Stop	%	Subtype-Specific Genes in Segment	CNA	Mouse Group	Solimini *et al* 2012 GO/STOP RNAi	Subtype
1	154270185	154318564	0.62	ATP8B2	GAIN	C3Tag	ATP8B2	Basal
1	154807935	155175657	0.62	PMVK, PYGO2, SHC1, CKS1B, FLAD1, LENEP, ZBTB7B, DCST2, DCST1, DPM3, KRTCAP2, TRIM46, MUC1, MIR92B	GAIN	C3Tag	SHC1, TRIM46	Basal
1	155936658	156321154	0.60	SSR2, UBQLN4, RAB25, MEX3A, LMNA, SEMA4A, SLC25A44, PMF1, BGLAP, PAQR6, SMG5, TMEM79, CCT3, C1orf182	GAIN	C3Tag	CCT3, C1ORF182	Basal
1	156321154	156545720	0.59	RHBG, MEF2D, IQGAP3	GAIN	C3Tag	RHBG	Basal
1	160043165	160439014	0.57	KCNJ9, IGSF8, ATP1A2, ATP1A4, CASQ1, DCAF8, PEX19, COPA, SUMO1P3, NCSTN, NCSTN, NHLH1, VANGL2	GAIN	C3Tag, p53null_Basal	CASQ1, COPA, SUMO1P3, NCSTN, NCSTN	Basal
1	160197660	160372346	0.51	PEX19, COPA, SUMO1P3, NCSTN, NCSTN, NHLH1	GAIN	C3Tag, p53null_Basal, p53null_Luminal	COPA, SUMO1P3, NCSTN	LumA
1	160906176	163790065	0.18	F11R, USF1, ARHGAP30, PVRL4, KLHDC9, NIT1, DEDD, DEDD, UFC1, PPOX, B4GALT3, ADAMTS4, NDUFS2, FCER1G, APOA2, TOMM40L, NR1I3, NR1I3, PCP4L1, MPZ, SDHC, C1orf192, FCGR2B, FCRLA, FCRLB, DUSP12, ATF6, OLFML2B, NOS1AP, MIR556, UHMK1, UAP1, DDR2, HSD17B7, RGS4, RGS5, NUF2	GAIN	p53null_Luminal	F11R, TOMM40L, NR1I3, FCGR2B, FCRLA	Claudin
1	182988016	184909056	0.18	LAMC1, LAMC2, NMNAT2, SMG7, NCF2, ARPC5, RGL1, APOBEC4, GLT25D2, TSEN15, EDEM3	GAIN	p53null_Luminal	GLT25D2	Claudin
1	205333969	206253777	0.73	LEMD1, MIR135B, CDK18, MFSD4, ELK4, SLC45A3, NUCKS1, RAB7L1, SLC41A1, PM20D1, SLC26A9, FAM72A, AVPR1B	GAIN	ClaudinLow	CDK18	LumB

Table 2.4 Chromosome 1 conserved CNAs with frequency >= 15%, concordant with gene expression, RNAi identified essential gene, and top DawnRank score

Chr	Start	Stop	Seg Freq	CNA	Mouse Model	Solimini et al 2012 GO/STOP RNAi	Concordant DNA/RNA	Subtype	Dawn Rank	ALL 3
1	78405135	78535584	0.20	GAIN	C3Tag	FUBP1	FUBP1	Basal		
1	113655140	116683656	0.22	GAIN	C3Tag	DCLRE1B, HIPK1	RSBN1, BCL2L15, AP4B1, DCLRE1B, HIPK1, SLC22A15	Basal	NGF	
1	145792064	150025833	0.50	GAIN	C3Tag	SF3B4	BOLA1, SF3B4, MTMR11, OTUD7B	LumA		
1	146101240	148205520	0.57	GAIN	C3Tag, Myc	PRKAB2	PRKAB2, FMO5, CHD1L, BCL9, ACP6	Basal		
1	149850351	149935164	0.59	GAIN	C3Tag	SF3B4	HIST2H2AC, BOLA1, SF3B4, MTMR11	Basal		
1	150929687	151773763	0.62	GAIN	C3Tag	GABPB2, PI4KB, PSMB4, TDRKH	LASS2, FAM63A, PRUNE, BNIPL, CDC42SE1, GABPB2, SEMA6C, LYSMD1, SCNM1, TMOD4, VPS72, PIP5K1A, PI4KB, RFX5, PSMB4, POGZ, CGN, TUFT1, MIR554, SNX27, MRPL9, TDRKH	Basal	PI4KB	PI4KB
1	151062957	151321770	0.50	GAIN	C3Tag	PI4KB	SEMA6C, LYSMD1, SCNM1, TMOD4, VPS72, PIP5K1A, PI4KB, RFX5	LumA	PI4KB	PI4KB
1	151321770	151409843	0.50	GAIN	C3Tag	PSMB4	PSMB4	LumA		
1	151659023	151789548	0.50	GAIN	ClaudinLow, C3Tag	TDRKH	MRPL9, TDRKH, LINGO4	LumA		
1	151880754	152292453	0.50	GAIN	ClaudinLow, C3Tag	RPTN	S100A10, S100A11	LumA		
1	153354347	153576396	0.64	GAIN	C3Tag	S100A5	S100A8, S100A6	Basal		
1	153576396	154012535	0.63	GAIN	C3Tag	NPR1	S100A16, S100A14, S100A13, C1ORF77, SNAPIN, ILF2, INTS3, GATAD2B, DENND4B, SLC39A1, CREB3L4, JTB, RPS27	Basal	GATAD2B, SNAPIN	

Table 2.4 Chromosome 1 conserved CNAs with frequency >= 15%, concordant with gene expression, RNAi identified essential gene, and top DawnRank score (Continued)

1	154807935	155175657	0.62	GAIN	C3Tag	SHC1, TRIM46	PMVK, PYGO2, SHC1, CKS1B, FLAD1, ZBTB7B, DCST2, DCST1, DPM3, KRTCAP2, TRIM46, MUC1	Basal	SHC1	SHC1
1	155936658	156321154	0.60	GAIN	C3Tag	CCT3, C1ORF182	SSR2, UBQLN4, RAB25, MEX3A, LMNA, SEMA4A, SLC25A44, PMF1, BGLAP, PAQR6, SMG5, TMEM79, CCT3, C1ORF182	Basal	SMG5	
1	156321154	156545720	0.59	GAIN	C3Tag	RHBG	MEF2D, IQGAP3	Basal		
1	160043165	160439014	0.57	GAIN	C3Tag, p53null_Basal	CASQ1, COPA, SUMO1P3, NCSTN, NCSTN	IGSF8, DCAF8, PEX19, COPA, SUMO1P3, NCSTN, NHLH1, VANGL2	Basal	NCSTN	NCSTN
1	160197660	160372346	0.51	GAIN	C3Tag, p53null_Basal, p53null_Luminal	COPA, SUMO1P3, NCSTN, NCSTN	PEX19, COPA, SUMO1P3, NCSTN, NHLH1	LumA	NCSTN	NCSTN
1	160906176	163790065	0.18	GAIN	p53null_Luminal	F11R, TOMM40L, NR1I3, NR1I3, FCGR2B, FCRLA	F11R, USF1, PVRL4, KLHDC9, NIT1, DEDD, UFC1, PPOX, B4GALT3, ADAMTS4, NDUFS2, TOMM40L, NR1I3, PCP4L1, SDHC, C1ORF192, DUSP12, ATF6, NOS1AP, MIR556, UHMK1, UAP1, HSD17B7, NUF2	Claudin	DEDD, F11R, FCER1G, FCGR2B	F11R
1	182988016	184909056	0.18	GAIN	p53null_Luminal	GLT25D2	LAMC2, SMG7, ARPC5, TSEN15, EDEM3	Claudin		
1	205333969	206253777	0.73	GAIN	ClaudinLow	CDK18	CDK18, MFSD4, ELK4, SLC45A3, NUCKS1, RAB7L1, SLC41A1, FAM72A	LumB		

FIGURES

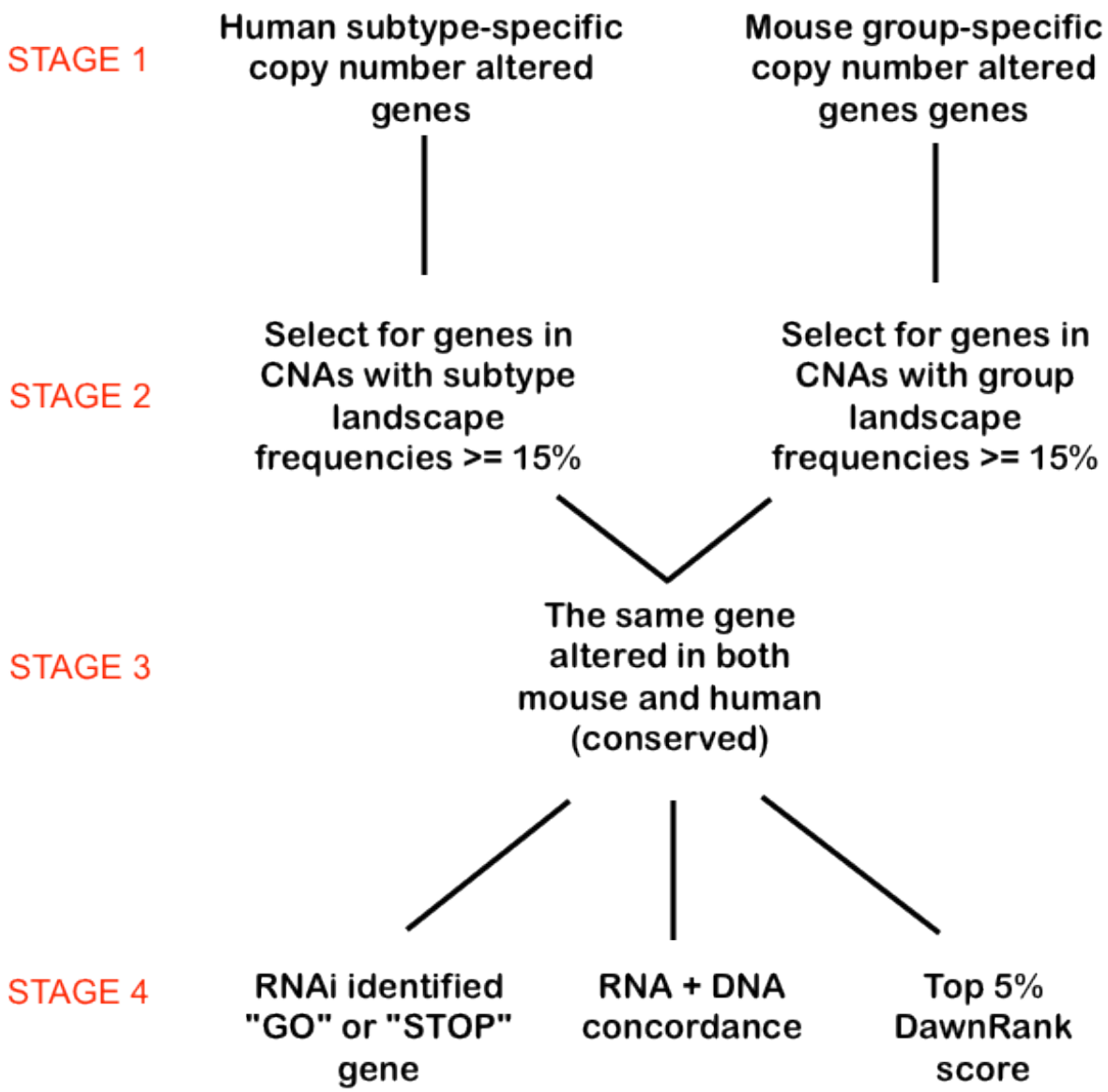

Figure 2.1 Data-analysis pipeline to identify subtype-specific CNA candidate driver genes.

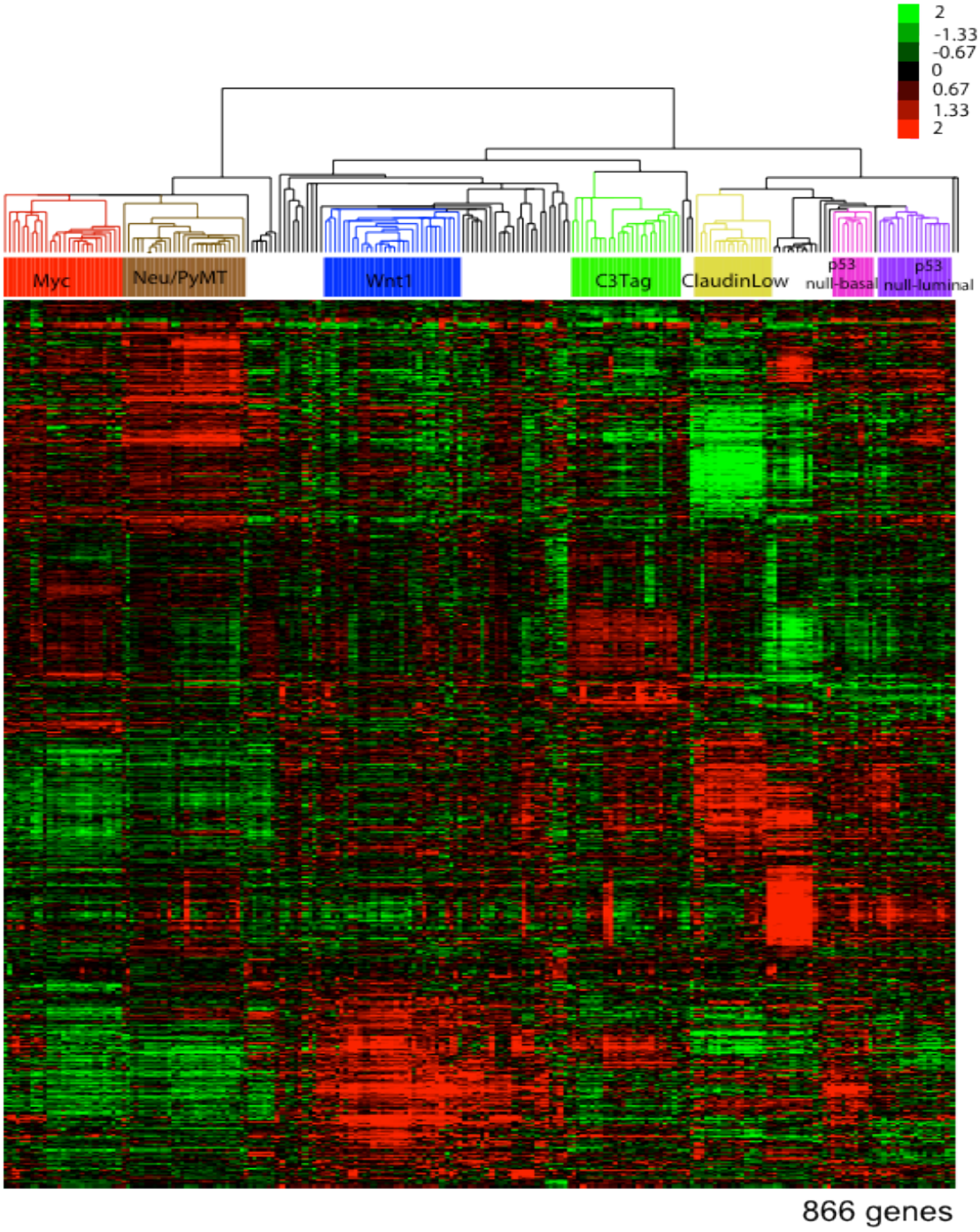

Figure 2.2 Supervised cluster of mouse gene expression data using an 866 intrinsic gene list. The cluster analysis identified 7 murine tumor subtypes, which were further used to supervise subsequent DNA copy number analyses. Each group is labeled according to the majority component mouse model within that group.

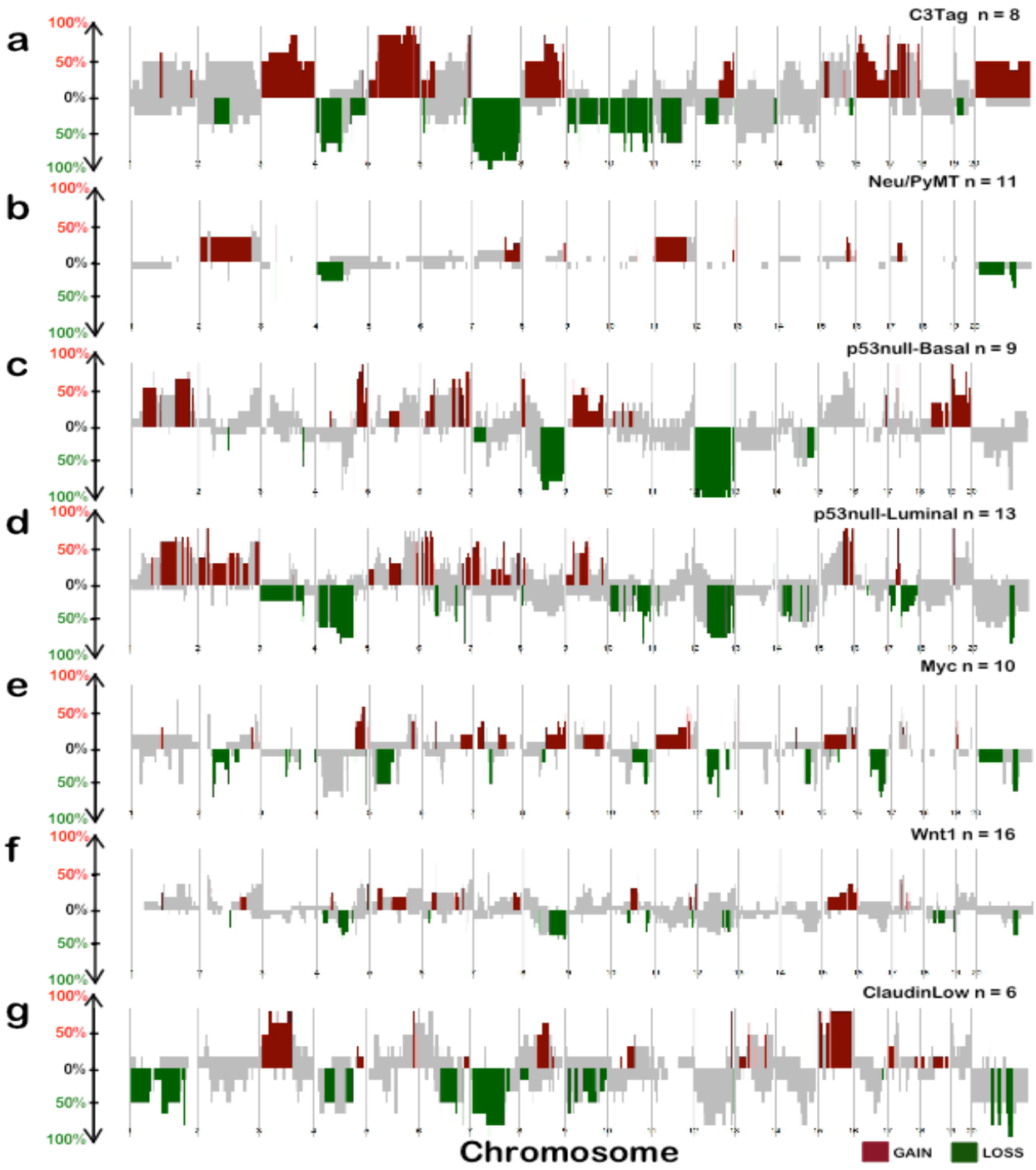

Figure 2.3 Copy number frequency plots from SWITCHplus showing mouse group-specific CNAs. Segments of group-specific copy number gains are plotted above the x-axis in red and segments of copy number loss plotted below the x-axis in green. Regions shaded gray indicate segments that are not group-specific or high frequent (greater than or equal to 15%). The frequency of alterations in each mouse group is indicated on the y-axis from 0-100%. **a** C3Tag, **b** Neu/PyMT, **c** p53null-Basal, **d** p53null-Luminal **e** Myc, **f** Wnt1, **g** ClaudinLow copy number landscapes.

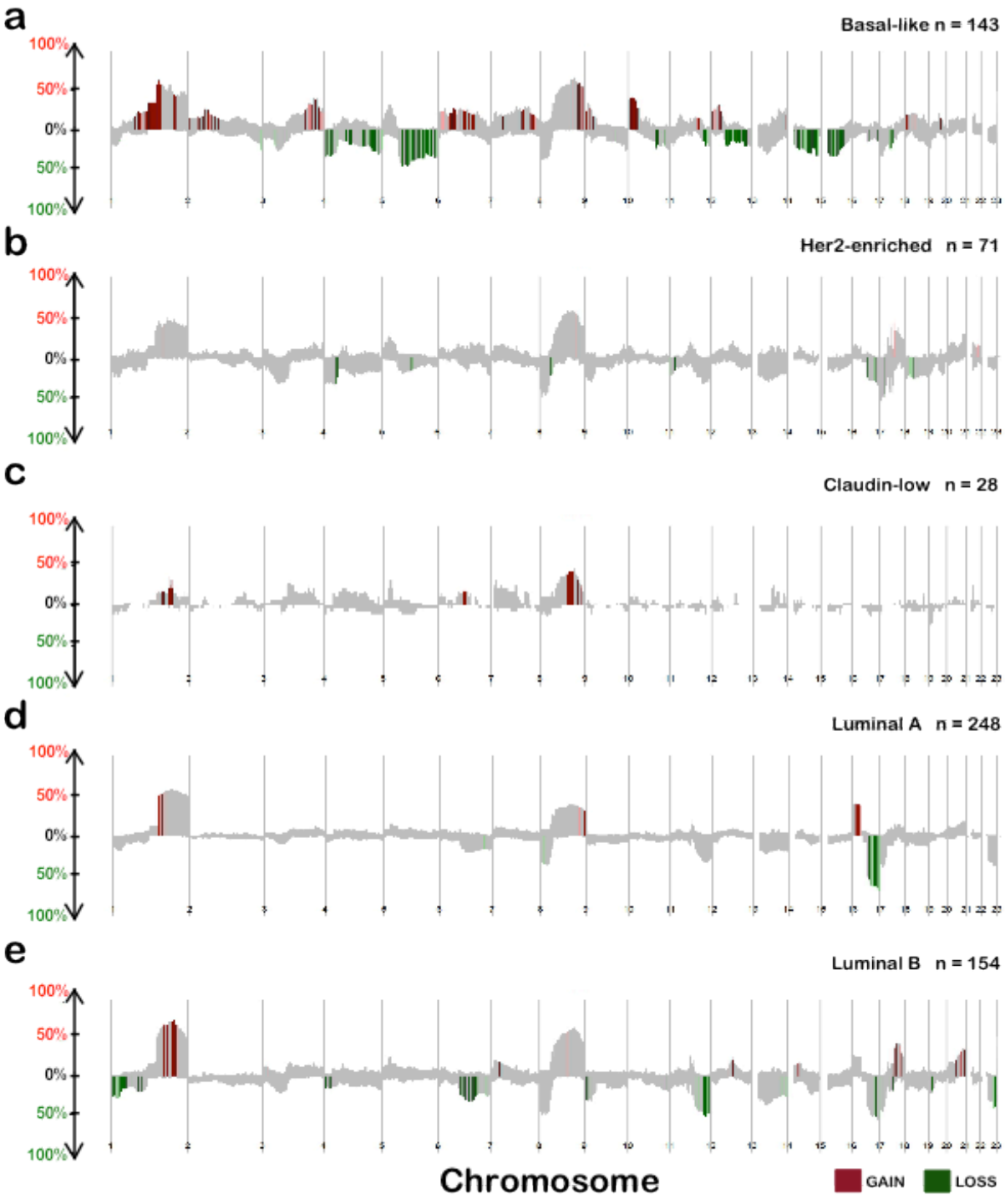

Figure 2.4 Copy number frequency plots from SWITCHplus showing human subtype-specific CNAs. Segments of subtype-specific copy number gains are plotted above the x-axis in red and segments of copy number loss plotted below the x-axis in green. Regions shaded gray indicate segments that are not subtype-specific or high frequent (greater than or equal to 15%). The frequency of alterations in each subtype is indicated on the y-axis from 0-100%. **a** Basal-like, **b** Her2-enriched, **c** Claudin-low, **d** Luminal A **e** Luminal B copy number landscapes.

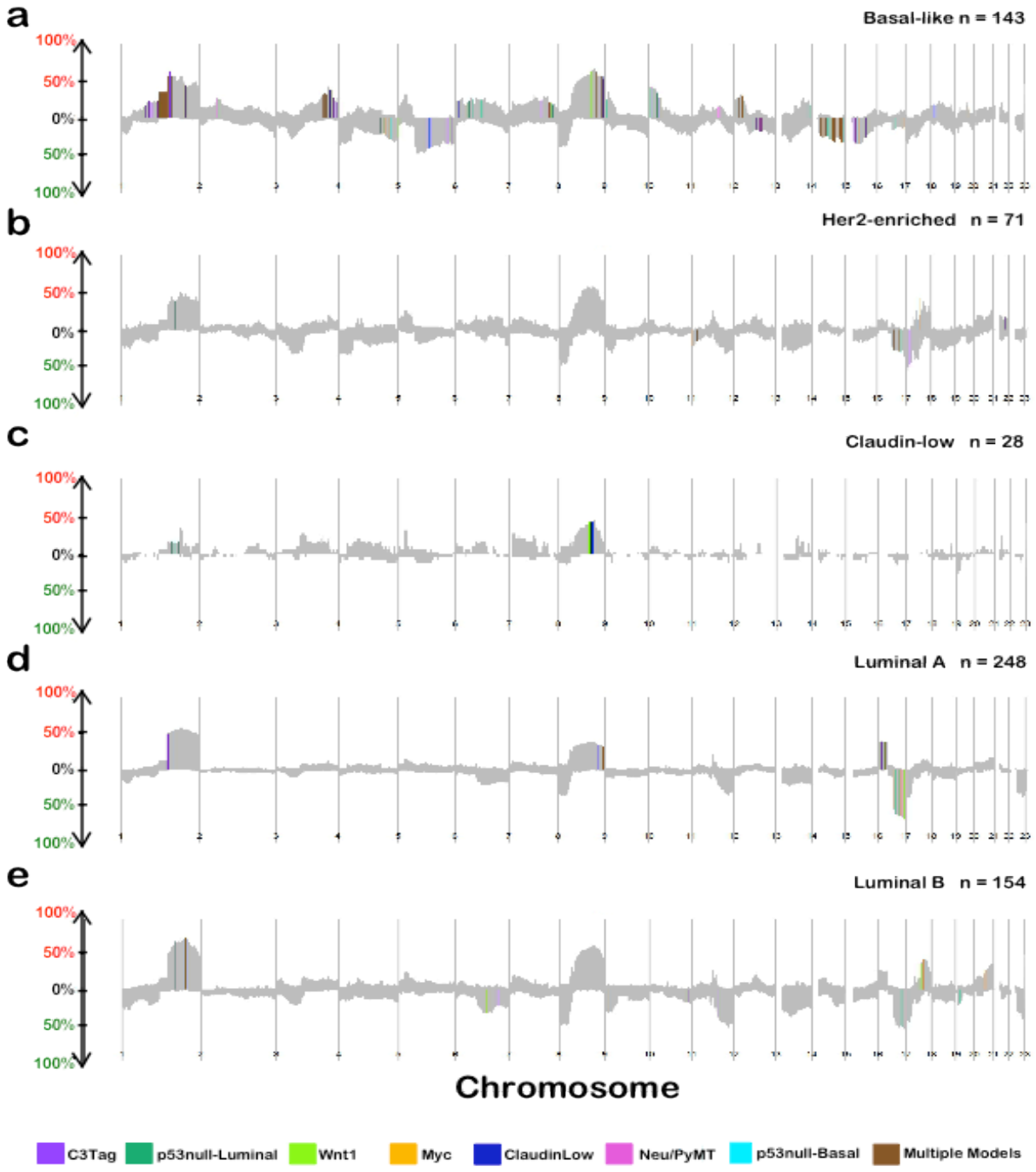

Figure 2.5 Copy number frequency plots from SWITCHplus showing conserved CNAs. Segments of copy number gains are plotted above the x-axis and segments of copy number loss plotted below the x-axis. Regions shaded gray indicate segments that are either not subtype-specific, mouse group-specific or high frequent (greater than or equal to 15%). The conserved segments are colored according to the mouse model(s) in which they appear. The frequency of alterations in each subtype is indicated on the y-axis from 0-100%. **a** Basal-like, **b** Her2-enriched, **c** Claudin-low, **d** Luminal A **e** Luminal B copy number landscapes.

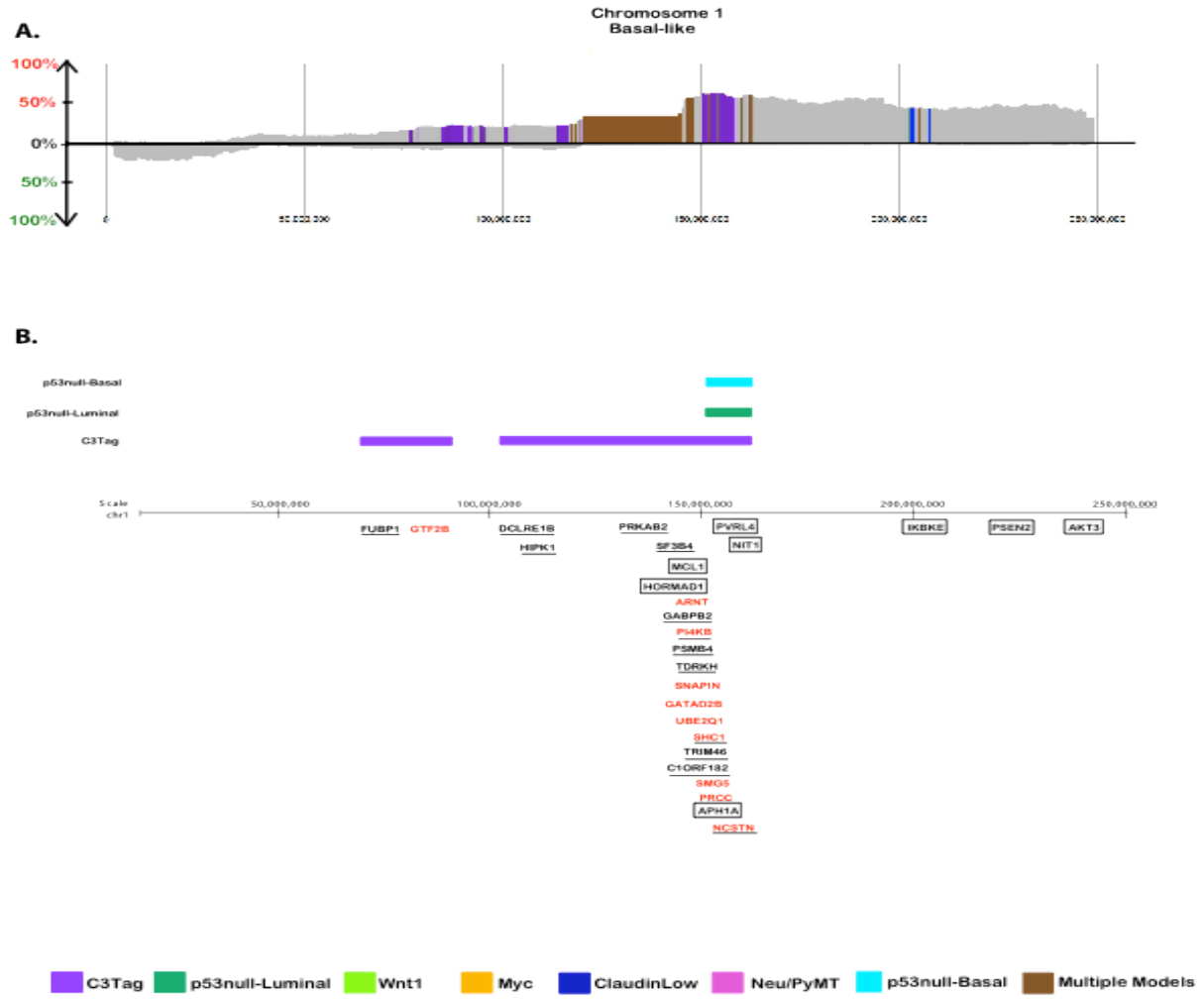

Figure 2.6 Expanded view of a chromosome 1 Basal-like conserved copy number frequency plots for SWITCHplus. Segments of copy number gains are plotted above the x-axis and segments of copy number loss plotted below the x-axis. Regions shaded gray indicate segments that are either not subtype-specific, mouse group-specific or high frequent (greater than or equal to 15%). The conserved segments are colored according to the mouse model(s) in which they appear. The frequency of alterations is indicated on the y-axis. **b** View of the genomic location of candidate chromosome 1 driver genes. Genes colored red are Basal-like subtype-specific or subtype-associated, demonstrate DNA and RNA concordance in human tumors and had a top DawnRank score; genes underlined are Basal-like subtype-specific or subtype-associated, demonstrate DNA and RNA concordance in human tumors and labeled as a growth enhancer and oncogene ("GO gene") in the Solimini *et al.* 2012 RNAi screen on Human Mammary Epithelial Cells; the remaining genes surrounded by a box are additional potential drivers. A color bar is placed above the genes conserved for a particular mouse group.

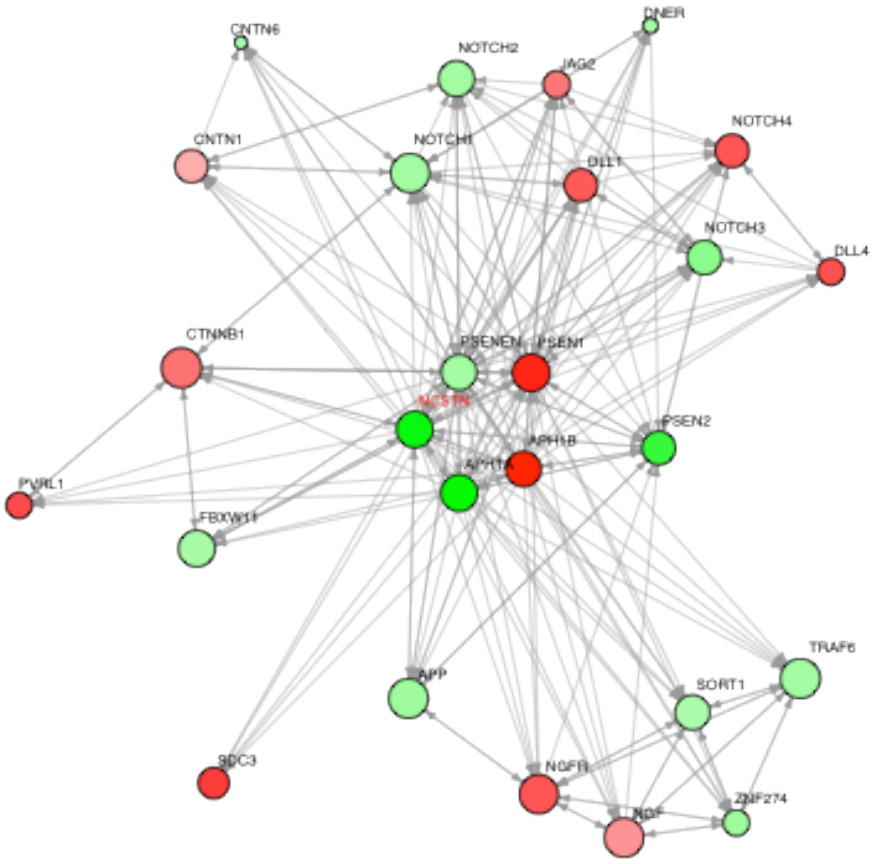

Figure 2.7 DawnRank identified *NCSTN* gene expression network

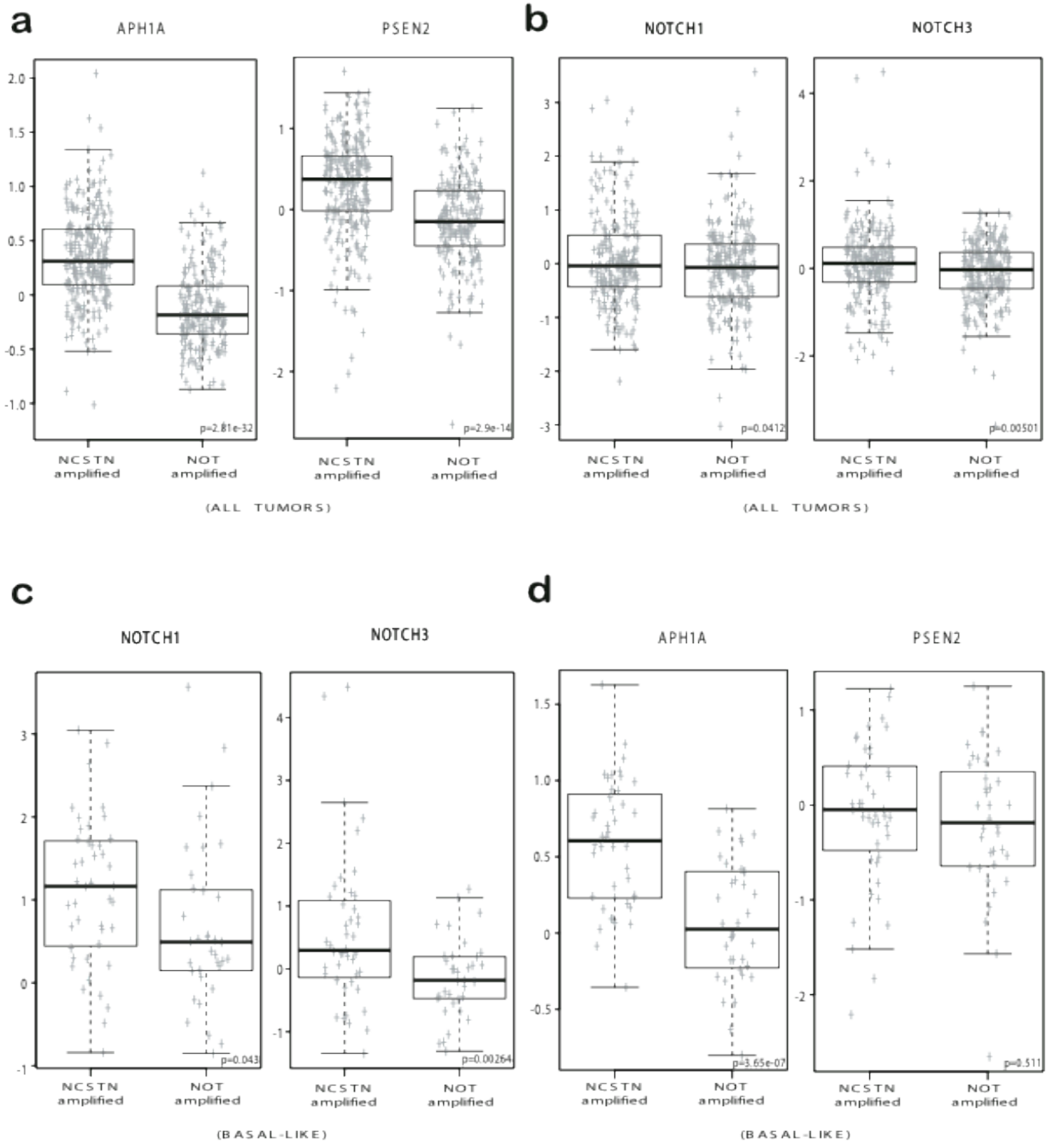

Figure 2.8 Box-and-whisker plots for expression of Notch signaling pathway targets. **a** *APH1A* and *PSEN2* mRNA expression across all tumors with *NCSTN* amplification versus all other tumors without *NCSTN* amplification. **b** *NOTCH1* and *NOTCH3* mRNA expression across all tumors with *NCSTN* amplification versus all tumors without *NCSTN* amplification. **c** *NOTCH1* and *NOTCH3* mRNA expression across all basal tumors with *NCSTN* amplification versus basal tumors without *NCSTN* amplification. **d** *APH1A* and *PSEN2* mRNA expression across all basal tumors with *NCSTN* amplification versus basal tumors without *NCSTN* amplification.

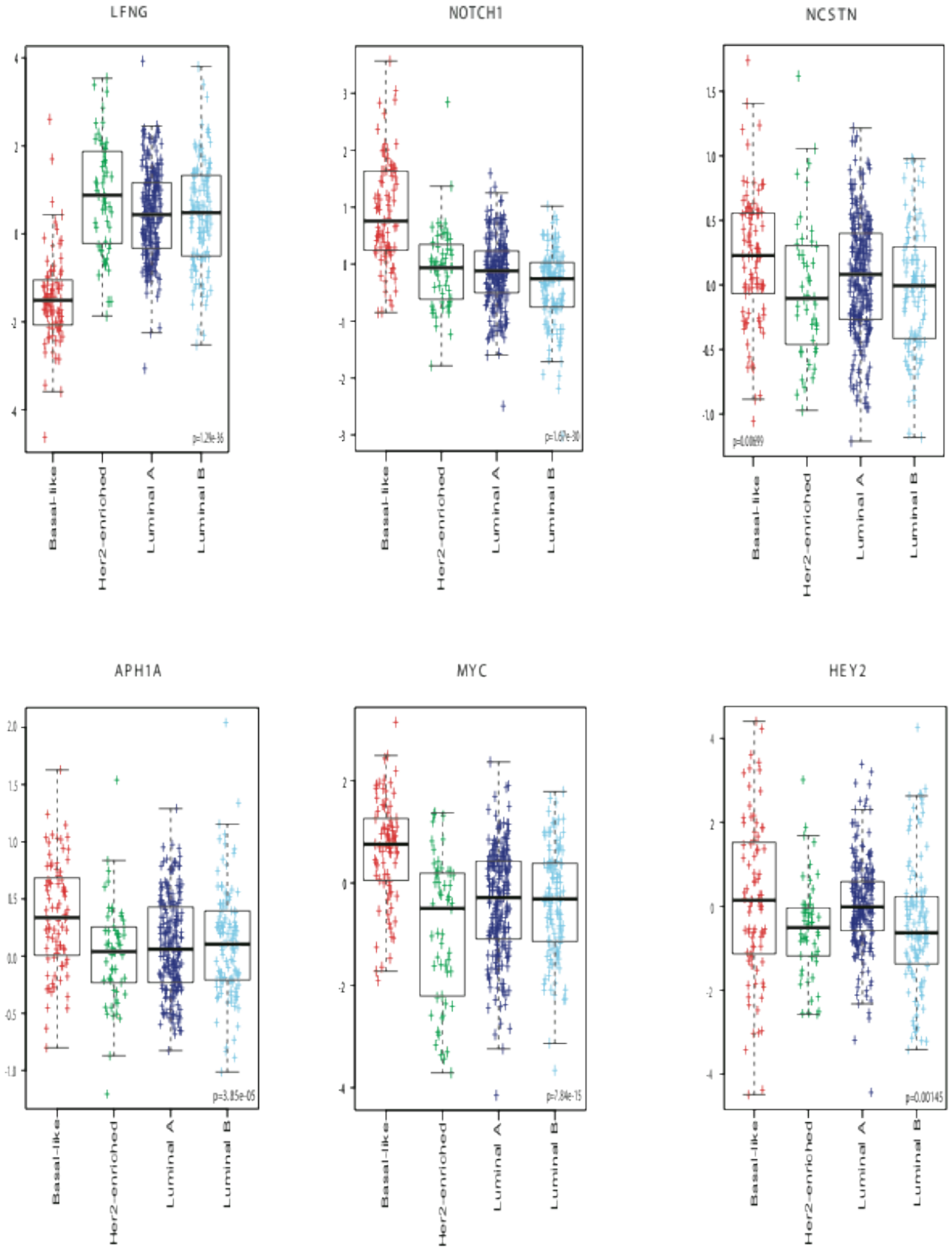

Figure 2.9 Box-and-whisker plots of the mRNA expression of *LFNG*, *NOTCH1*, *NCSTN*, *APH1A*, *MYC* and *HEY2* across breast cancer subtypes.

REFERENCES

1. Kravchenko J, Akushevich I, Seewaldt VL, et al. Breast cancer as heterogeneous disease: contributing factors and carcinogenesis mechanisms. *Breast Cancer Res Treat* **128**:483–93 (2011).

2. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. *Molecular Oncology* **5**:5–23 (2011).

3. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* **100**:8418–23 (2003).

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

5. Nordgard SH, Johansen FE, Alnaes GIG, et al. Genes harbouring susceptibility SNPs are differentially expressed in the breast cancer subtypes. *Breast Cancer Research* **9**:113 (2007).

6. Bergamaschi A, Kim Y. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene‐expression subtypes of breast cancer. *Genes, Chromosomes and Cancer* **1040**:1033–1040 (2006).

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

8. Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology* **27**:1160–7 (2009).

9. Prat A, Parker JS, Karginova O, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research* **12**:R68 (2010).

10. Hu Z, Fan C, Oh DS, et al. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* **7**:96 (2006).

11. Cheang MCU, Chia SK, Voduc D, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *Journal of the National Cancer Institute* **101**:736–50 (2009).

12. Cancer T, Atlas G Comprehensive molecular portraits of human breast tumours. *Nature* **490**:61–70 (2012).

13. Weigman VJ, Chao H-H, Shabalin AA, et al. Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. *Breast Cancer Res Treat* **113**:865–880 (2012).

14. Nishida N, Nishimura T, Ito T, Komeda T. Chromosomal instability and human hepatocarcinogenesis. *Histology and Histopathology* **18**:897–909 (2003).

15. Solimini NL, Xu Q, Mermel CH, et al. Recurrent hemizygous deletions in cancers may optimize proliferative potential. *Science* **337**:104–109 (2012).

16. Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. Genome Medicine **6**:56 (2014).

17. Liu Y, Hayes DN, Nobel A, Marron JS. Statistical Significance of Clustering for High-Dimension, Low–Sample Size Data. *Journal of the American Statistical Association* **103**:1281–1293 (2008).

18. Herschkowitz JI, Simin K, Weigman VJ, et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biology* **8**:R76 (2007).

19. Pfefferle AD, Herschkowitz JI, Usary J, et al. Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. *Genome Biology* **14**:R125 (2013).

20. Herschkowitz JI, Zhao W, Zhang M, et al. Comparative oncogenomics identifies breast tumors enriched in functional tumor-initiating cells. *Proc Natl Acad Sci U S A* **109**:2778–83 (2012).

21. Curtis C, Shah SP, Chin S-F, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**:346–52 (2012).

22. Xu K, Usary J, Kousis PC, et al. Lunatic fringe deficiency cooperates with the Met/Caveolin gene amplicon to induce basal-like breast cancer. *Cancer Cell* **21**:626–41 (2012).

23. Stoeck A, Lejnine S, Truong A, et al. Discovery of biomarkers predictive of GSI response in triple-negative breast cancer and adenoid cystic carcinoma. *Cancer Discovery* **4**:1154–1167 (2014).

24. Lombardo Y, Filipović A, Molyneux G, et al. Nicastrin regulates breast cancer stem cell properties and tumor growth in vitro and in vivo. *Proc Natl Acad Sci U S A* **109**:16558–63 (2012).

25. Murphy MP, Das P, Nyborg AC, et al. Overexpression of nicastrin increases Aβ production. *FASEB J* **17**:1138–40 (2003).

26. Sarajlić a, Filipović A, Janjić V, et al. The role of genes co-amplified with nicastrin in breast invasive carcinoma. *Breast Cancer Res Treat* **143**:393–401 (2014).

27. Shadeo A, Lam WL. Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Research* **8**:R9 (2006).

28. Taylor BS, Barretina J, Socci ND, et al. Functional copy-number alterations in cancer. *PloS one* **3**:e3179 (2008).

29. Kristensen VN, Lingjærde OC, Russnes HG, et al. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer* **14**:299–313 (2014).

30. Beroukhim R, Mermel CH, Porter D, et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**:899–905 (2010).

31. Katso R, Okkenhaug K. Cellular function of phosphoinositide 3-kinases: implications for development, immunity, homeostasis, and cancer. *Annual Rev Cell Dev Biol* **17**:615–675 (2001).

32. Cantley L. The phosphoinositide 3-kinase pathway. *Science* **296**:1655–1658 (2002).

33. Bunney TD, Katan M. Phosphoinositide signalling in cancer: beyond PI3K and PTEN. *Nature Reviews Cancer* **10**:342–52 (2010).

34. Altomare D a, Testa JR. Perturbations of the AKT signaling pathway in human cancer. *Oncogene* **24**:7455–64 (2005).

35. Chu KME, Minogue S, Hsuan JJ, Waugh MG. Differential effects of the phosphatidylinositol 4-kinases, PI4KIIα and PI4KIIIβ, on Akt activation and apoptosis. *Cell Death and Disease* **1**:e106 (2010).

36. Balla A, Balla T. Phosphatidylinositol 4-kinases: old enzymes with emerging functions. *TRENDS in Cell Biology* **16**:351–361 (2006).

37. Wagner K, Hemminki K, Grzybowska E, et al. The insulin-like growth factor-1 pathway mediator genes: SHC1 Met300Val shows a protective effect in breast cancer. *Carcinogenesis* **25**:2473–8 (2004).

38. Yu H, Rohan T. Role of the Insulin-Like Growth Factor Family in Cancer Development and Progression. *Journal of the National Cancer Institute* **92**:1472-89 (2000).

39. Zheng Y, Zhang C, Croucher DR, et al. Temporal regulation of EGF signalling networks by the scaffold protein Shc1. *Nature* **499**:166–71 (2013).

40. Duncan JS, Whittle MC, Nakamura K, et al. Dynamic Reprogramming of the Kinome in Response to Targeted MEK Inhibition in Triple-Negative Breast Cancer. *Cell* **149**:307–321 (2012).

41. Filipović A, Gronau JH, Green AR, et al. Biological and clinical implications of nicastrin expression in invasive breast cancer. *Breast Cancer Res Treat* **125**:43–53 (2011).

42. Kopan R, Ilagan M. γ-secretase: proteasome of the membrane? *Nature Reviews Molecular Cell Biology* **5**:499–504 (2004).

43. Hu Y, Ye Y, Fortini ME. Nicastrin Is Required for γ-secretase Cleavage of the Drosophila Notch Receptor. *Dev Cell* **2**:69–78 (2002).

44. D'Angelo RC, Ouzounova M, Davis A, et al. Notch Reporter Activity in Breast Cancer Cell Lines Identifies a Subset of Cells with Stem Cell Activity. *Molecular Cancer Therapeutics* **14**:779–787 (2015).

45. Sanchez-Garcia F, Villagrasa P, Matsui J, et al. Integration of Genomic Data Enables Selective Discovery of Breast Cancer Drivers. *Cell* **159**:1461–1475 (2014).

CHAPTER 3

AN INTEGRATED GENOMICS APPROACH IDENTIFIES DRIVERS OF PROLIFERATION IN LUMINAL-SUBTYPE HUMAN BREAST CANCER

Elucidating the molecular drivers of human breast cancers requires a strategy that is capable of integrating multiple forms of data and an ability to interpret the functional consequences of a given genetic aberration. Here we present an integrated genomic strategy based on the use of gene expression signatures of oncogenic pathway activity (n = 52) as a framework to analyze DNA copy number alterations in combination with data from a genome-wide RNA-mediated interference screen. We identify specific DNA amplifications and essential genes within these amplicons representing key genetic drivers, including known and new regulators of oncogenesis. The genes identified include eight that are essential for cell proliferation (*FGD5*, *METTL6*, *CPT1A*, *DTX3*, *MRPS23*, *EIF2S2*, *EIF6* and *SLC2A10*) and are uniquely amplified in patients with highly proliferative luminal breast tumors, a clinical subset of patients for which few therapeutic options are effective. This general strategy has the potential to identify therapeutic targets within amplicons through an integrated use of genomic data sets.[1]

[1] Gatza, M. L., Silva, G. O., Parker, J. S., Fan, C. & Perou, C. M. An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nature Genetics* **46**, 1051–9 (2014).

INTRODUCTION

Tumorigenesis is driven by a combination of inherited and acquired genetic alterations resulting in a complex and heterogeneous disease. The ability to dissect this heterogeneity is critical to understanding the relevance of these alterations for disease phenotypes but also to enable the development of rational therapeutic strategies that can match the characteristics of the individual patient's tumor. Many studies, including reports from The Cancer Genome Atlas (TCGA) project, have made use of the power of multiplatform genomic analyses to identify known and new genetic drivers of tumor phenotypes [1–3]. This has led to the identification of disease subgroups with distinct characteristics and, in some instances, distinct genetic mechanisms of disease [1, 2, 4]. The strength of this approach relies on the integration of large-scale genomic data to reveal biological covariation that cannot be identified when using a single technology. A weakness of this approach is in the interpretation of the underlying biology, which generally represents an inference about pathway activity based on prior knowledge concerning an individual gene mutation or protein alteration.

Altered signaling pathway activity is an important determinant of the biology of a tumor and may predict therapeutic response; therefore, identifying the mechanisms driving key tumorigenic pathways is essential to understanding the transformation process [2, 5–8]. To take advantage of the vast amounts of existing genomic data, we used a series of experimentally derived gene expression signatures that are capable of measuring oncogene or tumor suppressor pathway activity, aspects of the tumor microenvironment and other tumor characteristics, including proliferation rate, as a framework by which to integrate multiple forms of genomic data. Our results identify patterns of oncogenic signaling within each of the molecular subtypes of breast cancer, many of which correlate directly with DNA copy number aberrations. By

further analyzing functional data from a genome-wide RNA-mediated interference (RNAi) screen [9], we identified genes that are essential for cell viability in a pathway-dependent and, in some cases, subtype-dependent manner. Our results identify a small number of DNA amplifications as potential drivers of proliferation in poor-outcome luminal breast cancers, and in general terms, we outline an approach that could be applied to many other tumor types for which multiplatform genomic data exist.

MATERIALS AND METHODS

Gene expression data. Agilent custom 244K whole-genome gene expression microarray data for human breast cancer samples was acquired from the TCGA project [2] data portal. Samples were filtered to include only those 476 samples for which Affymetrix SNP 6.0 data was present. As previously described [2], (TCGA) data were median centered for each gene. Illumina HT-29 v3 expression data for the METABRIC project (n = 1,992 samples) were acquired from the European Genome-phenome Archive at the European Bioinformatics Institute, and data were median centered for each gene [3]. Expression data for a panel of 51 breast cancer cell lines were acquired from GEO (GSE12777) [41]. Affymetrix U133+2 data were MAS5.0 normalized using the Affymetrix Expression Console (ver1.2.1.20) and $\log_2$ transformed. Expression probes were collapsed using the median gene value with the GenePattern [56] module CollapseProbes.

Affymetrix SNP 6.0 data. DNA copy number values were determined in 490 TCGA primary breast tumors (476 of which had matched mRNA expression data) and 1,992 METABRIC primary breast tumors using Affymetrix 6.0 SNP arrays as described previously [2, 3]. Copy number segmentation and segment calls (i.e., NEUT, AMP, GAIN, HOMD or HETD) were performed using the circular binary segmentation (CBS) algorithm as described previously

[2, 3]. Using the hg19 build annotation from the UCSC genome browser, genes were selected if they fell completely within a CBS-identified copy number segment. Genes that were not found completely within a copy number segment across any sample were filtered out. In the METABRIC data set, the copy number call gene matrix was determined from genes that fell completely within a CBS-identified copy number segment. Out of the 12 genes of interest, *SNX21*, *ZBTB46* and *DNAJC5* were not found completely within a CBS-identified segment among the METABRIC samples and were excluded from further analyses.

Gene expression signatures. A panel of 52 previously published gene expression signatures was used to examine patterns of pathway activity and/or microenvironmental states (Table 3.1). To implement each signature, the methods detailed in the original studies were followed as closely as possible. Of these 52 signatures, 22 signatures [10, 11, 32] were originally developed using a Bayesian binary regression strategy and are comprised of Affymetrix probe sets with positive and negative regression weights. These signatures were translated to a form that could be applied to non-Affymetrix expression data. For each signature, we excluded those probe sets with a negative correlation coefficient. The remaining probe sets with a positive coefficient were then translated to the gene level, and replicate genes were merged. To apply a given signature to a new data set, the expression data were filtered to contain only those genes that met the previous criteria, and the mean expression value was calculated using all genes within a given signature that were present in more than 80% of samples. The list of genes in each modified signature is shown in the online version of the paper, along with the scores for the TCGA data and cell line data set.

Statistical analyses of signature scores. To quantify differences in patterns of signature scores across subtypes, ANOVA followed by Tukey's post-test for pairwise comparisons was

used (as shown in Figure 3.1b). To investigate the level of concordance between each of the 52 signatures, the pathway scores calculated for each sample in the TCGA data set (data available online) were analyzed. The *R* values calculated by Pearson correlation are reported in Figure 3.2 and the source data available online.

Identification of point mutations as a function of pathway activity. To compare the frequencies of mutations, the 35 genes identified as being significantly mutated in human breast cancer [2] were assessed in the context of the 11-gene PAM50 proliferation signature [31]. A Fisher's exact test (Bonferroni corrected) was used to compare the frequency of mutations in samples with high (top quartile) and low (all other samples) pathway activity in LumA, LumB and HER2E ($n = 388$) samples. The frequencies of mutations associated with each group for each signature are summarized in the online version of the paper.

Identification of CNAs as a function of pathway activity. To identify CNAs, two analysis methods were used independently. Spearman rank correlation, both positive and negative, was used to compare gene-level segment scores with predicted pathway activity. To compare the frequencies of amplifications and losses, a Fisher's exact test was used to compare the frequencies of either gene-specific copy number gains and amplifications or deletions (both LOH and deletions) against nonamplified or nondeleted samples. Samples in the top quartile of the calculated pathway activity were compared to those in the bottom three quartiles. For each analysis, the $-\log_{10}$ Bonferroni-adjusted *P* values are reported (Figure 3.3 and 3.4). To identify genes that were significant across both methods, a threshold of $q < 0.01$ (Bonferroni corrected) was set for validation (Figure 3.5) and $q < 0.05$ for discovery (Figure 3.6). The Bonferroni-corrected *P* values for the positive and negative Spearman rank correlation for each gene and each signature are reported in the online version of the paper. The frequency of copy number

gains in the top quartile compared to all other samples, as well as the Bonferroni-corrected *P* values calculated by Fisher's exact test, are reported for each gene and each signature and are available in the online version of the paper.

Analysis of genome-wide RNAi proliferation data. To identify genes that are required for cell viability in a signature-dependent manner, data from a previously published genome-wide RNAi screen carried out on a panel of breast cancer cell lines were analyzed [9]. The Gene Active Ranking Profile (GARP)-normalized data were obtained from the COLT database and filtered to include only those 27 cell lines for which gene expression data (GSE12777) were also available (acquired February 2013). To identify genes essential for pathway-dependent cell proliferation, a negative Spearman correlation was performed comparing predicted pathway activity and GARP score for each sample. A threshold of $P < 0.05$ was considered significant for all analyses.

Analysis of mRNA expression in copy number–neutral samples. To assess mRNA expression in luminal tumors lacking CNAs of each candidate gene, luminal and HER2E samples from the TCGA ($n = 388$) and METABRIC ($n = 1,333$) studies were grouped into those with high (top quartile) and low (all other samples) pathway activity. Samples with copy number gains (including high-level amplifications or gains) or losses (both LOH and homozygous deletions) were excluded, and a *t* test was used to examine statistical differences between the expression levels of genes in each cohort.

Survival analyses. To investigate the effect that candidate gene amplification has on disease-specific survival, clinical data for the 1,992 patients in the METABRIC study were obtained [3]. The 11-gene PAM50 proliferation signature [31] was applied to all 1,992 samples by calculating the median value of the signature for each sample. For survival analyses, patients

that died of causes unrelated to breast cancer and patients without a date of death were censored. We extracted patients with tumors classified as LumA, LumB or HER2E and for whom survival data were reported (n = 1,333). For survival analysis of the TCGA data set [2], we extracted patients with tumors classified as LumA, LumB or HER2E and for whom clinical data were available (September 2012). Disease-specific survival was calculated by comparing samples with amplification (including copy number gains and high-level amplification) of a candidate gene against those without. In each data set, patients without a CNA call for a specific gene were excluded from the survival analysis. For each analysis, significance was calculated by a log-rank test, and the hazard ratio (HR) is reported. To compare the effect of candidate gene copy number status on common prognostic markers, including proliferation (PAM50 proliferation signature), molecular subtype (PAM50), tumor stage, node status, ER status, HER2 status and age at diagnosis, a multivariate Cox model was used.

RESULTS

Subtype-specific patterns of oncogenic signaling

To objectively identify genetic drivers of breast cancer, we examined genomic-based patterns of oncogenic pathway activity, the tumor microenvironment and other important features in human breast tumors using a panel of 52 previously published gene expression signatures (Table 3.1 and the online version of the paper) [10 –32]. We applied each signature to the breast cancer gene expression microarray data (n = 476) from the TCGA project (online version of the paper), for which the molecular intrinsic subtype had been determined [2]. Consistent patterns of pathway activity emerged for each subtype (as illustrated in Figure 3.1a), and we quantitatively assessed these patterns using an analysis of variance (ANOVA) test

followed by Tukey's test for pairwise comparison (Figure 3.1b and Table 3.2). Analyzing differences across subtypes on the basis of these 52 features demonstrated that the strongest correlation between samples existed within each molecular subtype (Figure 3.7).

The patterns of pathway activity recapitulated known characteristics of each subtype, including dysregulation of pathways that can be linked to female hormone receptors, oncogenes and/or tumor suppressor mutation status (Figure 3.1). For example, basal-like tumors, which represent ~80% of triple-negative breast cancers, are characterized by low hormone receptor signaling, mutant p53 signaling and high expression of proliferation pathway activity (Figure 3.1). Likewise, HER2-enriched (HER2E) tumors show high expression of the HER2 [11] and HER2 amplicon (HER2-AMP) [12] signatures, whereas luminal A (LumA) tumors show high hormone receptor signaling and wild-type p53 signaling. Highly proliferative LumB tumors, which also show some hormone receptor signaling, are distinguished from less proliferative LumA samples by increased proliferation-associated pathways. Thus, these data robustly recapitulate many previously published pathway and subtype associations.

Calculating a Pearson correlation coefficient to assess the concordance between each of the 52 signatures (Figure 3.2 and data available online) identified strong relationships between independent signatures for a given pathway, as well as between related pathways. For example, two MYC signatures [11, 15, 32] demonstrated an *R* value of 0.72, whereas PIK3CA [18] and PTEN-deleted [27] signatures had an *R* value of 0.82. Signatures scoring different pathways were also concordant; for instance, MYC-mediated regulation of E2F signaling [33] was identified by the association between the RB loss of heterozygosity (RB-LOH) [16] and MYC [15] signatures ($R = 0.79$), whereas EGFR-mediated activation of STAT33 signaling [34] was recapitulated by the EGFR [11, 32] and STAT3 [11, 32] ($R = 0.72$) signatures. These results

provide a measure of validity for each signature, but because differences do exist between signatures for a specific pathway, the results suggest that each signature provides an opportunity to investigate a particular pathway, taking into account the genetic manipulation used to develop a given signature.

Characterization of pathway-specific copy number alterations

We next used DNA copy number data from the TCGA project (n = 476) to identify copy number alterations (CNAs) associated with pathway activity (Figure 3.5a). We first identified genes for which CNAs were positively (or inversely) correlated with pathway activity using a Spearman rank correlation (Bonferroni corrected to control the familywise error rate) to assess the relationship between pathway score and gene-level DNA segment score (Figure 3.3 and data available online). Second, we used a Fisher's exact test (Bonferroni corrected) to calculate the frequency of CNA gains (including high-level amplifications and gains) or losses (including LOH and deletions) in samples with high (top quartile) pathway activity compared to all other samples (low activity) (Figure 3.4 and data available online). To reduce potential false-positive results associated with either strategy alone, for each signature we focused on those genes that were significant in both analyses (Figure 3.5a); potential drivers of pathway activity had a positive correlation and a higher amplification frequency in samples with high pathway activity, whereas potential repressors had a negative correlation and increased frequency of copy number losses. Mapping genes that met these criteria to chromosomal loci identified pathway-specific patterns of CNAs (Figure 3.5b). Consistent with previous studies reporting that basal-like tumors have a higher incidence and larger spectrum of CNAs [2, 35], pathways associated with basal-

like tumors had more complex patterns of CNAs when compared to luminal-associated pathways.

To further assess the validity of this strategy, we investigated the relationship between pathway activity and a chromosomal alteration of known causative activity. We first focused on the HER2-AMP signature [12], as this signature is comprised of genes located at the 17q loci and the *ERBB2*/17q amplification is the dominant driver of this pathway. *ERBB2* was amplified in 84.9% of samples with high (top quartile) pathway activity compared to in 7.3% of low-scoring samples ($q = 1.1 \times 10^{-55}$); likewise, this relationship had a positive Spearman rank correlation ($q = 2.4 \times 10^{-108}$) (Figure 3.5c and data available online). Although several other alterations, including *MYC* amplification ($q = 1.1 \times 10^{-2}$ and $q = 6.3 \times 10^{-3}$), were also associated with this signature, thus identifying a previously known relationship [36], *ERBB2*/17q amplification was the dominant alteration identified, providing a robust positive control for this strategy. As expected, we observed similar results when analyzing the HER2 pathway using the independently developed HER2 [11, 32] signature (data available online).

We further validated this strategy by assessing the relationship between CNAs and pathways that are associated with a more complex genomic landscape. Previous studies from our group have suggested that the HER1-C2 [13] signature measures predominantly the RAS-RAF-MEK arm of the EGFR pathway [13]. Consistent with this observation, we detected a correlation between the HER1-C2 signature ($q < 0.01$) and *GRB2*, *SOS1*, *KRAS*, *BRAF*, *PIK3CA*, *PIK3CB* and *MYC* genomic DNA amplifications, as well as a negative correlation ($q < 0.01$) with loss of *NF1* and the PI3K repressors *INPP4B* and *PTEN* (Figure 3.5d and data available online). We then analyzed CNAs associated with the RB-LOH [16] signature (Figure 3.5e and data available online) and identified associations between it and CNAs of known RB-E2F components,

including loss of *RB1* and gains of *E2F1* and/or *E2F3*. Consistent with the role of the RB-E2F pathway in mediating cell cycle progression and proliferation [37], *CCND2*, *CCND3* and *MYC* amplification also correlated with this signature. Collectively these results demonstrate that this strategy is able to link CNAs with pathway activity and does so by focusing on all aspects of the pathway, often beyond the dominant regulator, potentially allowing for the identification of new regulatory components.

Identification of amplified genes linked to pathway activity

Given the ability of this strategy to identify known CNAs of pathway activity, we next used this approach to identify new drivers of pathway activity. Because highly proliferative luminal tumors have a poor prognosis and poor responses to existing therapies [38, 39], we sought to identify amplified genes and/or CNAs associated with our previously published 11-gene PAM50 proliferation signature with the hope that these might represent targetable drivers of oncogenesis.

To identify those genes that are altered specifically in highly proliferative luminal tumors while excluding those that are associated with proliferation irrespective of subtype, we performed analyses on two subsets of samples: all tumors and all non–basal like tumors (henceforth called luminal tumors). Some rationale for this binary distinction comes from recent TCGA studies in which 12 tumor types were studied simultaneously, and the results showed that breast tumors formed two groups, namely basal-like and all other breast tumors (called luminal and including $HER2^+$ tumors), suggesting that breast cancer might be considered broadly as two main disease types [40].

Examining the TCGA breast cancer data set using the PAM50 proliferation signature [31], we found that basal-like, LumB and HER2E tumors had the highest proliferation levels (Figures 3.8a, b), with the top quartile (Figure 3.8c) comprised of patients with basal-like (49.6%), LumB (33.6%) and HER2E (16.8%) tumors, whereas the top quartile of proliferative luminal tumors (Figure 3.8d) contained patients with LumB (68.0%) and HER2E (32.0%) tumors. Using the PAM50 proliferation signature, we examined the frequency of CNA gains and losses in highly proliferative (top quartile) tumors relative to less proliferative samples irrespective of subtype using the statistical strategies discussed previously (Figures 3.8e, f and data available online). To identify genes that are specifically amplified in highly proliferative luminal breast cancer, we repeated these analyses using the luminal tumor subset (Figures 3.8g, h and data available online). Analyzing both populations of tumors identified three classes of proliferation-associated regions ($q < 0.05$): (i) CNAs associated irrespective of subtype, (ii) CNAs altered in basal-like tumors, and (iii) CNAs altered in highly proliferative luminal tumors. These results allowed us to focus our analyses on those genes within regions that are uniquely altered in highly proliferative luminal tumors by censoring proliferation-associated genes that are altered in basal-like breast cancer (e.g., *TP53* or *INPP4B* loss) or that are altered irrespective of molecular subtype (e.g., *RB1* loss or *MYC* amplification). These analyses identified a number of regions, including 3p25, 5p15, 11q13, 17q22 and 20q11-13, that were uniquely amplified in highly proliferative luminal tumors.

Identification of pathway-specific essential genes

To distinguish essential from nonessential genes in amplified regions that are associated with proliferation in luminal tumors, we next examined data from a genome-wide RNAi screen

of multiple breast tumor–derived cell lines [9]. We applied the 52 gene expression signatures to a panel (GSE12777) [41] of breast cancer cell lines (Figure 3.9 and data available online), 27 of which had mRNA expression data and were also part of an RNAi proliferation screen in which a genome-wide shRNA library (~16,000 genes) had been used to identify essential genes (Figure 3.10a) [9]. For each signature, we used a negative Spearman rank correlation to identify pathway-specific essential genes (Figure 3.10b and data available online) by comparing the pathway score against the normalized shRNA score across the panel of 27 cell lines. These analyses identified inverse relationships between the abundance of shRNAs targeting key regulatory genes and pathway scores. For instance, examining the ER [11, 32], HER2 [11, 32] or STAT1 [42] signatures as controls (Figures 3.10c–e) showed a negative correlation between pathway score and shRNA against *ESR1* ($P = 0.0143$), *ERBB2* ($P = 0.0227$) and *STAT1* ($P = 0.0049$) or *JAK3* ($P = 0.00013$), respectively. These associations were expected for the ER and HER2 pathways given the relationship between HER2 or ER-α mRNA and/or protein expression and the response of cell lines or tumors to trastuzumab or anti-estrogen therapies, respectively. These results confirm that this approach is able to identify essential genes that are known to be functionally associated with pathway activity, thereby suggesting that these data can serve as a biological filter to distinguish pathway-specific essential from nonessential genes.

Amplified essential genes linked to luminal tumor proliferation

We next sought to distinguish between essential and nonessential genes within regions amplified specifically in highly proliferative luminal tumors. For each subset of tumors, we identified genes in amplified regions that were positively correlated with proliferation and showed an increased amplification frequency ($q < 0.05$). We next examined the RNAi data in all

breast cancer cell lines (Figure 3.11a) and in luminal $HER2^+$ cell lines (Figure 3.11b) in the context of the PAM50 proliferation signature (data available online). Comparing the results of these four analyses (Figure 3.6a) identified 19 genes that were uniquely essential for cell viability in luminal cell lines and that were amplified in highly proliferative luminal tumors (Figure 3.6b). Two additional genes, *DNAJC5* and *SNX21*, were identified by RNAi analysis but were initially overlooked in the CNA analyses, as they were located at the cusp of two segmented regions; however, because genes overlapping both 5′ and 3′ of these genes were amplified, we included them in further investigations. Of these 21 candidate genes, 12 showed a significant relationship ($P < 0.05$) between DNA copy number levels and mRNA expression in luminal tumors (Figure 3.12). Notably, half of these genes were located at 20q11-13 (*EIF2S2*, *EIF6*, *SLC2A10*, *SNX21*, *ZBTB46* and *DNAJC5*), with two located at 3p25.1 (*FGD5* and *METTL6*) and the remaining genes located at 5p15 (*TRIO*), 11q13 (*CPT1A*), 12q13 (*DTX3*) and 17q22-23 (*MRPS23*). In contrast, permuting the data labels 1,000 times for each analysis, in all samples and in luminal samples alone, identified no gene that met this statistical threshold, suggesting that the 21 candidate genes could not have been identified by chance alone.

Validation of identified candidate genes

We next confirmed that the majority of the identified genes were significantly amplified in highly proliferative luminal breast tumors by analyzing an independent breast tumor data set (Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), $n = 1{,}992$) for which both mRNA expression and genomic DNA CNA data were available [3]. Of the 12 genes identified, 9 (*FGD5*, *METTL6*, *TRIO*, *CPT1A*, *DTX3*, *MRPS23*, *EIFS2S*, *EIF6* and *SLC2A10*) were present on both platforms used in the METABRIC study. Each of these genes (Figure 3.13)

showed a significant ($P < 0.05$) relationship between CNA status and mRNA expression in luminal breast tumors ($n = 1{,}333$). Notably, eight of the nine genes, the exception being *TRIO*, also showed an increased amplification frequency ($P < 0.05$) in highly proliferative (top quartile) luminal tumors (Figure 3.14), thus recapitulating one of our main findings.

To confirm that DNA mutations of genes associated with proliferation in luminal tumors did not confound these results, we examined the relationship between the 11-gene proliferation score and the mutation frequency of the 35 previously identified significantly mutated genes in human breast cancers reported by TCGA [2]. Using a Fisher's exact test (Bonferroni corrected), we determined that only *TP53* ($q = 7.0 \times 10^{-10}$) and *MAP3K1* ($q = 5.0 \times 10^{-3}$) mutations occurred at significantly different frequencies in highly proliferative (top quartile) luminal tumors compared to all other samples; *TP53* mutations occurred more frequently (51.6% compared to 18.6%) and *MAP3K1* (2.1% compared to 12.4%) mutations occurred less frequently in highly proliferative luminal tumors (Table 3.3). Moreover, we found no significant relationship between *MAP3K1* or *TP53* mutation status (Bonferroni-corrected Fisher's exact test, $q > 0.05$) and the amplification status of each candidate gene (Table 3.4) in highly proliferative luminal tumors.

We then investigated whether expression of the candidate genes, independent of CNA status, was associated with proliferation in luminal breast tumors. By comparing the mRNA expression patterns of each candidate gene in highly proliferative luminal tumor samples (top quartile) against all other samples, we found that tumors lacking CNAs of each candidate gene fell into three categories: those that exhibited a positive relationship between mRNA expression and the PAM50 proliferation signature (*EIF2S2*, *EIF6*, *CPT1A* and *MRPS23*), those that were anticorrelated with the signature (*DTX3*) and those that showed no correlation (*FGD5*, *METTL6*

and *SLC2A10*) between data sets (Figure 3.15). These data suggest that amplification is a key mechanism driving the expression of these genes. However, our data also suggest, not surprisingly, that overall high expression may be the driver for some genes, which can be accomplished by amplification or through other unknown means.

Candidate gene amplification correlates with poor prognosis

Previous studies have shown that highly proliferative luminal tumors have a poor prognosis [38, 39]; therefore, we investigated what impact amplification of each candidate gene had on overall survival. From the TCGA (n = 388) [2] and METABRIC (n = 1,333) [3] data sets, we extracted the subset of patients with LumA, LumB or HER2E tumors for which survival data were available online. We first analyzed data from the TCGA project (Figures 3.16a–e), and despite the relatively short follow-up time (median, 1.7 years), we determined that amplification of *FGD5* ($P < 0.0001$; hazard ratio (HR), 8.0), *METTL6* ($P = 0.0003$; HR, 5.9), *DTX3* ($P = 0.0387$; HR, 2.6) and *MRPS23* ($P = 0.0078$; HR, 2.9) predicted a significantly worse outcome in patients with luminal breast cancer, whereas *CPT1A* amplification had no effect on patient survival ($P = 0.3738$). Extending these analyses to the METABRIC data set (Figures 3.16f–j), which had a longer median survival time (7.2 years), confirmed that *FGD5* ($P = 0.0170$; HR, 2.0), *METTL6* ($P = 0.0081$; HR, 2.1), *DTX3* ($P = 0.0098$; HR, 1.8) and *MRPS23* ($P = 0.0020$; HR, 1.5) amplification correlated with a poor prognosis, whereas gain of *CPT1A* had no effect ($P = 0.099$) on the survival of patients with luminal breast cancer. The remaining three genes showed no consistent effect on prognosis (Figure 3.17). Although it is possibly that other genes within these chromosomal loci are also prognostic, these amplified genes were associated with

proliferation *in vivo*, were prognostic in multiple patient cohorts and are essential for cell viability *in vitro*.

We likewise determined that for most of the identified candidate genes that failed to meet all our predetermined criteria, amplification alone, without a coordinate increase in mRNA expression, was not sufficient to affect prognosis, as only one (*TMEM117*) of these genes showed a consistently poor prognosis in the TCGA and METABRIC data sets (Table 3.5). We then investigated whether the 12 initial candidate genes were predictive of poor prognosis when compared with standard prognostic markers, including molecular subtype, tumor stage, node status, ER status, HER2 status, age at diagnosis and the 11-gene proliferation score, when tested using a multivariate analysis (Cox model). We determined that amplification of a single candidate gene did not consistently outperform or improve the prognostic capacity of these clinical and genomic variables (Appendix 4). However, these candidate genes were not identified to be prognostic markers, especially given that they correlate with proliferation, but instead were selected as likely drivers of proliferation, a highly important prognostic feature.

DISCUSSION

Numerous studies, including many that have focused on human breast cancer, used large-scale analyses to investigate the genomic landscape of human cancers in order to identify molecular heterogeneity and define new tumor subtypes not previously recognized [2, 3, 6, 11]. The challenge presented by these studies, and by the enormous amount of genomic data available from resources such as the TCGA and METABRIC projects, is how to integrate multiple forms of genomic data to investigate the biology of disease and how to interpret the relevance of

identified genomic alterations without relying on inferences of 'known' biology to determine the role that these alterations have in tumorigenesis.

In this study we utilized gene expression signatures of signaling pathways to identify patterns that can distinguish the known subtypes of breast cancer. These signatures were developed largely from controlled manipulations of the relevant pathways *in vitro* and are thus based on experimental evidence for pathway activation as opposed to extrapolations of pathway activity achieved from analyses of annotated gene lists. Therefore, the use of an experimentally derived pathway signature, as opposed to an analysis of a single genomic alteration, provides a measure of pathway activity irrespective of how the pathway may have been activated. For instance, a given pathway can be active in a subset of tumors as a result of either an activating alteration (i.e., *E2F1* or *E2F3* amplification) or an independent event that inactivates a negative regulator of the pathway (i.e., *RB1* loss and/or mutation), which nevertheless achieves the same end result (i.e., DNA replication and cell proliferation); notably, we identified these four genetic events as being statistically associated with the RB-LOH signature [16], which is dominated by E2F-regulated genes and is a strong indicator of cell proliferation and prognosis.

Proliferation is one of the most powerful prognostic features in breast cancers, especially for ER^+ cancers [38, 39]. Because proliferation is so important, we used a gene expression signature of proliferation as a means to integrate the DNA copy number data, along with data from a genome-wide RNAi screen of luminal breast cancer cell lines, to identify luminal-specific genetic drivers of proliferation. We identified 12 genes that were amplified uniquely in highly proliferative luminal tumors in the TCGA data set, have a correlation between mRNA expression and DNA copy number and have been shown to be essential for luminal breast cancer cell line viability; we validated 8 of these genes using the independent METABRIC data set. Whereas

FGD5, *METTL6*, *DTX3* and *MRPS23* amplification was prognostic in luminal tumors, these and many of the other identified genes have been reported previously to regulate tumorigenic characteristics, albeit not necessarily in human breast cancer. For example, *FGD5* has been shown to regulate the proangiogenic function of *VEGF* [43], potentially leading to increased proliferation. *DTX3* purportedly promotes Notch signaling [44, 45], whereas *EIF6* is a Notch-dependent regulator of cell invasion and migration [46], and its inhibition restricts lymphomagenesis and tumor progression [47]. *MRPS23* expression is associated with proliferation, oxidative phosphorylation, invasiveness and tumor size in uterine cervical cancer [48]. *METTL6* has been reported to contribute to cytotoxic chemotherapy sensitivity in lung cancers [49].

Several previous studies have identified chromosomal regions altered specifically in subsets of breast cancer, including 3p25 (encompassing *METTL6* and *FGD5*) [2] and 11q13 (*CPT1A*) [3] in luminal breast tumors; however, these studies neither discriminated between essential and nonessential genes within a specific amplicon nor identified the functional consequences of these alterations. In contrast, we have shown that these regions are amplified uniquely in highly proliferative luminal tumors, and we distinguish between amplified genes that are essential for cell proliferation and are thus likely contribute to tumorigenesis and those that are amplified but are not essential. For instance *SRC* (20q12-13), which is co-amplified with *EIF6,* is similarly amplified in a significant ($q < 0.01$) percentage of highly proliferative luminal tumors (data available online) but was not identified as being essential in highly proliferative luminal breast cancer cell lines in the RNAi screen (data available online). Notably, in addition to its role in regulating translation [50] and Notch signaling [46], *EIF6* has been reported to link integrin-β4 to the intermediate filament cytoskeleton [51], potentially leading to downstream

activation of *SRC* signaling. These results may explain some of the paradoxical findings of *SRC* in that it may contribute to proliferation status but may not be essential, whereas a gene very near it, which is also linked to proliferation, is essential for cell viability *in vitro*. Clearly, additional experiments are needed to address this issue, but these results highlight the complex nature and importance of this specific amplicon.

A major challenge to translating these findings into the clinic is the identification of genes within amplicons that are therapeutically targetable. One such event may be amplification of 11q13-14 (*CPT1A*), which was recently reported [3] to be a defining feature of a high-risk ER^+ subgroup (integrative cluster 2) and correlates with a poor prognosis in esophageal squamous cell carcinoma [52]. We identified *CPT1A* as the only gene within the amplified 11q13 locus that is required for cell viability within the confines of the proliferation signature and luminal cell lines, suggesting that repression of *CPT1A* could affect the proliferative phenotype of these tumors. Consistent with this hypothesis, it was recently reported that RNAi-mediated downregulation, or drug-mediated inhibition, of *CPT1A* inhibited cancer cell line proliferation, migration and metastasis [53-55], although not in breast cancer cell lines. In addition, a specific inhibitor of *CPT1A* (ST-1326) repressed tumor formation and proliferation in an Eμ-Myc mouse model of Burkett's lymphoma [55].

Collectively these data demonstrate the ability of this cross-platform genomics approach to identify new oncogenes that are essential for cell viability and are amplified in a subset of patients with highly proliferative luminal breast cancer. These data suggest that not only are these identified genes potential drivers of oncogenesis and that an emphasis should be placed on

elucidating their role in breast tumorigenesis but also that they, or their associated pathways, may serve as new therapeutic targets in a subset of human breast cancers for which limited therapeutic opportunities currently exist.

TABLES

Table 3.1 Summary of Gene Expression Signatures

Signature	Pathway	References
ACIDOSIS	Acidosis response	PMID: 21672245, PMID: 22078435
ACTIVE ENDOTHELIUM	Activated endothelium	PMID: 23975155
AKT	Akt signaling	PMID: 20335537, PMID: 22078435
βCATENIN	Beta catenin activation	PMID: 20335537, PMID: 22078435
BMYB	BMYB signaling	PMID: 19043454
BRCA1	BRCA1 signaling	PMID:11823860
CMYB	CMYB signaling	PMID:20949095
E2 ACTIVATED (IE)	Estrogen activated signaling	PMID:16505416
E2 REPRESSED (IIE)	Estrogene repressed signaling	PMID:16505416
E2F1	E2F1 signaling	PMID: 20335537, PMID: 22078435
EGFR	EGFR activation	PMID: 20335537, PMID: 22078435
ER	Estrogen receptor signaling	PMID: 20335537, PMID: 22078435
ESC HUMAN	Human Embryonic Stem Cell	PMID:18397753
FOS JUN	Fos-Jun kinase signaling	PMID: 21214954
GATA3	Wild-type GATA3-mediated signaling	PMID:15361840
GLUCOSE DEPLETION	Glucose depletion response	PMID: 21672245, PMID: 22078435
GYCOLYSIS	Glycolysis response	PMID:19291283
HER1 C1	HER1/EGFR Cluster 1	PMID:17663798
HER1 C2	HER1/EGFR Cluster 2	PMID:17663798
HER1 C3	HER1/EGFR Cluster 3	PMID:17663798
HER2	HER2/ERBB2 overexpression	PMID: 20335537, PMID: 22078435
HER2 AMP	HER2/ERBB2 amplification	PMID: 21214954
HYPOXIA	Hypoxia response	PMID: 21672245, PMID: 22078435
IFNA	Interferon alpha response	PMID: 20335537, PMID: 22078435
IFNG	Interferon gamma response	PMID: 20335537, PMID: 22078435
LACTIC ACID	Lactic acidosis response	PMID: 21672245, PMID: 22078435
LKB1	LKB1 signaling	PMID:17676035
MYC DUKE	Myc activation	PMID: 20335537, PMID: 22078435
MYC UNC	Myc signaling	PMID:19690609
P53	p53 signaling	PMID: 20335537, PMID: 22078435
P53 MUT	Mutant p53 signaling	PMID:17150101
P53 WT	Wild-type p53 signaling	PMID:17150101
P63	p63 activation	PMID: 20335537, PMID: 22078435
PI3K	PI3 kinase signaling	PMID: 20335537, PMID: 22078435
PIK3CA	PI3 kinase signaling	PMID:22552288
PR	Progesterone Receptor	PMID: 20335537, PMID: 22078435
PROLIFERATION	Proliferation	PMID: 21214954
PROLIFERATION (PAM50)	Proliferation	PMID:19204204
PTEN WT	Wild-type PTEN signaling	PMID: 17452630
PTEN DEL	Mutant PTEN signaling	PMID: 17452630
RAS	Ras activation	PMID: 20335537, PMID: 22078435
RB LOH	Loss of RB expression	PMID:18782450
SRC	Src kinase	PMID: 20335537, PMID: 22078435
STAT1	Stat1 activation	PMID: 19272155
STAT3	Stat3 activation	PMID: 20335537, PMID: 22078435
STEM CELL	Stem cell associated expression	PMID:15931389
STROMAL DOWN	Low stromal cellularity	PMID:19648928
STROMAL UP	High stromal cellularity	PMID:19648928
TGFB	Tumor growth factor beta	PMID: 20335537, PMID: 22078435
TNFA	Tumor necrosis factor alpha	PMID: 20335537, PMID: 22078435
VEGF/HYPOXIA	Vascular endothelial growth factor / hypoxia signaling	PMID:19291283
WOUND RESPONSE	Wound response in breast cancer microenvironment	PMID:19887484

Table 3.2 Summary of ANOVATukey test analysis of subtype pathway score

Pathway	ANOVA	Basal-HER2	Basal-LumA	Basal-LumB	HER2-LumA	HER2-LumB	LumA-LumB
BRCA1	<0.0001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
HER1 C2	<0.0001	<0.001	<0.001	<0.001	<0.001	<0.05	<0.001
P53 Mut	<0.0001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
E2 Repressed (IIE)	<0.0001	>0.05	<0.001	<0.001	<0.001	>0.05	<0.001
Proliferation (PAM50)	<0.0001	<0.01	<0.001	<0.001	<0.001	>0.05	<0.001
ESC Human	<0.0001	<0.01	<0.001	<0.001	<0.001	>0.05	<0.001
B-catenin	<0.0001	<0.001	<0.001	<0.001	<0.001	>0.05	<0.001
BMYB	<0.0001	<0.001	<0.001	<0.001	<0.001	>0.05	<0.001
PROLIFERATION	<0.0001	<0.001	<0.001	<0.001	<0.001	>0.05	<0.001
PTEN DEL	<0.0001	<0.001	<0.001	<0.001	<0.001	>0.05	<0.001
RB LOH	<0.0001	<0.001	<0.001	<0.001	<0.001	>0.05	<0.001
HER1 C3	<0.0001	<0.001	<0.001	<0.001	<0.001	>0.05	<0.001
MYC (DUKE)	<0.0001	<0.001	<0.001	<0.001	<0.001	>0.05	<0.001
Activated Endothelium	<0.0001	<0.001	<0.001	<0.001	<0.001	>0.05	<0.001
MYC (UNC)	<0.0001	<0.001	<0.001	<0.001	<0.001	<0.05	<0.001
PIK3CA	<0.0001	<0.001	<0.001	<0.001	<0.001	>0.05	<0.001
RAS	<0.0001	<0.001	<0.001	<0.001	<0.001	>0.05	<0.001
HER1 C1	<0.0001	<0.01	<0.001	<0.001	<0.001	<0.001	<0.001
GLYCOLYSIS	<0.0001	>0.05	<0.001	<0.001	<0.001	<0.001	<0.001
VEGF/Hypoxia	<0.0001	<0.001	<0.001	<0.001	<0.001	>0.05	>0.05
E2F1	<0.0001	>0.05	<0.001	<0.05	<0.001	>0.05	<0.001
PI3k	<0.0001	<0.001	<0.001	<0.001	<0.001	>0.05	<0.001
Stem Cell	<0.0001	<0.01	<0.001	<0.05	<0.001	>0.05	<0.001
P63	<0.0001	<0.001	<0.01	<0.001	<0.001	>0.05	<0.001
IFNA	0.0008	>0.05	<0.01	>0.05	<0.05	>0.05	<0.05
IFNG	0.0005	>0.05	<0.001	>0.05	>0.05	>0.05	>0.05
STAT1	<0.0001	>0.05	<0.001	<0.01	<0.001	>0.05	<0.01
TNFA	<0.0001	<0.05	<0.001	<0.001	<0.01	<0.001	>0.05
Glucose Depeletion	<0.0001	<0.05	<0.001	<0.001	<0.001	<0.001	>0.05

Table 3.2 Summary of ANOVATukey test analysis of subtype pathway score (Continued)

Pathway	ANOVA	Basal-HER2	Basal-LumA	Basal-LumB	HER2-LumA	HER2-LumB	LumA-LumB
CMYB	0.0957	>0.05	>0.05	>0.05	>0.05	>0.05	>0.05
HYPOXIA	<0.0001	>0.05	<0.01	<0.001	<0.05	<0.001	>0.05
AKT	0.0004	>0.05	<0.001	<0.05	>0.05	>0.05	>0.05
FOS JUN	<0.0001	>0.05	<0.001	>0.05	<0.001	>0.05	<0.001
ACID	<0.0001	<0.01	>0.05	<0.001	<0.001	>0.05	<0.001
Stromal UP	<0.0001	>0.05	<0.001	>0.05	<0.01	>0.05	<0.001
Wound Response	<0.0001	>0.05	>0.05	<0.001	<0.01	<0.05	<0.001
LKB1	<0.0001	>0.05	>0.05	<0.001	>0.05	<0.001	<0.001
HER2	<0.0001	<0.001	>0.05	<0.001	<0.001	<0.001	<0.001
EGFR	<0.0001	>0.05	>0.05	<0.001	>0.05	<0.001	<0.001
STAT3	<0.0001	>0.05	<0.001	<0.001	>0.05	<0.001	<0.001
TGFB	<0.0001	>0.05	>0.05	<0.001	>0.05	<0.001	<0.001
GATA3	<0.0001	<0.001	>0.05	<0.001	<0.01	<0.001	<0.001
HER2AMP	<0.0001	<0.001	>0.05	>0.05	<0.01	<0.001	<0.05
Lactic Acidosis	<0.0001	<0.001	<0.001	<0.001	>0.05	>0.05	>0.05
Stromal DOWN	<0.0001	>0.05	<0.001	<0.001	>0.05	>0.05	>0.05
ER	<0.0001	<0.001	<0.001	<0.01	<0.001	<0.001	>0.05
PR	<0.0001	<0.001	<0.001	<0.001	<0.001	<0.001	>0.05
P53	<0.0001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
E2 Activated (IE)	<0.0001	>0.05	<0.001	<0.001	<0.001	<0.001	<0.001
P53 WT	<0.0001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
PTEN WT	<0.0001	<0.001	<0.001	<0.001	<0.001	>0.05	<0.001
SRC	0.0016	>0.05	>0.05	>0.05	>0.05	>0.05	<0.001

Table 3.3 Summary of SMG mutation frequency associated with Proliferation (PAM50) in luminal/HER2E samples

GENE	%PROLIFERATION (PAM50) LOW	%PROLIFERATION (PAM50) HIGH	PROLIFERATION (PAM50) Q
MUT.7157_TP53	18.56	51.55	7.00E-10
MUT.5290_PIK3CA	44.67	34.02	7.17E-02
MUT.2625_GATA3	12.71	12.37	7.73E-01
MUT.4214_MAP3K1	12.37	2.06	4.98E-03
MUT.58508_MLL3	7.90	5.15	9.76E-01
MUT.999_CDH1	9.28	5.15	3.85E-01
MUT.6416_MAP2K4	6.53	1.03	1.14E-01
Mut.861_RUNX1	4.47	2.06	6.58E-01
MUT.5728_PTEN	4.12	3.09	9.70E-01
MUT.6926_TBX3	1.72	5.15	9.39E-01
MUT.5295_PIK3R1	2.06	6.19	4.55E-01
Mut.207_AKT1	3.44	2.06	9.89E-01
Mut.865_CBFB	2.06	2.06	9.10E-01
MUT.79718_TBL1XR1	2.06	2.06	1.00E+00
MUT.9611_NCOR1	3.78	3.09	9.66E-01
MUT.10664_CTCF	3.44	1.03	4.05E-01
Mut.677_ZFP36L1	1.37	1.03	1.00E+00
MUT.2874_GPS2	1.72	0.00	6.35E-01
MUT.23451_SF3B1	2.41	1.03	9.84E-01
MUT.1027_CDKN1B	1.03	2.06	7.24E-01
MUT.7399_USH2A	3.09	5.15	4.23E-01
MUT.6103_RPGR	1.03	0.00	1.00E+00
MUT.5925_RB1	0.34	3.09	1.15E-01
MUT.2334_AFF2	1.72	4.12	1.00E+00
MUT.4763_NF1	3.09	2.06	1.00E+00
MUT.26191_PTPN22	1.03	4.12	2.39E-01
MUT.6262_RYR2	4.12	8.25	8.61E-01
MUT.5789_PTPRD	2.06	4.12	8.99E-01
MUT.8590_OR6A2	1.37	0.00	1.00E+00

Table 3.3 Summary of SMG mutation frequency associated with Proliferation (PAM50) in luminal/HER2E samples (Continued)

GENE	%PROLIFERATION (PAM50) LOW	%PROLIFERATION (PAM50) HIGH	PROLIFERATION (PAM50) Q
MUT.8347_HIST1H2BC	0.69	1.03	1.00E+00
MUT.2854_GPR32	1.03	1.03	1.00E+00
MUT.100310847_CLEC19A	0.00	1.03	1.00E+00
Mut.896_CCND3	0.69	1.03	1.00E+00
MUT.641977_SEPT13	0.69	1.03	1.00E+00
MUT.138009_DCAF4L2	1.37	1.03	1.00E+00

Table 3.4 Summary of the association between TP53 and MAP3K1 mutations and gene amplification status in highly proliferative luminal breast tumors

Gene	% AMP TP53 WT	% AMP TP53 MUT	FISHER Q	Gene	% AMP MAP3K1 WT	% AMP MAP3K1 MUT	FISHER Q
aCGH.152273_FGD5	23.40	22.00	0.988	aCGH.152273_FGD5	23.16	0.00	1.000
aCGH.131965_METTL6	23.40	20.00	0.988	aCGH.131965_METTL6	22.11	0.00	1.000
aCGH.196403_DTX3	23.40	30.00	0.988	aCGH.196403_DTX3	26.32	50.00	1.000
aCGH.51649_MRPS23	55.32	54.00	0.988	aCGH.51649_MRPS23	55.79	0.00	1.000
aCGH.1374_CPT1A	31.91	48.00	0.902	aCGH.1374_CPT1A	40.00	50.00	1.000
aCGH.3692_EIF6	36.17	60.00	0.711	aCGH.3692_EIF6	49.47	0.00	1.000
aCGH.8894_EIF2S2	34.04	52.00	0.828	aCGH.8894_EIF2S2	44.21	0.00	1.000
aCGH.81031_SLC2A10	34.04	70.00	0.465	aCGH.81031_SLC2A10	53.68	0.00	1.000

*** only 2 samples in this cohort have a MAP3K1 mutation***

Table 3.5 Summary of overall survival associated with candidate gene amplification

Gene Name	mRNA/DNA Correlation (TCGA dataset)	mRNA/DNA Correlation and Increased Amplification frequency (METABRIC)	P-value (log rank) TCGA	Hazard Ratio (TCGA)	P-value (log rank) METABRIC	Hazard Ratio (METABRIC)	AMP Gene Prognosis Luminal Breast Tumors
FGD5	**YES**	**YES**	**<0.0001**	**8**	**0.017**	**2**	**Significant in both datasets**
METTL6	**YES**	**YES**	**0.0003**	**5.9**	**0.0081**	**2.1**	**Significant in both datasets**
DTX3	**YES**	**YES**	**0.0387**	**2.6**	**0.0098**	**1.8**	**Significant in both datasets**
MRPS23	**YES**	**YES**	**0.0078**	**2.9**	**0.002**	**1.5**	**Significant in both datasets**
CPT1A	**YES**	**YES**	**0.3738**	**1.4**	**0.0991**	**1.2**	**Not significant**
EIF2S2	**YES**	**YES**	**0.7651**	**1.1**	**0.0053**	**1.5**	**Significant in METABRIC**
EIF6	**YES**	**YES**	**0.6286**	**1.2**	**0.0001**	**1.8**	**Significant in METABRIC**
SLC2A10	**YES**	**YES**	**0.9482**	**1**	**0.0103**	**1.4**	**Significant in METABRIC**
TRIO	YES	NO	0.0436	2.4	0.0594	1.4	Significant in TCGA
SNX21	YES	N/T[#]	0.1009	1.1	N/A	N/A	Not significant
DNAJC5	YES	N/T[#]	0.4064	1.3	N/A	N/A	Not significant
ZBTB46	YES	N/T[#]	0.6409	1.2	N/A	N/A	Not significant
CD200R1	NO	N/A	0.0129	3.3	0.8645	1	Significant in TCGA
PRDM9	NO	N/A	0.2114	1.7	0.0477	1.4	Significant in METABRIC
FGF3	NO	N/A	0.2899	1.4	0.0006	1.5	Significant in METABRIC
FGF19	NO	N/A	0.6053	1.3	0.0029	1.4	Significant in METABRIC
TMEM117	NO	N/A	0.0415	2.9	0.0022	2.6	Significant in both datasets
SEMA5A	NO	N/A	0.082	2.2	0.2211	1.2	Not significant
PMEPA1	NO	N/A	0.5068	1.2	0.0011	1.5	Significant in METABRIC
ANKRD56	N/T[†]	N/A	0.1827	2.3	0.1259	1.5	Not significant
TMEM189	N/T[†]	N/A	0.8868	1	0.0076	1.4	Significant in METABRIC

Twenty-one candidate genes were identified by the integrative analysis to have an increased amplification frequency in highly proliferative luminal tumors and were essential in the RNAi screen.
Of these 12 had a positive correlation between mRNA expression and DNA copy number status;
8 were found to have the same characteristics in the METABRIC dataset.
Genes that did not meet each criteria in the TCGA discovery dataset
were filtered out prior to testing in the METABRIC validation dataset.

BOLD: Candidate gene (n=8)
RED: Significant in both datasets
BLUE: Not significant
Black: Significant in one dataset

FIGURES

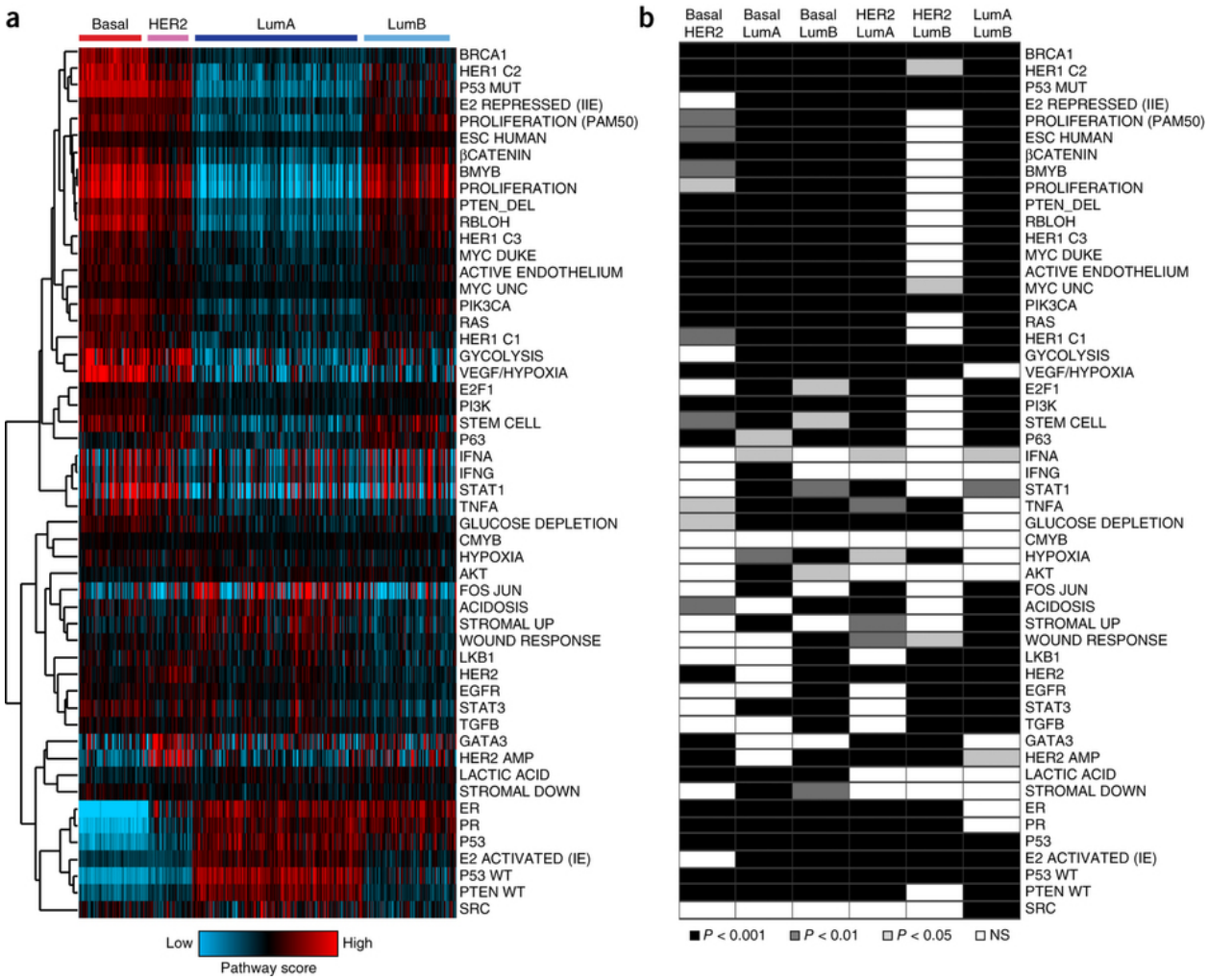

Figure 3.1 Patterns of genomic signature pathway activity in breast cancer. (a) Patterns of pathway activity (n = 52) were determined for each sample in the published TCGA breast cancer cohort (n = 476). Expression signature scores (y axis) are median centered and clustered by complete linkage hierarchical clustering. (b) ANOVA ($P < 0.0001$) for all signatures according to PAM50 subtype followed by Tukey's test for pairwise comparison demonstrates statistically significant differences in the levels of pathway expression between molecular subtypes. Box colors indicate the level of significance between subtypes, as indicated in the legend. NS, not significant.

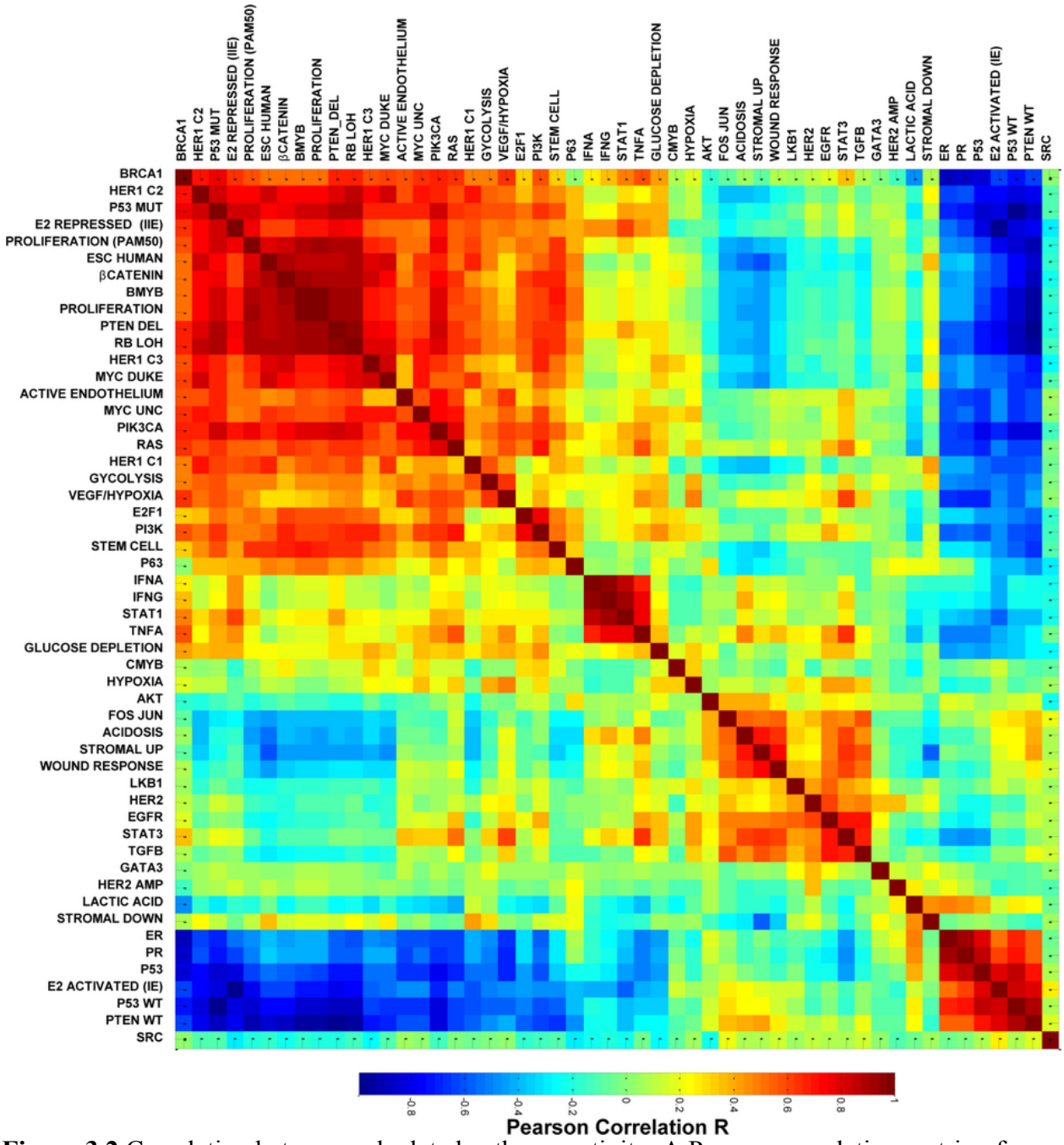

Figure 3.2 Correlation between calculated pathway activity. A Pearson correlation matrix of each signature versus all other signatures (including itself as the diagonal line) demonstrates a high degree of concordance amongst independently developed gene expression signatures measuring similar or associated pathways. Red indicates high positive correlation and blue a strong anti-correlation.

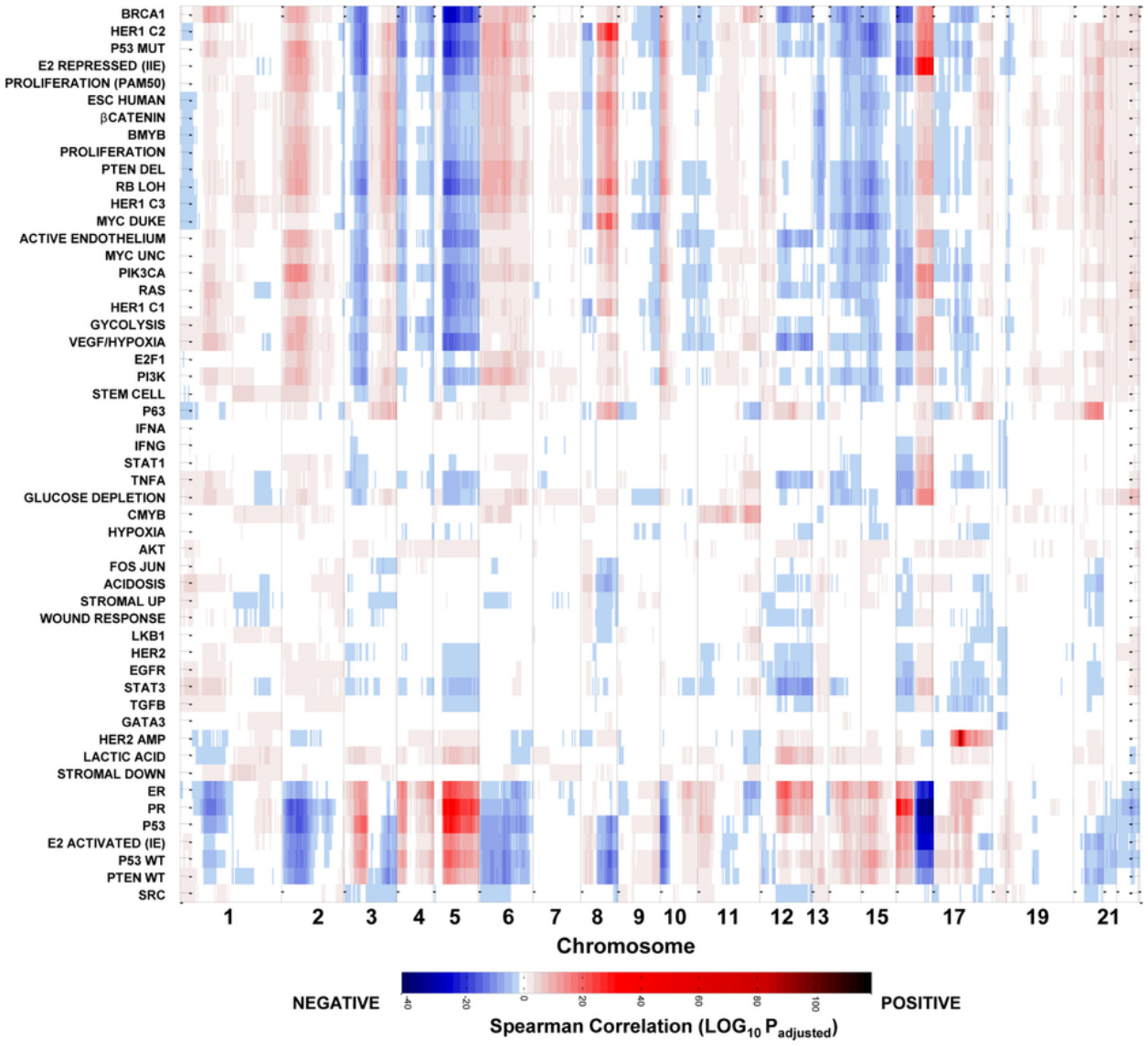

Figure 3.3 Identification of pathway-specific copy number alterations by Spearman Rank Correlation. A Spearman rank correlation, both positive (red) and negative (blue) were used to identify associations between predicted genomic signature pathway activity and gene-level DNA copy number content (n=476). The negative $\log_{10}$ Bonferroni adjust p-values are plotted according to chromosomal position. Chromosomal borders are delineated by vertical black lines.

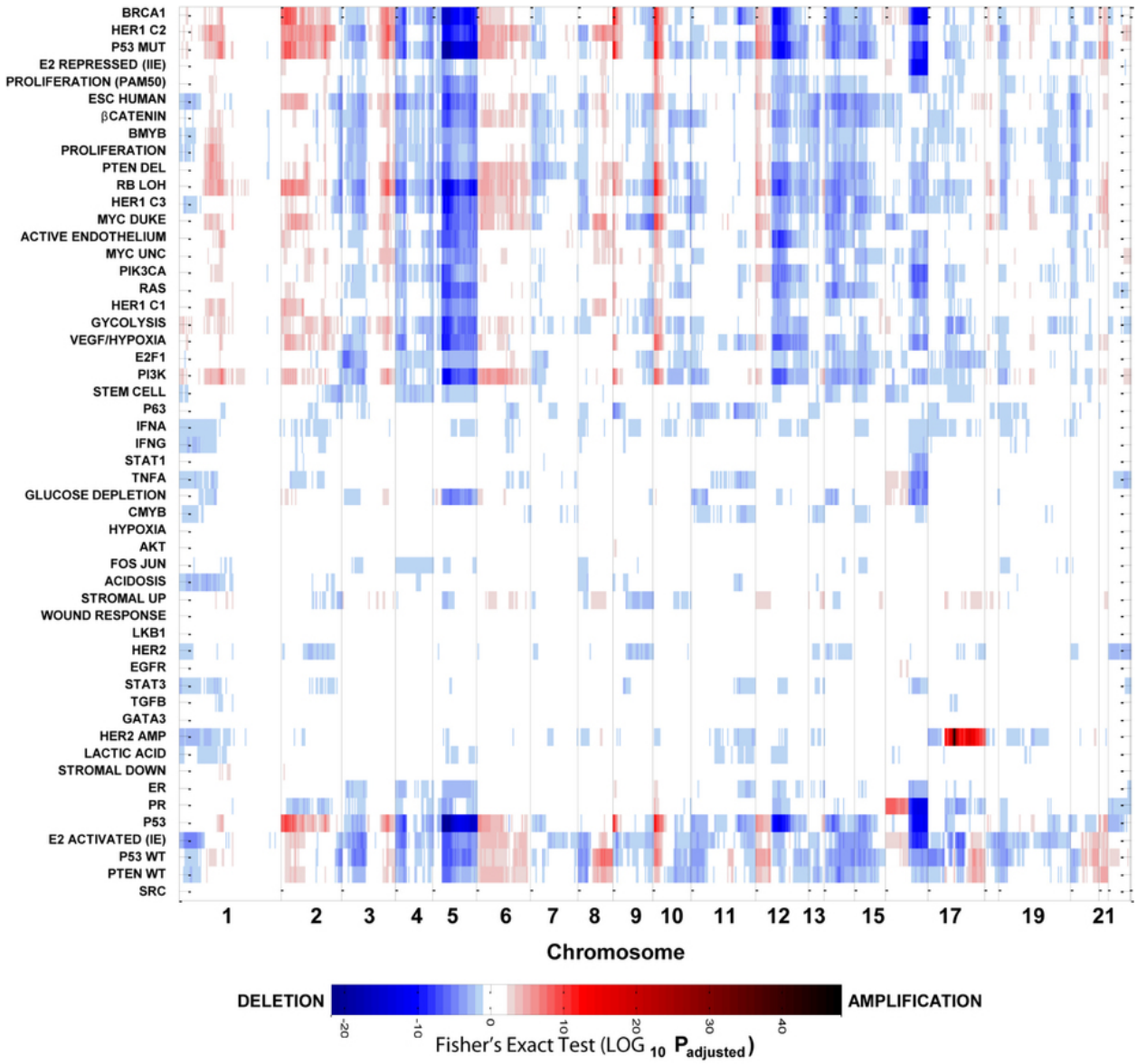

Figure 3.4 Identification of pathway-specific copy number alterations based on frequency of gains or losses calculated by Fisher's Exact test. A Fisher's exact test was used to calculate the statistical significance of the frequency of copy number gains (red) or losses (blue) in samples with the highest (top quartile) pathway signature activity relative to all other samples (n=476). The negative $\log_{10}$ Bonferroni adjust p-values are plotted according to chromosomal position. Vertical black lines indicate chromosomal borders.

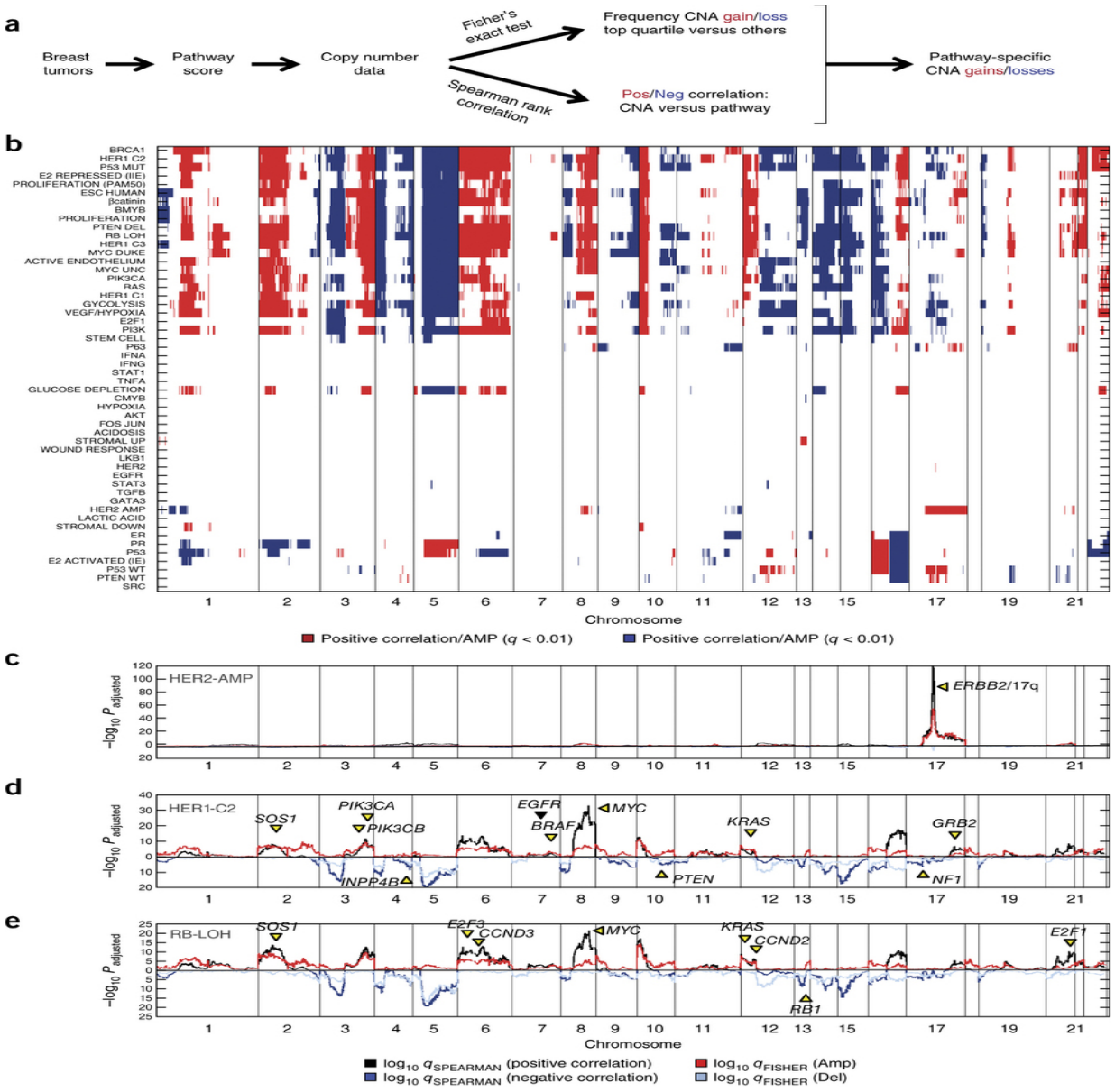

Figure 3.5 Identification of genomic pathway–specific CNAs. (a) Schematic outlining the strategy used to identify CNAs associated with pathway activity. Gain/loss indicates gains or losses; Pos/Neg indicates positive or negative. (b) For each signature, significant copy number gains and losses were calculated. The plot identifies those genes that had a positive Spearman rank correlation and increased amplification frequency ($q < 0.01$) (red) and those that had a negative Spearman rank correlation and an increased frequency of copy number losses in the top-scoring (top quartile) samples with pathway activity ($q < 0.01$) (blue). (c–e) Spearman rank correlation was used to identify genes positively (black line) or negatively (dark blue) associated with pathway activity, and Fisher's exact test was used to compare the frequency of copy number gains (Amp, red) or losses (Del, light blue) for the HER2-AMP (c), HER1-C2 (d) and RB-LOH (e) signatures. Yellow arrowheads indicate known pathway drivers with $q < 0.01$ for each analysis; the black arrowhead indicates $q < 0.01$ for a single analysis. In each figure, chromosomal boundaries are indicated by vertical black lines.

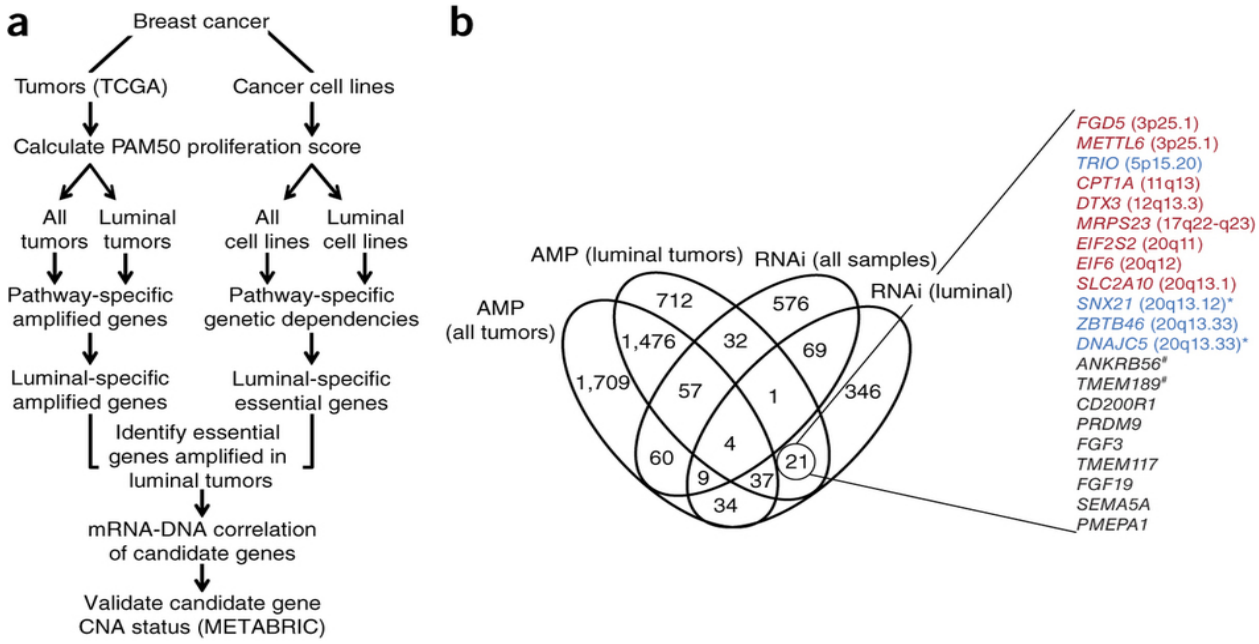

Figure 3.6 Identification of essential genes amplified in highly proliferative luminal tumors. (A) Schematic outlining the integrated genomic strategy to identify essential genes amplified in highly proliferative luminal breast tumors. (B) Identification of 21 genes in amplified loci that are unique to highly proliferative luminal tumors and are specifically required for luminal cell line proliferation *in vitro*. mRNA expression of genes in red and blue were significantly associated with CNA status, with the subset highlighted in red being further validated in the METABRIC dataset; genes in black do not show a significant mRNA-DNA correlation. Candidate genes demarcated by (*) are located at cusp of a CNA segment and were originally excluded, but mentioned here. Genes identified by (#) were not included on mRNA expression microarrays, and the correlation between DNA and mRNA expression was not assessed.

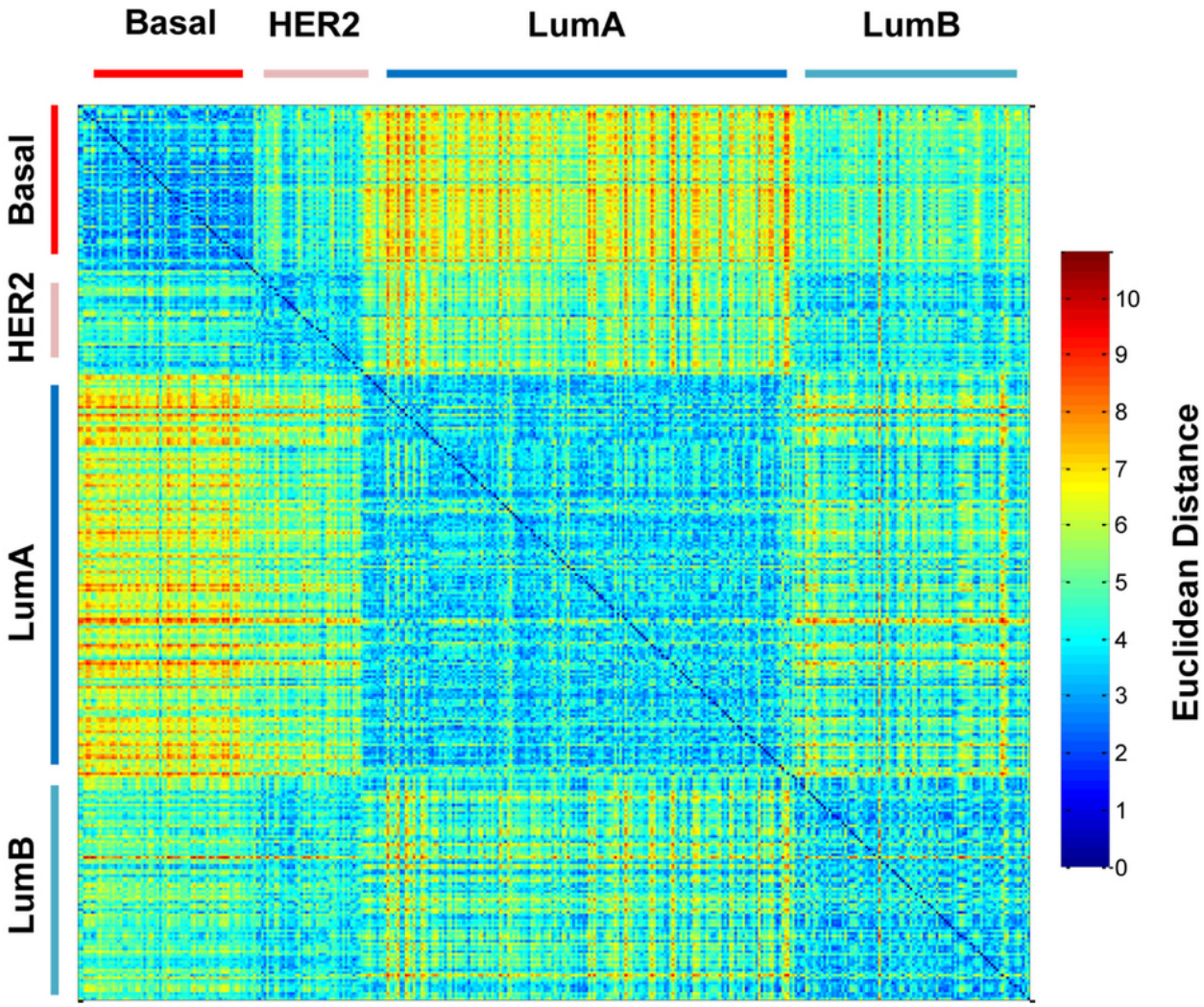

Figure 3.7 Patterns of pathway activity correspond with molecular subtypes of breast cancer. Analysis of molecular subtypes of breast cancer based on 52 gene expression signature scores. Euclidean distance was used to calculate the relationship between samples based on scores of 52 gene express signatures. Samples are commonly ordered on the X and Y axis according to molecular subtype. These results demonstrate high concordance within a subtype (dark blue), and lower concordance across subtypes; each sample versus itself is the blue diagonal line

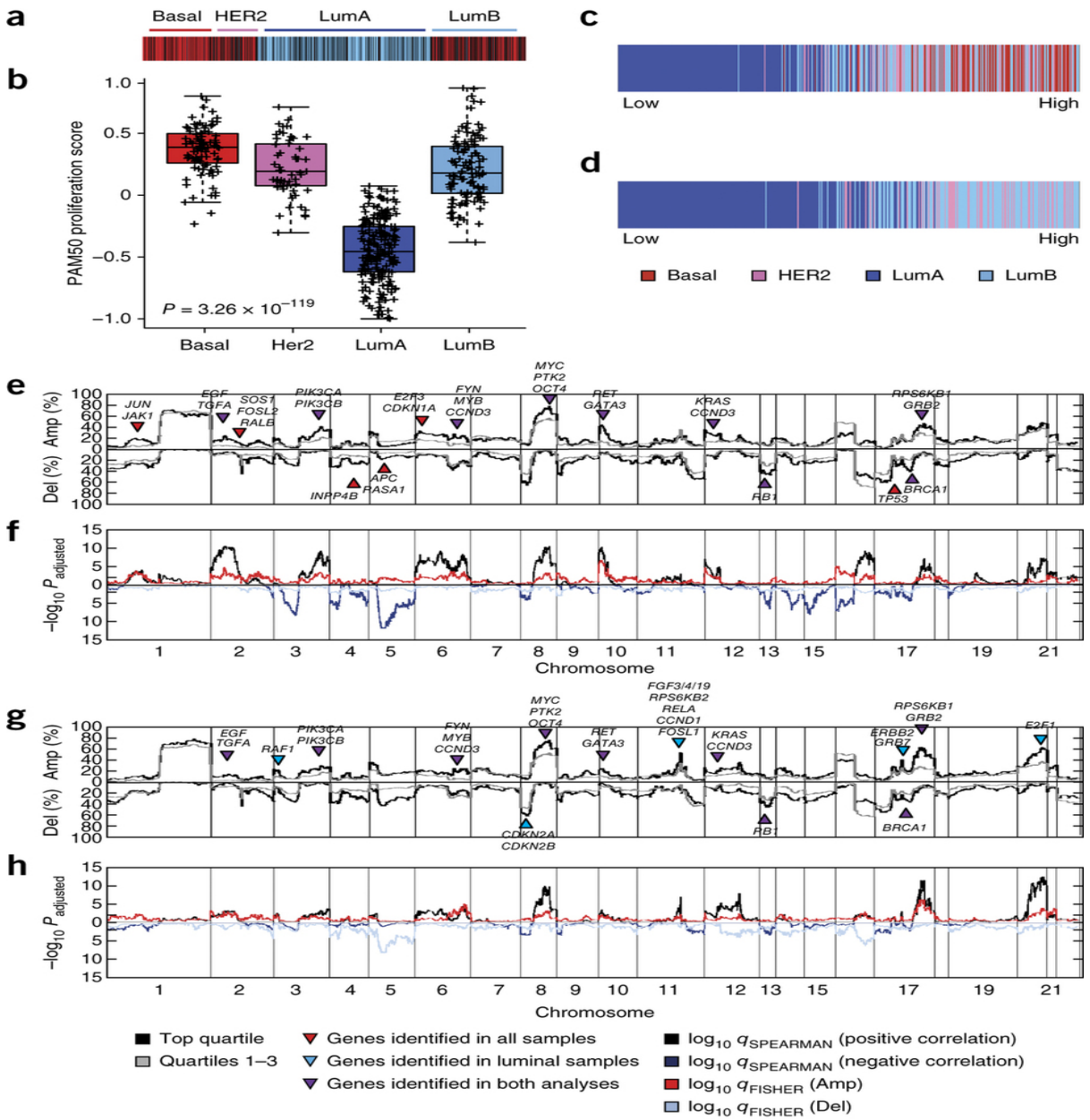

Figure 3.8 Identification of DNA copy number alterations in highly proliferative breast tumors. (A) Distribution of proliferation scores across all tumors and (B) by subtype. (B) Box and whisker plots indicate median score and the upper and lower quartile. Basal-like (n=88), HER2E (n=55), LumA (n=214) and LumB (n=119). (C) Highly proliferative tumors (top quartile) are comprised of Basal-like (49.6%), LumB (33.6%) and HER2E (16.8%). (D) Highly proliferative luminal tumors are restricted to LumB (68.0%) and HER2E (32.0%) samples. (E) Frequency of CNA in highly proliferative (black line) and all other samples (gray line). (F) Statistical analyses of CNA: positive correlation (black) and negative (dark blue) Spearman rank correlation and Fisher's exact test of amplification (red) or deletion (light blue) frequency. (G) Frequency of CNA in highly proliferative luminal tumors; color key same as (E). (H) Statistical analyses of CNA in proliferative luminal tumors; color key same as (F). Chromosomal boundaries in (E–H) are defined by vertical black lines.

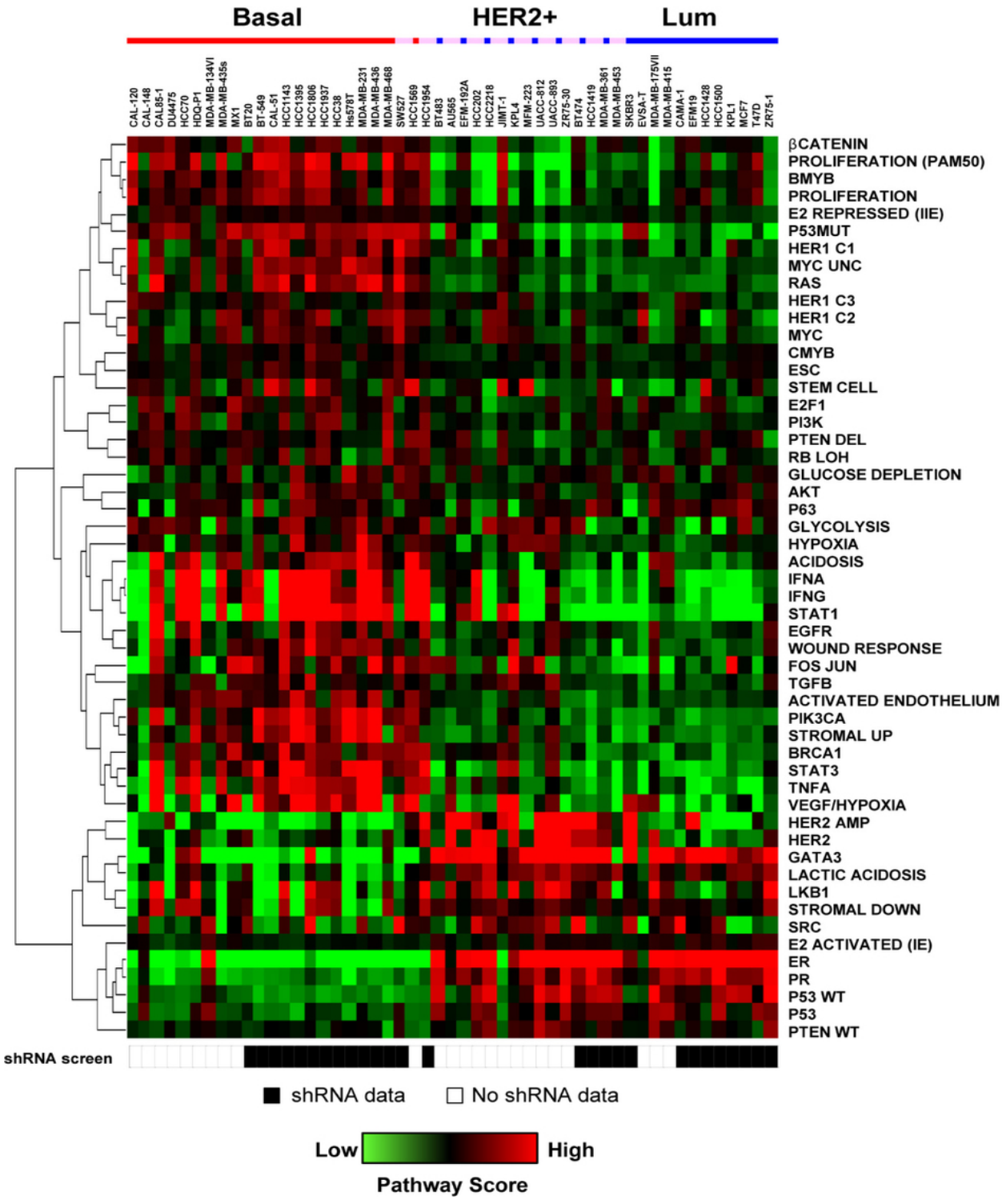

Figure 3.9 Patterns of pathway activity in human breast cancer cell lines. The scored pathway activity for a panel of 51 breast cancer cell lines (GSE12777) was calculated for the 52 pathway signatures. Of these cell lines, 27 which are denoted by black squares in lower panel were subjected to a genome-wide RNAi screen.

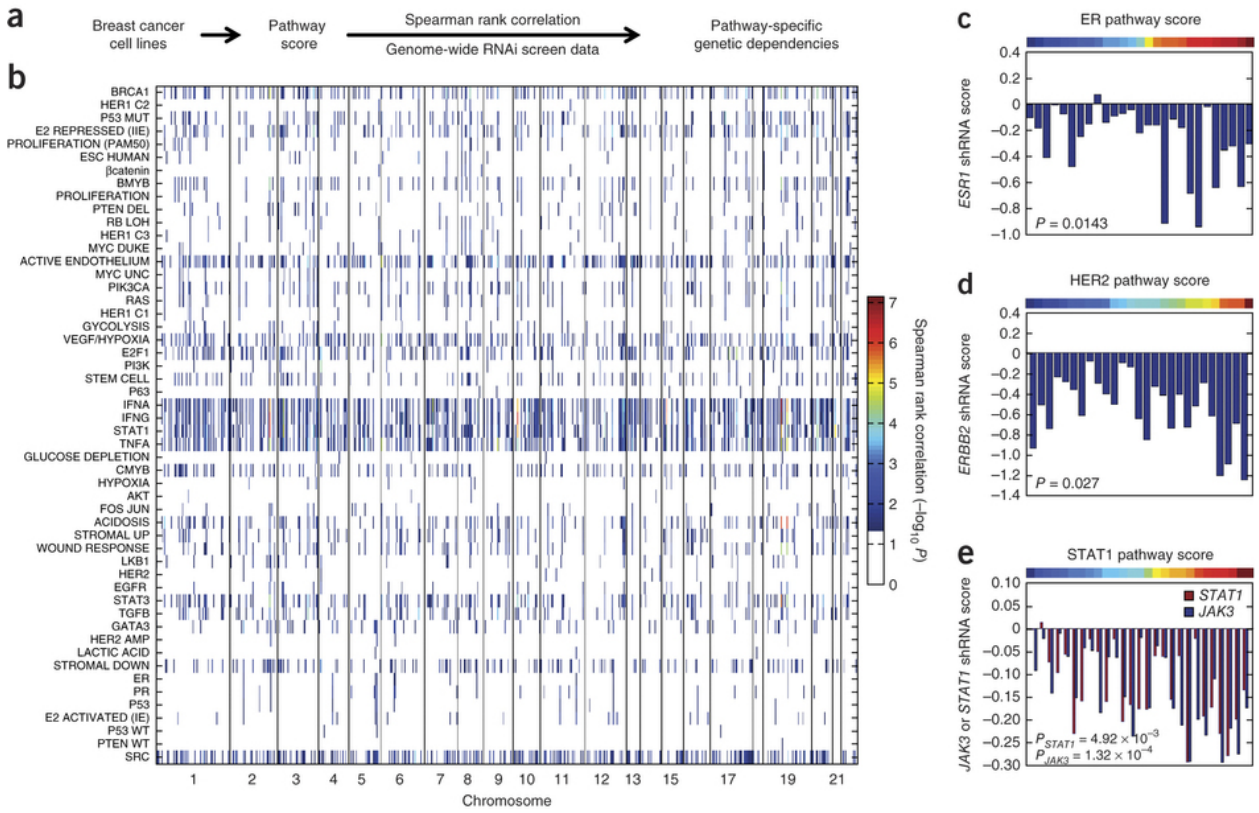

Figure 3.10 Identification of genomic pathway-associated essential genes in cell lines. (A) Schematic outlining strategy used to identify pathway-specific genetic dependencies. (B) A panel of 27 breast cancer cell lines with both expression data and data from a genome-wide RNAi screen was used to identify pathway-specific genes required for cell viability using a negative Spearman rank correlation (-log10 P-values plotted); significant genes (P<0.05) are shown according to chromosome location. Vertical black lines indicate chromosomal boundaries. (C) *ESR1* (D) *ERBB2* and (E) *STAT1* or *JAK3* shRNA levels are inversely associated with the ER, Her2 or Stat1pathway scores.

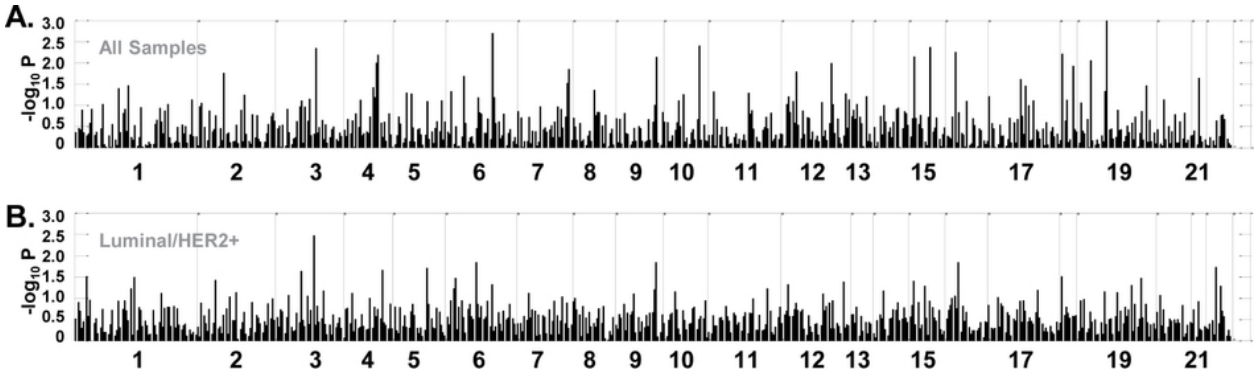

Figure 3.11 Identification of essential genes in proliferative breast cancer cell lines. Identification of genes essential for cell viability *in vitro* in the context of the 11-gene PAM50 Proliferation signature in (A) all cell line samples and (B) in luminal and HER2+ breast cancer cell lines. The negative $\log_{10}$ Spearman rank correlation p values are plotted for each gene relative to chromosomal position.

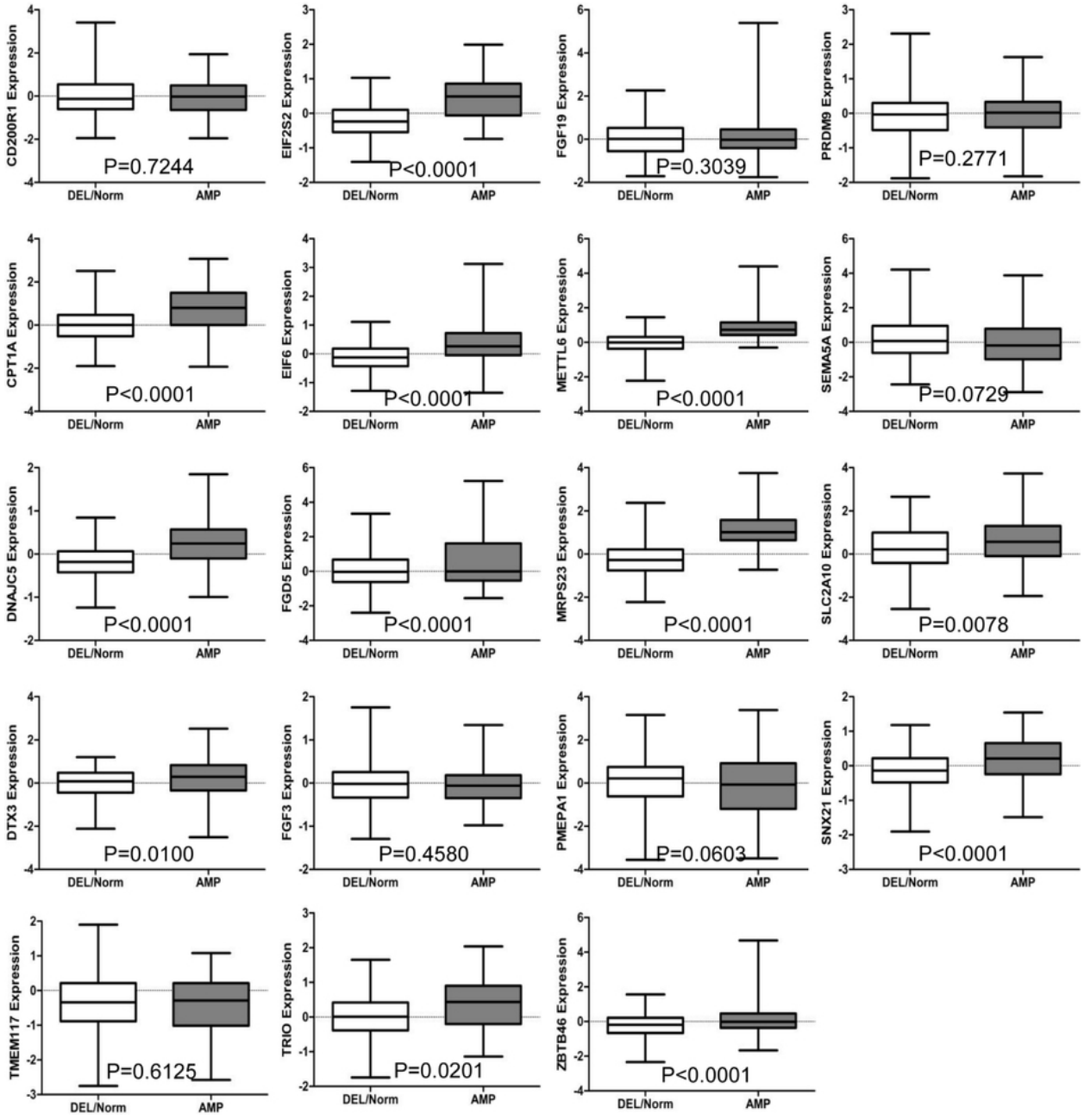

Figure 3.12 Correlation between candidate gene mRNA expression and DNA copy number status in TCGA samples. The mRNA expression levels of the 21 identified candidate genes that are required for cell viability and are uniquely amplified in highly proliferative luminal breast tumors. In each plot, the mRNA levels from the TCGA data are compared in those tumors with amplifications versus all others. Of the 21 genes, two (*ANKRD56* and *TMEM189*) were not present on the mRNA expression array and are not included here. Of the remaining 19 genes, 12 had a significant relationship ($p<0.05$) between copy number status and mRNA expression levels.

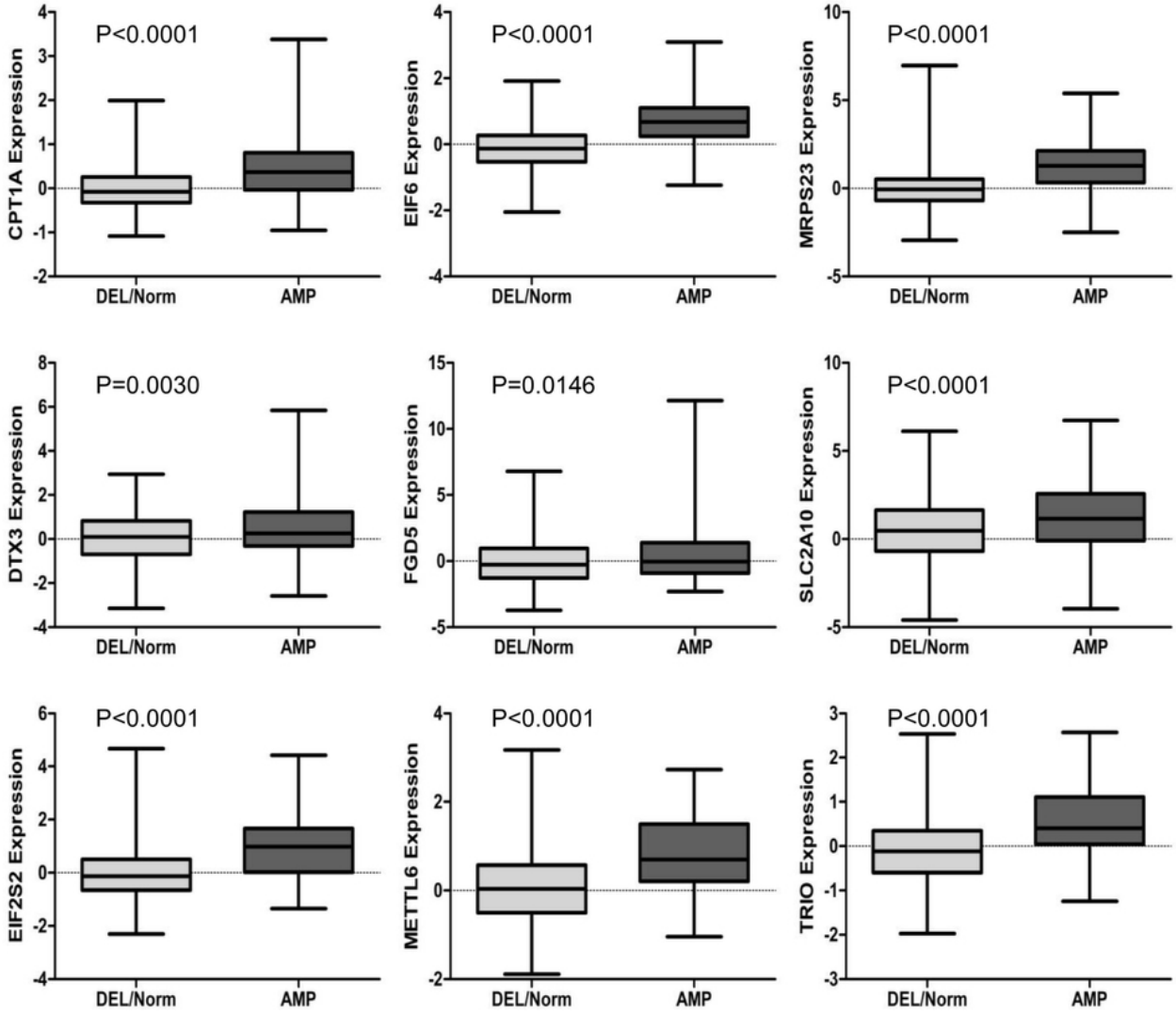

Figure 3.13 Correlation between candidate gene mRNA expression and DNA copy number status in METABRIC samples. The mRNA expression levels of the 12 identified candidate genes that are required for cell viability and are uniquely amplified in highly proliferative luminal breast tumors were analyzed within the context of copy number level in the METABRIC dataset. Of these 12 genes, three (*SNX21*, *ZBTB46* and *DNAJC5*) were not present on both of the METABRIC data platforms (mRNA expression and copy number). Of the remaining 9 genes, all had a significant relationship ($p<0.05$) between copy number status and mRNA expression levels

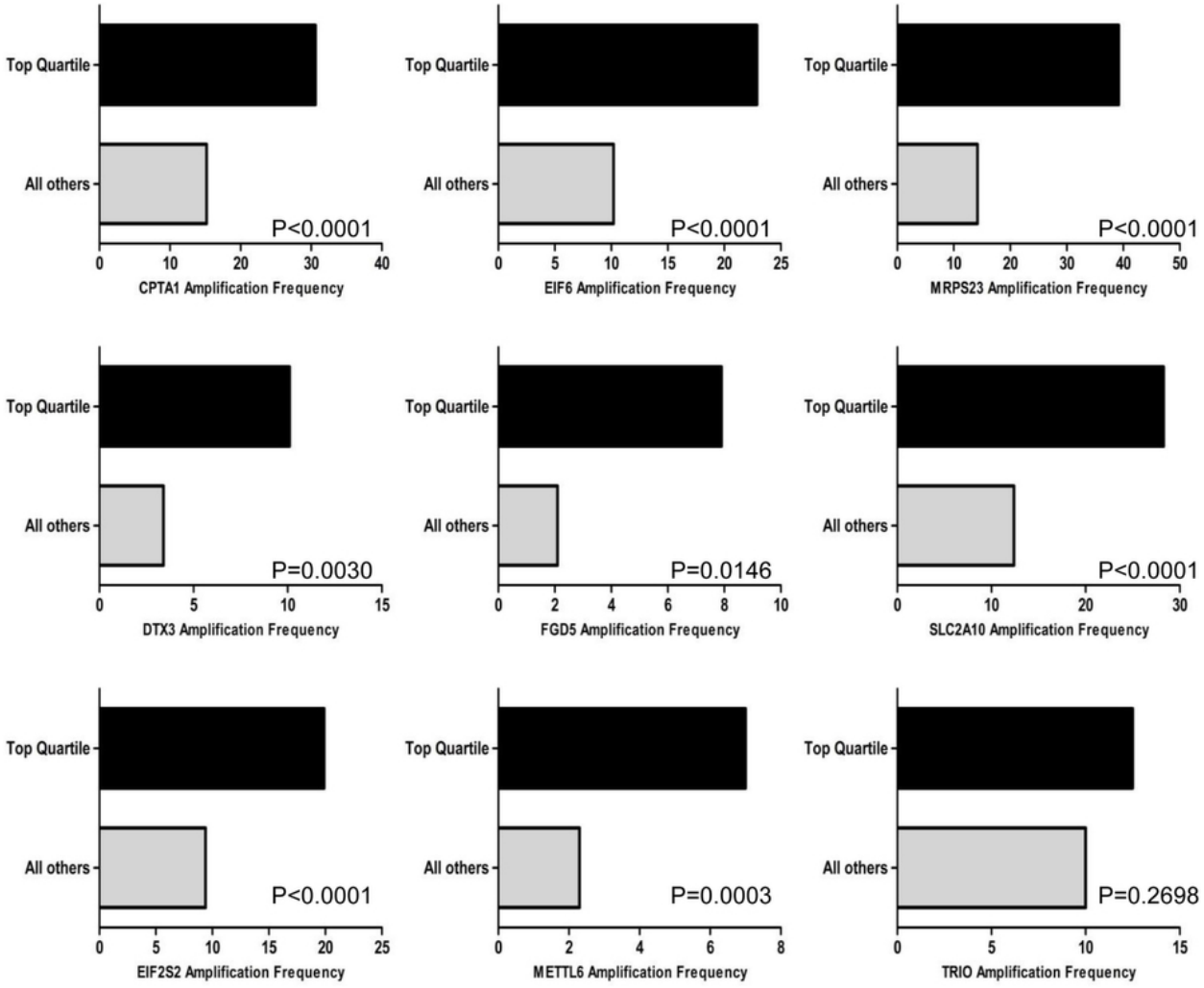

Figure 3.14 Validation of increased candidate gene copy number status in highly proliferative luminal breast tumors in METABRIC samples. The relationship between amplification of each candidate gene within the context of highly proliferative luminal breast tumors was examined in the METABRIC dataset. Of the nine candidate genes, eight showed a significant enrichment in highly proliferative (top quartile) luminal breast tumors.

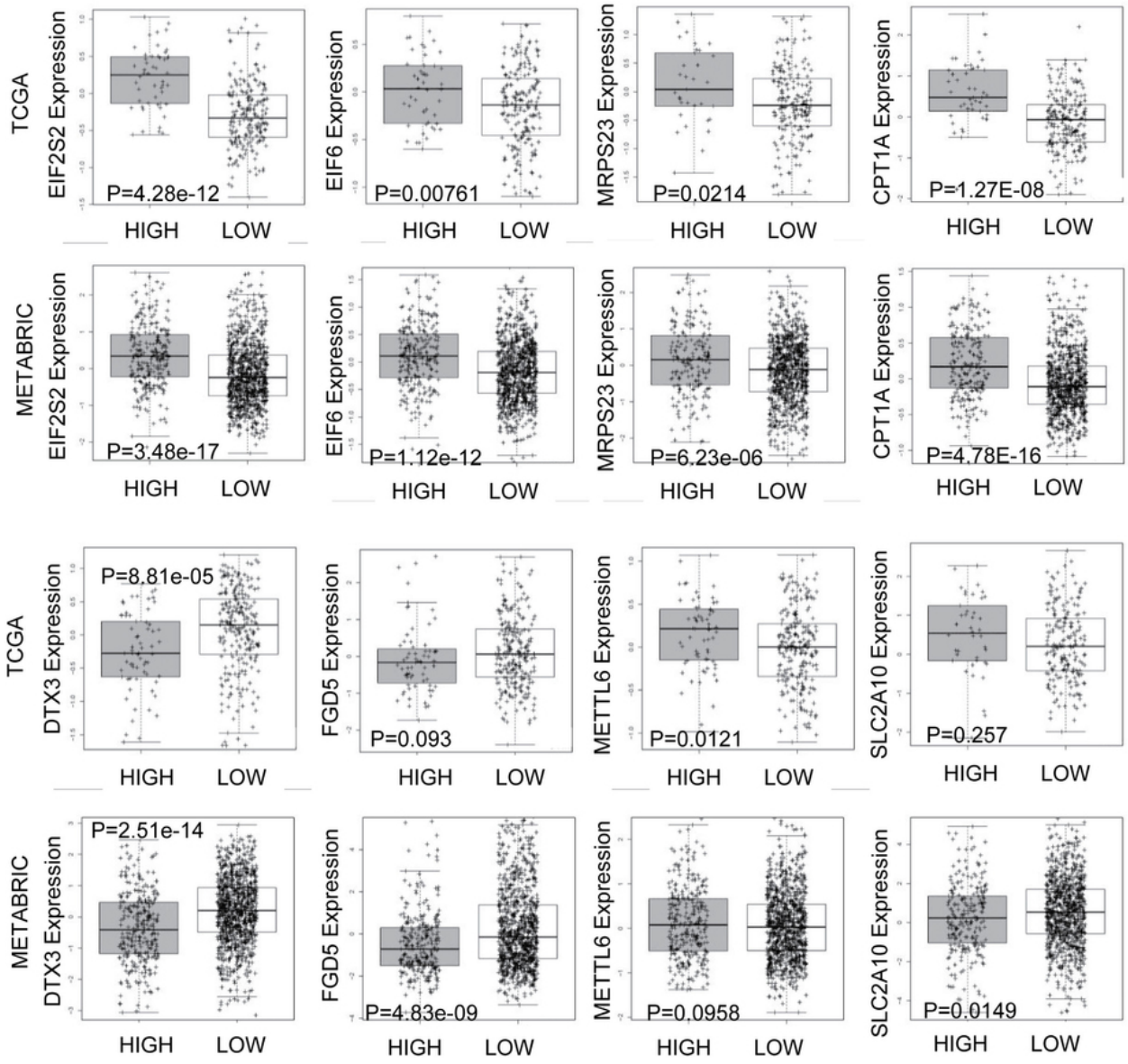

Figure 3.15 Candidate gene expression correlation with PAM50 Proliferation score independent of copy number status. The relationship between mRNA expression and the PAM50 proliferation signature was determined independent of copy number status (t-test) in the TCGA (n=388) and METABRIC (n=1,333) luminal/ ER+ subset of patients. Three classes of genes were identified (top rows) those that have a positive correlation with the signature score irrespective of CN status (EIF2S2, EIF6, MRPS23, CPT1A), those that have an inverse correlation (DTX3) and those that do not show a consistent pattern between datasets (FGD5, METTL6, SLC2A10)

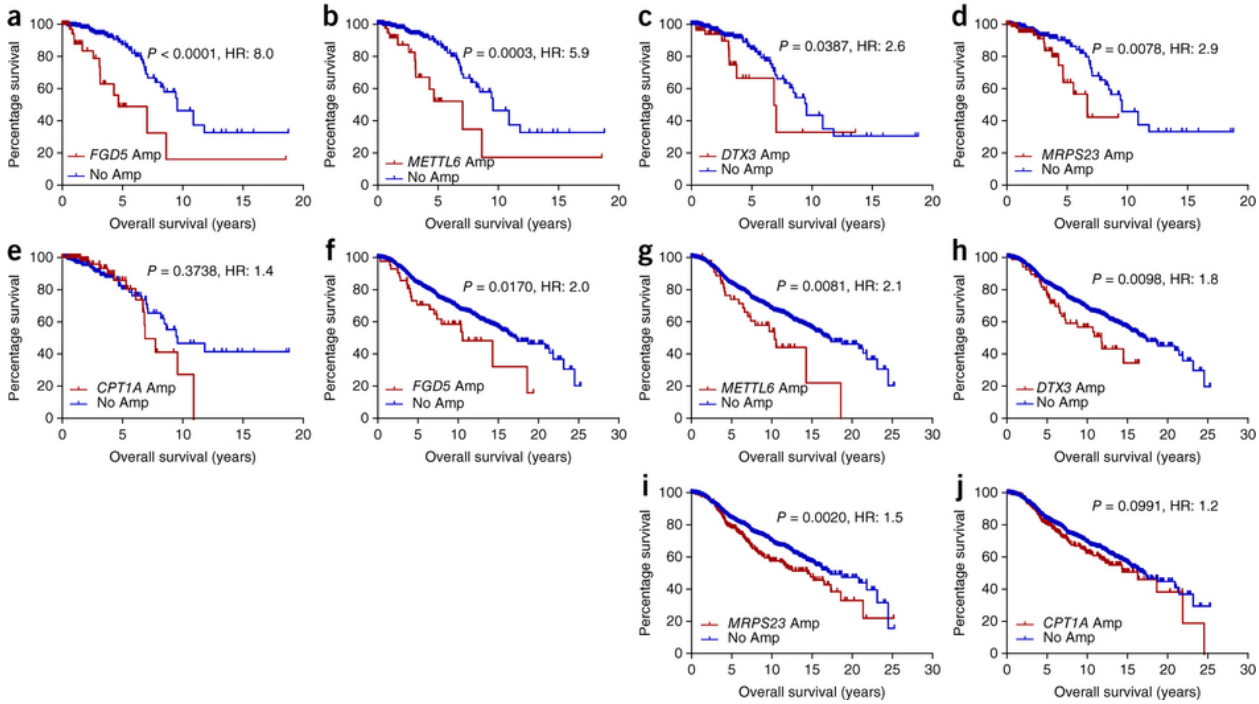

Figure 3.16 Candidate gene amplification correlates with a poor prognosis. Amplification of (A) *FGD5* (NAMP=51, NNoAMP=337), (B) *METTL6* (NAMP=51, NNoAMP=337), (C) *DTX3* (NAMP=71, NNoAMP=317) and (D) *MRSP23* (NAMP=127, NNoAMP=261) correlated with poor disease-specific outcome in the luminal breast cancer patients in the TCGA dataset (n=388) while (E) *CPT1A* (NAMP=111, NNoAMP=277) amplification had no effect on prognosis. Consistent results were observed in the METABRIC dataset (n=1,333) for (F) *FGD5* (NAMP=42, NNoAMP=1,218), (G) *METTL6* (NAMP=44, NNoAMP=1,278), (H) *DTX3* (NAMP=67, NNoAMP=1,266), (I) *MRPS23* (NAMP=266, NNoAMP=1,062) and (J) *CPT1A* (NAMP=241, NNoAMP=1,029). Samples in the METABRIC dataset missing CNA calls were excluded. For each analysis, P-value determined by log-rank test and Hazard Ratio (HR) are reported.

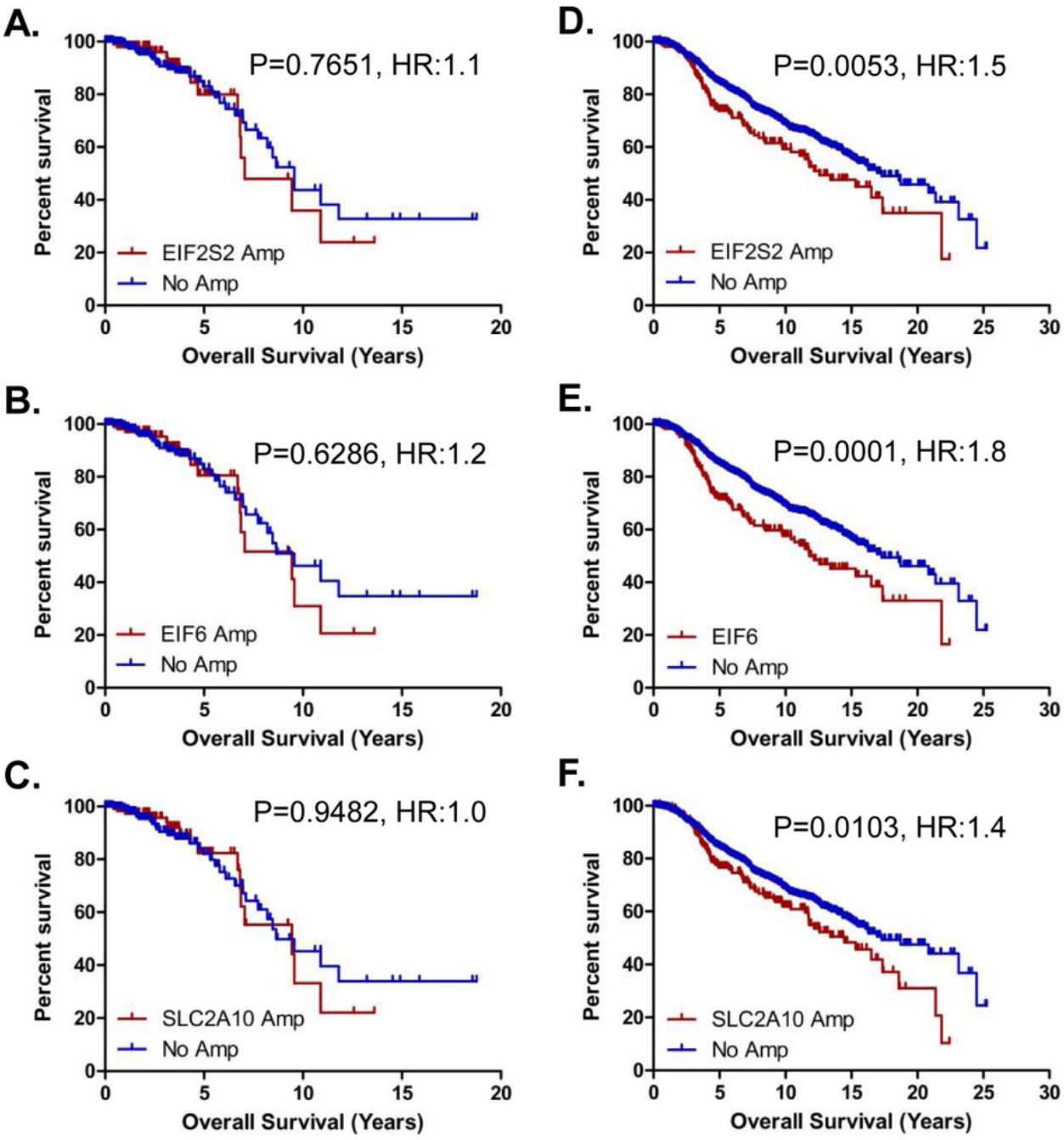

Figure 3.17 Amplification status of a subset of candidate genes has no reproducible effect on prognosis. Kaplan-Meier survival analysis based upon the amplification status of highly proliferative luminal tumor genes. No consistent difference in disease specific survival was observed for *EIF2S2* (A, D), *EIF6* (B, E) or *SLC2A10* (C, F) when comparing luminal tumors characterized by amplification of each candidate gene relative to luminal tumors without an amplification (log rank p>0.05) in the TCGA (A-C) and METABRIC (D-F) datasets.

REFERENCES

1. Perou, C.M. *et al.* Molecular portraits of human breast tumors. *Nature* **406**, 747–752 (2000).

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

3. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

4. Wood, L.D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113 (2007).

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

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

7. Rhodes, D.R. *et al.* Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. *Neoplasia* **9**, 443–454 (2007).

8. Vogelstein, B. & Kinzler, K.W. Cancer genes and the pathways they control. *Nat. Med.* **10**, 789–799 (2004).

9. Marcotte, R. *et al.* Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov.* **2**, 172–189 (2012).

10. Gatza, M.L. *et al.* Analysis of tumor environmental response and oncogenic pathway activation identifies distinct basal and luminal features in HER2-related breast tumor subtypes. *Breast Cancer Res.* **13**, R62 (2011).

11. Gatza, M.L. *et al.* A pathway-based classification of human breast cancer. *Proc. Natl. Acad. Sci. USA* **107**, 6994–6999 (2010).

12. Fan, C. *et al.* Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. **BMC Med. Genomics** *4*, 3 (2011).

13. Hoadley, K.A. *et al.* EGFR associated expression profiles vary with breast tumor subtype. *BMC Genomics* **8**, 258 (2007).

14. Troester, M.A. *et al.* Gene expression patterns associated with p53 status in beast cancer. *BMC Cancer* **6**, 276 (2006).

15. Chandriani, S. *et al.* A core MYC gene expression signature is prominent in basal- like breast cancer but only partially overlaps the core serum response. *PLoS ONE* **4**, e6693 (2009).

16. Herschkowitz, J.I., He, X., Fan, C. & Perou, C.M. The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas. *Breast Cancer Res.* **10**, R75 (2008).

17. Hu, Z. *et al.* A compact VEGF signature associated with distant metastases and poor outcomes. *BMC Med.* **7**, 9 (2009).

18. Hutti, J.E. *et al.* Oncogenic PI3K mutations lead to NF-κB–dependent cytokine expression following growth factor deprivation. *Cancer Res.* **72**, 3260–3269 (2012).

19. Oh, D.S. *et al.* Estrogen-regulated genes predict survival in hormone receptor– positive breast cancers. *J. Clin. Oncol.* **24**, 1656–1664 (2006).

20. Thorner, A.R. *et al.* In vitro and in vivo analysis of B-Myb in basal-like breast cancer. *Oncogene* **28**, 742–751 (2009).

21. Thorner, A.R., Parker, J.S., Hoadley, K.A. & Perou, C.M. Potential tumor suppressor role for the c-Myb oncogene in luminal breast cancer. *PLoS ONE* **5**, e13073 (2010).

22. Troester, M.A. *et al.* Activation of host wound responses in breast cancer microenvironment. *Clin. Cancer Res.* **15**, 7020–7028 (2009).

23. Usary, J. *et al.* Mutation of GATA3 in human breast tumors. **Oncogene** *23*, 7669–7678 (2004).

24. Harrell, J.C. *et al.* Endothelial-like properties of claudin-low breast cancer cells promote tumor vascular permeability and metastasis. *Clin. Exp. Metastasis* **31**, 33–45 (2014).

25. Wong, D.J. *et al.* Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell Stem Cell* **2**, 333–344 (2008).

26. Ji, H. *et al.* LKB1 modulates lung cancer differentiation and metastasis. *Nature* **448**, 807–810 (2007).

27. Saal, L.H. *et al.* Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proc. Natl. Acad. Sci. USA* **104**, 7564–7569 (2007).

28. Glinsky, G.V., Berezovska, O. & Glinskii, A.B. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *J. Clin. Invest.* **115**, 1503–1521 (2005).

29. Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat. Med.* **15**, 907–913 (2009).

30. van 't Veer, L.J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).

31. Parker, J.S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).

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

33. Leone, G. *et al.* Myc requires distinct E2F activities to induce S phase and apoptosis. *Mol. Cell* **8**, 105–113 (2001).

34. Grandis, J.R. *et al.* Requirement of Stat3 but not Stat1 activation for epidermal growth factor receptor–mediated cell growth in vitro. *J. Clin. Invest.* **102**, 1385–1392 (1998).

35. Weigman, V.J. *et al.* 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 (2012).

36. Park, K., Kwak, K., Kim, J., Lim, S. & Han, S. c-Myc amplification is associated with HER2 amplification and closely linked with cell proliferation in tissue microarray of nonselected breast cancers. *Hum. Pathol.* **36**, 634–639 (2005).

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

38. Wirapati, P. *et al.* Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res.* **10**, R65 (2008).

39. Perreard, L. *et al.* Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. *Breast Cancer Res.* **8**, R23 (2006).

40. Hoadley, K.A. *et al.* Multi-platform integration of 12 cancer types reveals cell-of- origin classes with distinct molecular signatures. *Cell* **158**, 1–16 (2014).

41. Hoeflich, K.P. *et al.* In vivo antitumor activity of MEK and phosphatidylinositol 3-kinase inhibitors in basal-like breast cancer models. *Clin. Cancer Res.* **15**, 4649–4664 (2009).

42. Rody, A. *et al.* T-cell metagene predicts a favorable prognosis in estrogen receptor–negative and HER2-positive breast cancers. *Breast Cancer Res.* **11**, R15 (2009).

43. Kurogane, Y. *et al.* FGD5 mediates proangiogenic action of vascular endothelial growth factor in human vascular endothelial cells. *Arterioscler. Thromb. Vasc. Biol.* **32**, 988–996 (2012).

44. Kishi, N. *et al.* Murine homologs of deltex define a novel gene family involved in vertebrate Notch signaling and neurogenesis. *Int. J. Dev. Neurosci.* **19**, 21–35 (2001).

45. Matsuno, K., Diederich, R.J., Go, M.J., Blaumueller, C.M. & Artavanis-Tsakonas, S. Deltex acts as a positive regulator of Notch signaling through interactions with the Notch ankyrin repeats. *Development* **121**, 2633–2644 (1995).

46. Benelli, D., Cialfi, S., Pinzaglia, M., Talora, C. & Londei, P. The translation factor eIF6 is a Notch-dependent regulator of cell migration and invasion. *PLoS ONE* **7**, e32047 (2012).

47. Miluzio, A. *et al.* Impairment of cytoplasmic eIF6 activity restricts lymphomagenesis and tumor progression without affecting normal growth. *Cancer Cell* **19**, 765–775 (2011).

48. Lyng, H. *et al.* Gene expressions and copy numbers associated with metastatic phenotypes of uterine cervical cancer. *BMC Genomics* **7**, 268 (2006).

49. Tan, X.L. *et al.* Genetic variation predicting cisplatin cytotoxicity associated with overall survival in lung cancer patients receiving platinum-based chemotherapy. *Clin. Cancer Res.* **17**, 5801–5811 (2011).

50. Gandin, V. *et al.* Eukaryotic initiation factor 6 is rate-limiting in translation, growth and transformation. *Nature* **455**, 684–688 (2008).

51. Biffo, S. *et al.* Isolation of a novel β4 integrin–binding protein (p27(BBP)) highly expressed in epithelial cells. *J. Biol. Chem.* **272**, 30314–30321 (1997).

52. Shi, Z.Z. *et al.* Genomic alterations with impact on survival in esophageal squamous cell carcinoma identified by array comparative genomic hybridization. *Genes Chromosom. Cancer* **50**, 518–526 (2011).

53. Liu, L., Wang, Y.D., Wu, J., Cui, J. & Chen, T. Carnitine palmitoyltransferase 1A (CPT1A): a transcriptional target of PAX3-FKHR and mediates PAX3-FKHR–dependent motility in alveolar rhabdomyosarcoma cells. *BMC Cancer* **12**, 154 (2012).

54. Samudio, I. *et al.* Pharmacologic inhibition of fatty acid oxidation sensitizes human leukemia cells to apoptosis induction. *J. Clin. Invest.* **120**, 142–156 (2010).

55. Pacilli, A. *et al.* Carnitine-acyltransferase system inhibition, cancer cell death, and prevention of myc-induced lymphomagenesis. *J. Natl. Cancer Inst.* **105**, 489–498 (2013)

56. Reich, M. *et al*. GenePattern 2.0. *Nat. Genet.* **38**, 500–501 (2006).

CHAPTER 4

LOSS OF HETEROZYGOSITY AT THE CYP2D6 LOCUS IN BREAST CANCER: IMPLICATIONS FOR GERMLINE PHARMACOGENETIC STUDIES[2]

Background: Controversy exists regarding the impact of CYP2D6 genotype on tamoxifen responsiveness. We examined loss of heterozygosity (LOH) at the CYP2D6 locus and determined its impact on genotyping error when tumor tissue is used as a DNA source.

Methods: Genomic tumor data from the adjuvant and metastatic settings (The Cancer Genome Atlas [TCGA] and Foundation Medicine [FM]) were analyzed to characterize the impact of CYP2D6 copy number alterations (CNAs) and LOH on Hardy Weinberg equilibrium (HWE). Additionally, we analyzed CYP2D6 *4 genotype from formalin-fixed paraffin-embedded (FFPE) tumor blocks containing nonmalignant tissue and buccal (germline) samples from patients on the North Central Cancer Treatment Group (NCCTG) 89-30-52 tamoxifen trial. All statistical tests were two-sided.

Results: In TCGA samples (n =627), the CYP2D6 LOH rate was similar in estrogen receptor (ER)–positive (41.2%) and ER-negative (35.2%) but lower in HER2-positive tumors (15.1%) ($P < .001$). In FM ER+ samples (n = 290), similar LOH rates were observed (40.8%). In 190 NCCTG samples, the agreement between CYP2D6 genotypes derived from FFPE tumors

[2] Goetz, M. P. *et al.* Loss of heterozygosity at the CYP2D6 locus in breast cancer: implications for germline pharmacogenetic studies. *Journal of the National Cancer Institute* **107**, 2–9 (2015).

and FFPE tumors containing nonmalignant tissue was moderate (weighted Kappa = 0.74; 95% CI = 0.63 to 0.84). Comparing CYP2D6 genotypes derived from buccal cells to FFPE tumor DNA, CYP2D6*4 genotype was discordant in six of 31(19.4%). In contrast, there was no disagreement between CYP2D6 genotypes derived from buccal cells with FFPE tumors containing nonmalignant tissue.

Conclusions: LOH at the CYP2D6 locus is common in breast cancer, resulting in potential misclassification of germline CYP2D6 genotypes. Tumor DNA should not be used to determine germline CYP2D6 genotype without sensitive techniques to detect low frequency alleles and quality control procedures appropriate for somatic DNA.

INTRODUCTION

The CYP2D6 enzyme metabolizes tamoxifen to its active metabolites (4-hydroxy-tamoxifen and 4-hydroxy-N-desmethyl-tamoxifen [endoxifen]), and numerous studies have demonstrated that CYP2D6 genetic variants are associated with steady state endoxifen concentrations [1–2]. However, there is substantial controversy on the validity of CYP2D6 genotype as a predictor of benefit from tamoxifen therapy in the adjuvant setting (reviewed in [3]). Secondary analyses of adjuvant trials administering five years of tamoxifen (the North Central Cancer Treatment Group [NCCTG] 89-30-52 [4], Arimidex, tamoxifen, alone or in combination (ATAC) [5], BIG1-98 [6], and the Austrian Breast and Colorectal Cancer Study Group [ABCSG] 8 [7] have reached discrepant conclusions). Multiple investigators have voiced concern regarding the unprecedented departure of CYP2D6 allele frequencies from Hardy-Weinberg equilibrium (HWE) in the BIG 1-98 study [8–10]. While substantial departure from HWE was not observed in the ABCSG 8 analysis [7], some departure from HWE was observed with the CYP2D6*4 allele frequencies reported in the NCCTG 89-39-52 [4] and ATAC [5, 9] CYP2D6 analyses. Given previous demonstration of genomic instability at the chromosomal segment where CYP2D6 is located [11–12], it has been hypothesized that when tumor DNA is used for genotyping, the presence of tumor loss of heterozygosity (LOH) at the CYP2D6 locus distorts the frequencies of observed alleles, which could lead to an excessive homozygous assignment of the germline genotype [8–10]. To address this question, we undertook a detailed evaluation of whether somatic LOH occurs at the CYP2D6 locus by analyzing genomic tumor data from the adjuvant (The Cancer Genome Atlas [TCGA]) [13] and metastatic settings. Furthermore, we sought to determine whether CYP2D6 LOH could affect the accuracy of calling germline CYP2D6 genotypes. Finally, in a limited number of adjuvant cases in which both

formalin-fixed paraffin-embedded (FFPE) tumor blocks and buccal samples were available, we compared CYP2D6 *4 genotypes obtained from each DNA source.

MATERIALS AND METHODS

Samples. Three previously published data sets were analyzed. The first data set included tumors collected and annotated within The Cancer Genome Atlas breast dataset [13]. TCGA collected breast tumors from newly diagnosed patients who underwent surgical resection. Extensive quality control was employed to verify the presence of both tumor DNA and germline DNA. Briefly, each frozen primary tumor specimen had a companion normal tissue DNA specimen that was derived from blood components (including DNA extracted at the tissue source site) (n = 684), adjacent normal tissue taken from greater than 2 cm from the tumor (n = 76), or both (n = 65). Each hematoxylin and eosin (H&E) stained case was reviewed by a board-certified pathologist to confirm that the tumor specimen was histologically consistent with breast adenocarcinoma and the adjacent normal specimen contained no tumor cells. The tumor sections were required to contain an average of 60% of tumor cell nuclei with less than 20% necrosis for inclusion in the study per TCGA protocol requirements. The clinical characteristics of this cohort and the process for informed consent have been previously described [13].

The second set included paraffin-embedded blocks from 360 patients, with relapsed and metastatic ER+ (n = 261) or ER- (n = 99) breast cancers derived from a subset of patients from the NCT00780676 trial and from pathology departments of several medical centers, as recently described [14]. From these samples, CYP2D6 sequencing was performed by Foundation Medicine (FM). In addition, samples were stained for ER, progesterone receptor (PR), and human epidermal growth factor receptor–2 (HER2) and reviewed by a pathologist to confirm ER

positivity. All tissue collections were done with the approval of the corresponding institutional review boards, and the process for informed consent was previously published [14].

The third set included specimens from 190 ER-positive breast cancer case patients from the NCCTG 89-30-52 clinical trial [4]. In the initial reported CYP2D6 analysis, an H&E section was obtained from FFPE tumors and a board-certified pathologist identified the invasive component and DNA was extracted from a 1 cm area of highest tumor cellularity for both DNA [4] and RNA [15] studies. At a later date, the same tissue block was accessed and whole tissue sections containing both invasive and benign tissue were processed for DNA extraction as previously described [16–17]. Additionally, germline DNA from a buccal sample was collected and reported initially on 17 patients [4] and an additional 21 patients later provided buccal samples. All tissue collections were done with the approval of the corresponding institutional review boards, and the process for informed consent was previously published [4].

Genomic Analysis. For the TCGA cohort, DNA copy number at the CYP2D6 locus (Chr. 22: 42 522 501 – 42 525 911) was determined using the Affymetrix 6.0 single-nucleotide polymorphism (SNP) arrays [13] and copy number segmentation was performed using the Circular Binary Segmentation (CBS) algorithm version 1.12.0, as previously described [13]. Copy number segments of interest were identified as regions with intensity values greater than |0.3|. Frequency landscape plots of these segments were created using the SWITCHdna R-package plotting function [18]. Exome sequencing was performed as previously described [13]. Regions of LOH were identified using the Broad Institute's ABSOLUTE method on exome sequencing data and Affymetrix 6.0 SNP arrays [19]. LOH landscape frequency plots were created using modifications of SWITCHdna's plotting function. The percentage of overlap

between breast TCGA samples analyzed on SNP arrays and those through exome sequencing was 86%.

For the FM cohort, genomic DNA was extracted from 40 μm of FFPE tissue and up to 200ng of extracted DNA was sheared by sonication, followed by ligation of Illumina sequencing adaptors. Sequencing libraries were hybridization captured using RNA-based baits (Agilent), targeting a total of 3320 exons of 182 cancer-related genes and 78 polymorphisms in 34 ADME-related genes. Deep (>500x) paired-end sequencing (49 x 49 cycles) was performed using the HiSeq2000 (Illumina). Sequence reads were mapped to the reference human genome (hg19), analyzed for all classes of genomic alterations (substitutions, indels, and copy number alterations), using custom methods optimized for clinical tumor specimens with stromal admixture. Variant calls at the CYP2D6 locus were resolved into genotypes according to the star (*) allele nomenclature [20]. If the minor allele frequency was greater than 5%, the patient was considered to have germline heterozygosity. To determine tumor LOH at CYP2D6, a genome-wide copy number model was fitted to the coverage data at all sequenced exons and more than 1800 SNPs. This profile was segmented and interpreted using allele frequencies of sequenced SNPs to estimate tumor purity and copy number at each segment. Fitting was performed using Gibbs sampling, assigning total copy number and minor allele count to all segments. LOH was called if total copy number at the CYP2D6 locus was 1 (copy loss LOH), or if copy number was 2 or more with a minor allele count of 0 (copy neutral LOH). The distortion of the germline alternate allele frequency from 50% because of LOH is calculated. To assess the impact of LOH, we simulated low-sensitivity genotyping assays by requiring minor allele frequencies to have minimum levels of 10% and 20% before assigning genotypes as heterozygous. The estimate of potential error impact on genotyping methods was then estimated using the HWE test.

For the NCCTG samples, CYP2D6 genotyping (*3, *4, *6, *10, *41) was performed at the Mayo Clinic using the Applied Biosystems' Taqman Allelic Discrimination Assay (Foster City, CA), as previously described and reported in the context of a pooled analysis of NCCTG and Stuttgart patients [16] and as submitted to the International Tamoxifen Pharmacogenomics Consortium [21]. Analyses were performed irrespective of ethnicity.

Statistical Methods. Within the TCGA, a Pearson's Chi Square Test was used to determine whether LOH rates differed across intrinsic subtypes. Within the FM cohort, a two-sided Fisher's exact test was used to assess whether copy loss rate differed with respect to ER status. Within the NCCTG cohort, the extent of agreement between CYP2D6 genotypes derived from FFPE tumor and FFPE tumors containing nonmalignant tissue was assessed using weighted Kappa statistics and the corresponding 95% confidence interval. HWE tests were calculated using an exact test (the Simple Hardy-Weinberg Calculator by Michael H Court) (http://www.tufts.edu/~mcourt01/Documents/Court%20lab%20-%20HW%20 calculator.xls) by comparing the observed and expected genotype frequencies for case patients and control patients. All statistical tests were two-sided, and a P value of less than .05 was considered statistically significant.

RESULTS

TCGA Samples

Using SNP array data (n = 728) [13], evaluation of the CYP2D6 locus at chromosome 22 demonstrated copy number alterations (CNA) in 29.0% (n = 211) (Figure 4.1A). Among the 627 case patients with exome sequencing data, 219 case patients (34.9%) had LOH at the CYP2D6 locus (Figure 4.1B). While the CNA were higher for the ER-positive (35.0%) (Figure 4.2A)

compared with the ER-negative (12.0%) (Figure 4.2B), LOH rates were similar comparing ER-positive (41.2%) (Figure 4.3A) and ER-negative (35.2%) (Figure 4.3B). Analyzing according to intrinsic subtypes, LOH rates among the ER+ (luminal A [40.3%] luminal B [42.7%]) and basal-like (43.4%) subsets were similar but greater than that in the HER2-enriched subtype (15.1%) ($P < .001$, Pearson's Chi Square Test). For each of these subtypes, a "zoomed-in plot" of the region containing the CYP2D6 gene is indicated (Figure 4.4 and Figure 4.5). A further analysis within the clinically defined HER2+ subset demonstrated that LOH rates were lower within the ER-/HER2+ (14.3%) compared with ER+/ HER2+ (26.6%).

Foundation Medicine Samples

The findings among the case patients comprising the FM cohort were similar to those from the TCGA cohort, where 82 of 201(40.8%) and 23 of 89 (25.8%) of the ER+ and ER- case patients, respectively, had LOH at the CYP2D6 locus (Figure 4.6). While copy- neutral LOH was similar in both ER+ and ER- (18.9% and 19.1%, respectively), the copy loss rate among ER+ case patients was statistically significantly greater relative to ER- case patients (21.9% vs 6.7%; $P = .001$, two-sided Fisher's exact test). Given that standard genotyping assays (eg, Taqman) may not be able to detect an allele that is present at low frequency because of LOH, CYP2D6 genotypes were determined using next generation sequencing (Table 4.1) and the potential effect of LOH on CYP2D6 genotype was assessed (Table 4.2). Among the 105 case patients with LOH, a substantial fraction had a low frequency of one of the germline alleles: under 20% ($n = 27$), under 10% ($n = 7$). If such samples were assumed to be homozygous, this would result in excessive number of homozygotes and, statistically, departure from HWE (Table 4.2).

NCCTG 89-30-52 Samples

The original CYP2D6 *4 genotyping results were derived from tumor FFPE (FFPE-T) and demonstrated departure from HWE (chi square = 16.1, $P \leq .001$) [4]. These case patients (n = 190) were reassessed using FFPE sections containing nonmalignant tissue (FFPE-NM) [16]. For CYP2D6 *4, the agreement was moderate comparing CYP2D6 *4 genotypes derived from FFP-T with FFPE-NM (weighted Kappa 0.74; 95% CI = 0.63 to 0.84), resulting in excess homozygous genotypes and departure from HWE ($P < .001$). Specifically, 15 original homozygous wild-type (Wt/Wt) cases were reclassified as heterozygous for *4 (Wt/*4) and three homozygous variant (*4/*4) were reclassified as (Wt/*4). The *4 discrepancies among the remaining five cases were likely unexplained by LOH (Table 4.3). An evaluation for HWE using the genotyping data derived from FFPE-NM demonstrated that CYP2D6*4 is within HWE (chi square = 1.34, P = .25).

To further investigate the observed discrepancy between these results, the CYP2D6 genotypes derived from FFPE-T tumor [4] and FFPE-NM [16] were compared with CYP2D6*4 genotype derived from buccal cells (germline). Among the 31 case patients with both FFPE-T and buccal cells available for CYP2D6*4 genotyping, there were six (19.4%) cases of disagreement. In four of these six case patients, CYP2D6 *4 genotypes classified as homozygous wild-type using FFPE-T were determined to be heterozygous for *4 (Wt/*4) using DNA derived from buccal cells, and, in another case, a homozygous variant (*4/*4) based on FFPE-T was classified as (Wt/*4) using DNA from buccal cells. One of the errors appeared to be unrelated to LOH, as the tumor-derived genotype of *4/*4 was classified as Wt/Wt using buccal cells. In contrast, among the 35 case patients with DNA from both FFPE-NM and buccal cells, there was 100% agreement comparing CYP2D6 *4 genotypes from each source.

DISCUSSION

Using two large breast cancer datasets, we have demonstrated the presence of extensive LOH at the CYP2D6 locus in breast cancer. Furthermore, our data demonstrate that determination of germline CYP2D6 genotype using cancer tissue can result in substantial departure from HWE, as was seen in the original NCCTG CYP2D6 analysis [4], ATAC [5], and BIG 1-98 [6] studies. In the cohorts examined in this study, CYP2D6 genotyping using DNA extracted from FFPE-T blocks resulted in erroneous classification of up to 40% of CYP2D6*4 heterozygotes (intermediate metabolizers) as either extensive metabolizers or poor metabolizers.

Recently, Rae et al., in a cohort of 122 patients, extracted DNA from three 0.6-mm diameter cores obtained from FFPE breast tumor blocks as well as DNA derived from either normal lymph nodes or leukocytes [22]. Rae et al. used DNA from these sources to genotype for CYP2D6 and demonstrated a concordance rate of over 94% between these different sources, concluding that this modest quality control study was sufficient to support the use of breast cancer tissue for germline genotyping of CYP2D6 [22]. The results of our studies in this report clearly refute the conclusions of Rae and colleagues and provide further confirmation of the concerns raised by multiple authors [8–10] regarding the fidelity of the CYP2D6 genotyping performed in the context of the BIG 1-98 study [6].

Quality control procedures are critical for accurate genotyping. This includes a requirement to develop assays for all relevant variants, particularly for a locus as complex as CYP2D6 [23]. An additional critical aspect of quality control relates to the source of DNA used for germline genotyping. In ATAC [5], FFPE tumor blocks from the trans-ATAC tumor collection were used for DNA extraction. In BIG 1-98 [6], DNA was extracted from one or two 1

mm cores that were punched into an area of the FFPE block most representative of the invasive tumor component.

Given our observation of LOH at the chromosomal locus containing CYP2D6, it was critical to understand whether the use of tumor DNA could contribute to the observed departures from HWE. In the FM cohort, nearly one-third of the tumors with LOH had a frequency of the germline allele under 20%, suggesting that use of a low-sensitivity polymerase chain reaction (PCR) assay could result in misclassification of heterozygous CYP2D6 genotypes as homozygous. Therefore, we directly compared CYP2D6 genotyping results from different laboratories using DNA from the same patients. In the original publication of the NCCTG 89-30-52 clinical trial, CYP2D6 genotyping (using DNA extracted from tumors) was performed in the laboratory of Rae et al. at the University of Michigan [4]. When CYP2D6 genotyping was repeated at the Mayo Clinic using DNA derived from the same FFPE blocks but using whole tissue sections containing benign tissue, genotyping errors were identified, which appeared to be partially related to the lack of detection of low-frequency alleles in the 2005 analysis; however, additional discrepancies were observed that appear to be unrelated to LOH (Table 4.3). A full reanalysis of the NCCTG data set demonstrated that CYP2D6 genotypes met HWE, with complete agreement (35/35) between the updated genotype results with the germline (buccal) cells in those patients that provided a buccal sample. Furthermore, as previously reported, CYP2D6 genotype was statistically significantly associated with the risk of recurrence [16, 21].

In ATAC [5], the departure from HWE with regard to CYP2D6 *4 was similar in magnitude as observed in the original NCCTG CYP2D6 analysis (HWE $\chi 2 = 18.1$, $P = .000021$). While we are confident in our conclusions that LOH at the CYP2D6 locus is common in breast cancer and that the use of tumor DNA for CYP2D6 genotype results in misclassification of

germline CYP2D6 genotype, we were unable to reproduce the extreme departure from HWE observed in BIG 1-98 (P = 10-92) [6]. Stanton noted that if LOH was the sole cause of deviation from HWE in BIG 1-98, the distorted genotype frequencies could be normalized by adjusting for LOH [9]. Therefore, we agree with Stanton that the extreme departure from HWE in BIG 1-98 may be related to other factors, such as the use of nonstandard PCR techniques (use of upwards of 60 PCR cycles) [6].

Following the simultaneous publication of the CYP2D6 analyses of the ATAC and BIG 1-98 data sets, the authors of these studies argued that testing for CYP2D6 has no value in clinical practice, and an accompanying editorial concluded that this matter can be likely laid to rest [24]. However, our findings have validated the initial concerns raised by multiple investigators regarding genotyping error [8–10] and the conclusions that were generated based on these erroneous data. It is now clear that data from ongoing prospective clinical trials will be necessary to settle the debate on whether or not CYP2D6 genotyping can identify patients in whom tamoxifen would be would be an inferior therapy. However, until such data are available, clinicians and patients should be aware of the data generated from secondary analyses of prospective clinical trials that support the importance of both CYP2D6 genotype [7, 16] and endoxifen concentrations [25] and that these data fulfill the basic criteria of Simon et al. for a "prospective-retrospective" design in which the biomarker test is analytically and preanalytically validated for use with archived tissue [26].

An important finding within the TCGA CYP2D6 analysis was the observation of a substantially higher rate of LOH within the luminal A (40%), luminal B (43%), and basal-like subsets (40%), compared with the HER2-enriched (15%) and normal-like (8%) subtypes. Within the clinically defined HER2+ subset, LOH rates were lower within the ER-/HER2+ (14%)

compared with ER+/ HER2+ (27%). Within the FM cohort, the CYP2D6 loss rate among ER+ case patients was statistically significantly greater relative to ER- case patients While the biological relevance of these findings is unknown, the demonstration of substantial LOH at chromosome 22q13, the cytogenetic segment which contains the CYP2D6 gene, has been implicated in breast [11], colon [11, 27], and insulinomas [28], suggesting that a putative tumor suppressor gene in this region may be important in the pathogenesis of cancer, and particularly in the luminal and basal-like subtypes of breast cancer.

There are some limitations to our study. While we have demonstrated that the use of tumor-derived DNA contributes to CYP2D6 genotyping error (analytical validity), this is unlikely to be the only factor contributing to the heterogeneity in the tamoxifen CYP2D6 literature. In addition to "analytical validity," Simon et al. pointed out that an "adequate number of patients with archived tissue must be present," and suggested that the correlative study "include at least two-thirds of the total accrued patients" [26]. It should be noted that in the ATAC study, less than 19% of the patients receiving tamoxifen were analyzed with regard to CYP2D6 genotype. Lastly, Simon et al. pointed out the critical nature of "clinical validity" [26]. Here, it should be noted that the tamoxifen CYP2D6 literature contains variability in tamoxifen dosing (20-40 mg/day), duration of therapy (one to 10 years), ER status of the primary tumor, use of CYP2D6 inhibiting medications, and, finally, lack of control for drugs that alter the hazard for recurrence (chemotherapy and aromatase inhibitors) [21]. Therefore, we recommend careful control for each of these factors when analyzing and interpreting the tamoxifen CYP2D6 literature.

In summary, we have provided definitive data from independent data sets that over 40% of primary and metastatic breast tumors exhibit tumor LOH at the CYP2D6 locus and that the

use of standard PCR (eg, Taqman) genotyping techniques applied to purified tumor DNA to detect germline CYP2D6 variation results in genotyping error because of an excess number of homozygotes and departure from HWE. Based on these results, we recommend that CYP2D6 genotyping be repeated in those studies in which the use of tumor DNA to derive germline CYP2D6 genotype resulted in substantial departure from HWE. Furthermore, recommendations and/or guidelines for the use of CYP2D6 genotyping should not be derived from studies with evidence for genotyping error.

TABLES

Table 4.1 CYP2D6 allele frequencies determined by NGS in the Foundation Medicine cohort*

CYP2D6 allele	Enzyme activity	Count in FM cohort	Frequency in FM cohort, %	Expected frequency, % (20) (for Europeans)
*1 or *2	Normal (wild-type)	461	64.0	63.1
*4	None	120	16.7	17.2
*41	Reduced	65	9.0	7.0
*9	Reduced	15	2.1	2.5
*10	Reduced	14	1.9	2.9
*5(deletion of allele)	None	11	1.5	3.2
*6	None	7	1.0	0.6
*29	Reduced	7	1.0	0
*3	None	6	0.8	0.3
*17	Reduced	4	0.6	0
Other rare alleles	Various	10	1.4	0.4
Tandem duplications	Increased	Not assessed	Not assessed	2.8

* CYPD6 = cytochrome P450 2D6; FM = Foundation Medicine; NGS = next generation sequencing.

Table 4.2 The potential effects of CYP2D6 tumor LOH on Hardy Weinberg equilibrium

	Clinical subtype	HWE test: CYP2D6 *4	HWE test: CYP2D6 *41
NGS-based calls	ER+	0.005	0.19
NGS-based calls	ER-	0.95	0.75
NGS-based calls	All	0.02	0.11
*Germline allele <10%	ER+	7.2 x 10-4	0.16
*Germline allele <10%	ER-	0.78	0.75
*Germline allele <10%	All	3.2 x 10-4	0.09
*Germline allele <10%	ER+	1.2 x 10-6	0.04
*Germline allele <10%	ER-	0.018	0.75
*Germline allele <10%	All	8.3 x 10-8	0.02

* Hardy Weinberg equilibrium calculation assuming a low-sensitivity genotyping assay would misclassify the low frequency allele as homozygous. CYPD6 = cytochrome P450 2D6; ER = estrogen receptor; HWE = Hardy Weinberg equilibrium; LOH = loss of heterozygosity; NGS = next generation sequencing.

Table 4.3 CYP2D6*4 genotypes obtained from FFPE blocks enriched for tumor or benign tissues

CYP2D6 *4 genotype using tumor-enriched DNA(4)	CYP2D *4 genotype using DNA from tumors containing benign tissues (16)			No call on updated analysis	Total
	Wt/Wt	Wt/*4	*4/*4		
Wt/Wt	121	15	1		
Wt/*4	2	34	0		
*4/*4	2	3	8		
Total	125	52	9		

* CYPD6 = Cytochrome P450 2D6; FFPE = formalin-fixed paraffin-embedded; Wt = wild-type.

FIGURES

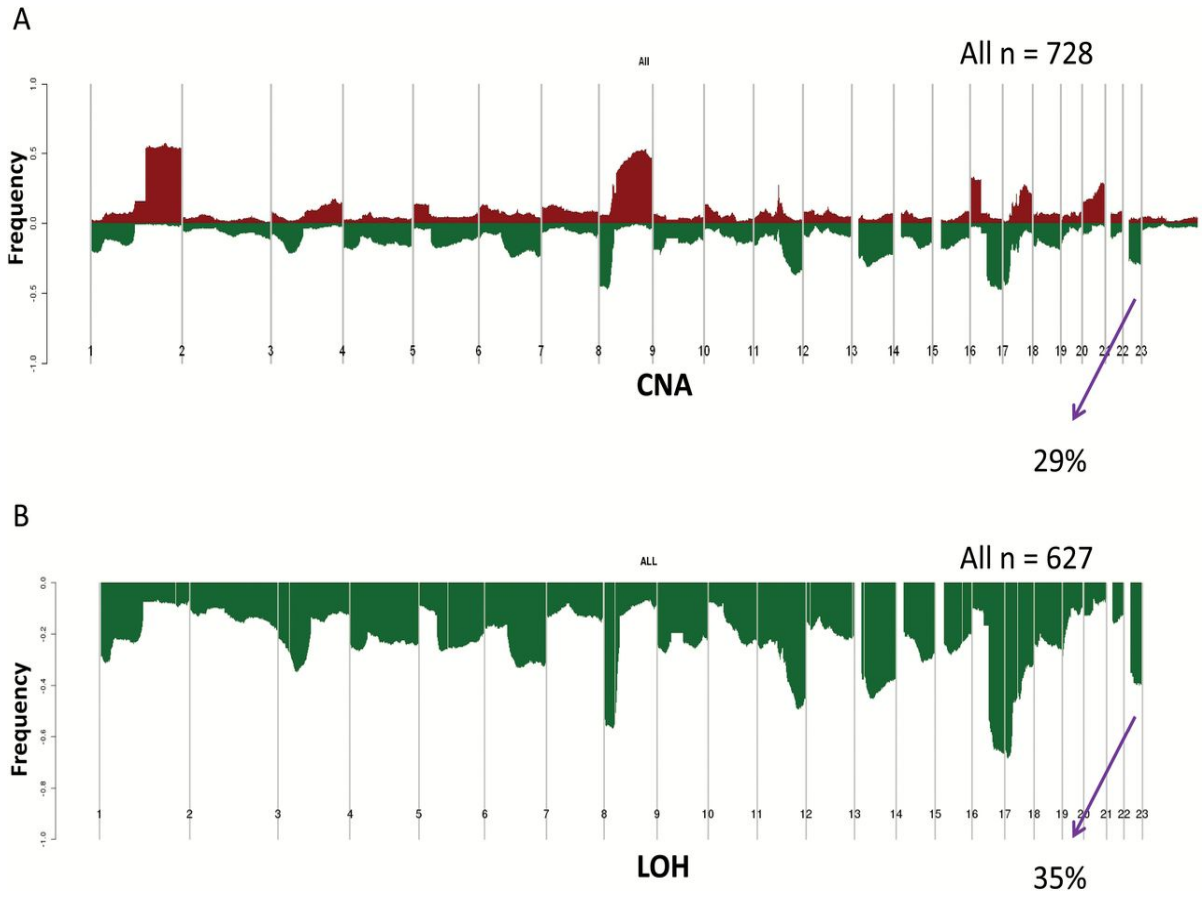

Figure 4.1 Cytochrome P450 2D6 (CYP2D6) copy number alterations (A) and loss of heterozygosity (B) within the entire Cancer Genome Atlas cohort. CNA = copy number alteration; LOH = loss of heterozygosity.

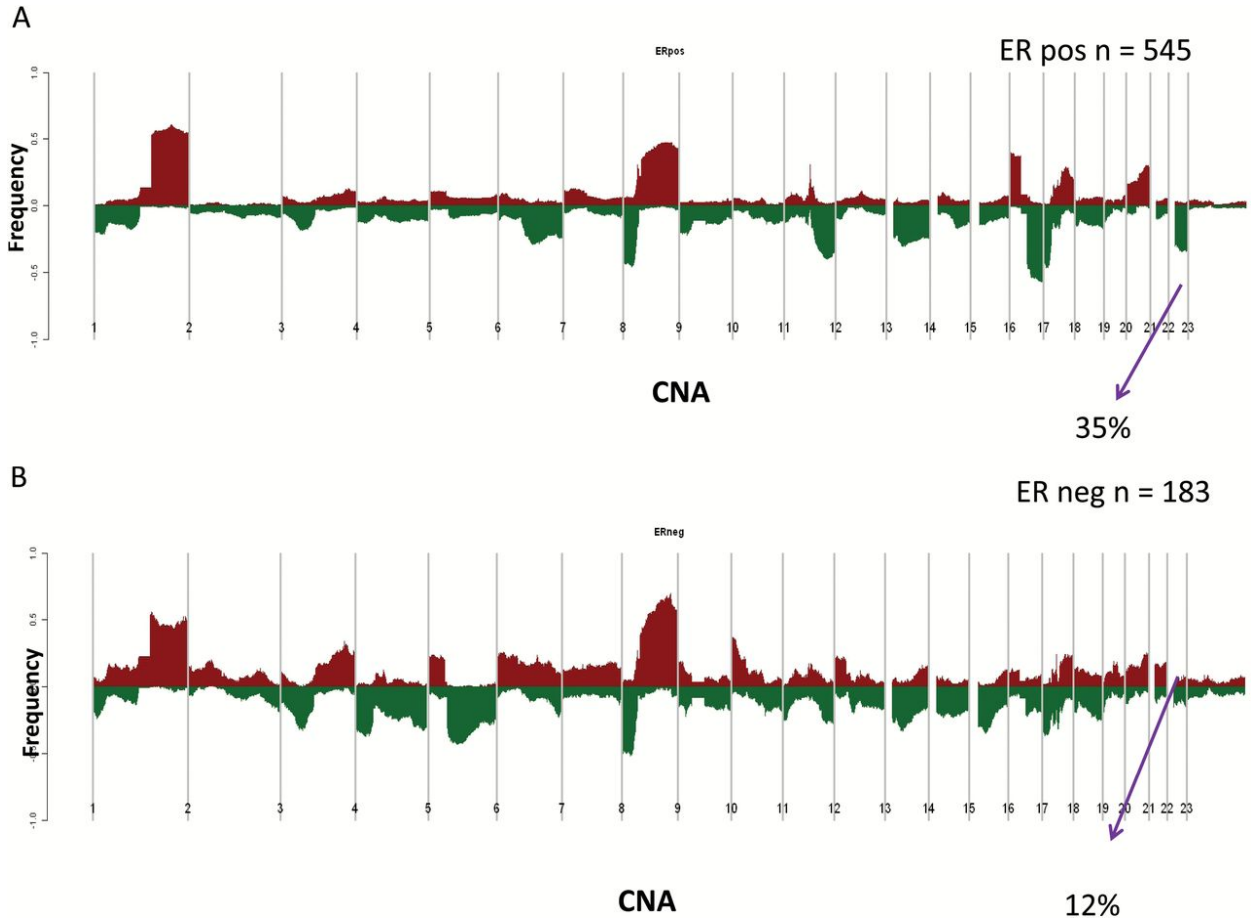

Figure 4.2. Cytochrome P450 2D6 (CYP2D6) copy number alterations within The Cancer Genome Atlas estrogen receptor (ER)–positive (A) and ER-negative (B) cohorts. CNA = copy number alteration; ER = estrogen receptor.

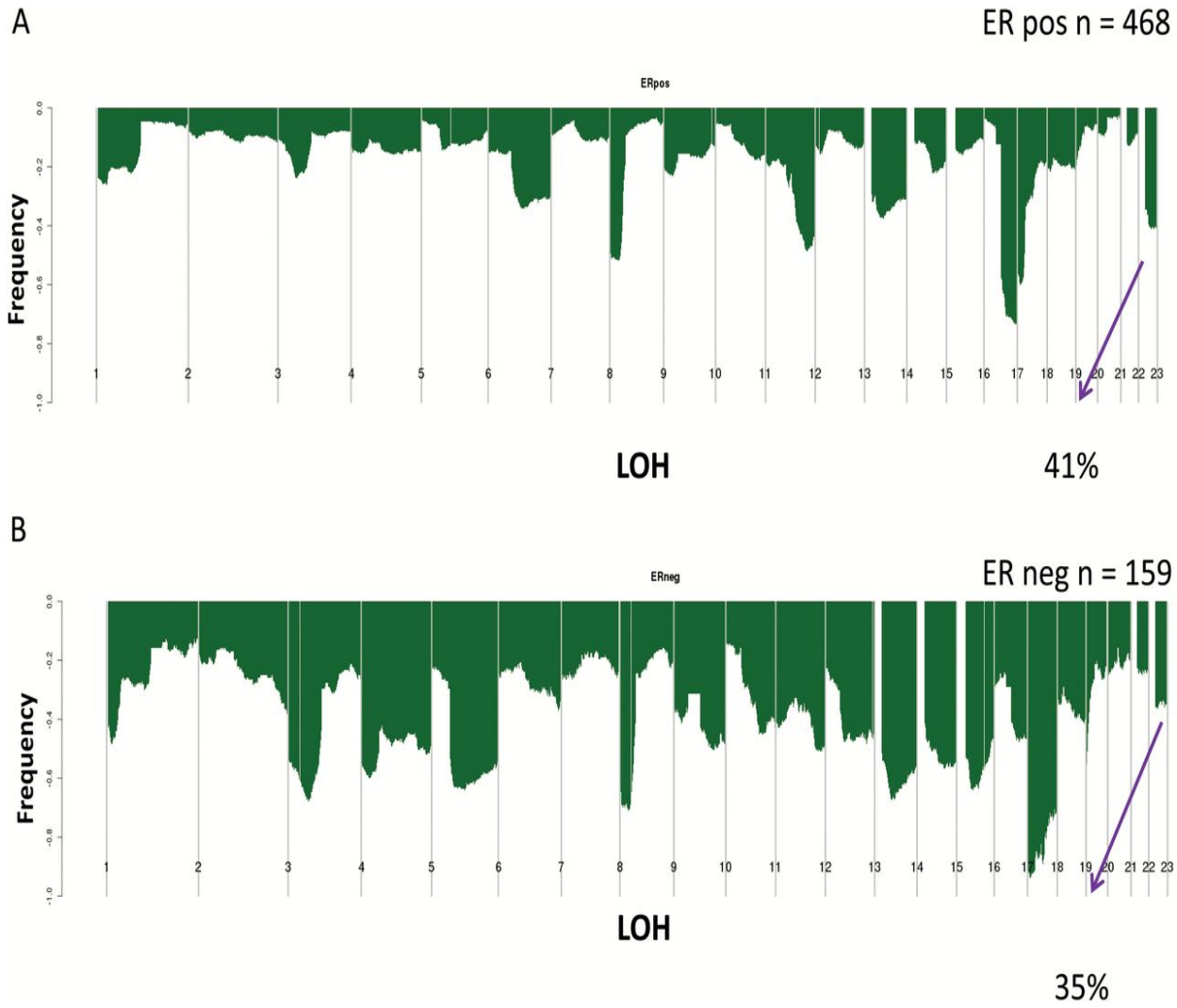

Figure 4.3 Cytochrome P450 2D6 (CYP2D6) loss of heterozygosity within The Cancer Genome Atlas estrogen receptor (ER)–positive (A) and ER-negative (B) cohorts. ER = estrogen receptor; LOH = loss of heterozygosity.

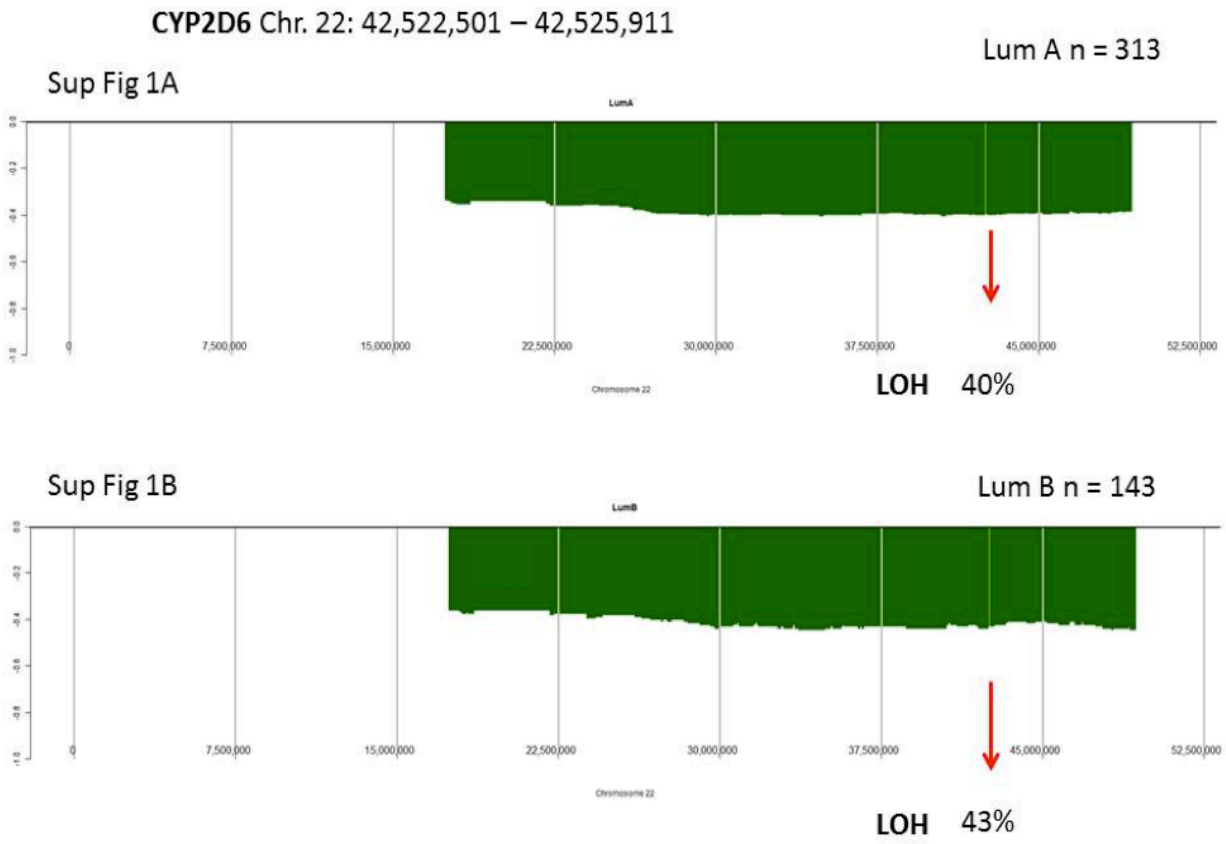

Figure 4.4 Cytochrome P450 2D6 (*CYP2D6)* Loss of Heterozygosity (LOH) rates according to intrinsic subtypes: Luminal A (A), Luminal B (B) within The Cancer Genome Atlas (TCGA) samples. The *CYP2D6* gene is indicated by a yellow line.

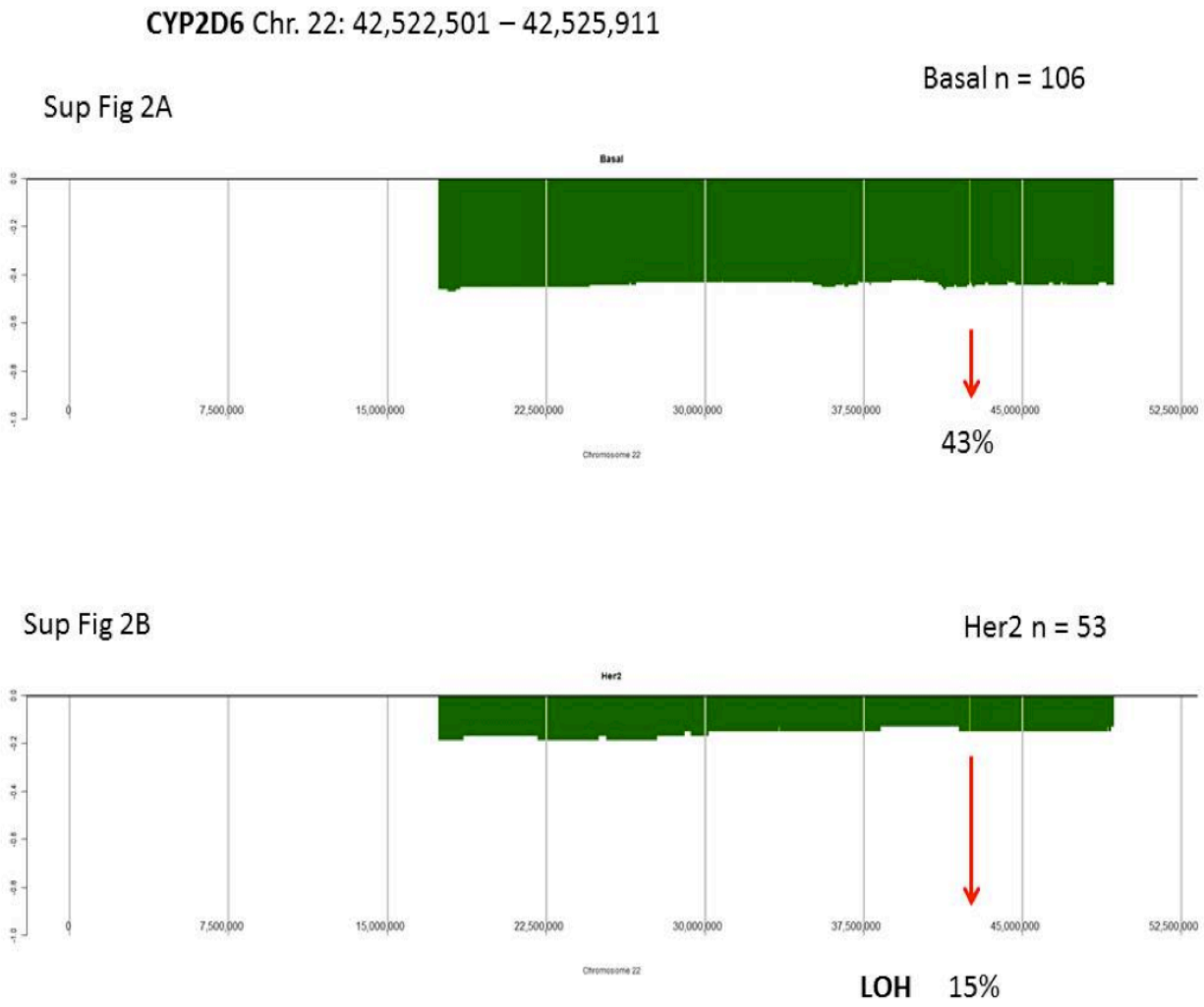

Figure 4.5 Cytochrome P450 2D6 (*CYP2D6)* Loss of Heterozygosity (LOH) rates according to intrinsic subtypes Basal (A) and HER2 (B) within The Cancer Genome Atlas (TCGA) samples. The *CYP2D6* gene is indicated by a yellow line.

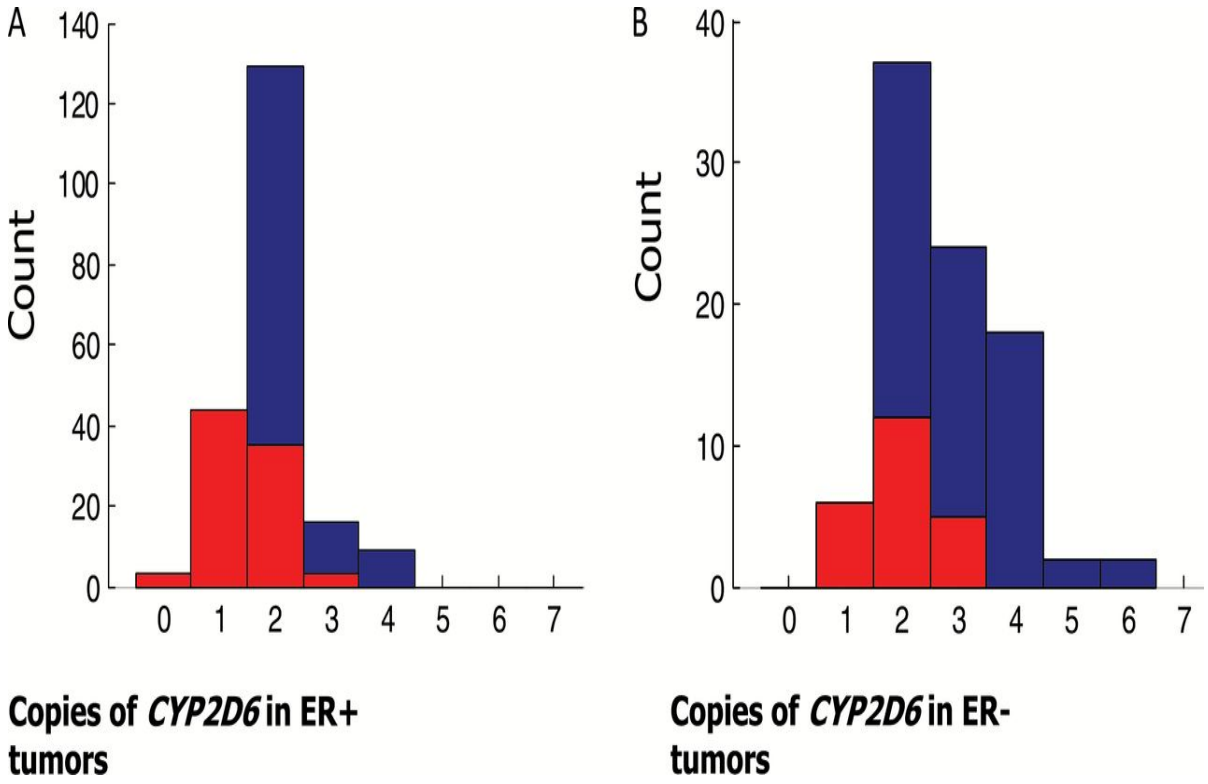

Figure 4.6 Frequency of Cytochrome P450 2D6 (CYP2D6) loss of heterozygosity within the Foundation Medicine estrogen receptor (ER)–positive (A) and ER-negative (B) cohorts. Tumor LOH is denoted in red. ER = estrogen receptor.

REFERENCES

1. Stearns V, Johnson MD, Rae JM, et al. Active tamoxifen metabolite plasma concentrations after coadministration of tamoxifen and the selective serotonin reuptake inhibitor paroxetine. *J Natl Cancer Inst.* 2003; **95(23)**:1758–1764.

2. Murdter TE, Schroth W, Bacchus-Gerybadze L, et al. Activ- ity levels of tamoxifen metabolites at the estrogen receptor and the impact of genetic polymorphisms of phase I and II enzymes on their concentration levels in plasma. *Clin Pharmacol Ther.* 2011; **89(5)**:708–717.

3. Brauch H, Schwab M. Prediction of tamoxifen outcome by genetic variation of CYP2D6 in post-menopausal women with early breast cancer. *Br J Clin Pharmacol.* 2014; **77(4)**:695–703.

4. Goetz MP, Rae JM, Suman VJ, et al. Pharmacogenetics of tamox- ifen biotransformation is associated with clinical outcomes of efficacy and hot flashes. *J Clin Oncol.* 2005; **23(36)**:9312–9318.

5. Rae JM, Drury S, Hayes DF, et al. CYP2D6 and UGT2B7 geno- type and risk of recurrence in tamoxifen-treated breast cancer patients. *J Natl Cancer Inst.* 2012; **104(6)**:452–460.

6. Regan MM, Leyland-Jones B, Bouzyk M, et al. CYP2D6 geno- type and tamoxifen response in postmenopausal women with endocrine-responsive breast cancer: the breast interna- tional group 1-98 trial. *J Natl Cancer Inst.* 2012; **104(6)**:441–451.

7. Goetz MP, Suman VJ, Hoskin TL, et al. CYP2D6 metabolism and patient outcome in the Austrian Breast and Colorectal Cancer Study Group trial (ABCSG) 8. *Clin Cancer Res.* 2013; **19(2)**:500–507.

8. Nakamura Y, Ratain MJ, Cox NJ, et al. Re: CYP2D6 genotype and tamoxifen response in postmenopausal women with endocrine-responsive breast cancer: the Breast International Group 1-98 trial. *J Natl Cancer Inst.* 2012; **104(16)**:1264; author reply 1266-1268.

9. Stanton V Jr. Re: CYP2D6 genotype and tamoxifen response in postmenopausal women with endocrine-responsive breast cancer: the Breast International Group 1-98 trial. *J Natl Cancer Inst.* 2012; **104(16)**:1265–1266; author reply 1266-1268.

10. Pharoah PD, Abraham J, Caldas C. Re: CYP2D6 genotype and tamoxifen response in postmenopausal women with endocrine-responsive breast cancer: the Breast International Group 1-98 trial and Re: CYP2D6 and UGT2B7 genotype and risk of recurrence in tamoxifen-treated breast cancer patients. *J Natl Cancer Inst.* 2012; **104(16)**:1263–1264; author reply 1266-1268.

11. Castells A, Gusella JF, Ramesh V, et al. A region of deletion on chromosome 22q13 is common to human breast and colorectal cancers. *Cancer Res.* 2000; **60(11)**:2836–2839.

12. Hirano A, Emi M, Tsuneizumi M, et al. Allelic losses of loci at 3p25.1, 8p22, 13q12, 17p13.3, and 22q13 correlate with postoperative recurrence in breast cancer. *Clin Cancer Res. 2001*; **7(4)**:876–882.

13. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012; **490(7418)**:61–70.

14. Jeselsohn R, Yelensky R, Buchwalter G, et al. Emergence of constitutively active estrogen receptor-alpha mutations in pretreated advanced estrogen receptor-positive breast cancer. *Clin Cancer Res.* 2014; **20(7)**:1757–1567.

15. Goetz MP, Suman VJ, Ingle JN, et al. A two-gene expression ratio of homeobox 13 and interleukin-17B receptor for pre- diction of recurrence and survival in women receiving adjuvant tamoxifen. *Clin Cancer Res.* 2006; **12(7)**:2080–2087.

16. Schroth W, Goetz MP, Hamann U, et al. Association between CYP2D6 polymorphisms and outcomes among women with early stage breast cancer treated with tamoxifen. *JAMA.* 2009; **302(13)**:1429–1436.

17. Brauch H, Schroth W, Goetz MP, et al. Tamoxifen use in post-menopausal breast cancer: CYP2D6 matters. *J Clin Oncol.* 2013; **31(2)**:176–180.

18. Weigman VJ, Chao HH, Shabalin AA, et al. Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. *Breast Cancer Res Treat.* 2012; **133(3)**:865–880.

19. Pinto D, Darvishi K, Shi X, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotech.* 2011; **29(6)**:512–520.

20. Sistonen J, Sajantila A, Lao O, et al. CYP2D6 worldwide genetic variation shows high frequency of altered activity variants and no continental structure. *Pharmacogenet Genomics.* 2007; **17(2)**:93–101.

21. Province MA, Goetz MP, Brauch H, et al. CYP2D6 Genotype and Adjuvant Tamoxifen: Meta-analysis of Heterogeneous Study Populations. *Clin Pharmacol Ther.* 2013; **95(2)**:216–227.

22. Rae JM, Regan MM, Thibert JN, et al. Concordance Between CYP2D6 Genotypes Obtained From Tumor-Derived and Germline DNA. *J Natl Cancer Inst.* 2013; **105(17)**: 1332–1334.

23. Committee THCPCAN. The human cytochrome P450 (CYP) allele nomenclature database. Available at: http://www.cyp- alleles.ki.se/. Accessed November 21, 2014.

24. Kelly CM, Pritchard KI. CYP2D6 genotype as a marker for benefit of adjuvant tamoxifen in postmenopausal women: les- sons learned. *J Natl Cancer Inst.* 2012; **104(6)**:427–428.
25. Madlensky L, Natarajan L, Tchu S, et al. Tamoxifen metabolite concentrations, CYP2D6 genotype, and breast cancer outcomes. *Clin Pharmacol Ther.* 2011; **89(5)**:718–725.

26. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J Natl Cancer Inst.* 2009; **101(21)**:1446–1452.

27. Zheng HT, Peng ZH, Zhou CZ, et al. Detailed deletion mapping of loss of heterozygosity on 22q13 in sporadic colorectal cancer. *World J Gastroenterol.* 2005; **11(11)**:1668– 1672.

28. Jonkers YM, Claessen SM, Feuth T, et al. Novel candidate tumour suppressor gene loci on chromosomes 11q23-24 and 22q13 involved in human insulinoma tumourigenesis. *J Pathol. 2006*; **210(4)**:450–458.

CHAPTER 5

CONCLUSION

Numerous hallmarks of cancers have been proposed including sustaining proliferative signaling, activating metastasis and invasion, evading growth suppressors, resisting cell death, and inducing angiogenesis [1]. The function of these hallmarks are fostered, and at times accelerated, by genomic instability. In tumors, increases in genomic instability and aneuploidy are usually associated with poor patient prognosis. Early on, the focus of copy number analyses research favored whole arm/whole chromosome changes that were traditionally studied by spectral karyotyping. However, using high-density arrays, this work demonstrated that many of the sub-chromosomal changes, which are less characterized, are also significant and detrimental in complex human disease. Understanding how DNA Copy Number Alterations (CNAs) promote disease is an important challenge and is further complicated by the complexity of cancer. We demonstrate that one single strategy, which has normally been presented in copy number analyses, is not sufficient to fully understand the role of CNAs in breast cancer progression.

In breast cancer, numerous genetic alterations have been identified and demonstrate clinical implications, with the best example being the biological and clinical importance of HER2 amplification. We highlighted another class of potential drivers that have yet to be established in the clinical setting. This copy number based approach identified targets that were rarely mutated, and drivers where the genetic alteration is used as a mechanism for differential

gene expression and cellular changes. In addition, we demonstrated that numerous genes could be targeted by a single alteration, where in essence a given amplification in fact has at least three driver genes as opposed to the more conventional expectation that a given CNA has a single target gene. More importantly, we demonstrated that a single alteration could lead to disruptions of multiple pathways with similar, or diverse, downstream biological processes that promote the cancer phenotype. These findings support the use of combinatorial-targeted drug therapies and can further help in the development of new 'personalized' cancer treatment options.

Previous studies have demonstrated the heterogeneity of breast cancers, and used histopathological features as prognostics for clinical outcomes. More recently, various molecular-profiling methods are used to identify and characterize new clinically relevant features. The increase in high-throughput molecular data from cancer-profiling projects using microarrays (and now next-generation sequencing) platforms creates great research opportunities, but also significant computational challenges for analysis and interpretation. These challenges emphasized the necessity of integrative analyses, and the need to investigate multi-dimensional interactions across multi-level '-omics' data types. For example, integrative analyses can incorporate data from different sample types (i.e. germline DNA genotypes, tumor, or model system samples), different disease states (i.e. normal or malignant), and multi-level classified data that may include histopathology, genetics, transcriptomics and/or proteomics features.

Given the growing importance of integrative analyses, it was also important to point out necessary caveats. From a statistical perspective, as we increased the dimensionality in the analyses we also increased the amount of unknown parameters. Therefore, we selected data where the underlying relationships between the levels are known (i.e. similar copy number

changes are present and spontaneously occurring cross-species). Furthermore, various normalization steps were performed, including numerous data checks to test and correct for platform batch effects, and to adjust the data to the same scale (both technically or biologically). From a biological perspective, this work demonstrated that multiple potential drivers can exists within a single region of CNA, and as a result, different strategies may identify different drivers within the same altered region.

The integrative analyses in Chapter 2 showed a novel way to comprehensively compare CNAs observed in human breast cancer subtypes and genetically engineered mouse (GEM) mammary tumor models. We highlighted regions of shared CNAs between human breast tumors and GEM models that also shared similar gene expression features, with a particular emphasis on basal-like breast cancers. We demonstrated that there was not a single GEM mammary model that shared an overall defining pattern of CNAs with any single human breast intrinsic subtype; however, numerous sub-chromosomal commonalities were identified, and were the focus of our additional analyses. These analyses incorporated additional functional resources using data from a RNAi screen, DNA to RNA correlations, and a DawnRank network analysis [2] to develop a comprehensive map of essential breast cancer driver genes on a few selected conserved CNAs. Importantly, this work provided a preclinical resource for selecting GEM mammary models for therapeutic response testing based upon the genetics shared between human tumors and mouse models.

Interestingly, the majority of conserved CNAs between humans and mice were identified in the Basal-like breast cancer subtype. We suggest this is due to the fact these CNAs are altering pathways that promote the Basal-like phenotype such as PIK3CA/AKT and NOTCH signaling pathways. Additionally, we demonstrated that chromosome 1 is a rich region of CNAs across all

subtypes, especially for the Basal-like subtype. And unlike previous works that could not narrow down the driving genes of 1q arm amplification, we've highlighted a more target-gene concentrated region of 1q21-23 that contains multiple drivers. Here we highlighted *PI4KB*, *SHC1,* and *NCSTN* as candidate Basal-like-specific driving genes. These results suggested that the strategy presented in Chapter 2 was sensitive enough to identify regions of CNAs that harbor multiple drivers within an amplicon while still filtering out "passenger" genes.

Another example of using integrative analyses to highlight driving genes within CNAs was shown in Chapter 3. In this study, experimentally derived gene expression signatures of signaling pathway activity, and a RNAi screen dataset, was used to identify the driver genes within CNAs that were associated with high proliferation among luminal/ER+ breast cancers. Gene expression signatures provided an essential tool by highlighting patterns of pathway deregulation that reflect activation of various oncogenic pathways, like proliferation. Using experimentally derived gene-expression signatures, we were able to define subtype-specific patterns of oncogenic signaling and identified a novel drug target that regulates fatty acid oxidation (i.e. *CTP1A*).

In Chapter 3, we highlighted known characteristics of each subtype such as low hormone receptor signaling, mutant p53 signaling and high proliferation activity in Basal-like/triple-negative breast cancers. In addition numerous significant correlations were highlighted between gene expression signatures of shared pathways such as between separate MYC signatures, and between PIK3CA and PTEN-deleted signatures. Significant correlations were also observed between signatures from separate pathways but that shared similar associations such as EGFR-mediated activation of STAT3 signaling resulting in the observed concordance between STAT3 and EGFR signatures.

We highlighted expected finding such as an association between the RB-LOH signature and *RB1* DNA copy number loss and also with HER2/ERBB2 amplicon and the HER2-AMP signature. In addition, we presented many novel correlations, including a set of amplified genes associated with the PAM50 proliferation signature [3] that was specific to proliferative luminal tumors. This association is particularly interesting as defects in cell-cycle regulation are hallmarks of cancer progression. Additional RNAi screen data was used to filter out non-essential genes in the region highlighting *FGD5*, *METTL6*, *CPT1A*, *DTX3*, *MRPS23*, *EIF2S2*, *EIF6* and *SLCA10* as essential genes, and which also demonstrate prognostic characteristics.

Another clinically relevant implication of CNAs analyses was highlighted in Chapter 4. In this chapter, we aimed to address the controversy regarding the impact of *CYP2D6* genotype on tamoxifen responsiveness. Tamoxifen, an effective breast cancer treatment in the subgroup of ER+ patients, demonstrates anti-estrogen properties and inhibits estrogen-dependent breast cancer growth. As a result, tamoxifen is often used to treat all stages of ER+ breast cancer. The most potent metabolite of tamoxifen is produced as a result of a cyotchrome p450 enzymes [4], which is encoded by the *CYP2D6* gene. Previously, pharmacodynamics studies demonstrated varying results as to whether *CYP2D6* genotype is associated with a patient's ability to metabolize tamoxifen with the potential high metabolizing variants predicting response, and the low tamoxifen metabolizing alleles predicting resistance.

This work addressed previous concerns that somatic deletion and/or LOH at *CYP2D6* distorts genotype calls and lead to excessive homozygous assignments, and thus, incorrect genotype calls. We evaluated the frequency of LOH and copy number loss across multiple cohorts. We identified frequent copy number loss and LOH at *CYP2D6* in breast cancer patients (~30%). This region of copy number loss was more frequently observed in ER-positive patients

(i.e. patients that are offered Tamoxifen therapies) compared to ER-negative, whereas, the rate of LOH was comparable across both patient group. In addition, we compared genotyping results in the same group of patients across different labs and observed that when tumor DNA was used to infer germline *CYP2D6* genotype, this resulted in genotyping error and the misclassification of some patients which affects whether or not, and potentially at what dose, a patient might receive Tamoxifen treatment.

In conclusion, my dissertation demonstrated two separate clinical utilities of CNA analyses in breast cancer; one includes highlighting driver genes on frequent regions of CNA, and the other highlights the effect of CNAs on the genotype calls used to predict a patients' therapeutic responsiveness. The two integrative analyses presented in Chapter 2 and Chapter 3 were successful in identifying driver genes within regions of CNAs and discovering the functional implications of identified CNAs in subtype-specific breast cancer progression. However, the analyses highlighted in Chapter 2 and Chapter 3 does share some limitations. In both chapters we were confounded by the amount that either mouse models or gene expression signatures recapitulate driving human breast cancer features. To validate all potential targets identified in these chapters will require new experiments, both at the bench and computationally. In some cases published functional analyses were already performed (i.e. *NCSTN* in Chapter 2). In these cases a comprehensive data-mining approach to match identified genes with relevant published functional experiments is feasible and provides for quick functional validation of our computational findings.

For the cases where published functional information is unavailable, we suggest additional biological experiments, which likely include forward genetics (cDNA overexpression), or reverse genetic (RNAi or CRISPR) type experiments. These experiments

would likely incorporate breast cancer cell lines such as ME16C, SUM102, SUM149 and MCF7 to act as a model system. In these experiments, CNAs can be identified as was described in Chapter 2 or through publically available resources such as the Cancer Cell Line Encyclopedia [5]. Using cell lines, an identified altered gene can be functionally characterized through additional experiments that over-expresses the gene, alters the protein function via drug-delivery, or compare growth in soft agar (which only cancer cells can do) with and without the gene of interest. Any biologically validated genes can then be compared with the clinical outcomes of patients who, for example, highly express that given amplified gene, to see if expression and/or alteration demonstrate any prognostic or predictive benefits. The work presented provides important insights and strategies to understand and characterize the genetic and cellular defects that promote the cancer phenotype using a novel analysis of DNA CNAs and a logical computational strategy, which can be applied to multiple tumor types.

REFERENCES

1. Hanahan, D., & Weinberg, R. Hallmarks of cancer: the next generation. *Cell* **144**, 646–74 (2011)

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

3. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology* **27**, 1160–7 (2009).

4. Crewe, H., Notley, L. & Wunsch, R. Metabolism of Tamoxifen by Recombinant Human Cytochrome P450 Enzymes: Formation of the 4-Hydroxy, 4′-Hydroxy and N-Desmethyl Metabolites and Isomerization of trans-4-Hydroxytamoxifen. *Drug Metabolism and Disposition* **30**, 869–874 (2002).

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

APPENDIX 1

Basal-specific conserved 1q segments of CNAs with frequency >= 15%

Chr	Start	Stop	%	Subtype-Specific Segment Genes	CNA	Mouse Groups
1	120315523	144109392	0.35	REG4, ADAM30, NOTCH2	GAIN	C3Tag, Myc
1	144369072	145390170	0.37	PDE4DIP, SEC22B	GAIN	C3Tag, Myc
1	145406799	145431022	0.52	HFE2	GAIN	C3Tag
1	145431022	145464501	0.53	TXNIP	GAIN	C3Tag
1	145464501	145737051	0.54	ANKRD34A, LIX1L, RBM8A, GNRHR2, PEX11B, PIAS3, NUDT17, POLR3C, RNF115, CD160	GAIN	C3Tag, Myc
1	146101240	148205520	0.57	PRKAB2, FMO5, CHD1L, BCL9, ACP6	GAIN	C3Tag, Myc
1	149850351	149935164	0.59	HIST2H2BE, HIST2H2AC, HIST2H2AB, BOLA1, SV2A, SF3B4, MTMR11	GAIN	C3Tag
1	150102064	150184759	0.61	PLEKHO1	GAIN	C3Tag
1	150184759	150372866	0.62	ANP32E, MRPS21, PRPF3	GAIN	C3Tag
1	150372866	150673482	0.62	TARS2, ECM1, ADAMTSL4, ENSA, GOLPH3L	GAIN	C3Tag
1	150673606	150763841	0.62	CTSS	GAIN	C3Tag
1	150763841	150908906	0.62	CTSK, ARNT	GAIN	C3Tag
1	150929687	151773763	0.62	LASS2, ANXA9, FAM63A, PRUNE, BNIPL, CDC42SE1, MLLT11, GABPB2, SEMA6C, TNFAIP8L2, LYSMD1, SCNM1, TMOD4, VPS72, PIP5K1A, PI4KB, RFX5, PSMB4, POGZ, CGN, TUFT1, MIR554, SNX27, CELF3, MRPL9, TDRKH	GAIN	C3Tag
1	151773763	151814405	0.61	C2CD4D, LOC100132111	GAIN	ClaudinLow, C3Tag
1	151880387	152067728	0.62	S100A10, S100A11	GAIN	ClaudinLow, C3Tag
1	152067728	152208144	0.62	TCHH, RPTN, HRNR	GAIN	ClaudinLow, C3Tag
1	152233280	152643406	0.64	FLG2, CRNN, CRCT1, LCE3C, LCE3B, LCE3A	GAIN	C3Tag
1	152661380	153346263	0.64	KPRP, LCE1F, LCE1E, LCE1C, LCE1B, LCE6A, SMCP, IVL, SPRR4, SPRR1A, SPRR3, SPRR1B, SPRR2D, SPRR2B, SPRR2E, SPRR2F, SPRR2G, LELP1, LOR, PGLYRP3, PGLYRP4	GAIN	C3Tag
1	153354347	153576396	0.64	S100A8, S100A7A, S100A6, S100A5, S100A4, S100A3	GAIN	C3Tag
1	153576396	154012535	0.63	S100A16, S100A14, S100A13, C1orf77, SNAPIN, ILF2, NPR1, INTS3, SLC27A3, GATAD2B, DENND4B, SLC39A1, CREB3L4, JTB, JTB, RPS27	GAIN	C3Tag
1	154105694	154234864	0.63	TPM3, MIR190B	GAIN	C3Tag
1	154270185	154318564	0.62	ATP8B2	GAIN	C3Tag
1	154318564	154670097	0.63	UBE2Q1, ADAR	GAIN	C3Tag
1	154807935	155175657	0.62	PMVK, PYGO2, SHC1, CKS1B, FLAD1, LENEP, ZBTB7B, DCST2, DCST1, DPM3, KRTCAP2, TRIM46, MUC1, MIR92B	GAIN	C3Tag
1	155175657	155269798	0.63	GBA, FAM189B, SCAMP3	GAIN	C3Tag
1	155269798	155551091	0.62	FDPS, C1orf104, RUSC1, ASH1L, MIR555, LOC645676	GAIN	C3Tag
1	155556216	155699118	0.62	MSTO1	GAIN	C3Tag
1	155808785	155922229	0.60	SYT11, RIT1, RXFP4	GAIN	C3Tag
1	155936658	156321154	0.60	SSR2, UBQLN4, RAB25, MEX3A, LMNA, SEMA4A, SLC25A44, PMF1, BGLAP, PAQR6, SMG5, TMEM79, CCT3, C1orf182	GAIN	C3Tag
1	156321154	156545720	0.59	RHBG, MEF2D, IQGAP3	GAIN	C3Tag

1	156545720	156587032	0.60	TTC24, APOA1BP, GPATCH4	GAIN	C3Tag
1	156587032	157009095	0.59	HAPLN2, BCAN, NES, CRABP2, ISG20L2, MRPL24, HDGF, PRCC, SH2D2A, NTRK1, NTRK1, INSRR, PEAR1	GAIN	C3Tag
1	157009095	157324089	0.59	ETV3L, ETV3, CYCSP52	GAIN	C3Tag
1	157324089	157657927	0.58	FCRL5	GAIN	C3Tag
1	157657927	158585877	0.57	FCRL1, CD5L	GAIN	C3Tag
1	160043165	160439014	0.57	KCNJ9, IGSF8, ATP1A2, ATP1A4, CASQ1, DCAF8, PEX19, COPA, SUMO1P3, NCSTN, NCSTN, NHLH1, VANGL2	GAIN	C3Tag, p53null_Basal
1	160439014	160626897	0.58	SLAMF6, SLAMF1	GAIN	p53null_Luminal
1	160630352	160735536	0.59	CD48, SLAMF7	GAIN	p53null_Luminal
1	162023809	162694593	0.61	NOS1AP, MIR556, UHMK1, UAP1	GAIN	p53null_Basal, p53null_Luminal
1	162695676	162802177	0.61	HSD17B7	GAIN	p53null_Basal, p53null_Luminal
1	162926083	163194263	0.59	RGS4, RGS5	GAIN	p53null_Basal, p53null_Luminal
1	202542202	202901068	0.45	KDM5B	GAIN	p53null_Luminal
1	202992792	203977548	0.45	FMOD	GAIN	ClaudinLow
1	204164507	204187207	0.45	GOLT1A	GAIN	ClaudinLow, p53null_Luminal
1	204674026	205386068	0.45	NFASC	GAIN	ClaudinLow, p53null_Luminal
1	207257303	207792896	0.43	CD55	GAIN	ClaudinLow

APPENDIX 2

Chromosome 1 subtype-specific conserved CNAs with frequency >= 15% and concordant with gene expression in Humans

Chr				CNA		Concordant DNA/RNA Gene	Subtype
	Start	Stop	%		Mouse Group		
1	76174149	77155253	0.17	GAIN	MixedRb_C3Tag	, ST6GALNAC3	Basal
1	78405135	78535584	0.20	GAIN	MixedRb_C3Tag	, FUBP1	Basal
1	84505777	84732529	0.21	GAIN	MixedRb_C3Tag	, PRKACB	Basal
1	84915300	85310085	0.21	GAIN	MixedRb_C3Tag	, GNG5, GNG5, CTBS, SSX2IP	Basal
1	85311192	85537890	0.21	GAIN	MixedRb_C3Tag	, MCOLN2	Basal
1	86037551	86753596	0.22	GAIN	MixedRb_C3Tag	, CYR61	Basal
1	86753596	87297758	0.23	GAIN	MixedRb_C3Tag	, ODF2L, SH3GLB1	Basal
1	87436385	89979180	0.23	GAIN	MixedRb_C3Tag	, LMO4, GTF2B, GBP2, GBP6	Basal
1	91482031	92144933	0.22	GAIN	MixedRb_C3Tag	, CDC7	Basal
1	92332865	92754726	0.21	GAIN	MixedRb_C3Tag	, EPHX4, BTBD8	Basal
1	94323020	94402263	0.23	GAIN	MixedRb_C3Tag	, DNTTIP2, GCLM	Basal
1	94402263	94847424	0.22	GAIN	MixedRb_C3Tag	, ABCA4	Basal
1	94847424	95095305	0.22	GAIN	MixedRb_C3Tag	, ABCD3, F3	Basal
1	95095305	95481339	0.21	GAIN	MixedRb_C3Tag	, SLC44A3, CNN3	Basal
1	113655140	116683656	0.22	GAIN	MixedRb_C3Tag	, RSBN1, BCL2L15, AP4B1, DCLRE1B, HIPK1, TSPAN2, CASQ2, SLC22A15	Basal
1	116947174	117783083	0.24	GAIN	MixedRb_C3Tag, p53null2, Wap_Myc	, IGSF3, CD101, TRIM45, VTCN1	Basal
1	118147744	118546081	0.25	GAIN	MixedRb_C3Tag, p53null2	, FAM46C, GDAP2	Basal
1	119419556	119735326	0.29	GAIN	MixedRb_C3Tag, p53null2	, WARS2	Basal
1	120315523	144109392	0.35	GAIN	MixedRb_C3Tag, Wap_Myc	, NOTCH2	Basal
1	144369072	145390170	0.37	GAIN	MixedRb_C3Tag, Wap_Myc	, PDE4DIP, SEC22B	Basal
1	145464501	145737051	0.54	GAIN	MixedRb_C3Tag, Wap_Myc	, ANKRD34A, GNRHR2, PEX11B, PIAS3, NUDT17, POLR3C, RNF115, CD160	Basal
1	145792064	150025833	0.50	GAIN	MixedRb_C3Tag	, BOLA1, SF3B4, MTMR11, OTUD7B	LumA
1	146101240	148205520	0.57	GAIN	MixedRb_C3Tag, Wap_Myc	, PRKAB2, FMO5, CHD1L, BCL9, ACP6	Basal
1	149850351	149935164	0.59	GAIN	MixedRb_C3Tag	, HIST2H2AC, BOLA1, SF3B4, MTMR11	Basal
1	150184759	150372866	0.62	GAIN	MixedRb_C3Tag	, ANP32E, MRPS21, PRPF3	Basal
1	150189284	150401522	0.50	GAIN	MixedRb_C3Tag	, ANP32E, MRPS21, PRPF3	LumA
1	150372866	150673482	0.62	GAIN	MixedRb_C3Tag	, TARS2, ECM1, ENSA, GOLPH3L	Basal
1	150402288	150677017	0.50	GAIN	MixedRb_C3Tag	, TARS2, ECM1, ENSA, GOLPH3L	LumA
1	150677017	151008852	0.50	GAIN	MixedRb_C3Tag	, ARNT, SETDB1, LASS2, FAM63A, PRUNE	LumA
1	150763841	150908906	0.62	GAIN	MixedRb_C3Tag	, ARNT	Basal
1	150929687	151773763	0.62	GAIN	MixedRb_C3Tag	, LASS2, FAM63A, PRUNE, BNIPL, CDC42SE1, GABPB2, SEMA6C, LYSMD1, SCNM1, TMOD4, VPS72, PIP5K1A, PI4KB, RFX5, PSMB4, POGZ, CGN, SNX27, MRPL9, TDRKH	Basal
1	151062957	151321770	0.50	GAIN	MixedRb_C3Tag	, SEMA6C, LYSMD1, SCNM1, TMOD4, VPS72, PIP5K1A, PI4KB, RFX5	LumA

1	151321770	151409843	0.50	GAIN	MixedRb_C3Tag	, PSMB4	LumA
1	151409843	151575155	0.50	GAIN	MixedRb_C3Tag	, CGN	LumA
1	151659023	151789548	0.50	GAIN	MixedGroup, MixedRb_C3Tag	, MRPL9, TDRKH, LINGO4	LumA
1	151773763	151814405	0.61	GAIN	MixedGroup, MixedRb_C3Tag	, LOC100132111	Basal
1	151789548	151880754	0.50	GAIN	MixedGroup, MixedRb_C3Tag	, LOC100132111	LumA
1	151880387	152067728	0.62	GAIN	MixedGroup, MixedRb_C3Tag	, S100A10, S100A11	Basal
1	151880754	152292453	0.50	GAIN	MixedGroup, MixedRb_C3Tag	, S100A10, S100A11	LumA
1	153354347	153576396	0.64	GAIN	MixedRb_C3Tag	, S100A8, S100A6	Basal
1	153576396	154012535	0.63	GAIN	MixedRb_C3Tag	, S100A16, S100A14, C1ORF77, SNAPIN, ILF2, INTS3, GATAD2B, DENND4B, SLC39A1, JTB, JTB, RPS27	Basal
1	154105694	154234864	0.63	GAIN	MixedRb_C3Tag	, TPM3	Basal
1	154318564	154670097	0.63	GAIN	MixedRb_C3Tag	, UBE2Q1, ADAR	Basal
1	154807935	155175657	0.62	GAIN	MixedRb_C3Tag	, PMVK, PYGO2, SHC1, CKS1B, FLAD1, ZBTB7B, DCST2, DPM3, KRTCAP2, TRIM46, MUC1	Basal
1	155175657	155269798	0.63	GAIN	MixedRb_C3Tag	, GBA, FAM189B, SCAMP3	Basal
1	155269798	155551091	0.62	GAIN	MixedRb_C3Tag	, RUSC1, LOC645676	Basal
1	155556216	155699118	0.62	GAIN	MixedRb_C3Tag	, MSTO1	Basal
1	155808785	155922229	0.60	GAIN	MixedRb_C3Tag	, RIT1	Basal
1	155936658	156321154	0.60	GAIN	MixedRb_C3Tag	, SSR2, UBQLN4, RAB25, MEX3A, LMNA, SEMA4A, SLC25A44, PMF1, BGLAP, PAQR6, SMG5, TMEM79, C1ORF182	Basal
1	156321154	156545720	0.59	GAIN	MixedRb_C3Tag	, MEF2D, IQGAP3	Basal
1	156545720	156587032	0.60	GAIN	MixedRb_C3Tag	, APOA1BP, GPATCH4	Basal
1	156587032	157009095	0.59	GAIN	MixedRb_C3Tag	, CRABP2, ISG20L2, MRPL24, HDGF, PRCC	Basal
1	157009095	157324089	0.59	GAIN	MixedRb_C3Tag	, ETV3L	Basal
1	160043165	160439014	0.57	GAIN	MixedRb_C3Tag, p53null1	, IGSF8, DCAF8, PEX19, NCSTN, NCSTN, VANGL2	Basal
1	160197660	160372346	0.51	GAIN	MixedRb_C3Tag, p53null1, p53null2	, PEX19, NCSTN, NCSTN	LumA
1	160906176	163790065	0.18	GAIN	p53null2	, F11R, USF1, PVRL4, KLHDC9, DEDD, DEDD, UFC1, PPOX, B4GALT3, ADAMTS4, NDUFS2, NR1I3, NR1I3, PCP4L1, C1ORF192, DUSP12, ATF6, UHMK1, UAP1, HSD17B7, NUF2	Claudin-Low
1	162023809	162694593	0.61	GAIN	p53null1, p53null2	, UHMK1, UAP1	Basal
1	162695676	162802177	0.61	GAIN	p53null1, p53null2	, HSD17B7	Basal
1	167718720	170486625	0.18	GAIN	p53null2	, ADCY10, DCAF6, DCAF6, GPR161, TIPRL, TBX19, NME7, NME7, SLC19A2, SCYL3, KIFAP3	Claudin-Low
1	168137580	168255773	0.67	GAIN	p53null1, p53null2	, TIPRL	LumB
1	169049881	169622927	0.37	GAIN	p53null1, p53null2	, NME7	Her2
1	169622927	169957961	0.38	GAIN	p53null1, p53null2	, SCYL3	Her2
1	169728459	170031491	0.68	GAIN	p53null1, p53null2	, SCYL3	LumB
1	170453516	170634305	0.68	GAIN	p53null2	, GORAB	LumB
1	182988016	184909056	0.18	GAIN	p53null2	, SMG7, ARPC5, TSEN15, EDEM3	Claudin-Low
1	202542202	202845388	0.73	GAIN	p53null2	, SYT2, KDM5B	LumB

1	202542202	202901068	0.45	GAIN	p53null2	, KDM5B	Basal
1	202845388	203372186	0.73	GAIN	MixedGroup	, RABIF, KLHL12, ADIPOR1, CYB5R1, PPFIA4, ADORA1, MYBPH, BTG2	LumB
1	203372186	204393038	0.73	GAIN	MixedGroup, p53null2	, SOX13, KISS1, GOLT1A, PLEKHA6, PPP1R15B	LumB
1	204164507	204187207	0.45	GAIN	MixedGroup, p53null2	, GOLT1A	Basal
1	204393582	204863970	0.73	GAIN	MixedGroup, p53null2	, MDM4, LRRN2	LumB
1	205035520	205333969	0.73	GAIN	MixedGroup	, TMEM81, RBBP5, DSTYK, NUAK2	LumB
1	205333969	206253777	0.73	GAIN	MixedGroup	, CDK18, MFSD4, ELK4, SLC45A3, NUCKS1, RAB7L1, SLC41A1, FAM72A	LumB
1	206617766	206976274	0.72	GAIN	MixedGroup	, IKBKE, DYRK3, MAPKAPK2	LumB
1	206976766	207086194	0.72	GAIN	MixedGroup	, IL20	LumB
1	207086194	207153759	0.71	GAIN	MixedGroup	, PIGR	LumB
1	207205030	207256205	0.72	GAIN	MixedGroup	, YOD1, PFKFB2	LumB
1	207257303	207792896	0.43	GAIN	MixedGroup	, CD55	Basal
1	207260606	207864198	0.72	GAIN	MixedGroup	, CD55	LumB

APPENDIX 3

Top 5% DawnRank Scores

Gene	DawnRankPvalue	DawnRankRawScore
PRCC	0.0956	0.8326
STX6	0.0954	0.8328
UBE2Q1	0.0944	0.8341
SNAPIN	0.0942	0.8343
KLHL12	0.0914	0.8379
RAP1A	0.0893	0.8407
CASP9	0.0882	0.8421
PI4KB	0.0877	0.8428
GATAD2B	0.0870	0.8437
GTF2B	0.0867	0.8441
BPNT1	0.0861	0.8449
MAPKAPK2	0.0850	0.8464
SMG5	0.0842	0.8475
RAB4A	0.0800	0.8533
DEDD	0.0800	0.8534
VPS45	0.0776	0.8569
PSEN2	0.0743	0.8617
ZBTB17	0.0740	0.8621
ACTN2	0.0725	0.8643
NGF	0.0703	0.8677
LCK	0.0671	0.8727
JAK1	0.0664	0.8739
HNRNPU	0.0648	0.8765
HDAC1	0.0614	0.8822
GNAI3	0.0596	0.8853
F11R	0.0577	0.8886
FCER1A	0.0572	0.8895
JUN	0.0557	0.8924
SETDB1	0.0523	0.8987
CDC42	0.0506	0.9020
FCGR2B	0.0476	0.9082
DHX9	0.0465	0.9105
IKBKE	0.0399	0.9251
NCSTN	0.0394	0.9264
FCER1G	0.0390	0.9273
PTPRC	0.0383	0.9290
FASLG	0.0377	0.9306
AKT3	0.0336	0.9411
ACTA1	0.0315	0.9469
SHC1	0.0286	0.9558
CD247	0.0274	0.9593

POU2F1	0.0265	0.9624
ARNT	0.0246	0.9688
ARF1	0.0238	0.9715

APPENDIX 4

Multivariate survival analaysis of prognostic markers

TCGA Variable	Univariate P-value	Hazard Ratio	Multivariate FGD5	Hazard Ratio	Multivariate METLL6	Hazard Ratio	Multivariate CPT1A	Hazard Ratio
FGD5	**8.91E-05**	3.6120	**0.0175**	2.6298				
METTL6	**0.0006**	3.1500			**0.0256**	2.4917		
CPT1A	0.3720	1.3340					0.4430	1.3574
DTX3	**0.0435**	2.0720						
MRPS23	**0.0097**	2.4250						
EIF2S2	0.7690	1.0990						
EIF6	0.6250	1.1630						
SLC2A10	0.9440	1.0220						
Age	**0.0205**	1.0280	**0.0006**	1.0524	**0.0006**	1.0519	**0.0005**	1.0537
Stage	0.5780	1.1040	0.6387	1.0524	0.6645	0.8970	0.6998	0.9104
ER.Statuspos	0.3850	0.6612	0.1195	0.3805	0.1352	0.3944	0.1728	0.4163
HER2.Statuspos	0.8210	0.8972	**0.0241**	0.2007	**0.0219**	0.1913	**0.0305**	0.1957
Node.Status	**0.0542**	1.3830	0.1907	1.2983	0.2152	1.2787	0.2069	1.2799
PAM50LumB	**0.0078**	2.4860	0.7205	1.2335	0.7297	1.2234	0.7911	0.8527
PAM50Her2	**0.0098**	2.8210	0.1174	3.3879	0.1210	3.3601	0.2495	2.4861
Proliferation.Score	**0.0006**	3.5540	0.1105	2.9367	0.1025	3.0054	**0.0182**	4.8883

TCGA Variable (Proliferation Excluded)	Univariate P-value	Hazard Ratio	Multivariate FGD5	Hazard Ratio	Multivariate METLL6	Hazard Ratio	Multivariate CPT1A	Hazard Ratio
FGD5	**8.91E-05**	3.6120	**0.0023**	3.2201				
METTL6	**0.0006**	3.1500			**0.0038**	3.0597		
CPT1A	0.3720	1.3340					0.4840	1.3285
DTX3	**0.0435**	2.0720						
MRPS23	**0.0097**	2.4250						
EIF2S2	0.7690	1.0990						
EIF6	0.6250	1.1630						
SLC2A10	0.9440	1.0220						
Age	**0.0205**	1.0280	**0.0019**	1.0466	**0.0018**	1.0462	**0.0029**	1.0466
Stage	0.5780	1.1040	0.5047	0.8443	0.5305	0.8535	0.5893	0.8765
ER.Statuspos	0.3850	0.6612	0.1580	0.4361	0.1760	0.4505	0.1969	0.4592
HER2.Statuspos	0.8210	0.8972	**0.0242**	0.2016	**0.0223**	0.1934	**0.0351**	0.2101
Node.Status	0.0542	1.3830	0.1318	1.3517	0.1512	1.3303	0.1380	1.3424
PAM50LumB	**0.0078**	2.4860	**0.0252**	2.4405	**0.0248**	2.4433	0.0451	2.3183
PAM50Her2	**0.0098**	2.8210	**0.0053**	6.4303	**0.0055**	6.4608	**0.0088**	6.1336

METABRIC Variable	Univariate P-value	Hazard Ratio	Multivariate FGD5	Hazard Ratio	Multivariate METLL6	Hazard Ratio	Multivariate CPT1A	Hazard Ratio
FGD5	**1.84E-02**	1.719	0.64934	1.2143				
METTL6	**0.0090**	1.7730			0.5551	1.2620		
CPT1A	0.0996	1.2220					0.3278	1.2281
DTX3	**0.0105**	1.6250						
MRPS23	**0.0021**	1.4080						
EIF2S2	**0.0056**	1.4530						
EIF6	**0.0002**	1.6180						
SLC2A10	**0.0106**	1.3690						
Age	**0.0004**	1.0160	0.3437	1.0079	0.2761	1.0088	0.1682	1.0114
Stage	**2.83E-05**	1.279	0.83096	0.9786	0.6986	1.0376	0.8082	1.024
ER.Statuspos	**1.15E-07**	0.4953	0.42202	0.777	0.3262	0.7409	0.2005	0.6674

HER2.Statuspos	**0.0004**	1.4010	0.2750	1.1612	0.2034	1.1799	0.2999	1.1503
Node.Status	**<2e-16**	1.0720	**2.37E-07**	1.1188	**1.83E-07**	1.1138	**3.31E-08**	1.1238
PAM50LumB	**3.24E-08**	1.9570	0.6493	1.1546	0.3329	1.3418	0.5780	1.1914
PAM50Her2	**4.00E-15**	2.7760	0.7051	1.1592	0.5954	1.2251	0.8446	1.0801
Proliferation.Score	**1.55E-15**	2.4800	**0.0056**	2.3635	**0.0164**	2.0314	**0.0111**	2.1539

METABRIC Variable (Proliferation Excluded)	**Univariate P-value**	**Hazard Ratio**	**Multivariate FGD5**	**Hazard Ratio**	**Multivariate METLL6**	**Hazard Ratio**	**Multivariate CPT1A**	**Hazard Ratio**
FGD5	**1.84E-02**	1.719	0.449884	1.3768				
METTL6	**0.00903**	1.773			0.358892	1.4305		
CPT1A	0.0996	1.222					0.28784	1.2496
DTX3	**0.0105**	1.625						
MRPS23	**0.00212**	1.408						
EIF2S2	**0.00557**	1.453						
EIF6	**0.000158**	1.618						
SLC2A10	**0.0106**	1.369						
Age	**0.000415**	1.016	0.430909	1.0066	0.378172	1.0071	0.23022	1.0099
Stage	**2.83E-05**	1.279	0.902639	0.9877	0.627168	1.0474	0.73455	1.0336
ER.Statuspos	**1.15E-07**	0.4953	0.35883	0.7486	0.307927	0.7314	0.17769	0.6521
HER2.Statuspos	**0.000352**	1.401	0.261251	1.1657	0.196853	1.1825	0.2907	1.1529
Node.Status	**<2e-16**	1.072	**1.15E-07**	1.1231	**1.25E-07**	1.1164	**3.56E-08**	1.125
PAM50LumB	**3.24E-08**	1.957	**0.000957**	2.1096	**0.000282**	2.2106	**0.00173**	2.045
PAM50Her2	**4.00E-15**	2.776	**0.017171**	2.141	**0.019139**	2.0753	**0.04356**	1.9082

Multivariate survival analaysis of prognostic markers (Continued)

TCGA Variable	Multivariate DTX3	Hazard Ratio	Multivariate MRPS23	Hazard Ratio	Multivariate EIF2S2	Hazard Ratio	Multivariate EIF6	Hazard Ratio	Multivariate SCL2A10	Hazard Ratio
FGD5										
METTL6										
CPT1A										
DTX3	0.3516	1.4869								
MRPS23			0.2727	1.5937						
EIF2S2					0.5275	0.7736				
EIF6							0.6294	0.8239		
SLC2A10									0.4468	0.7336
Age	**0.0011**	1.0519	**0.0006**	1.0521	**0.0003**	1.0552	**0.0004**	1.0546	**0.0004**	1.0547
Stage	0.6113	0.8864	0.6346	0.8907	0.6168	0.8875	0.6205	0.8886	0.6091	0.8857
ER.Statuspos	0.1493	0.4052	0.1515	0.4053	0.1965	0.4401	0.2014	0.4422	0.2133	0.4521
HER2.Statuspos	**0.0461**	0.2345	**0.0272**	0.1952	**0.0388**	0.2184	**0.0367**	0.2155	**0.0404**	0.2218
Node.Status	0.1668	1.3057	0.1278	1.3422	0.1074	1.3849	0.1186	1.3706	0.0982	1.3993
PAM50LumB	0.9972	1.0021	0.8679	0.9091	0.9298	1.0531	0.9568	1.0324	0.8814	1.0926
PAM50Her2	0.2355	2.4946	0.2503	2.4449	0.2016	2.7165	0.2005	2.7363	0.1916	2.7956
Proliferation.Score	**0.0355**	4.1506	**0.0219**	4.5151	**0.0218**	4.6018	0.0217	4.6041	**0.0237**	4.5122

TCGA Variable (Proliferation Excluded)	Multivariate DTX3	Hazard Ratio	Multivariate MRPS23	Hazard Ratio	Multivariate EIF2S2	Hazard Ratio	Multivariate EIF6	Hazard Ratio	Multivariate SCL2A10	Hazard Ratio
FGD5										
METTL6										
CPT1A										
DTX3	0.1572	1.8168								
MRPS23			0.2256	1.6704						
EIF2S2					0.4275	0.7213				
EIF6							0.5018	0.7613		
SLC2A10									0.3306	0.6733
Age	**0.0063**	1.0432	**0.0023**	1.0459	**0.0017**	1.0490	**0.0021**	1.0480	**0.0021**	1.0480
Stage	0.4412	0.8331	0.5365	0.8599	0.5627	0.8702	0.5663	0.8710	0.5632	0.8711
ER.Statuspos	0.1583	0.4316	0.1772	0.4500	0.2261	0.4835	0.2384	0.4910	0.2513	0.5002
HER2.Statuspos	0.0550	0.2518	**0.0326**	0.2105	**0.0424**	0.2246	**0.0396**	0.2203	**0.0421**	0.2242
Node.Status	0.0909	1.3824	0.0715	1.4170	0.0532	1.4760	0.0585	1.4625	**0.0464**	1.4921
PAM50LumB	**0.0195**	2.4954	**0.0441**	2.2690	**0.0109**	2.7622	**0.0120**	2.7181	**0.0095**	2.8503

PAM50Her2	**0.0150**	5.2160	**0.0100**	5.7788	**0.0066**	6.4947	**0.0063**	6.5935	**0.0061**	6.6658

METABRIC Variable	**Multivariate DTX3**	**Hazard Ratio**	**Multivariate MRPS23**	**Hazard Ratio**	**Multivariate EIF2S2**	**Hazard Ratio**	**Multivariate EIF6**	**Hazard Ratio**	**Multivariate SCL2A10**	**Hazard Ratio**
FGD5										
METTL6										
CPT1A										
DTX3	0.4702	1.2737								
MRPS23			0.3804	0.8397						
EIF2S2					0.2185	1.3234				
EIF6							**0.0431**	1.5433		
SLC2A10									0.9790	1.0059
Age	0.2733	1.0088	0.2061	1.0102	0.2951	1.0084	0.3199	1.0080	0.2432	1.0094
Stage	0.7224	1.0346	0.64839	1.0447	0.7164	1.0351	0.7323	1.033	0.7164	1.0354
ER.Statuspos	0.267	0.7149	0.33076	0.7419	0.2628	0.7126	0.2549	0.7093	0.2765	0.7175
HER2.Statuspos	0.2412	1.1642	0.2044	1.1789	0.2529	1.1602	0.2635	1.1565	0.2356	1.1667
Node.Status	**1.13E-07**	1.1159	**2.14E-07**	1.1129	**1.13E-07**	1.114	**9.30E-08**	1.1142	**1.41E-07**	1.1142
PAM50LumB	0.3754	1.3103	0.4105	1.2833	0.3482	1.3308	0.3580	1.3241	0.3470	1.3305
PAM50Her2	0.6461	1.1898	0.7015	1.1567	0.5505	1.2547	0.5112	1.2832	0.6319	1.1996
Proliferation.Score	**0.0146**	2.0493	**0.0054**	2.3158	**0.0262**	1.9365	**0.0315**	1.8944	**0.0154**	2.0549

METABRIC Variable (Proliferation Excluded)	**Multivariate DTX3**	**Hazard Ratio**	**Multivariate MRPS23**	**Hazard Ratio**	**Multivariate EIF2S2**	**Hazard Ratio**	**Multivariate EIF6**	**Hazard Ratio**	**Multivariate SCL2A10**	**Hazard Ratio**
FGD5										
METTL6										
CPT1A										
DTX3	0.426052	1.3061								
MRPS23			0.745102	0.9385						
EIF2S2					0.097623	1.4482				
EIF6							**0.016634**	1.6597		
SLC2A10									0.651409	1.1045
Age	0.375404	1.0071	0.295207	1.0084	0.388197	1.0069	0.413615	1.0065	0.332118	1.0078
Stage	0.672986	1.0413	0.601474	1.0514	0.665224	1.0419	0.677309	1.0402	0.652646	1.0439
ER.Statuspos	0.246021	0.7036	0.262833	0.7091	0.232072	0.6962	0.226318	0.6938	0.236349	0.6961

HER2.Status pos	0.222193	1.171	0.212743	1.1762	0.243994	1.1632	0.254022	1.1597	0.223382	1.1712
Node.Status	**7.71E-08**	1.1186	**1.21E-07**	1.1163	**8.63E-08**	1.1157	**7.36E-08**	1.1157	**1.04E-07**	1.1162
PAM50LumB	**0.000377**	2.1815	**0.000252**	2.2549	**0.000699**	2.1115	**0.000917**	2.0756	**0.000398**	2.1875
PAM50Her2	**0.022442**	2.0194	**0.02065**	2.0738	**0.020796**	2.0406	**0.019391**	2.0548	**0.023053**	2.0247