MOLECULAR CHARACTERIZATION AND CLINICAL IMPLEMENTATION OF BREAST CANCER GENOMICS USING MASSIVE PARALLEL SEQUENCING AND MICROARRAY

Wei Zhao

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

2014

Approved by:

Charles M. Perou

J. S. Marron

D. Neil Hayes

Yufeng Liu

Jan F. Prins

ABSTRACT

WEI ZHAO: Molecular Characterization and Clinical Implementation of Breast Cancer Genomics
using Massive Parallel Sequencing and Microarray
(Under the direction of Charles M. Perou)

Genomic studies have revealed the heterogeneity of breast cancer and identified “intrinsic molecular subtypes” with significant difference in incidence, survival and therapeutic response. Investigation of their clinical implications is critical for personalized therapeutics and drug development. The characteristics of cancer genomics require special considerations in the application of laboratory and computational approaches. Therefore, my research explored the use of two technologies, Genetically Engineered Mouse Model (GEMM) and RNA-sequencing (RNA-seq), to facilitate the translation of cancer biology into clinical knowledge.

One powerful GEMM, the p53-null transplant model, was identified as a heterogeneous model that gave rise to multiple subtypes, including Basal-like, Luminal and Claudin-low. Molecular characterization identified genetic signatures of GEMM and its human counterpart and distinct genomic DNA copy number changes associated with each subtype. The analysis on the Claudin-low p53-null tumors showed that they have high expression of epithelial-to-mesenchymal transition genes and are enriched for tumor initiating cells, therefore revealing the stem-cell characteristics of Claudin-low.

The utility of GEMM also involves preclinical drug efficacy testing. We evaluated the efficacy of four chemotherapeutic and/or targeted anti-cancer drugs using three well-established mouse models that recapitulate three human subtypes: Basal-like, Luminal B and Claudin-low. Additionally, we identified two gene signatures that predicted pathological complete response to neoadjuvant anthracycline/taxane therapy in humans. The predictive significance was further validated in two large, independent cohorts of human patients, suggesting the possibility of deriving new biomarkers for humans from analysis of GEMM genomic data.

Another resource of cancer genomics is the formalin-fixed paraffin-embedded (FFPE) samples. Though RNA-seq has been adopted by many studies, the mRNA enrichment protocol (mRNA-Seq) to remove rRNA restricted its utility in FFPE. We examined two rRNA depletion protocols on paired fresh-frozen (FF) and FFPE samples, and compared them with mRNA-seq and DNA microarray. We demonstrated that Ribo-Zero-Seq provides equivalent rRNA removal efficiency and coverage uniformity. Both protocols have consistent transcript quantification using FF and FFPE, suggesting that RNA-seq can be performed on FFPE.

My work uses multiple genomic data types to identify murine models and to develop new protocols for the development and evaluation of new biomarkers for human breast cancer patients.

ACKNOWLEDGEMENTS

I would like to thank:

My advisor, Charles M. Perou

The Perou Lab, all past and present members

The Lineberger Bioinformatics Group

My committee members:

J. S. Marron
D. Neil Hayes
Yufeng Liu
Jan F. Prins

My Family: Jun Zhao and Dan Zhu

PREFACE

Chapter II represents a previously published paper in Proceedings of the National Academy of Sciences of the United States of America. I performed data analysis of genetic signatures and copy number aberrations, and contributed to the writing of the manuscript. I would like to thank the following scientists for their collaborative help on this project:

Concept and design: Jason I. Herschkowitz, Daniel Medina, Charles M. Perou, and Jeffrey M. Rosen

Collection and assembly of data: Jason I. Herschkowitz, Mei Zhang, Jerry Usary, George Murrow, David Edwards, Jana Knezevic, Stephanie B. Greene, David Darr, Melissa A. Troester, Susan G. Hilsenbeck, Daniel Medina

Provision of study materials or analytic tools: Jason I. Herschkowitz, Mei Zhang, David Edwards, Susan G.Hilsenbeck, Daniel Medina, Charles M. Perou, Jeffrey M. Rosen

Analysis and interpretation of data: Jason I. Herschkowitz, Mei Zhang, Jerry Usary, Jana Knezevic, Stephanie B. Greene, Melissa A. Troester, Susan G. Hilsenbeck, Daniel Medina, Charles M. Perou, Jeffrey M. Rosen

Manuscript writing: Jason I. Herschkowitz, Charles M. Perou, and Jeffrey M. Rosen

Chapter III was previously published in Clinical Cancer Research. My role in this project include data analysis to derive the genetic signatures from the *C3(1)-T-antigen* mouse cohort and to test their predictive potential in human patients. And I wrote the computational part of the manuscript. I would like to thank the following scientists for their collaborative help on this project:

Conception and design: Jerry Usary, Norman E. Sharpless, Charles M. Perou

Development of methodology: Jerry Usary, Olga Karginova, Austin Combest, Aleix Prat, Charles M. Perou

Acquisition of data: Jerry Usary, David Darr, Patrick J. Roberts, Mei Liu, Olga Karginova, Austin Combest, Arlene Bridges, Aleix Prat, Maggie C. U. Cheang, Jason I. Herschkowitz , Jeffrey M. Rosen

Analysis and interpretation of data: Jerry Usary, Patrick J. Roberts, Jamie Jordan, Austin Combest, Aleix Prat, Maggie C.U. Cheang, Norman E. Sharpless, Charles M. Perou

Manuscript writing: Jerry Usary, Patrick J. Roberts, Austin Combest, Aleix Prat, Jason I. Herschkowitz, Norman E. Sharpless, Charles M. Perou

Administrative, technical, or material support: Jerry Usary, David Darr, Lorraine Balletta, Olga Karginova, Maggie C.U. Cheang

Study supervision: Jerry Usary, William Zamboni, Charles M. Perou

Chapter IV was previously publish.in BMC Genomics. I contributed to the study design, performed data analysis and initiated all aspects of manuscript preparation. I would like to thank the following scientists for their collaborative help on this project:

Concept and design: D. Neil Hayes and Charles M. Perou

Acquisition of data: Xiaping He

Analysis and interpretation of data: Katherine A. Hoadley, D. Neil Hayes, Joel S. Parker, Charles M. Perou

Manuscript writing: Katherine A. Hoadley and Charles M. Perou

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

aCGH	array Comparative Genome Hybridization
ANOVA	Analysis of variance
AUC	area under curve
CNA	Copy Number Aberration
CT	carboplatin/paclitaxel
CV	coefficient of variation
DSN-Seq	Duplex-Specific Nuclease
EMT	epithelial-to-mesenchymal transition
ER	Estrogen Receptor
FDR	false discovery rate
FF	fresh-frozen
FFPE	formalin-fixed and paraffin-embedded
GEMM	genetically engineered mouse model
HER2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2
LumA	Luminal A

LumB	Luminal B
MPS	Massive Parallel Sequencing
mRNA-Seq	mRNA enrichment RNA-Seq protocol
MVA	multivariate analysis
OST	orthotopic syngeneic murine transplant
PAM50	Subtype classification algorithm consisting of 50 genes
RD	residual disease
RNA-Seq	RNA sequencing
SAM	Significance Analysis of Microarrays
TCGA	The Cancer Genome Atlas
TIC	tumor initiating cells
TNBC	triple-negative breast cancer

CHAPTER I

INTRODUCTION

Breast cancer is the most common cancer in women in the United States. In 2013, NCI SEER Program reported that the estimation of new cases of breast cancer was 232,340, which represented 14.1% of all new cancer cases in women in the U.S. The lifetime risk of developing breast cancer is approximately 12.3% in women [1]. On the other hand, breast cancer mortality has been remarkably declining in the last 20 years as a result of the application of improved screening and improved adjuvant systemic therapy, the latter of which was in part fueled by the development of genomic profiling technology.

Breast cancer is a heterogeneous disease with respect to the incidence, histology, baseline prognosis and response to treatment. A small set of biomarkers have been used for many years in clinical practice and provide substantial prognostic and predictive information. Estrogen receptor (ER) status is associated with good prognosis [2] and is predictive of response to endocrine therapy (both for tamoxifen and aromatase inhibitors). Human epidermal growth factor receptor 2 (HER2) overexpression and/or amplification predicts a benefit for trastuzumab, a monoclonal antibody against HER2 [3]. Ki67 is a proliferation marker and has various applications as a biomarker. Baseline Ki67 is associated with poor outcomes and predicts to good response to chemotherapy [4]. Changes of Ki67 measurement in neoadjuvant setting predicts benefit from endocrine treatment [5, 6]. However, much variation has been observed within the subpopulations defined by classical clinical-pathological markers. For instance, a subgroup of HER2-positive tumor that also expresses p95HER2, a cytoplasmic amino terminally truncated receptor that has kinase activity, has a worse response to trastuzumab but is sensitive to the HER2 tyrosine kinase inhibitor lapatinib [7]. Likewise, while breast tumors expressing high levels of ER are typically responsive to endocrine therapy [8], around 40% ER+ patients fail to respond to tamoxifen or a prolonged treatment leads to drug resistance [9].

Molecular Intrinsic Subtypes

Over the past decade, genomic studies have revealed the heterogeneity of breast cancer. Global gene expression-based analysis has identified four molecular intrinsic subtypes of breast cancer, Luminal A, Luminal B, HER2-enriched, Basal-like and a subgroup of normal-like tumors. Each subtype is characterized by the expression of a set of gene signatures (Figure 1.1). More recently, large-scale genomic research primarily by massive parallel sequencing on several platforms have further demonstrated the complexity of the intrinsic subtypes and highlighted the valuable insight provided by genomic data into personalized treatment. The genomic features of the intrinsic subtypes will be described further below and were a major emphasis of my thesis.

Luminal subtypes. The majority of ER+ and/or PR+ tumors are of the Luminal subtypes. RNA expression profiling revealed that the Luminal tumors have high expression of GATA3, Cyclin D1 as well as Keratin 8 and 18 (Figure 1.1) [10, 11]. In addition, at least two subtypes, Luminal A and Luminal B, have been identified within this subpopulation. Luminal A is the most common breast cancer subtype and represents 30-40% of breast cancers [12, 13]. Compared with Luminal B, Luminal A tumors are characterized by low expression of HER2 cluster and proliferation markers such as Ki67, and are typically associated with a better prognosis, although the risk of late mortality after 10 years persists [14].

Recent studies have shown that a diverse spectrum of copy number aberration (CNA) is observed in Luminal A subtype, potentially indicating the presence of additional substructure within this subtype [15]. Four major CNA patterns have been identified; these subgroups are characterized by (a) chromosome 1q gain and 16q loss, (b) a quiet copy number spectrum, (c) chromosome 8p gain and 8q loss and (d) high level of genomic instability (CNH pattern). Intriguingly, the CNH pattern is associated with over-expression of regulators of mitosis and Aurora kinase pathway components, which have previously been identified as gene markers of proliferation [16] and have been found to be associated with 5q loss in Basal-like subtype [17]. The CNA patterns are also correlated with distinct mutation profiles and carry clinical implications with the CNH Luminal A patients having a worse outcome.

The mutation profile of Luminal A is markedly diverse and with many recurrent genes. Luminal A has the highest number of frequently mutated genes despite a low mutation rate per tumor [18]. Among them, PIK3CA is the most common significantly mutated gene, which has drawn much attention as a therapeutic target because these mutations are gain-of-function. Other mutations in PI3K pathway has also been observed at a lower frequency including PTEN inactivating, and AKT1 activating mutations. MAP3K1 is the most common mutated tumor suppressor in Luminal A. Several studies have confirmed that inactivation of MAP3K1 and MAP2K4 are mutually exclusive, suggesting the reduced p38-JNK stress kinase pathway in this subtype [18, 19].

Basal-like. The Basal-like subtype is notable for low expression of hormone receptors and HER2, high expression of proliferative genes (e.g. Ki67), and high expression of a set of genes called the Basal gene cluster including basal epithelial cytokeratin (CK) such as CK5, 6 and 17, epidermal growth factor receptor (EGFR), αB-crystallin, P-cadherin, and c-Kit (Figure 1.1) [10, 11, 20]. Massive parallel sequencing studies have shown that Basal-like tumors may arise from “Luminal Progenitor Cells”, and are more similar to other high-grade epithelial tumors such as squamous carcinoma of lung, head and neck, serous ovarian carcinoma and serous endometrial cancers than they are to ER+/luminal breast cancers [21].

Basal-like tumors account for 10-25% of all breast tumors, and represent 50-75% of the triple-negative cancers [22]. This subtype is generally associated with high histologic and nuclear grade, high proliferation indices and poor prognosis[23]. Basal-like tumors are also mostly aneuploidy and show high level of genomic instability. Comparison of this subtype and ovarian tumors have revealed some common features in the DNA copy number landscape, including gains of 1q, 3q, 8q and 12p, loss of 4q, 5q, and 8p, and focal amplification of MYC [18].

The majority of Basal-like tumors have TP53 somatic mutations (>80%), which is similar to HER2-enriched subtype (72%), and mostly are nonsense or frame shift. In marked contrast, the TP53 mutations occur in Luminal subtypes at a much lower frequency. In addition to TP53, the loss of function of RB1, another tumor suppressor, is also common in the Basal-like subtype. Activation of PI(3)K pathway has also been identified in Basal-like tumors; however, in contrast with Luminal subtypes, the PIK3CA mutation occurs at a much lower frequency (9%), while alternative aberrations were identified including loss of INPP4B and/or PTEN and/or amplification of PIK3CA [17–19, 24, 25]. BRCA1 pathway is also associated with Basal-like in a more complex fashion. The BRCA1 mutation carriers if and when they develop breast cancer, the majority is Basal-like (~80%). And similar to Basal-like, BRCA1-related cancer shows high level of genomic instability [26] and early relapse [27]. However, most Basal-like are sporadic and with the intact BRCA1 gene and protein. Nevertheless, some studies suggested that the BRCA1 pathway is dysfunctional in at least a subset of sporadic Basal-like tumors [23]. The Cancer Genome Atlas (TCGA) data suggested ~20% of Basal-like tumors have germline and/or somatic BRCA1 or BRCA2 variants.

HER2-enriched subtype. The HER2-enriched subtype is characterized by DNA amplification of HER2 and over expression of HER2 protein and a subset of genes located in the same region of chromosome 17 [10]. These tumors share a few common features with Basal-like such as the low expression of ER and hormone receptor-related gene cluster. Notably, ~60% of clinically HER2+ tumors fall into this subtype and not all tumors within this subtype are clinically HER2+ or HER2-amplified [22]. Non-HER2-enriched but clinical HER2+ tumors tend to be ER+, have high expression of Luminal cluster and are predominantly of the Luminal A subtype [18, 28, 29]. Population-based study revealed that HER2-enriched subtype constitutes 5-10% of all breast cancers and accounts for 7.8% Triple-negative subpopulation [30].

In the TCGA study, HER2-enriched is mostly aneuploidy and shows high genome instability. The frequent somatic mutation alterations in this subtype include TP53 mutation (75%) and PIK3CA (39%). Other mutations occur at a much lower rate. Gene expression and reverse-phase protein array (RPPA) data both confirmed that clinical HER2+/HER2-enriched subgroup is associated with high expression and high level of phosphorylation of EGFR, which provides another therapeutic target for this typically aggressive tumor subtype.

Genetically Engineered Mouse Model (GEMM) in cancer study

Efforts of genomic, epigenomic, transcriptomic and proteomic studies have led to remarkable advances in our knowledge of cancer biology in the past decade. In recognition of diverse genetic aberrations in human tumors, nearly 1,000 small molecules drug inhibitors are being tested and under development for cancer treatments [31]. However, about 95% of anti-cancer compounds that enter preclinical testing fail to gain FDA approval [29]. The high attrition rate of anti-cancer compound candidates clearly reveals the significant challenges in drug development, and suggests the need for improved pre-clinical testing.

The high investment in clinical trials has further suggested for a better way of compound screening in the early stage of drug development. In addition, a more optimized efficacy testing must account for the heterogeneity of breast cancer. Indeed, many of these compounds might only be very effective in subpopulations that express specific biological targets or harbor genetic alterations in specific pathways. However, molecular marker-based clinical trials in which sensitive patient subpopulations are identified early during the trial are expensive and take longer time to recruit patients. A few large clinical trials used unselected patient populations, which is not effective in terms of evaluating the efficacy of the drug on the responsive patient subset.

Murine models, as the most experimentally tractable mammalian system, have contributed significantly to the basic scientific discoveries in cancer biology. Using transgenic and knockout technologies, numerous genetically engineered mouse models (GEMMs) have been generated. Early GEMMs are conventional models that are driven by overexpression of oncogenes or carry germline mutations in tumor suppressor genes [28]. With the development of spatiotemporally controlled induction of mutations (i.e. those controlled by inducible expression of Cre recombinase), conditional models are used to possible better mimic sporadic cancers [30, 32].

Although GEMMs have been designed to emulate genetic lesions found in human tumors, it is not always clear to what extent the mouse models faithfully recapitulates features of human subtypes because of the different physiology across species. Therefore, many efforts have been made to identify the conserved features shared by human cancer and mouse models using genomic profiling on large datasets [33, 34]. It has been shown that although no single mouse models embraced all the expression features of specific human subtypes, for each human subtype, multiple GEMMs expressed a few shared signatures. This basic biological understanding would be especially valuable for preclinical testing, as it suggested the choice of mouse models that are most consistent with a subpopulation of patients in terms of the targeted pathways.

Interestingly, similar to human patients, while some mouse models have homogeneous gene expression patterns, others show ‘semi-homogeneous’ or even heterogeneous patterns, which suggested the existence of a higher level of genetic diversity. Indeed, the discovery of the heterogeneity in both human and mouse highlights the importance of selecting the appropriate GEMM to model specific human subgroups, and to use multiple GEMMs to provide a comprehensive portrait of human diversity.

On the other hand, despite of the progress in our knowledge of the genetic characteristics of GEMMs in the context of cross-species comparison, their use in preclinical assessment of drug development is still underappreciated. Most studies examined a small number of candidate compounds in a small cohort of GEMMs [35]. The low-throughput approach has been successfully applied to validate drug targets, especially their role in tumor maintenance. For example, GEMM experiments demonstrated that the inhibition of tumor growth induced by Farnysl-transferase Inhibitor (FTI) is not solely mediated by K-Ras [36]. But this approach is flawed in that it only recapitulates a handful of features shared by human tumors and the selected GEMM, rather than providing a comprehensive evaluation. As consequence, the capability to predict response is attenuated by the cross-species complexity, and the limited number of GEMMs used.

More recently, the efforts were extended to medium-throughput testing using larger numbers of mice [37]. These ‘co-clinical’ trial studies are a promising approach to inform the clinical trials in several aspects. Firstly, the results from *in vivo* models may directly predict the response of cognate human subpopulation to the therapeutics. Secondly, as an experimentally tractable system, the underlying biological basis or clinical hypothesis could be investigated. Lastly, biological signatures identified from mice could be applied to guide the design of clinical trials, either to identify patients that most likely to benefit from the therapy, or to predict to response in early trial. Noteworthy, identification and validation of signatures have only been made possible with the use of large GEMM cohorts and the availability of the large data sets of human clinical trials, and the power of this application still remains largely untapped; my work has directly addressed this issue.

Massive Parallel Sequencing (MPS) in cancer study

Another technology revolution that had profound impact on cancer genomics is the advance of massive parallel sequencing (MPS), also known as deep sequencing. Today, rapid, accurate and relatively affordable genome sequencing has become feasible. The diverse application of MPS has contributed remarkably to the cancer biology studies, and provided opportunity for improvement in diagnosis, prognosis and treatment. Cancer genomes, however, have a few distinct characteristics that affect the cancer genome-sequencing study design. To reveal the underlying mechanism of cancer biology, a generally accepted method is to identify somatic mutations by comparing the matched normal samples and the tumor counterpart. For solid tumors, the matched normal samples are frequently peripheral blood samples [18, 38, 39], and in some studies surgical margins and proximal lymph nodes [40, 41]. However, cancer samples are distinct in their quality, quantity and tumor cell purity, all of which pose biological and technical challenges.

Firstly, a large number of cancer samples are archived formalin-fixed and paraffin-embedded (FFPE). The nucleic acid extracted from FFPE blocks are likely to have low nucleic acid quality due to the cross-linkage caused by fixation process and partial degradation [42]. Overcoming this challenge is especially critical for RNA-Sequencing (RNA-Seq), in which the standard protocol requires intact RNA to deplete the highly abundant ribosomal RNA (rRNA). Besides, the necrosis and apoptosis of cancer cells also contribute to the lower quality [43].

Secondly, for safety consideration, the biopsy size from patients with disseminated disease is typically small in size. Consequently, the nucleic acid from tumors is of limited quantity. Though it is possible to perform whole-genome amplification prior to sequencing, this procedure might produce artifacts [44]. Hence, cancer genome sequencing requires decreasing the minimum input of nucleic acid.

Lastly, the purity of the cancer specimens can be low for two main reasons, which include normal tissue contamination and intra-tumoral heterogeneity. Indeed, matched normal samples potentially also contain a mixture of malignant and non-malignant genomes, as residual disease could exist in surrounding tissues. A few MPS-based genomic studies have demonstrated that intra-tumoral heterogeneity is a common feature of multiple tumor types [18, 45–47]. Of note, characterization of the clonality provides the potential mechanism by which tumors acquire drug resistance to targeted therapy. For instance, the emergence of KRAS mutated subclones conveys resistance to anti-EGFR treatment in colorectal cancer [48]. On the other hand, it highlights the requirements that experimental and computational methods for cancer genome sequencing should account for this heterogeneous nature of tumor samples.

The MPS-based cancer genomic studies vary by the input materials (DNA, RNA, chromatin) and the targeted regions (whole genome, exome, transcriptome, or targeted genes). MPS of the transcriptome, also known as RNA-Sequencing (RNA-Seq), has been proved to be a powerful approach for studies of a variety of goals. RNA-Seq provides an efficient way for the identification of novel transcripts [49], alternative splicing [49, 50], and gene fusion events [49–52]. Given proper matched normal samples, it has also been applied, alone or in combination with whole-genome sequencing or exome sequencing, to detect somatic mutations [52].

Another major goal of RNA-Seq studies is to characterize the overall gene expression profile of a tumor or normal. Compared to array-based technology, RNA-Seq demonstrated its superiority in accuracy and comprehensiveness for several reasons. First, RNA-Seq provides a near digital measure of gene expression levels, which enables the comparison across genes, samples, experiments and platforms. While using array-based strategies, considerable difference were observed in terms of the hybridization properties of probes [52]. In addition, array-based technologies are limited in its dynamic range due to the background hybridization level [53] and saturation of signals. While RNA-Seq is superior in detecting transcripts with low expression level, accordingly identifies differentially expressed genes in higher sensitivity. More importantly, RNA-Seq is not limited to detection of known genes. With the significant improvement in cost and efficiency, RNA-Seq is becoming the predominant tools for transcriptome measurement.

Of note, despite of all the favorable properties, there exist several types of bias or limitations unique to RNA-Seq. For instance, to allow for cost-efficient detection of genes/mRNAs, highly abundant rRNA must be removed from total RNA before sequencing. The standard rRNA removal strategy relies on enrichment of poly(A) RNAs. This procedure requires intact RNAs and restricts the detection of non-poly(A) RNA species. Moreover, it is not applicable for samples with small input quantity or archived as FFPE. In addition, the selection based on 3’-end of each transcript introduces 3’ bias and causes inaccurate quantification in partially degraded samples.

Other challenges arise from computational analysis. Mapping reads to transcripts with complicated splicing patterns, large introns, low complexity, or to homologous genes could cause ambiguity. Hence, mapping algorithms should account for these scenarios. Likewise, the uniformity of sequence coverage across transcripts could vary by protocol [54, 55], which affects the sensitivity and accuracy of quantification. Therefore, a careful normalization procedure is required to minimize the bias introduced by uneven coverage. Also, as RNA-Seq-based studies span a wide range of interests, it is critical to determine the sequencing depth needed for each specific purpose. Studies whose goal is to comprehensively catalogue transcripts or to investigate transcripts of low abundance would require more sequencing depth. Nevertheless, with the advent of new experimental protocols, this standard has not always been available. Developments in experimental and computational techniques contributes to leverage the importance of RNA-Seq. Studies that evaluate the features of these new techniques and determine their suitability for distinct research interests would guide study design and facilitate the application of RNA-Seq in cancer genomic research, and it is for these reasons that my thesis has focused much attention on improvements in RNA-Seq methods.

Research introduction

In order to span these many key topics discussed above, my thesis work has also covered a broad range of topics spanning computational analyses of whole transcriptome data to the validation of mouse models of breast cancer. The studies in Chapter 2 and 3 identify and detail the heterogeneity of several mouse models with respect to gene expression, DNA copy number and clinical response, and eventually utilize these models to increase our knowledge of human tumors and to facilitate drug development. In particular in Chapter 2, the characterization of fifty p53 null transplant mouse tumors revealed that this model gives rise to multiple molecular subtypes, including

Basal-like, Luminal and Claudin-low. These subtypes also show distinct DNA copy number changes, some of which recapitulate their human counterpart subtype. Analysis of this heterogeneous murine tumor model provided insights into the stem-cell characteristics of a rare human subtype, Claudin-low. In Chapter 3, we extended our efforts to utilizing GEMMs to examine the efficacy of chemotherapeutics or targeted anti-cancer drugs. In particular, the response of *C3(1)-T-antigen* model to the treatment using carboplatin/paclitaxel(CT) is reminiscent of the bi-phasic response pattern observed in human Basal-like tumors. Therefore, gene expression signatures that predict the response to cytotoxic chemotherapeutics were derived from analysis of mouse genomic data and validated in human neoadjuvant data sets.

In Chapter 4, the focus turns to exploration of using RNA-seq to extract meaningful information from clinical materials archived as FFPE. Two protocols for performing RNA-Seq using FFPE were extensively tested and compared with the results of RNA-Seq from fresh-frozen tumors and DNA-microarray. The features of each protocol were evaluated by objective statistical analysis, and these results should impact experimental design and cost. Through all of this work we have now set the stage for an improved means of pre-clinical drug testing using animal models, and laid the foundation for how to translate these data from mice into humans using the materials that are common to human clinical studies.

FIGURES

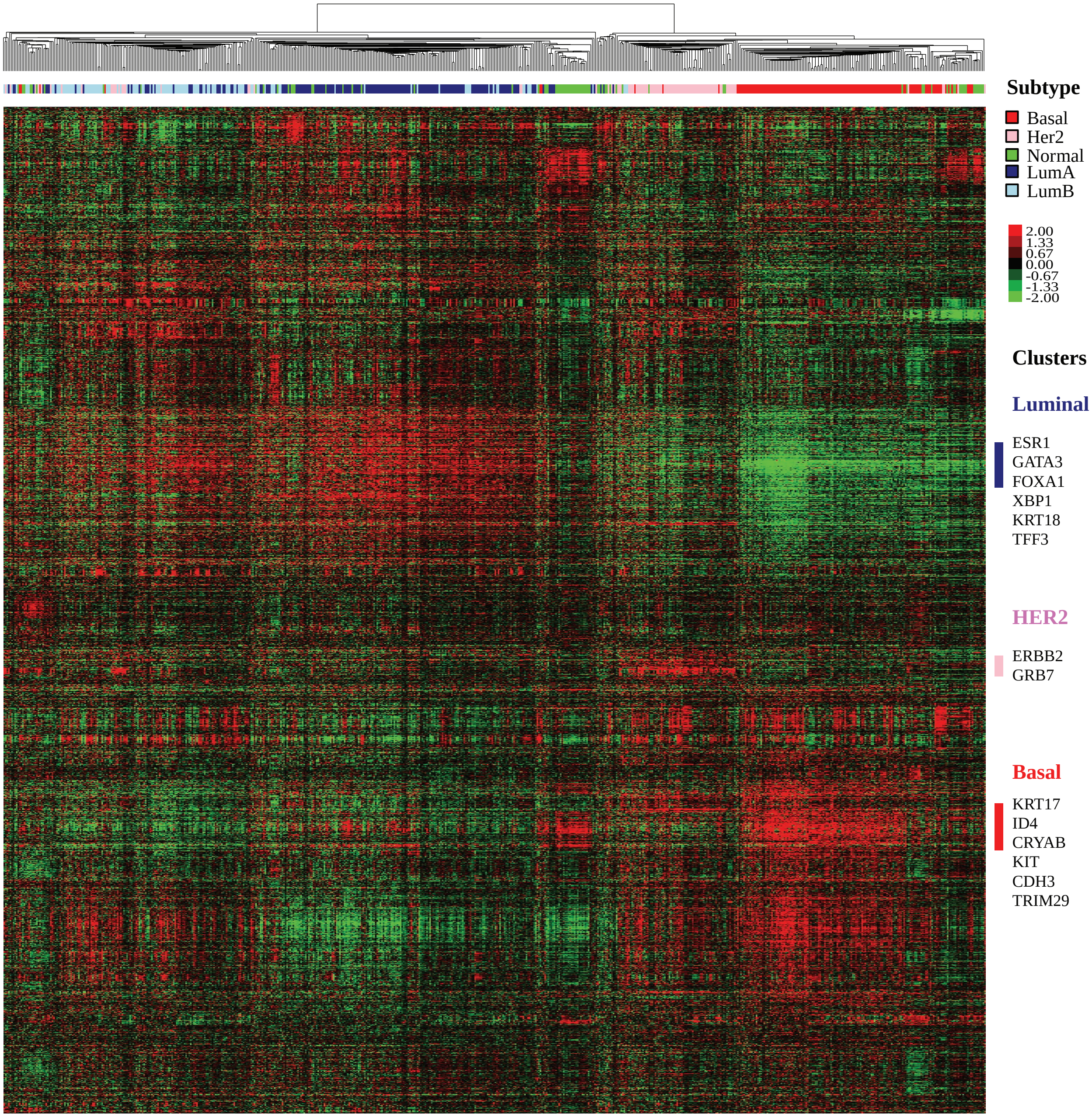


Figure 1.1 Hierarchical clustering of human breast tumors.
A combined dataset of 337 samples collected from UNC-Chapel Hill and 295 samples from the NKI was clustered using an intrinsic gene list comprised of 1800 genes published in four previous studies [56]. Clustering identified the five intrinsic subtypes of Luminal A, Luminal B, Normal-like, Basal-like and HER2-enriched. Gene clusteres with high expression levels associated each subtype are labeled on the right.

REFERENCES

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

2. Onitilo AA, Engel JM, Greenlee RT, Mukesh BN: **Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival.** *Clin Med Res* 2009, **7**:4–13.

3. Mass RD, Press MF, Anderson S, Cobleigh MA, Vogel CL, Dybdal N, Leiberman G, Slamon DJ: **Evaluation of clinical outcomes according to HER2 detection by fluorescence in situ hybridization in women with metastatic breast cancer treated with trastuzumab.** *Clin Breast Cancer* 2005, **6**:240–246.

4. De Azambuja E, Cardoso F, de Castro G, Colozza M, Mano MS, Durbecq V, Sotiriou C, Larsimont D, Piccart-Gebhart MJ, Paesmans M: **Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12,155 patients.** *Br J Cancer* 2007, **96**:1504–1513.

5. Dowsett M, Ebbs SR, Dixon JM, Skene A, Griffith C, Boeddinghaus I, Salter J, Detre S, Hills M, Ashley S, Francis S, Walsh G, Smith IE: **Biomarker changes during neoadjuvant anastrozole, tamoxifen, or the combination: influence of hormonal status and HER-2 in breast cancer--a study from the IMPACT trialists.** *J Clin Oncol* 2005, **23**:2477–92.

6. Dowsett M, Smith IE, Ebbs SR, Dixon JM, Skene A, Griffith C, Boeddinghaus I, Salter J, Detre S, Hills M, Ashley S, Francis S, Walsh G: **Short-term changes in Ki-67 during neoadjuvant treatment of primary breast cancer with anastrozole or tamoxifen alone or combined correlate with recurrence-free survival.** *Clin Cancer Res* 2005, **11**:951s–8s.

7. Scaltriti M, Rojo F, Ocaña A, Anido J, Guzman M, Cortes J, Di Cosimo S, Matias-Guiu X, Ramon y Cajal S, Arribas J, Baselga J: **Expression of p95HER2, a truncated form of the HER2 receptor, and response to anti-HER2 therapies in breast cancer.** *J Natl Cancer Inst* 2007, **99**:628–638.

8. Harvey JM, Clark GM, Osborne CK, Allred DC: **Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer.** *J Clin Oncol* 1999, **17**:1474–1481.

9. Johnston SR: **Acquired tamoxifen resistance in human breast cancer--potential mechanisms and clinical implications.** *Anticancer Drugs* 1997, **8**:911–930.

10. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406**:747–752.

11. Sotiriou C, Neo S-Y, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET: **Breast cancer classification and prognosis based on gene expression profiles from a population-based study.** *Proc Natl Acad Sci U S A* 2003, **100**:10393–10398.

12. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L, Nobel A, Parker J, Ewend MG, Sawyer LR, Wu J, Liu Y, Nanda R, Tretiakova M, Ruiz Orrico A, Dreher D, Palazzo JP, Perreard L, Nelson E, Mone M, Hansen H, Mullins M, Quackenbush JF, Ellis MJ, Olopade OI, Bernard PS, et al.: **The molecular portraits of breast tumors are conserved across microarray platforms.** *BMC Genomics* 2006, **7**:96.

13. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS a, Nobel AB, van't Veer LJ, Pero CM: **Concordance among Gene-Expression– Based Predictors for Breast Cancer**. *N Engl J Med* 2006, **355**:560–569.

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

15. Ciriello G, Sinha R, Hoadley K a, Jacobsen AS, Reva B, Perou CM, Sander C, Schultz N: **The molecular diversity of Luminal A breast tumors.** *Breast Cancer Res Treat* 2013, **141**:409–20.

16. Whitfield ML, George LK, Grant GD, Perou CM: **Common markers of proliferation.** *Nat Rev Cancer* 2006, **6**:99–106.

17. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale A-L, Brenton JD, Tavaré S, Caldas C, et al.: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.** *Nature* 2012, **486**:346–52.

18. The Cancer Genome Atlas Network: **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490**:61–70.

19. Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, Van Tine BA, Hoog J, Goiffon RJ, Goldstein TC, Ng S, Lin L, Crowder R, Snider J, Ballman K, Weber J, Chen K, Koboldt DC, Kandoth C, Schierding WS, McMichael JF, Miller CA, Lu C, Harris CC, McLellan MD, Wendl MC, DeSchryver K, Allred DC, Esserman L, Unzeitig G, et al.: **Whole-genome analysis informs breast cancer response to aromatase inhibition.** *Nature* 2012, **486**:353–60.

20. Mullins M, Perreard L, Quackenbush JF, Gauthier N, Bayer S, Ellis M, Parker J, Perou CM, Szabo A, Bernard PS: **Agreement in breast cancer classification between microarray and quantitative reverse transcription PCR from fresh-frozen and formalin-fixed, paraffin-embedded tissues.** *Clin Chem* 2007, **53**:1273–1279.

21. Tabchy A, Ma CX, Bose R, Ellis MJ: **Incorporating genomics into breast cancer clinical trials and care.** *Clin Cancer Res* 2013, **19**:6371–9.

22. Perou CM: **Molecular stratification of triple-negative breast cancers.** *Oncologist* 2011, **16 Suppl 1**:61–70.

23. Schneider BP, Winer EP, Foulkes WD, Garber J, Perou CM, Richardson A, Sledge GW, Carey L a: **Triple-negative breast cancer: risk factors to potential targets.** *Clin Cancer Res* 2008, **14**:8010–8018.

24. Ellis MJ, Perou CM: **The genomic landscape of breast cancer as a therapeutic roadmap.** *Cancer Discov* 2013, **3**:27–34.

25. Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick a M, Lawrence MS, Sivachenko a Y, Sougnez C, Zou L, Cortes ML, Fernandez-Lopez JC, Peng S, Ardlie KG, Auclair D, Bautista-Pi ña V, Duke F, Francis J, Jung J, Maffuz-Aziz a, Onofrio RC, Parkin M, Pho NH, Quintanar-Jurado V, Ramos a H, Rebollar-Vega R, Rodriguez-Cuevas S, Romero-Cordoba SL, Schumacher SE, Stransky N, et al.: **Sequence analysis of mutations and translocations across breast cancer subtypes**. *Nature* 2012, **486**:405–409.

26. Turner N, Tutt A, Ashworth A: **Hallmarks of "BRCAness" in sporadic cancers.** *Nat Rev Cancer* 2004, **4**:814–819.

27. Kriege M, Seynaeve C, Meijers-Heijboer H, Collee JM, Menke-Pluymers MBE, Bartels CCM, Tilanus-Linthorst MMA, van den Ouweland A, van Geel B, Brekelmans CTM, Klijn JGM: **Distant disease-free interval, site of first relapse and post-relapse survival in BRCA1- and BRCA2-associated compared to sporadic breast cancer patients.** *Breast Cancer Res Treat* 2008, **111**:303–311.

28. Jonkers J, Berns A: **Conditional mouse models of sporadic cancer.** *Nat Rev Cancer* 2002, **2**:251–265.

29. Kola I, Landis J: **Can the pharmaceutical industry reduce attrition rates?** *Nat Rev Drug Discov* 2004, **3**:711–715.

30. Lewandoski M: **Conditional control of gene expression in the mouse.** *Nat Rev Genet* 2001, **2**:743–755.

31. **PhRMA database** [http://www.phrma.org/innovation/meds-in-development]

32. Heyer J, Kwong LN, Lowe SW, Chin L: **Non-germline genetically engineered mouse models for translational cancer research.** *Nat Rev Cancer* 2010, **10**:470–480.

33. Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z, Rasmussen KE, Jones LP, Assefnia S, Chandrasekharan S, Backlund MG, Yin Y, Khramtsov AI, Bastein R, Quackenbush J, Glazer RI, Brown PH, Green JE, Kopelovich L, Furth PA, Palazzo JP, Olopade OI, Bernard PS, Churchill GA, Van Dyke T, Perou CM: **Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors.** *Genome Biol* 2007, **8**:R76.

34. Pfefferle AD, Herschkowitz JI, Usary J, Harrell JC, Spike BT, Adams JR, Torres-Arzayus MI, Brown M, Egan SE, Wahl GM, Rosen JM, Perou CM: **Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts.** *Genome Biol* 2013, **14**:R125.

35. Sharpless NE, Depinho RA: **The mighty mouse: genetically engineered mouse models in cancer drug development**. *Nat Rev Drug Discov* 2006, **5**:741–754.

36. Omer CA, Chen Z, Diehl RE, Conner MW, Chen HY, Trumbauer ME, Gopal-Truter S, Seeburger G, Bhimnathwala H, Abrams MT, Davide JP, Ellis MS, Gibbs JB, Greenberg I, Koblan KS, Kral AM, Liu D, Lobell RB, Miller PJ, Mosser SD, O'Neill TJ, Rands E, Schaber MD, Senderak ET, Oliff A, Kohl NE: **Mouse mammary tumor virus-Ki-rasB transgenic mice develop mammary carcinomas that can be growth-inhibited by a farnesyl:protein transferase inhibitor.** *Cancer Res* 2000, **60**:2680–2688.

37. Roberts PJ, Usary JE, Darr DB, Dillon PM, Pfefferle AD, Whittle MC, Duncan JS, Johnson SM, Combest AJ, Jin J, Zamboni WC, Johnson GL, Perou CM, Sharpless NE: **Combined PI3K/mTOR and MEK Inhibition Provides Broad Antitumor Activity in Faithful Murine Cancer Models**. *Clin Cancer Res* 2012, **18**:5290–5303.

38. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J, Steidl C, Holt RA, Jones S, Sun M, Leung G, Moore R, Severson T, Taylor GA, Teschendorff AE, Tse K, Turashvili G, Varhol R, Warren RL, Watson P, Zhao Y, Caldas C, Huntsman D, Hirst M, Marra MA, Aparicio S: **Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution.** *Nature* 2009, **461**:809–813.

39. Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL, Abbott RM, Hoog J, Dooling DJ, Koboldt DC, Schmidt H, Kalicki J, Zhang Q, Chen L, Lin L, Wendl MC, McMichael JF, Magrini VJ, Cook L, McGrath SD, Vickery TL, Appelbaum E, Deschryver K, Davies S, Guintoli T, Lin L, et al.: **Genome remodelling in a basal-like breast cancer metastasis and xenograft.** *Nature* 2010, **464**:999–1005.

40. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, Ha C, Johnson S, Kennemer MI, Mohan S, Nazarenko I, Watanabe C, Sparks AB, Shames DS, Gentleman R, de Sauvage FJ, Stern H, Pandita A, Ballinger DG, Drmanac R, Modrusan Z, Seshagiri S, Zhang Z: **The mutation spectrum revealed by paired genome sequences from a lung cancer patient.** *Nature* 2010, **465**:473–477.

41. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordonez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A, Carter RJ, Chen L, Cox AJ, Edkins S, Kokko-Gonzales PI, Gormley NA, Grocock RJ, Haudenschild CD, Hims MM, James T, Jia M, Kingsbury Z, Leroy C, Marshall J, Menzies A, et al.: **A comprehensive catalogue of somatic mutations from a human cancer genome**. *Nature* 2010, **463**:191–196.

42. Granato A, Giantin M, Ariani P, Carminato A, Baratto C, Zorzan E, Vascellari M, Bozzato E, Dacasto M, Mutinelli F: **DNA and RNA isolation from canine oncologic formalin-fixed, paraffin-embedded tissues for downstream "-omic" analyses: possible or not?** *J Vet Diagn Invest* 2014, **26**:117–24.

43. Meyerson M, Gabriel S, Getz G: **Advances in understanding cancer genomes through second-generation sequencing.** *Nat Rev Genet* 2010, **11**:685–696.

44. Pugh TJ, Delaney AD, Farnoud N, Flibotte S, Griffith M, Li HI, Qian H, Farinha P, Gascoyne RD, Marra MA: **Impact of whole genome amplification on analysis of copy number variants.** *Nucleic Acids Res* 2008, **36**:e80.

45. Jones SJ, Laskin J, Li YY, Griffith OL, An J, Bilenky M, Butterfield YS, Cezard T, Chuah E, Corbett R, Fejes AP, Griffith M, Yee J, Martin M, Mayo M, Melnyk N, Morin RD, Pugh TJ, Severson T, Shah SP, Sutcliffe M, Tam A, Terry J, Thiessen N, Thomson T, Varhol R, Zeng T, Zhao Y, Moore RA, Huntsman DG, et al.: **Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors.** *Genome Biol* 2010, **11**:R82.

46. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, Kalicki-Veizer J, Magrini VJ, Cook L, McGrath SD, Vickery TL, Wendl MC, Heath S, Watson MA, Link DC, Tomasson MH, et al.: **Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing.** *Nature* 2012, **481**:506–10.

47. Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, Kamiyama M, Hruban RH, Eshleman JR, Nowak MA, Velculescu VE, Kinzler KW, Vogelstein B, Iacobuzio-Donahue CA: **Distant metastasis occurs late during the genetic evolution of pancreatic cancer.** *Nature* 2010, **467**:1114–1117.

48. Misale S, Yaeger R, Hobor S, Scala E, Janakiraman M, Liska D, Valtorta E, Schiavo R, Buscarino M, Siravegna G, Bencardino K, Cercek A, Chen C-T, Veronese S, Zanon C, Sartore-Bianchi A, Gambacorta M, Gallicchio M, Vakiani E, Boscaro V, Medico E, Weiser M, Siena S, Di Nicolantonio F, Solit D, Bardelli A: **Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer.** *Nature* 2012, **486**:532–6.

49. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, et al.: **Landscape of transcription in human cells**. *Nature* 2012, **489**:101–108.

50. Eswaran J, Horvath A, Godbole S, Reddy SD, Mudvari P, Ohshiro K, Cyanam D, Nair S, Fuqua S a W, Polyak K, Florea LD, Kumar R: **RNA sequencing of cancer reveals novel splicing alterations.** *Sci Rep* 2013, **3**:1689.

51. Frattini V, Trifonov V, Chan JM, Castano A, Lia M, Abate F, Keir ST, Ji AX, Zoppoli P, Niola F, Danussi C, Dolgalev I, Porrati P, Pellegatta S, Heguy A, Gupta G, Pisapia DJ, Canoll P, Bruce JN, McLendon RE, Yan H, Aldape K, Finocchiaro G, Mikkelsen T, Privé GG, Bigner DD, Lasorella A, Rabadan R, Iavarone A: **The integrated landscape of driver genomic alterations in glioblastoma.** *Nat Genet* 2013, **45**:1141–9.

52. Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, Corbett RD, Johnson NA, Severson TM, Chiu R, Field M, Jackman S, Krzywinski M, Scott DW, Trinh DL, Tamura-Wells J, Li S, Firme MR, Rogic S, Griffith M, Chan S, Yakovenko O, Meyer IM, Zhao EY, Smailus D, Moksa

M, Chittaranjan S, Rimsza L, Brooks-Wilson A, Spinelli JJ, Ben-Neriah S, et al.: **Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma.** *Nature* 2011, **476**:298–303.

53. Casneuf T, Van de Peer Y, Huber W: **In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation.** *BMC Bioinformatics* 2007, **8**:461.

54. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**:621–628.

55. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby M a, Berlin AM, Sivachenko A, Thompson DA, Wysoker A, Fennell T, Gnirke A, Pochet N, Regev A, Levin JZ: **Comparative analysis of RNA sequencing methods for degraded or low-input samples.** *Nat Methods* 2013, **10**:623–9.

CHAPTER II

COMPARATIVE ONCOGENOMICS IDENTIFIES BREAST TUMORS ENRICHED IN FUNCTIONAL TUMOR INITIATING CELLS[1]

The claudin-low subtype is a recently identified rare molecular subtype of human breast cancer that expresses low levels of tight and adherens junction genes and shows high expression of epithelial-to-mesenchymal transition (EMT) genes. These tumors are enriched in gene expression signatures derived from human tumor-initiating cells (TIC) and human mammary stem cells. Through cross-species analysis, we discovered mouse mammary tumors that have similar gene expression characteristics as human claudin-low tumors and were also enriched for the human TIC signature. Such claudin-low tumors were similarly rare, but came from a number of distinct mouse models, including the p53 null transplant mouse model. Here we present a molecular characterization of fifty p53 null mammary tumors compared with other mouse models and human breast tumor subtypes.

Similar to human tumors, the murine p53null tumors fell into multiple molecular subtypes including two basal-like, a luminal, a claudin- low, and a subtype unique to this model. The claudin-low tumors also showed high gene expression of EMT inducers, low expression of the miR-200 family, and low to absent expression of both claudin 3 and E-cadherin. These murine subtypes also contained distinct genomic DNA copy number changes, some of which are similarly altered in their cognate human subtype counterpart. Finally, limiting dilution transplantation revealed that p53 null

[1] Herschkowitz JI, Zhao W, Zhang M, Usary J, Murrow G, Edwards D, Knezevic J, Greene SB, Darr D, Troester MA, Hilsenbeck SG, Medina D, Perou CM, Rosen JM: **Comparative oncogenomics identifies breast tumors enriched in functional tumor-initiating cells**. *Proc Natl Acad Sci* 2012, **109**(8):2778–2783.

claudin-low tumors are highly enriched for TICs as compared to the more common adenocarcinomas arising in the same model, thus providing a novel preclinical mouse model to investigate the therapeutic response of TICs.

INTRODUCTION

Breast cancer is the second leading cause of cancer-related deaths among women in the United States [1]. The large compendium of underlying genetic alterations and the resulting histological and molecular subtypes illustrate the heterogeneous nature of this disease. Both this inter-tumor heterogeneity and the cellular heterogeneity found within a breast tumor (intra-tumor heterogeneity) are major obstacles towards effective treatments. One common feature of breast cancers (and most cancers) is the loss of the tumor suppressor p53 function. p53 has been shown to be mutated in ~40% of breast cancers, associated with poor clinical outcomes, and a higher frequency of mutations occurs in more aggressive molecular subtypes including the basal-like subtype of human breast cancers [2].

Mice homozygous for p53 loss have been shown to develop lymphomas and sarcomas with a short latency [3, 4]. When crossed into the BALB/c background mammary tumors were observed in p53$^{+/-}$ mice [5]. To circumvent the appearance of other tumor types that occurred with short latency, the model was further modified [6]; namely, 6-wk-old p53$^{-/-}$ glands were removed and transplanted into 3-wk-old wild-type BALB/c recipients. These mice develop mammary tumors stochastically with an average latency of about 12 months. Interestingly, the p53 null epithelium initially forms normal ductal growths, which exhibit few genetic changes compared with wild-type outgrowths [7]. Unlike many transgenic mouse models, the p53 null tumor model exhibits histological heterogeneity reminiscent of human breast cancers, including a subset of the tumors expressing the estrogen receptor (ER). In addition anti-estrogens are able to significantly delay tumor formation in this model [8]. Lastly, these p53-deficient tumors exhibit genetic instability and/or aneuploidy, which likely play a critical role in progression [6].

Using gene expression profiling for classification, we show that like human tumors, p53 null mouse mammary tumors fall into multiple molecular groups including basal-like, luminal, and claudin-low subtypes. The claudin-low tumors contain a majority of spindle-shaped cells, a histology originally described for carcinosarcomas, which are now called EMT tumors [9]. Like their human counterparts, the p53 null claudin-low tumors exhibit high expression of EMT inducers and a core EMT signature [10]. Unlike many other mouse tumor models, p53 null tumors show extensive genomic instability. Accordingly, we determined that these different p53-deficient murine subtypes contain distinct genomic DNA copy number changes, as assessed by array comparative genomic hybridization (aCGH). We also show by limiting dilution transplantation, that p53 null claudin-low tumors have a marked enrichment of functional tumor-initiating cells (TICs). These data show the utility of this heterogeneous tumor model and provide for the first time functional data further demonstrating the stem-cell characteristics of the claudin-low subtype.

MATERIALS AND METHODS

Mice. All animal protocols were reviewed and approved by the Animal Protocol Review Committee at Baylor College of Medicine and University of North Carolina, Chapel Hill.

Preparation of single mammary tumor cells. Tumors were processed and digested into single cells as previously described [11]. The cells were resuspended in HBSS (Invitrogen) containing 2% FBS and 10 mM Hepes buffer (Invitrogen) before labeling with antibodies.

Flow cytometry. Cells were labeled with antibodies (Dataset S5) at a concentration of 10 × 10^6 cells/mL under optimized conditions and were subjected to FACS analysis and sorting on an ARIA II sorter (BD Biosciences). Data analysis was performed using FlowJo (v9.1).

Transplantation. Clearance of MECs and transplantation procedures were performed as previously described [12]. Following FACS, the designated number of cells were washed once with PBS and transplanted into the cleared fat pads of 21-day-old female BALB/c mice (Harlan).

Immunostaining. Paraffin-embedded sections (5 µm thick) were processed using standard immunostaining methods. Briefly, slides were deparaffinized and hydrated through a series of xylenes and graded ethanol steps. Heat-mediated epitope retrieval was performed in boiling citrate buffer (pH 6.0) for 15 min, then samples cooled to room temperature for 30 min. Secondary antibodies for immunofluorescence were conjugated with Alexa Fluor 488 or -594 fluorophores (1:200; Molecular Probes, Invitrogen). Immunofluorescent samples were mounted with VectaShield Hardset with DAPI mounting media (Vector Laboratories).

Real-Time PCR. Total RNA was prepared from tumors using the miRNeasy Kit (Qiagen). cDNA was synthesized from 10 ng of total RNA using the TaqMan MiRNA Reverse Transcription Kit with miRNA-specific RT primers (Applied Biosystems). miRNA levels were then measured using the miRNA-specific TaqMan probe provided in the MicroRNA Assays and the TaqMan Gene Expression Maser Mix (Applied Biosystems). miRNA levels were normalized to snoRNA55 and U6 (Applied Biosystems). Student's t test was used to compare claudin-low vs. the rest.

Microarray analysis. Total RNA was collected from 45 murine tumors and purified using the Qiagen RNeasy Mini Kit according to the manufacturer's protocol using ~25 mg tissue. RNA integrity was assessed using the RNA 6000 Nano LabChip kit followed by analysis using a Bioanalyzer (Agilent). Two ug of total RNA was reverse transcribed, amplified and labeled with Cy5 using a Low RNA Input Amplification kit (Agilent). The common reference RNA sample for these experiments was as previously described [13]. The reference RNA was reverse transcribed, amplified, and labeled with Cy3. The amplified sample and reference were co- hybridized overnight to Agilent Mouse Oligo 4x44K Microarrays. They were then washed and scanned on an Agilent scanner (G2505B), and uploaded into the database where a Lowess normalization is automatically performed. The genes for all analyses were filtered by requiring the Lowess normalized intensity values in both channels to be >10. The $\log_2$ ratio of Cy5/Cy3 was then reported for each gene. In the final dataset, only genes that reported values in 70% or more of the samples were included.

Microarray platform correction. Previously published data on 22K arrays can be found under accession no. GSE3165 in the Gene Expression Omnibus database. Platform correction (i.e., 44K vs. 22K arrays) was performed by making a systematic, gene-by-gene correction based on similar samples across platforms. For both 22K and 44K arrays, six to eight tumors from MMTV-Neu and C3(1)-Tag and two pairs of BALB/c p53$^{+/-}$ were assayed on each platform, and for each gene on each platform a median expression ratio was determined. The assumption is then made that the median expression ratio on each platform should be the same, so then an adjustment factor is determined for each gene using these similar tumor samples across platforms. Next, all samples on the 44K platform were adjusted using this factor. The complete data set contains the previous 122 arrays, the new 45 p53 null samples/arrays, and two p53$^{+/-}$, six MMTV-Neu, and eight C3(1)-Tag 44K arrays used for adjustment. Hierarchical clustering was then performed using the mouse 866 intrinsic gene list [14], which shows 669 genes in common across these two platforms. The genes were median centered and then hierarchical clustered using Centroid linkage with gene and array “correlation centered” using Cluster v3 [15], and cluster viewing and display was performed using JavaTreeview v1.0.8 [16]. SigClust was then performed as described by Liu et al. [17], to identify the significant clusters/groups of samples.

Gene expression signatures. A number of different signatures, and many individual genes, were also tested for associations with the five p53 null tumor subtypes. For these analyses, the signatures/modules used were taken from the set of 298 signatures/modules described previously by Fan et al. [18], which contains ~100 previously published signatures and ~200 signatures coming from newly performed unsupervised analyses. Using just the subset of tumors/arrays specific for each of the five p53 null SigClust-defined groups (34 arrays total), ANOVA were performed in R and the data displayed using a box-and-whisker diagram, with the statistical test asking whether a given gene (or signature) shows average class expression differences, when considering all classes simultaneously.

Array Comparative Genomic Hybridization. Genomic DNA was collected from 44 p53 null tumors and purified using the Qiagen DNeasy Blood and Tissue Kit. DNA was labeled according to the direct incorporation method, hybridized to Agilent 244k CGH arrays (G4415A), and scanned. As a control, DNA from FVB mice was collected and labeled as Cy3, and BALB/c p53 null tumors were labeled with Cy5, thus providing a ratio of Cy3/Cy5 for all 244,000 features. All of the aCGH probes were filtered for>10 normalized intensity in control channel. The log2 ratios of Cy5/Cy3 were reported. Probes that have missing values in greater than 30% of samples, probes mapped to ChrN_random or chromosome Y, as well as the unmapped probes, were excluded. Arrays that have missing values in greater than 60% of probes were excluded. Missing values were k-NN imputed within chromosome. The final dataset contained 231,894 probes. Chromosomal physical positions of probes were annotated in mouse genome (National Center for Biotechnology Information Build 36).

All primary microarray data and aCGH data is available from the UNC Microarray Database (UMD) https://genome.unc.edu/, and at the Gene Expression Omnibus (GEO) http://www.ncbi.nlm.nih.gov/geo/ under the series GSE27101.

Identification of subtype-specific DNA Copy Number Alterations. Copy number aberration events associated with each subtype (one subtype vs. the rest of the p53 null samples) were identified using 34 arrays representing the five subtypes only. Two levels of this analysis occurred next. First, at the genome level, SWITCHdna [19] was used to identify significantly altered regions/segments and to determine the frequency of each copy number event (by segment) within each subtype, which was then visualized in the copy number landscape plots. Briefly, the SWITCHdna method first identifies the transition point in chromosome/probe copy number assessments by calculating the F statistics recursively, then tests the significance of segments according to the segment's average intensity and segment size. With the default setting of SWITCHdna, 15,469 segments were defined in the 34 array comparative genomic hybridization slides. The copy number of each segment was calculated by taking the average value of probes in this region.

To analyze a specific small region, we performed a one subtype vs. the rest Student t test in a segment-by-segment manner. Mouse segments were rearranged in the order of human chromosome positions by chained alignments of human genome (National Center for Biotechnology Information Build 36) against mouse (University of California, Santa Cruz Genome Browser; http://genome.ucsc.edu) [20]. Last, we note that the control DNA was FVB, whereas all tumor samples were BALB/c, therefore, regions of 100% gain or 100% loss could be attributed to strain-specific germline copy number differences, and caution is needed in interpreting these data; however, the strain differences will not affect any of the subtype-specific analyses.

Statistical analysis of limiting dilution transplantation. Limiting dilution transplantation data were analyzed using a binomial generalized linear model with a complementary log–log link [21, 22](7, 8). After determining that assumptions of the Single Hit Poisson Model were not met [22], we used the more general model, fitting parameters for slope, intercept, and interaction. After verifying lack of significant interaction between dose and cell line, we tested for the main effect of cell line. Results were summarized as the “fold change in dose” (FCD) required for equal take rates. The FCD 95% confidence interval was computed from the covariance matrix of the model parameters using the delta method (p43 in [23]) and back-transformation by exponentiating.

RESULTS

p53 null tumors show variable histology

Previously, we hypothesized that the heterogeneity observed in human breast cancer might arise not only due to activation of different oncogenes or loss of specific tumor suppressor genes, but might also be dependent upon the cell of origin in which these genetic changes occur [24]. Initially transplanted p53 null mammary epithelial cells gave rise to phenotypically normal ductal outgrowths, which then stochastically developed mammary tumors. We, therefore, hypothesized that the deletion of this single tumor suppressor gene might give rise to a spectrum of heterogeneous tumors, depending upon the cell of origin in which additional stochastic changes occurred. To test this hypothesis, we collected 44 p53 null tumors that arose in wild-type BALB/c mice after transplantation of p53 null BALB/c mammary tissue into the cleared fat pads of 3-wk-old mice in addition to 6 tumors arising spontaneously in p53 null glands (without prior transplantation) [6]. Like some other genetically engineered mouse (GEM) mammary tumor models, the p53 null model gave rise to tumors with a diversity of histological phenotypes(Figure 2.1, Table 2.1) [9, 11]. Approximately 10% of the tumors contained a majority of spindle-shaped cells, a histology originally described for carcinosarcomas, now called EMT tumors [9].

p53 null tumors cluster into distinct tumor subtypes

In a previous study, we profiled 13 distinct mouse models including the p53 null model [25]. However, with only five p53 null tumor samples, we were not able to appreciate the full spectrum of molecular heterogeneity represented in this mouse model. Now, with a total of 50 tumors from the p53 null model, we see that these tumors cluster into five distinct tumor subtypes when performing hierarchical clustering analysis using our previously defined mouse intrinsic gene list [25] (Figure 2.2); furthermore, we used SigClust [17] to assess the significance of this clustering and objectively determined that the p53 null model did populate multiple statistically significant groups/subtypes, as follows.

Basal-like: Two groups of basal-like mouse mammary tumors were observed (Figure 2.2); in part, we define these groups as basal-like according to their high expression of known basal markers including keratin 5 (K5), ID4, and TRIM29 (Figure 2.2d), and selective high expression of the human basal-like tumor expression cluster (Figure 2.3). Basal 1 tumors (5/50, 10%) clustered along with a group of other mouse basal-like tumors including, BRCA1-deficient and MMTV-Wnt1 tumors. This group contained 4/6 spontaneous p53 null tumors. Basal 2 tumors, (8/50, 16%) clustered next to the Basal1 tumors, but showed a higher expression of the murine luminal cluster than did Basal 1 (Figure 2.2c). Basal 1 p53 null tumors showed an increased proliferation signature separating them from Basal 2 and the other p53 null subtypes (Figure 2.2g, Figure 2.3), and they also showed high p16 expression, which is a hallmark of impaired RB1 function [14]. Basal tumors (8/9) tested stained positively for K5 as expected (Figure 2.4 and Table 2.1); however, paradoxically, 5/8 tested stained positively for the estrogen receptor (ER) of which 4/5 were of the Basal 2 subtype.

Luminal: 8/50 (16%) of the p53 null tumors clustered close together and with the mouse luminal models MMTV-Neu and MMTV-PyMT. As we have seen for other luminal models, these tumors express luminal specific genes like XBP1, but are missing ER and estrogen responsive genes; accordingly, only 1/8 of the luminal tumors stained positively for ER. Interestingly, like human luminal tumors, p53 null luminal tumors showed lower levels of p18^{INK4C}, and p18 null mice develop predominantly luminal-type mammary tumors [26].

Claudin-low: 5/50 (10%) of the p53 null tumors showed the murine claudin-low expression phenotype (Figure 2.2f) and significantly clustered with the previously defined murine claudin-low tumors. These tumors had an EMT tumor histology (Figure 2.1) and showed expression of the human claudin-low signature (Figure 2.3). In agreement with the gene expression data and immunohistochemistry on human samples [27], we observed low to absent to absent expression of CLDN3 and CDH1 by immunofluorescent (IF) staining (Figure 2.4). Like human claudin-low tumors, these tumors highly express markers of EMT [27] and the previously determined EMT core signature (Figure 2.5a) [10]. All p53 null tumors tested stained positively for the luminal marker, keratin 8 (K8), including the claudin-low tumors (thus suggesting an epithelial origin), however they often exhibited comparatively less staining (Figure 2.4).

P53 null subtype: an additional 13/50 (26%) of p53 null tumors clustered into a unique group made up exclusively of tumors from this BALB/c p53 null model. Tumors of this subtype appear to not show high expression of any of the other tumor subtype defining clusters. Lastly, 11/50 p53 null tumors clustered separately from these 5 groups and without a consistent group signature. Thus, at least 5 expression subtypes/phenotypes can be found from this single murine model, three of which mimic previously known human tumor subtypes (luminal, basal- like, and claudin-low).

miRNAs

Because expression of a number of specific miRNAs has been associated previously with an EMT transition [28, 29], we took a candidate approach to identify miRNAs that were differentially expressed between p53 null claudin-low tumors and the other subtypes. First we evaluated the miR-200 family of miRNAs and miR-205; which are miRNAs that have been implicated in EMT and TICs [28, 30]. While a number of targets for these miRNAs have now been identified, important targets with respect to EMT are ZEB1 and ZEB2. ZEB1 and ZEB2 are expressed at high levels in claudin-low tumors (mouse and human), and as expected, these miRNAs were present at very low levels relative to the other p53 null tumors (Figure 2.5b). Another cluster of miRNAs expressed at lower levels in both cancer and normal mammary stem cells contains miRNAs 182, 96, and 183 [30]. Likewise, this cluster of miRNAs was expressed at low levels in murine p53 null claudin-low tumors. Additionally, miR-203, another stem-ness-inhibiting miRNA regulated by Zeb1 [31], was also expressed at low levels in claudin-low tumors. Marked decreases, however, were not seen for all miRNAs tested (e.g. miR-21).

It has been shown that human breast tumor subtype correlates with miRNA profiles [32, 33]. We re-analyzed the Blenkiron et al. dataset [32] to determine which tumors contained the claudin-low gene expression pattern using the Prat claudin-low predictor [27]. Using a supervised analysis, 17 miRNAs were identified that were significantly differentially expressed between claudin-low tumors versus the other breast cancers (Table 2.2). This included 7 of the miRNAs that we had observed including; miR-200a, 200b, 200c, 149, 182, 183, and 203. These results indicate that in addition to mRNA gene expression changes, mouse and human claudin-low tumors share common miRNA expression patterns.

p53 null tumor subtypes display distinct copy number alterations

Presumably stochastic genetic alterations selected during neoplastic progression collaborate with the loss of p53. It is also likely that different genetic events can cause tumors to show a given phenotype, or only sensitize one particular cell type to malignant transformation; thus specific copy number and/or mutations may be highly enriched within a specific tumor subtype, as shown for human breast tumors [34]. In order to investigate this on the genomic level, we performed aCGH on 44 p53 null tumors using Agilent 244,000 feature DNA microarrays to determine whether there were subtype specific Copy Number Alterations (CNAs) (Figure 2.6). In comparison with many mouse models [35, 36] the p53 null mammary tumors contained a fair amount of genomic instability. Interestingly, all five tumor subtypes contain distinct CNAs (data not shown). In the p53 null basal-like tumors (both Basal1 and Basal2 considered together), there was loss of the distal half of chromosome 8, including INPP4B, which has now been shown to be selectively lost in human basal-like/triple-negative tumors (4q31.22-q35.2(12)) [37, 38]. p53 null luminal tumors showed loss of chromosome 4 and gain of chromosome 7. The p53 null unique subtype showed very few subtype specific events, however, when converted to human genomic coordinates, these events identify amplification of human chromosome 17q12-q21.2(2), which is a common amplicon that is distal to the HER2 amplicon. Interestingly, one of these murine tumors (2304L) that clusters in the p53 null unique subtype, but which is not contained within the SigClust defined group (Figure 2.2b), showed high Her2 mRNA and protein expression, and was amplified for Her2 on mouse chromosome 11 (Figure 2.7); thus the p53 null model is even able to generate HER2-amplified tumors, albeit at a low frequency.

The copy number landscape of human claudin-low tumors is not known, but the p53 null claudin-low tumors showed numerous subtype-enriched CNAs. These changes included the near-complete loss of mouse chromosome 1, and frequent but smaller losses on 7, 12, and 14. There were also specific gains on 3, 8, and 13. Many of these map to common regions of copy number changes in human breast cancer; however, additional studies will be required to define the driving mutations/changes present in each region. Nonetheless, these claudin-low subtype- specific copy number changes do suggest the possible existence of driver mutations/changes.

The work of Bergamaschi et al. [34] identified numerous CNA associated with some of the intrinsic subtypes. We, therefore, converted our mouse CNA into human equivalent chromosome locations and determined that a number of significantly altered regions were in common between p53 null mouse tumors and human breast tumors (Table 2.3). Of note were the loss of two regions that occurred in both mouse and human basal-like tumors, human 4q31.22-q35.2(12) that contains INPP4B, and human 14q22.1-q23.1(4); the somatic loss of these two regions across species suggests that each contains a tumor suppressor(s) gene, and that the loss of these genes may sensitize cells to become the basal-like subtype, similar to germline inactivation of BRCA1 [39].

p53 null claudin-low tumors are enriched for tumor-initiating cells

Similar to their utility in the isolation of mouse mammary stem cells [40], CD29 and CD24 have been employed as markers that enrich for TICs in the p53 null tumor model, with the $CD29^{+}/CD24^{+}$ fraction showing the TIC capabilities [11]. Furthermore, an EMT program has been shown to correlate with stem-like properties, and the loss of miR-200 expression as well as a "claudin-low" signature has been suggested to characterize both normal and cancer stem cells [30]. By FACS analysis, in the p53 null claudin-low tumors tested, the percentage of double positive cells was 70-85% as compared a maximum of 14% in the other p53 null tumors analyzed(Figure 2.8b,c).

Interestingly, some luminal tumors exhibit very low levels of double-positive cells. This was suggestive, therefore, that there might be a high percentage of TICs in the claudin-low tumors.

To test this hypothesis, two different claudin-low p53 null tumors were FACS sorted for all four possible populations using CD29 and CD24, and limiting dilution transplantation was performed (Tables 2.4 and 2.5). The tumor initiating frequency was similar between the $CD29^+/CD24^+$ and $CD29^+/CD24^-$ fractions, and these two populations were highly enriched for TICs as compared to the other two fractions. In addition, by transplanting FACS-sorted lineage-negative cells in limiting dilution, we determined that the tumor repopulating ability of these claudin-low phenotype tumors was >38 times greater than that of two other p53 null adenocarcinomas (T1 and T7) performed using the same methods(Figure 2.8d) [41]. Thus, these data indicate that an expanded population of TICs exists within these murine claudin-low tumors.

DISCUSSION

GEMM have provided a rich resource for the study of different cancers; however, many individual models show significant molecular and histological heterogeneity [25]. This heterogeneity complicates studies as multiple disease types may actually be present within a given model. One way to address this heterogeneity is to genomically characterize each tumor, then group tumors together according to important features and, most importantly, perform functional studies. The p53 null mammary transplant model is one such heterogeneous model, and we have taken advantage of this feature and identified transplantable lines that represent at least three human breast tumor subtypes. In addition, since all these tumors develop subsequent to the same initial loss of p53, the question is whether this heterogeneity is due to different collaborating oncogenes/tumor suppressors and/or different cells of origins. The cell type of origin in cancer is a highly debated topic (reviewed recently in [42]). Although specific genetic lesions clearly play a major role in determining the tumor phenotype, growing evidence indicates that cancers of different subtypes within an organ system may also reflect distinct cells of origin. However, it is not apparent whether a given oncogenic lesion actually dictates the cell of origin or, conversely, whether the cell of origin determines which oncogenic lesions can occur. Both of these possibilities most likely exist. There is evidence that tumors generated using the same oncogene targeted to different cell lineages can be phenotypically distinct [43]. Recent studies have shown that BRCA1 mutant and basal-like human tumors are enriched in gene expression profiles and surface markers of luminal progenitors [44]. Similarly, inactivation of BRCA1 (and p53) in the luminal or basal cell population of the mouse mammary gland showed that only the luminal cells gave rise to tumors histologically resembling those of BRCA1 mutation carriers [45]. These results, however, fall short of actually proving that these tumor types originated in these cell types. Mouse models, like the heterogeneous one presented here, can provide an invaluable tool with which to decipher the cell of origin when genetics is combined with precise lineage tracing. At present, we cannot definitively answer the cell of origin question without performing lineage tracing experiments, as done recently using mouse models of intestinal cancer [46]. However, our experiments do provide several important insights: First, tumors of the basal-like, luminal, and claudin-low phenotypes clearly arise, although at different frequencies and with a predilection for basal-like; in particular, the Basal 1 group appears to most faithfully recapitulate human basal-like tumors in that it shows high expression of basal gene expression features, of the proliferation signature, and of p16 (a hallmark of RB1 loss), all of which are features of human basal-like tumors [14]. Second, the luminal tumors that do arise are largely ER-negative (as are the vast majority of murine tumors from other GEMMs) and thus fundamentally more similar to a luminal B than the ER+ luminal A human subtype.

Interestingly, claudin-low p53 null tumors were also seen, although at the lowest frequency (5 total). As was shown for human claudin-low tumors and cell lines, these murine tumors lack tight junction proteins including claudin 3and E-cadherin, and show expression features of mesenchymal cells, normal mammary stem cells, and TIC. In addition to previously defined subtypes, we also identified a new phenotype unique to this model, and noted that nearly 20% of the tumors were scattered throughout the cluster, indicating even greater heterogeneity within this model. For example, tumor 2304L showed clear amplification and high expression of HER2, thus even somatically HER2 amplified tumors occur within this model.

The presence of specific copy number alterations in different subtypes of tumors arising in the p53 null model suggests that different gains and losses are important for tumor progression subsequent to p53 loss, and these are possibly occurring within different cell types. In the p53 null basal-like tumors, there was specific loss of the distal half of chromosome 8, which is in conserved synteny with human chromosome 4. Recently, loss of this region has been seen specifically with human basal-like/triple-negative breast cancers, and it is thought that the target of this loss is INPP4B. This gene is selectively low in human and murine basal-like tumors, thus suggesting that this approach of finding common regions of loss/gains across species can identify putative important tumor and/or subtype causative events. Interestingly, p53 null luminal tumors showed loss of chromosome 4. Chromosome 4 deletions and loss of heterozygosity have been reported in other luminal mouse models, including MMTV-Neu, MMTV-Myc, and MMTV-Ras [36, 47–49]. Presumably there exist multiple luminal-specific tumor suppressor genes on chromosome 4. Although other subtypes showed gain of chromosome 1, p53 null claudin-low tumors showed large regions of loss on chromosome 1, which again hints at their uniqueness.

Several lines of evidence have suggested that claudin-low tumors are enriched in functional TICs, predominantly coming from expression analyses (Figure 2.5a, Figure 2.3). However, due to their rarity and limitations in procurement of primary human claudin-low tumors, this hypothesis has not been tested functionally using human clinical samples. We have, however, herein identified a counterpart of human claudin-low tumors in the mouse. Accordingly, we have taken advantage of this mouse model to test by limiting dilution, the gold standard functional stem cell assay, whether these tumors are enriched in TICs compared with other tumors arising in the same model. With the p53 null model we also have the advantage of being able to transplant these tumors into syngeneic mice with an appropriate microenvironment complete with normal immune function. We showed that the claudin-low murine tumors were significantly more enriched for surface markers of TICs as well as functional TICs as compared to other p53 null tumors. Recent studies have shown that minority subsets of tumors from MMTV-Myc and MMTV-MET tumors cluster with our claudin-low mouse tumors [50, 51]. It has not been determined if they too are enriched in functional TICs. However the MMTV-Myc EMT- like/claudin-low tumors were reported to show an increase in metastasis.

The murine claudin-low tumors show large percentages of $CD29^+/CD24^+$ cells, MaSC-like mRNA and miRNA expression profiles, and expression of other markers of MaSCs (e.g. high s-SHIP expression [52]). Therefore it is conceivable that these tumors might have arisen from the MaSC population. Alternatively, they may have resulted from dedifferentiation of a progenitor or even a more differentiated cell. Lineage tracing experiments will be required to definitively resolve this issue.

To effectively target cancer stem cells or TICs, one pressing need is a genetically defined and renewable preclinical model to identify and test new stem cell targeted therapies. To address this need, we now have identified a mouse model that develops claudin-low tumors, in which the bulk of the tumors cells appear to be TICs. This is an example of a spontaneously occurring breast tumor with a high proportion of TICs. Thus, we now have appropriate and validated models for the investigation of important signaling pathways and therapeutics. Due to their transplantability into syngeneic hosts, this panel of tumors provides a valuable resource for preclinical testing of novel therapeutics. These tumors should serve as excellent models for both the general study of breast cancer stem cells and preclinical models for testing stem cell targeted agents enabling translation into the clinic. Finally, the finding that claudin-low tumors have an enrichment of functional TICs challenges the popular paradigm that TICs always need be a rare subpopulation [53].

TABLES

Table 2.1 Summary of p53 null tumor samples.

Tumor	Model	Pathology	Latency	Molecular Subtype	Keratin 5	Keratin 8	Keratin 14	Vimentin
2331L	transplant	IDC - low grade (apocrine)	23 wk		positive	positive	positive	positive
2151R	transplant	mesenchymal	32 wk	CLOW	negative	low positive	negative	positive
2151L	transplant	IDC low grade	32 wk	TP53	negative	positive	<10% positive	negative
2412R	transplant	IDC low grade	23 wk	Basal 2	positive	positive	positive	stroma
2153L	transplant	IDC low grade	37 wk	Basal 2	positive	positive	positive	positive
2224L	transplant	IDC low grade	35 wk	Basal 2	positive	positive	positive	stroma
2243L	transplant	IDC low grade	37 wk	luminal	negative	positive	rare	stroma
2247R	transplant	mesenchymal	37 wk	CLOW	negative	positive	negative	positive
2336R	transplant	IDC low grade	30 wk	Basal 2	negative	positive	negative	stroma
2245R	transplant	IDC low grade	34 wk		negative	positive	positive	
2153R	transplant	IDC low grade	40 wk	TP53	negative	positive	negative	positive
2304L	transplant	IDC low grade	38 wk		negative	positive	negative	positive
2225L	transplant	IBC (myoepi)	38 wk	Basal 1	positive	positive	positive	stroma
2333R	transplant	IDC	32 wk	TP53	negative	positive	positive	stroma
2250L	transplant	IDC low grade (apocrine)	41 wk	luminal	negative	positive	negative	stroma
2208L	transplant	IDC	43 wk	luminal	positive	positive		stroma
2225R	transplant	IDC low grade	44 wk	TP53	negative	positive	negative	stroma
T11 (753R)	transplant	mesenchymal	25 wk	CLOW	negative	low positive	negative	positive
2228R	transplant	IDC low grade	48 wk		negative <1%	positive		
2249L	transplant	IBC - myoepi +	46 wk	Basal 2	positive	positive		
2356R	transplant	IDC high grade (very undiff.)	41 wk	Basal 2	positive	positive		
2374R	transplant	IBC - EMT?	38 wk	TP53	positive	positive		
2374L	transplant	EMT	39 wk	TP53	negative	positive		
2397L	transplant	IDC low grade	35 wk	TP53	negative	positive		
2211R	transplant	IDC high grade (undiff.)	47 wk	luminal	positive	positive		
2211L	transplant	IDC high grade	47 wk	luminal	negative	positive		
2209R	transplant	IDC high grade	48 wk		negative	positive		
2530R	transplant	IDC high grade	32 wk	Basal 2	positive	positive		
2349R	transplant	IDC high grade	47 wk	luminal	negative	positive		
2349L	transplant	IDC high grade	47 wk	luminal	positive	positive		
2350R	transplant	IDC high grade	47 wk	luminal	positive	positive		
2154L	transplant	IDC high grade	50 wk		positive	positive		
2210L	transplant	IDC high grade	52 wk	TP53	positive	positive		
2377R	transplant	IDC high grade	45 wk	TP53	negative	positive		
2396R	transplant	IDC high grade	47 wk	Basal 2	positive	positive		
2376R	transplant	IDC high grade	48 wk	NOT ARRAYED	positive	positive		
2393R	transplant	IDC high grade	52 wk		positive	positive		
T1	transplant	IDC	40 wks		positive	positive	positive	
T2	transplant	IDC	46 wks		positive	positive	positive	
T7	transplant	?	N/A	TP53	negative	positive	positive	
1634R	transplant	IDC high grade	50 wk					
2657R	transplant	mesenchymal	51 wk	CLOW				
3939R	transplant	N/A	45 wk	TP53				
3941R	transplant	IDC low grade	48 wk	TP53				
4304R	transplant	N/A	66 wk	TP53				
4706L	transplant	IDC high grade	59 wk	CLOW				
4100R	transplant	IDC	41 wk	Basal 1				
4127R	transplant	IDC	44 wk	Basal 1				
4702L	transplant	N/A	56 wk	Basal 1				
4729L	transplant	IDC high grade	47 wk	Basal 1				
2297R	transplant	IDC low grade	27 wk					

Table 2.1 Summary of p53 null tumor samples. (Continued)

Tumor	SMA	ERα	E-cadherin	K19	claudin 3	HER2	Keratin 6
2331L	very positive	negative	positive	negative	positive		
2151R	positive	negative	low positive	negative	negative	low (not membrane)	negative
2151L	positive	<1% negative	positive	negative	positive	low	
2412R	positive	positive	positive	positive	positive	negative	positive
2153L	positive			negative			
2224L	positive	negative	positive	negative			
2243L	stroma	negative	positive	negative		negative	
2247R	positive	positive only in epithelial	low positive		negative	negative	negative
2336R	positive	positive	positive				
2245R		positive (heterogeneous)	positive				negative
2153R	positive	<1% positive	positive				
2304L	stroma	negative	positive			positive	negative?
2225L	positive	negative	positive			negative	negative
2333R	stroma	negative	positive				
2250L	stroma	negative	positive				
2208L	stroma	negative	positive			negative	
2225R	stroma	negative	positive				
T11 (753R)	positive	negative	low positive	negative	negative/low	low	negative
2228R		negative	positive				
2249L		positive	positive				
2356R		positive ~10%	positive				
2374R		negative	positive				
2374L		negative	low				
2397L		negative	positive ?				
2211R		negative	positive				
2211L		positive	positive				
2209R		negative	positive				
2530R		negative	positive				
2349R		negative	positive				
2349L		negative	positive				
2350R		negative	positive				
2154L		negative	positive				
2210L		negative	negative				
2377R		negative	negative				
2396R		positive	positive				
2376R		positive	negative				
2393R		positive	positive				
T1		negative					
T2		positive					
T7		negative					
1634R							
2657R							
3939R							
3941R							
4304R							
4706L							
4100R							
4127R							
4702L							
4729L							
2297R							

Table 2.2 miRNAs differentially expressed in human claudin-low tumors

miRNA ID	FC	FC (log2)	Direction	Score	Numerator	Denominator	q-value(%)
hsa-miR-31	1.860	0.895	up in claudin	-2.548	-0.895	0.351	0.000
hsa-miR-221	1.732	0.792	up in claudin	-2.398	-0.792	0.330	0.000
hsa-miR-382	1.662	0.733	up in claudin	-2.419	-0.733	0.303	0.000
hsa-miR-146b	1.499	0.584	up in claudin	-1.891	-0.584	0.309	5.109
hsa-miR-224	0.721	-0.473	down in claudin	1.670	0.473	0.283	4.208
hsa-miR-200a	0.718	-0.478	down in claudin	1.509	0.478	0.317	4.208
hsa-let-7f	0.706	-0.503	down in claudin	1.708	0.503	0.295	4.208
hsa-miR-30d	0.697	-0.521	down in claudin	1.861	0.521	0.280	0.000
hsa-miR-149	0.680	-0.557	down in claudin	1.544	0.557	0.361	4.208
hsa-miR-203	0.677	-0.563	down in claudin	1.702	0.563	0.331	4.208
hsa-miR-183	0.649	-0.623	down in claudin	1.882	0.623	0.331	0.000
hsa-miR-30a-5p	0.638	-0.648	down in claudin	1.989	0.648	0.326	0.000
hsa-miR-200c	0.628	-0.670	down in claudin	2.584	0.670	0.259	0.000
hsa-miR-182	0.615	-0.700	down in claudin	2.265	0.700	0.309	0.000
hsa-miR-200b	0.606	-0.724	down in claudin	3.055	0.724	0.237	0.000
hsa-miR-187	0.587	-0.768	down in claudin	1.498	0.768	0.513	4.208
hsa-miR-375	0.482	-1.051	down in claudin	2.269	1.051	0.463	0.000

Table 2.3 Comparison with Bergamaschi et al. [90]

	LumA	LumB	ERBB2	Basal
1p34.1-p34.2(2)		▨		
1q12-q23.3(8)	▨			▨
1q25.2	🟥			▨
1q31.1	🟥			▨
1q31.3	▨			▨
1q32.1	▨			🟥
1q41	🟥			▨
3p11.2		▨		▨
3q12.1-q12.3(3)		▨		▨
3q21.1				▨
4p15.2-p15.32(3)				▨
4q31.22-q35.2(12)				🟩 Δ
5q11.1-q11.2(2)				▨
5q12.1				▨
5q12.3-q14.2(6)				▨
5q15				▨
5q21.1				▨
5q21.3				▨
5q22.1-q31.3(9)				▨
6p12.1-p25.3(18)				▨
6q22.33		▨		▨
7p22.1-p22.2(2)		🟥		
7q21.12				▨
7q22.1				▨
7q32.2-q34(4)				▨
7q36.1-q36.3(3)				▨
8q11.21		▨		
8q11.23		▨		
8q12.1-q24.3(24)		▨		
9q34.13		🟥		
10p12.33-p15.3(6)				▨
11p11.2				▨
12p12.3				▨
12q22				▨
14q22.1-q23.1(4)				🟩 Δ
15q22.2				▨
16p12.1-p12.2(2)	🟥			
16p13.2-p13.3(2)	▨			
17q12-q21.2(2)			🟥 Δ	
17q25.2-q25.3(2)				▨
19q13.32-q13.33(2)		🟥		
20p12.2		🟥		
20q13.13-q13.33(5)		🟥		
21q22.12-q22.3(4)				▨
16p12.1-p12.2(2)				

Table 2.4 Limiting dilution transplantation of adenocarcinomas (T1 and T7)

Cells injected	**5000**	**1500**	**1000**	**100**	**50**	**25**	**10**
Lin^{-}CD29^{H}CD24^{H}		4/4	2/2	12/12	12/12	4/12	3/12
Lin^{-}CD29^{H}CD24^{L}		2/4	4/6	4/12	2/12	0/8	0/6
Lin^{-}CD29^{L}CD24^{H}	4/5	2/8	2/7	0/8	0/6	0/2	0/6
Lin^{-}CD29^{L}CD24^{L}	2/6	2/8	0/7	0/8	0/6	0/2	0/6
Lin^{-}	2/3	6/10	4/9	2/12	1/10	0/4	

Data from Zhang et al. [97]

Table 2.5 Limiting dilution transplantation of claudin-low tumors (T11 and 2247R)

Cells injected	**500**	**250**	**100**	**50**	**10**
Lin^{-}CD29^{H}CD24^{H}	2/2	2/2	5/6	5/6	3/6
Lin^{-}CD29^{H}CD24^{L}	1/2	2/2	5/6	5/6	3/4
Lin^{-}CD29^{L}CD24^{H}	2/2	2/4	0/4	2/6	0/2
Lin^{-}CD29^{L}CD24^{L}	0/2	3/4	2/4	2/6	0/2
Lin^{-}	2/2	1/2	8/8	5/8	2/12

FIGURES

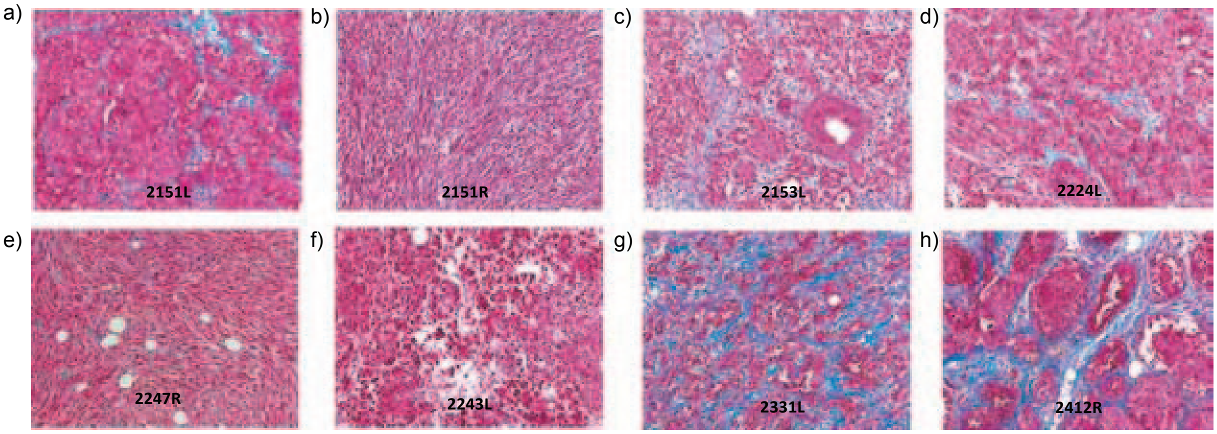


Figure 2.1 Morphological features of p53 null mammary tumors. p53 null tumors display variable histological features, including spindloid tumors (b and e).

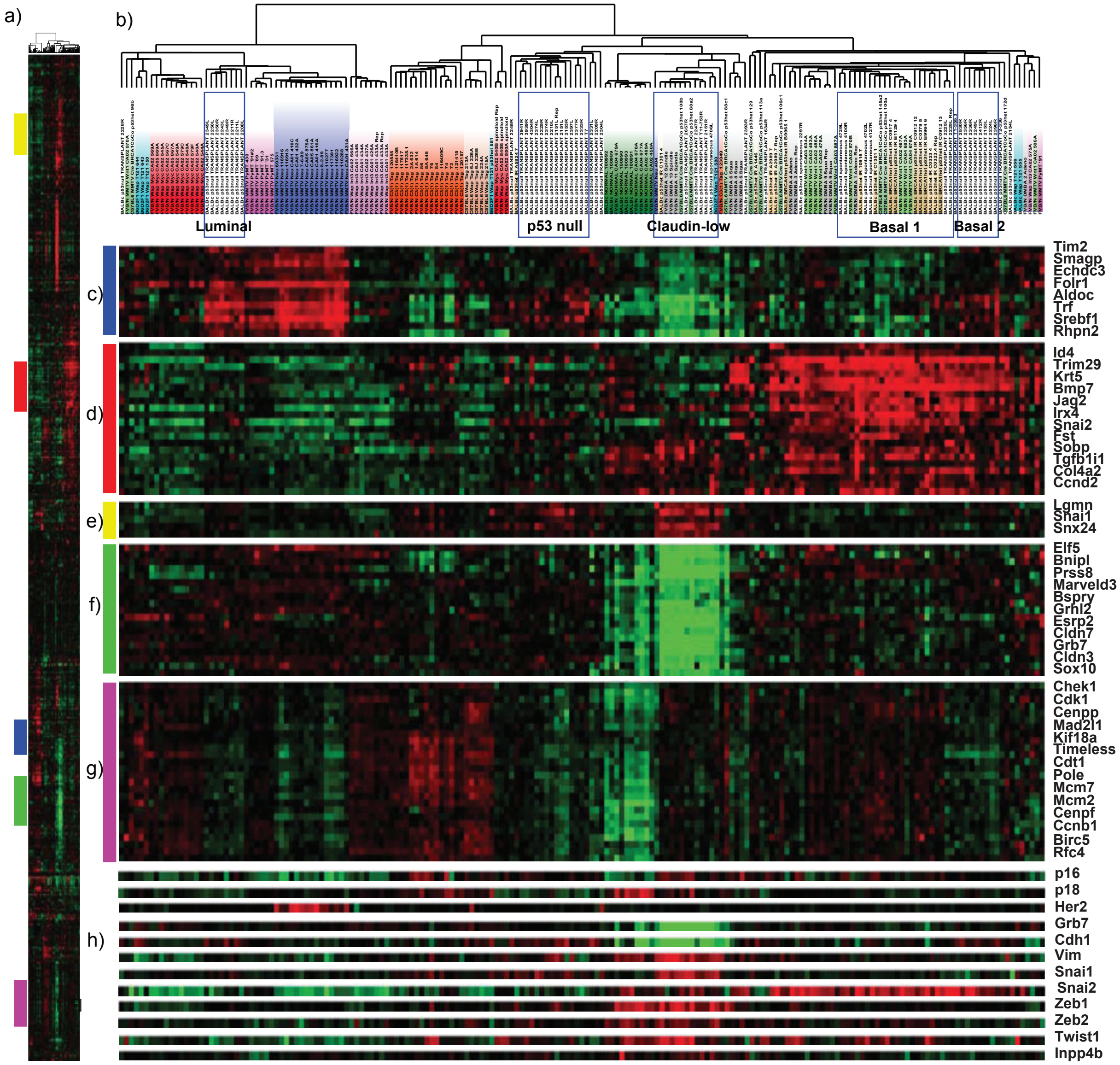


Figure 2.2 Intrinsic gene set clustering analysis of 50 p53 null tumors and 117 samples from 13 GEMM previously published in Herschkowitz et al. 2007. (a) Overview of the complete cluster diagram. (b) Experimental sample associated dendrogram. Boxes indicate the p53 null tumor subtypes based on SigClust analysis. (c) Luminal epithelial gene expression pattern that is highly expressed in luminal p53 null tumors, MMTV-Neu, and MMTV-PyMT tumors (d). Basal epithelial gene expression pattern including Keratin 5 and ID4, which are highly expressed in basal-like p53 null tumors (e) mesenchymal genes including snail homolog 1. (f) Genes expressed at low levels in claudin-low tumors including CLDN3, CLDN7, and ELF5. G) Proliferation signature genes, and H) Individual genes discussed within the text.

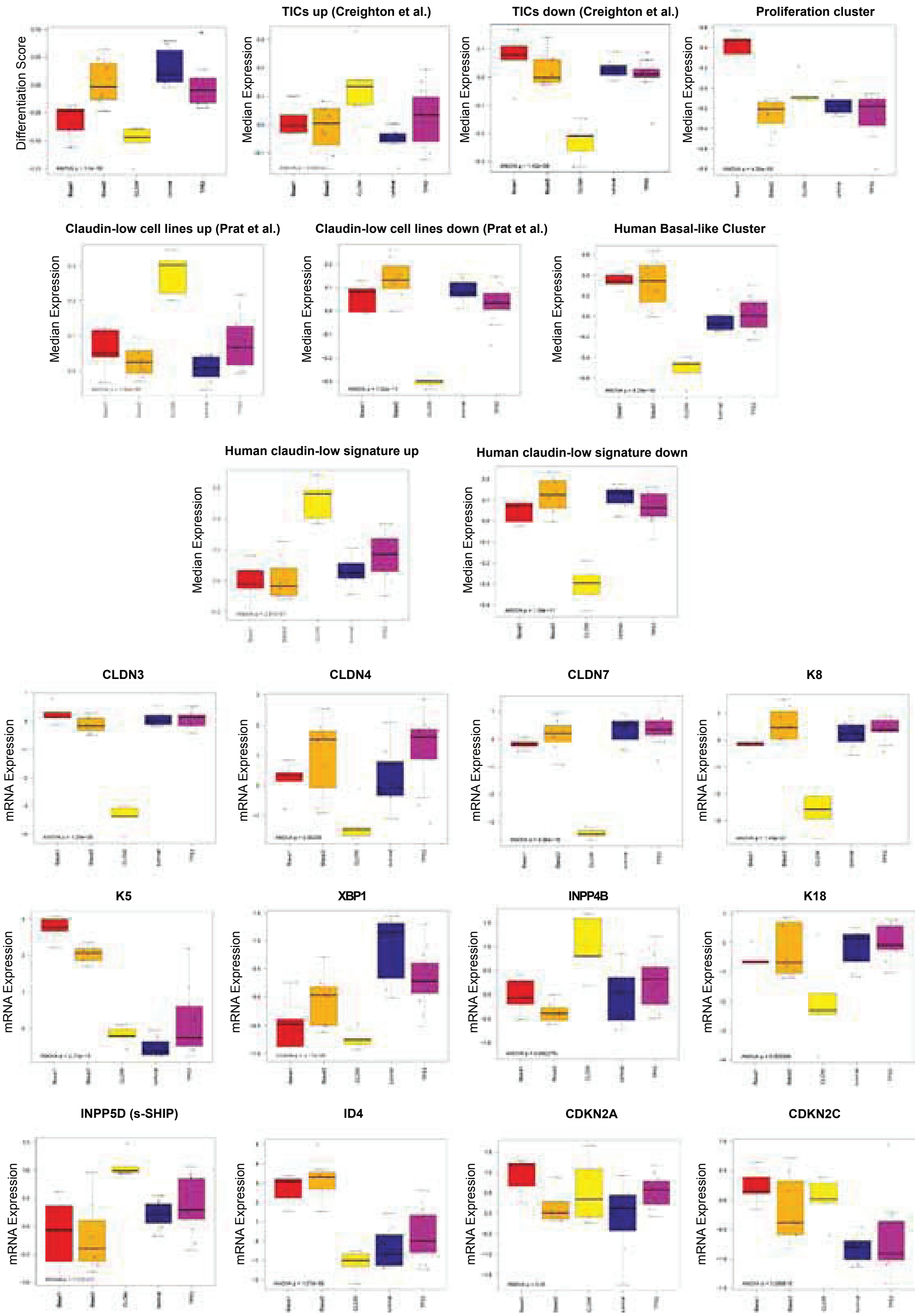


Figure 2.3 ANOVA of gene signatures and individual genes across five subtypes of p53 null tumors.

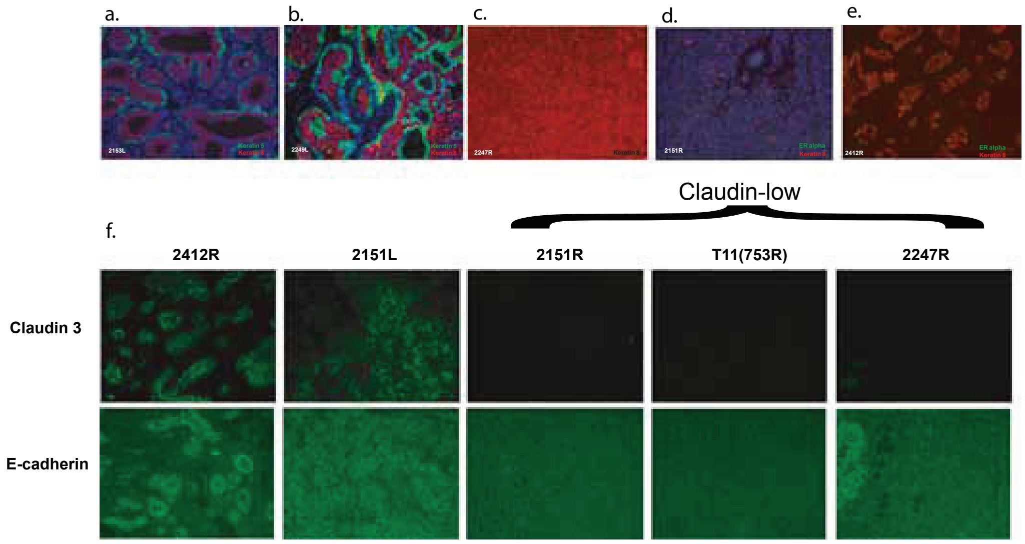


Figure 2.4 Immunofluorescent staining. p53 null basal tumors often show staining for both keratin 5 (K5) and keratin 8 (K8) (a and b). p53 null claudin-low tumors stain for K8 (c), some with less intensity (d). Some tumors are estrogen receptor positive (e). Claudin-low tumors show little staining for CLDN3 and CDH1 (f)

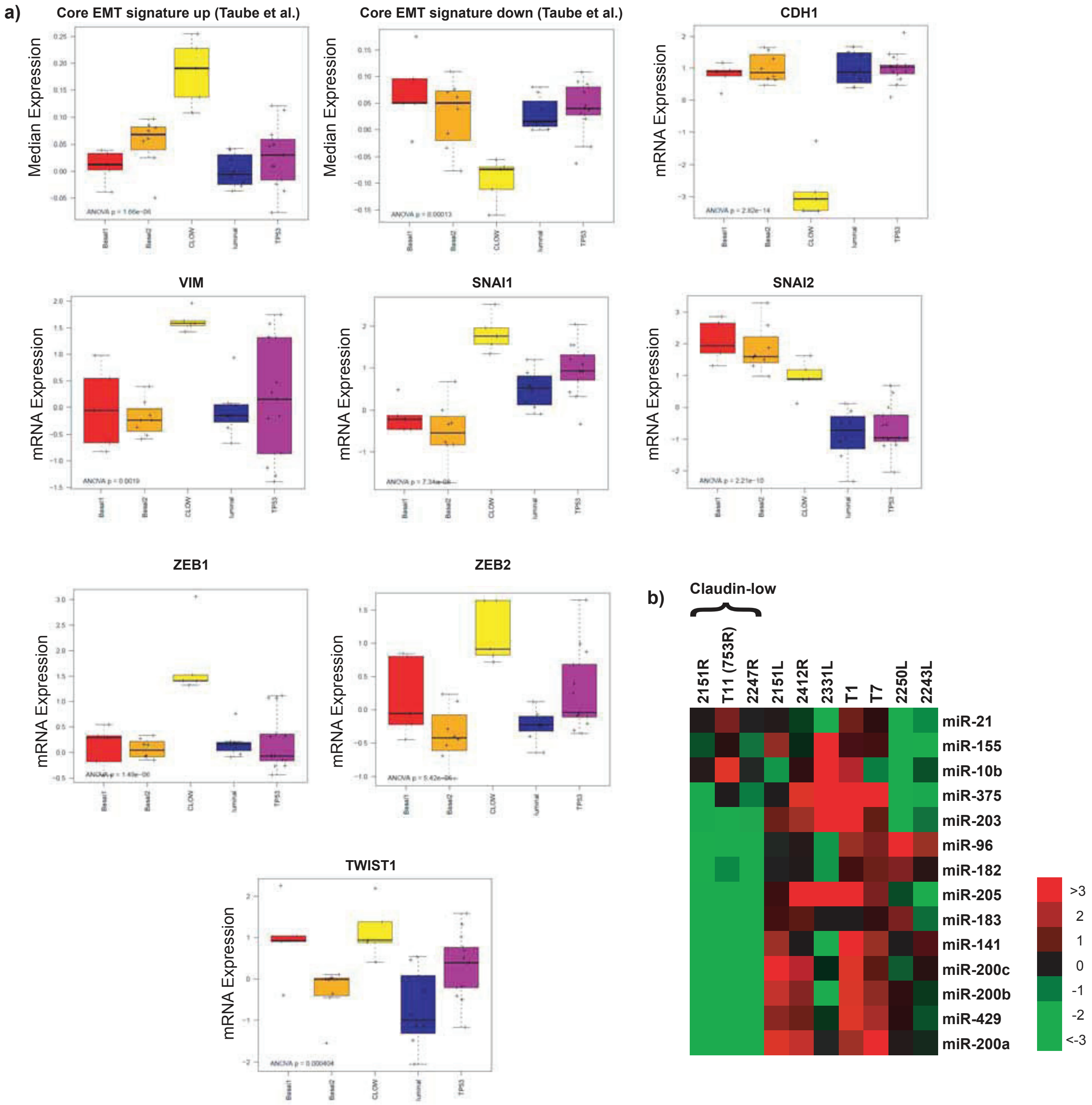


Figure 2.5 p53 null claudin-low tumors have features of EMT. These tumors express (a) core EMT signature, (b) EMT markers, and showed marked downregulation of miRNAs involved in negative regulation of stemness and EMT. Three technical replicates were averaged for each. 200a (P<0.0001), 200b (P=0.0002), 200c (P<0.0001), 141 (P=0.0005), 429 (P<0.0001), 205 (P=0.004), 182 (P=0.004), 96 (P=0.005), 183 (P<0.0001), 203 (P=0.04).

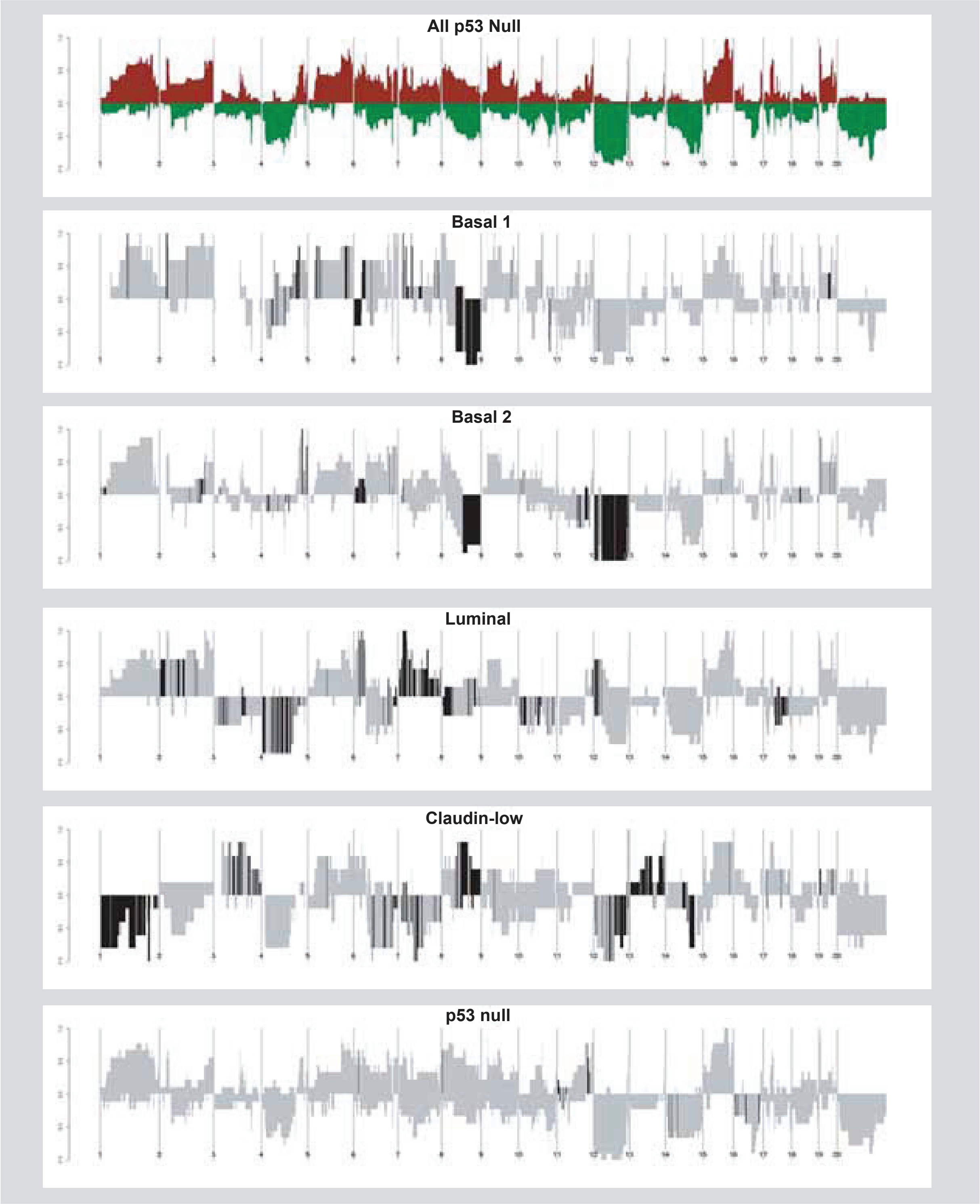


Figure 2.6 Tumor genomic DNA copy number landscape plots for mouse p53 null tumor classes. The top shows the overall pattern for all 34 tumors considered together, and then below are the landscape plots for each of the 5 expression defined subtypes. Gray shading indicate the overall frequency of aberrations seen in that subtype, and the black shading indicate the group specific CNA (p-value threshold 0.05).

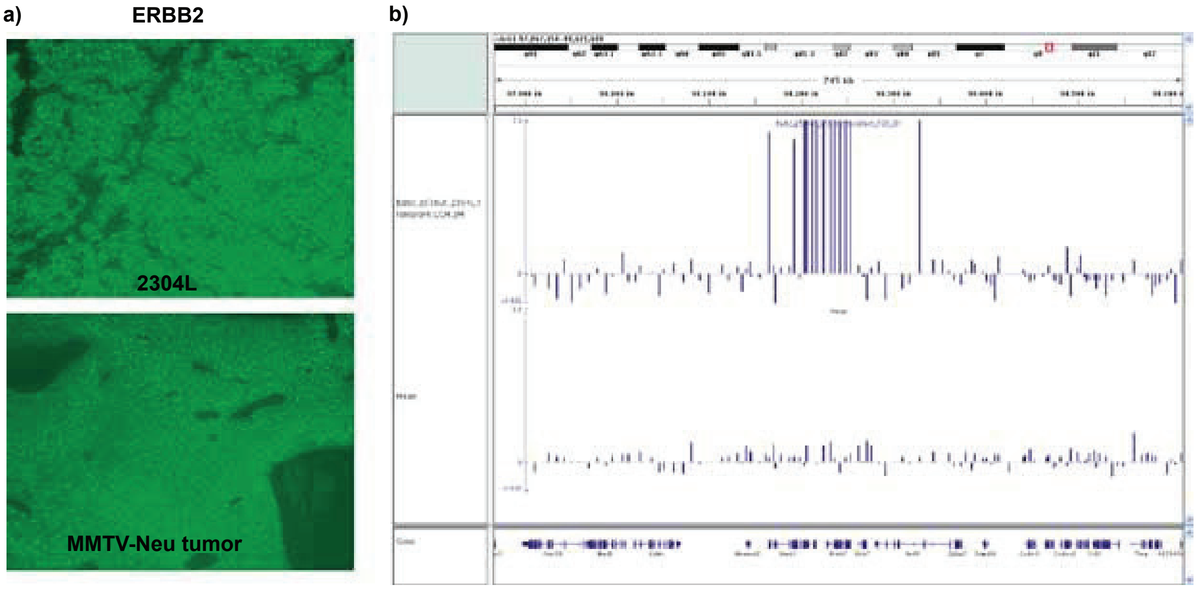


Figure 2.7 p53 null tumor 2304L has (a) overexpression and (b) amplification of HER2/ERBB2.

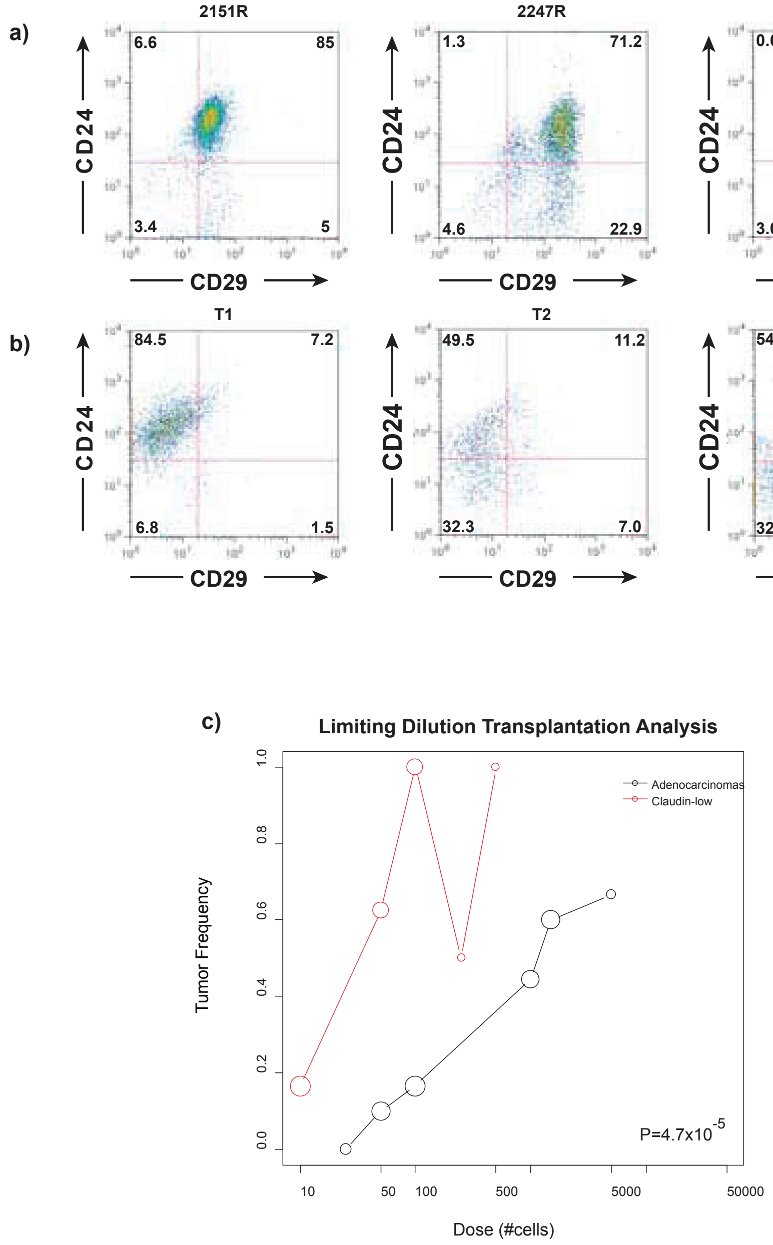


Figure 2.8 p53 null claudin-low tumors express markers of stem cells and are enriched for tumor initiating ability. a) Claudin-low tumors top have high percentages of double positive (CD29+/CD24+) cells compared to other p53 null tumors. b) Limiting dilution transplantation of claudin-low versus adenocarcinoma cells. Sample sizes are implied by the sizes of the circles (area is proportional to sample size).

REFERENCES

1. Jemal A, Siegel R, Xu J, Ward E: **Cancer statistics, 2010.** *CA Cancer J Clin* , **60**:277–300.

2. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lønning P, Børresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci U S A* 2001, **98**:10869–10874.

3. Donehower LA, Harvey M, Slagle BL, McArthur MJ, Montgomery CA, Butel JS, Bradley A: **Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours.** *Nature* 1992, **356**:215–221.

4. Jacks T, Remington L, Williams BO, Schmitt EM, Halachmi S, Bronson RT, Weinberg RA: **Tumor spectrum analysis in p53-mutant mice**. *Curr Biol* 1994, **4**:1–7.

5. Kuperwasser C, Hurlbut GD, Kittrell FS, Dickinson ES, Laucirica R, Medina D, Naber SP, Jerry DJ: **Development of spontaneous mammary tumors in BALB/c p53 heterozygous mice. A model for Li-Fraumeni syndrome.** *Am J Pathol* 2000, **157**:2151–2159.

6. Jerry DJ, Kittrell FS, Kuperwasser C, Laucirica R, Dickinson ES, Bonilla PJ, Butel JS, Medina D: **A mammary-specific model demonstrates the role of the p53 tumor suppressor gene in tumor development.** *Oncogene* 2000, **19**:1052–1058.

7. Aldaz CM, Hu Y, Daniel R, Gaddis S, Kittrell F, Medina D: **Serial analysis of gene expression in normal p53 null mammary epithelium.** *Oncogene* 2002, **21**:6366–6376.

8. Medina D, Kittrell FS, Hill J, Shepard A, Thordarson G, Brown P: **Tamoxifen inhibition of estrogen receptor-alpha-negative mouse mammary tumorigenesis.** *Cancer Res* 2005, **65**:3493–3496.

9. Cardiff RD: **The pathology of EMT in mouse mammary tumorigenesis.** *J Mammary Gland Biol Neoplasia* 2010, **15**:225–233.

10. Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, Yang J, Hartwell K, Onder TT, Gupta PB, Evans KW, Hollier BG, Ram PT, Lander ES, Rosen JM, Weinberg RA, Mani SA: **Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes.** *Proc Natl Acad Sci U S A* 2010, **107**:15449–15454.

11. Zhang M, Behbod F, Atkinson RL, Landis MD, Kittrell F, Edwards D, Medina D, Tsimelzon A, Hilsenbeck S, Green JE, Michalowska AM, Rosen JM: **Identification of tumor-initiating cells in a p53-null mouse model of breast cancer.** *Cancer Res* 2008, **68**:4674–4682.

12. Medina D: **The mammary gland: a unique organ for the study of development and tumorigenesis.** *J Mammary Gland Biol Neoplasia* 1996, **1**:5–19.

13. He X-R, Zhang C, Patterson C: **Universal mouse reference RNA derived from neonatal mice.** *Biotechniques* 2004, **37**:464–468.

14. Herschkowitz JI, He X, Fan C, Perou CM: **The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas.** *Breast Cancer Res* 2008, **10**:R75.

15. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95**:14863–14868.

16. Saldanha AJ: **Java Treeview--extensible visualization of microarray data.** *Bioinformatics* 2004, **20**:3246–3248.

17. Liu Y, Hayes DN, Nobel A, Marron JS: **Statistical Significance of Clustering for High-Dimension, Low–Sample Size Data**. *J Am Stat Assoc* 2008, **103**:1281–1293.

18. Fan C, Prat A, Parker JS, Liu Y, Carey LA, Troester MA, Perou CM: **Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures.** *BMC Med Genomics* 2011, **4**:3.

19. Weigman VJ, Chao H-H, Shabalin AA, He X, Parker JS, Nordgard SH, Grushko T, Huo D, Nwachukwu C, Nobel A, Kristensen VN, B ørresen-Dale A-L, Olopade OI, Perou CM: **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**:865–880.

20. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ: **The UCSC Genome Browser database: update 2011.** *Nucleic Acids Res* 2011, **39**:D876–D882.

21. Bonnefoix T, Bonnefoix P, Verdiel P, Sotto JJ: **Fitting limiting dilution experiments with generalized linear models results in a test of the single-hit Poisson assumption**. *J Immunol Methods* 1996, **194**:113–119.

22. Hu Y, Smyth GK: **ELDA: Extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays**. *J Immunol Methods* 2009, **347**:70–78.

23. Faraway JJ: *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. *Volume 66*; 2006:301.

24. Li Y, Welm B, Podsypanina K, Huang S, Chamorro M, Zhang X, Rowlands T, Egeblad M, Cowin P, Werb Z, Tan LK, Rosen JM, Varmus HE: **Evidence that transgenes encoding components of the Wnt signaling pathway preferentially induce mammary cancers from progenitor cells.** *Proc Natl Acad Sci U S A* 2003, **100**:15853–15858.

25. Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z, Rasmussen KE, Jones LP, Assefnia S, Chandrasekharan S, Backlund MG, Yin Y, Khramtsov AI, Bastein R, Quackenbush J,

Glazer RI, Brown PH, Green JE, Kopelovich L, Furth PA, Palazzo JP, Olopade OI, Bernard PS, Churchill GA, Van Dyke T, Perou CM: **Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors.** *Genome Biol* 2007, **8**:R76.

26. Pei XH, Bai F, Smith MD, Usary J, Fan C, Pai SY, Ho IC, Perou CM, Xiong Y: **CDK Inhibitor p18INK4c Is a Downstream Target of GATA3 and Restrains Mammary Luminal Progenitor Cell Proliferation and Tumorigenesis**. *Cancer Cell* 2009, **15**:389–401.

27. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, He X, Perou CM: **Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer.** *Breast Cancer Res* 2010, **12**:R68.

28. Gregory PA, Bert AG, Paterson EL, Barry SC, Tsykin A, Farshid G, Vadas MA, Khew-Goodall Y, Goodall GJ: **The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1.** *Nat Cell Biol* 2008, **10**:593–601.

29. Polyak K, Weinberg RA: **Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits.** *Nat Rev Cancer* 2009, **9**:265–273.

30. Shimono Y, Zabala M, Cho RW, Lobo N, Dalerba P, Qian D, Diehn M, Liu H, Panula SP, Chiao E, Dirbas FM, Somlo G, Pera RAR, Lao K, Clarke MF: **Downregulation of miRNA-200c Links Breast Cancer Stem Cells with Normal Stem Cells**. *Cell* 2009, **138**:592–603.

31. Wellner U, Schubert J, Burk UC, Schmalhofer O, Zhu F, Sonntag A, Waldvogel B, Vannier C, Darling D, zur Hausen A, Brunton VG, Morton J, Sansom O, Schüler J, Stemmler MP, Herzberger C, Hopt U, Keck T, Brabletz S, Brabletz T: **The EMT-activator ZEB1 promotes tumorigenicity by repressing stemness-inhibiting microRNAs.** *Nat Cell Biol* 2009, **11**:1487–1495.

32. Blenkiron C, Goldstein LD, Thorne NP, Spiteri I, Chin S-F, Dunning MJ, Barbosa-Morais NL, Teschendorff AE, Green AR, Ellis IO, Tavaré S, Caldas C, Miska EA: **MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype.** *Genome Biol* 2007, **8**:R214.

33. Greene SB, Herschkowitz JI, Rosen JM: **Small players with big roles: microRNAs as targets to inhibit breast cancer progression.** *Curr Drug Targets* 2010, **11**:1059–1073.

34. Bergamaschi A, Kim YH, Wang P, Sørlie T, Hernandez-Boussard T, Lonning PE, Tibshirani R, Børresen-Dale A-L, Pollack JR: **Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer.** *Genes Chromosomes Cancer* 2006, **45**:1033–1040.

35. Donehower LA, Godley LA, Aldaz CM, Pyle R, Shi YP, Pinkel D, Gray J, Bradley A, Medina D, Varmus HE: **Deficiency of p53 accelerates mammary tumorigenesis in Wnt-1 transgenic mice and promotes chromosomal instability.** *Genes Dev* 1995, **9**:882–895.

36. Hodgson JG, Malek T, Bornstein S, Hariono S, Ginzinger DG, Muller WJ, Gray JW: **Copy number aberrations in mouse breast tumors reveal loci and genes important in tumorigenic receptor tyrosine kinase signaling.** *Cancer Res* 2005, **65**:9695–9704.

37. Fedele CG, Ooms LM, Ho M, Vieusseux J, O'Toole SA, Millar EK, Lopez-Knowles E, Sriratana A, Gurung R, Baglietto L, Giles GG, Bailey CG, Rasko JEJ, Shields BJ, Price JT, Majerus PW, Sutherland RL, Tiganis T, McLean CA, Mitchell CA: **Inositol polyphosphate 4-phosphatase II regulates PI3K/Akt signaling and is lost in human basal-like breast cancers.** *Proc Natl Acad Sci U S A* 2010, **107**:22231–22236.

38. Gewinner C, Wang ZC, Richardson A, Teruya-Feldstein J, Etemadmoghadam D, Bowtell D, Barretina J, Lin WM, Rameh L, Salmena L, Pandolfi PP, Cantley LC: **Evidence that Inositol Polyphosphate 4-Phosphatase Type II Is a Tumor Suppressor that Inhibits PI3K Signaling**. *Cancer Cell* 2009, **16**:115–125.

39. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lønning PE, Brown PO, Børresen-Dale A-L, Botstein D: **Repeated observation of breast tumor subtypes in independent gene expression data sets.** *Proc Natl Acad Sci U S A* 2003, **100**:8418–8423.

40. Shackleton M, Vaillant F, Simpson KJ, Stingl J, Smyth GK, Asselin-Labat M-L, Wu L, Lindeman GJ, Visvader JE: **Generation of a functional mammary gland from a single stem cell.** *Nature* 2006, **439**:84–88.

41. Zhang M, Atkinson RL, Rosen JM: **Selective targeting of radiation-resistant tumor-initiating cells.** *Proc Natl Acad Sci U S A* 2010, **107**:3522–3527.

42. Visvader JE: **Cells of origin in cancer.** *Nature* 2011, **469**:314–322.

43. Du Z, Podsypanina K, Huang S, McGrath A, Toneff MJ, Bogoslovskaia E, Zhang X, Moraes RC, Fluck M, Allred DC, Lewis MT, Varmus HE, Li Y: **Introduction of oncogenes into mammary glands in vivo with an avian retroviral vector initiates and promotes carcinogenesis in mouse models.** *Proc Natl Acad Sci U S A* 2006, **103**:17396–17401.

44. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, Asselin-Labat M-L, Gyorki DE, Ward T, Partanen A, Feleppa F, Huschtscha LI, Thorne HJ, Fox SB, Yan M, French JD, Brown MA, Smyth GK, Visvader JE, Lindeman GJ: **Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers.** *Nat Med* 2009, **15**:907–913.

45. Molyneux G, Geyer FC, Magnay FA, McCarthy A, Kendrick H, Natrajan R, MacKay A, Grigoriadis A, Tutt A, Ashworth A, Reis-Filho JS, Smalley MJ: **BRCA1 basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells**. *Cell Stem Cell* 2010, **7**:403–417.

46. Barker N, Ridgway RA, van Es JH, van de Wetering M, Begthel H, van den Born M, Danenberg E, Clarke AR, Sansom OJ, Clevers H: **Crypt stem cells as the cells-of-origin of intestinal cancer.** *Nature* 2009, **457**:608–611.

47. Montagna C, Andrechek ER, Padilla-Nash H, Muller WJ, Ried T: **Centrosome abnormalities, recurring deletions of chromosome 4, and genomic amplification of HER2/neu define mouse mammary gland adenocarcinomas induced by mutant HER2/neu.** *Oncogene* 2002, **21**:890–898.

48. Radany EH, Hong K, Kesharvarzi S, Lander ES, Bishop JM: **Mouse mammary tumor virus/v-Ha-ras transgene-induced mammary tumors exhibit strain-specific allelic loss on mouse chromosome 4.** *Proc Natl Acad Sci U S A* 1997, **94**:8664–8669.

49. Weaver ZA, McCormack SJ, Liyanage M, du Manoir S, Coleman A, Schröck E, Dickson RB, Ried T: **A recurring pattern of chromosomal aberrations in mammary gland tumors of MMTV-cmyc transgenic mice.** *Genes Chromosomes Cancer* 1999, **25**:251–260.

50. Andrechek ER, Cardiff RD, Chang JT, Gatza ML, Acharya CR, Potti A, Nevins JR: **Genetic heterogeneity of Myc-induced mammary tumors reflecting diverse phenotypes including metastatic potential.** *Proc Natl Acad Sci U S A* 2009, **106**:16387–16392.

51. Ponzo MG, Lesurf R, Petkiewicz S, O’Malley FP, Pinnaduwage D, Andrulis IL, Bull SB, Chughtai N, Zuo D, Souleimanova M, Germain D, Omeroglu A, Cardiff RD, Hallett M, Park M: **Met induces mammary tumors with diverse histologies and is associated with poor outcome and human basal breast cancer.** *Proc Natl Acad Sci U S A* 2009, **106**:12903–12908.

52. Bai L, Rohrschneider LR: **s-SHIP promoter expression marks activated stem cells in developing mouse mammary tissue.** *Genes Dev* 2010, **24**:1882–1892.

53. Reya T, Morrison SJ, Clarke MF, Weissman IL: **Stem cells, cancer, and cancer stem cells.** *Nature* 2001, **414**:105–111.

CHAPTER III

PREDICTING DRUG RESPONSIVENESS IN HUMAN CANCERS USING GENETICALLY ENGINEERED MICE[2]

Purpose: To use genetically engineered mouse models (GEMMs) and orthotopic syngeneic murine transplants (OSTs) to develop gene-expression based predictors of response to anti-cancer drugs in human tumors. These mouse models offer advantages including precise genetics and an intact microenvironment/immune system.

Experimental Design: We examined the efficacy of four chemotherapeutic or targeted anti-cancer drugs, alone and in combination, using mouse models representing three distinct breast cancer subtypes: Basal-like (*C3(1)-T-antigen* GEMM), Luminal B (*MMTV-Neu* GEMM), and Claudin-low (*T11/TP53*$^{-/-}$ OST). We expression-profiled tumors to develop signatures that corresponded to treatment and response, then tested their predictive potential using human patient data.

Results: Although a single agent exhibited exceptional efficacy (i.e. lapatinib in the *Neu*-driven model), generally single-agent activity was modest, while some combination therapies were more active and life-prolonging. Through analysis of RNA expression in this large set of chemotherapy-treated murine tumors, we identified a pair of gene expression signatures that predicted pathological complete response to neoadjuvant anthracycline/taxane therapy in human patients with breast cancer.

[2] Usary J, Zhao W, Darr D, Roberts PJ, Liu M, Balletta L, Karginova O, Jordan J, Combest A, Bridges A, Prat A, Cheang MCU, Herschkowitz JI, Rosen JM, Zamboni W, Sharpless NE, Perou CM: **Predicting drug responsiveness in human cancers using genetically engineered mice.** *Clin Cancer Res* 2013, **19**:4889–99.

Conclusions: These results show that murine-derived gene signatures can predict response even after accounting for common clinical variables and other predictive genomic signatures, suggesting that mice can be used to identify new biomarkers for human cancer patients.

INTRODUCTION

Gene expression profiling has identified five molecular subtypes of breast cancer (Luminal A, Luminal B, Basal-like, HER2-Enriched, Claudin-low) and a normal-like group, which show significant differences in epidemiologic associations and clinical features including survival [1–3]. Mounting evidence suggests that these subtypes vary in their responsiveness to chemotherapeutics [2, 4–6] and to biologically targeted agents [7–9]. Methods for selecting the optimal chemotherapeutic agent for each breast tumor subtype have yet to be determined. Instead, chemotherapy choices for breast cancer patients have been mainly empiric and based upon large clinical trials using unselected patient populations, and population-based benefits. The Basal-like subtype of breast tumor, of which the majority are also "triple-negative" breast cancers, is particularly challenging due to its lack of validated biological targets (i.e. ER-, PR- and HER2 normal) [10, 11]. Other breast cancer subtypes with poor prognosis also exist including the Luminal B subtype [2, 5] and the recently discovered Claudin-low subtype, which exhibits high numbers of tumor initiating cells [12].

Genetically Engineered Mouse Models (GEMMs) have proven valuable for validating the causal role of oncogenes and tumor suppressor genes in cancer [13], but their use in efficacy testing is less mature, with most studies being low-throughput efforts examining model-specific compounds in small numbers of tumor-bearing mice (<50) [14]. Recently, academic and industry researchers have begun simultaneous efficacy testing at medium throughput, employing larger numbers of compounds (5-50) in larger numbers of GEMMs (100-1000) [15, 16]. In particular, these efforts have attempted to mirror and inform ongoing human clinical trials, by testing novel therapeutics in faithful murine models as "co-clinical trials" [17]. While this approach has been promising, we believe an additional untapped power of medium-throughput GEMM testing is the ability to use murine models to identify biomarkers of response for human cancer patients.

Previously, we performed RNA expression profiling of 13 distinct GEMMs of breast cancer [12, 18] and compared these signatures to human expression subtypes using an across-species expression analysis. These analyses identified murine models that faithfully represent multiple human breast tumor subtypes including Basal-like tumors (*C3(1)-T-antigen*) [19] and Luminal B tumors (*MMTV-Ne*u) [20]. No single Claudin-low GEMM was identified, but an orthotopic, transplantable syngeneic tumor from a *BALB/c TP53*$^{-/-}$ mouse was found to exhibit a stable Claudin-low expression phenotype [12]. In this work, we used these credentialed murine tumor models and determined their sensitivities to a variety of chemotherapeutic and biologically targeted agents in routine clinical use. This analysis identified a heterogeneity of responses to certain cytotoxics in the Basal-like model. We exploited this existence of sensitive and resistant tumors from GEMMs to develop genomic signatures of chemotherapy response, which we tested in a large, clinically annotated human cohort of breast cancer patients.

METHODS

Genetically Engineered Mouse Models. All work was done under protocols approved by the UNC Institutional Animal Care and Use Committee (IACUC). GEMMs of strain *FVB/n* carrying a transgene for *Tg(MMTVneu)202Mul/J* (*MMTV-Neu*) [20] and *C3(1)SV40 T-antigen (C3(1)-T-antigen* or *C3-TAg*) [21] were bred in-house and observed until the onset of a mammary tumor ~0.5 cm in any dimension. Tumors derived from *BALB/c TP53*$^{-/-}$ orthotropic mammary gland transplant line (T11) were passaged in *BALB/c* wild-type mice by subcutaneous injection of one half million cells resuspended in matrigel into the flank as previously described [22]. Mice were randomized into treatment groups and monitored with tumor growth measurements. Tumor volumes were measured by caliper in two dimensions and/or by ultrasound (Vevo 770 ultrasound imaging system (Visualsonics Inc.)). Chemotherapy was started at time zero and repeated weekly for a total of three injections over a twenty-one day period. The mice were further assessed for long-term survival as follows: if after a one week break from treatment a tumor increased in volume more than 1mm in any dimension, then an additional three cycles of therapy were initiated. This continued until either the mouse developed a tumor burden sufficient to warrant euthanasia (2 cm in any dimension or 3 tumors present) or until weight loss totaling 20% of the initial starting body mass was observed or because of any other severe health problems. Orally administered biological inhibitors were given continuously with no dose interruption. In the case of a chemotherapeutic plus an oral inhibitor, the chemotherapy agent was dosed once weekly for 21 days and stopped until progression, while the small molecule inhibitors were dosed continuously.

Compounds. Compounds were obtained from commercial sources: Carboplatin (Hospira, Inc), cyclophosphamide (Hospira, Inc), doxorubicin (Bedford Laboratories), paclitaxel (Ivax Pharmaceuticals, Inc), erlotinib (Genentech, Inc) and lapatinib (GlaxoSmithKline). Oral biological inhibitors (erlotinib and lapatinib) were milled into chow by Research Diets, Inc. while carboplatin and paclitaxel were delivered via intraperitoneal injection.

Treatments. The drug-specific approach to determine schedule and dose is described in Table 3.1. A minimum tumor volume of ~0.5cm in size was required for randomization into a treatment group (including a control group). Combination treatments were given at the same doses as the individual treatments. Chemotherapy was started at time zero and repeated weekly for over a 14-day (*T11/TP53-/-*) or 21-day (*C3(1)-T-antigen* and *MMTV-Neu*) period.

Pharmacokinetic (PK) Studies. PK studies were performed after administration of paclitaxel (Figure 3.1), erlotinib, and lapatinib (data not shown). For paclitaxel, seventeen transgenic *FVB/n* mice bearing the *MMTV-Neu* transgene were administered a single intraperitoneal dose of paclitaxel at 10 mg/kg. Plasma and tumor samples (3 mice used at each time point; 2 mice used for the 48 hour time point) were collected at 0.083, 1, 4, 8, 24, and 48 hours after administration and flash frozen in liquid nitrogen. The samples were analyzed via liquid chromatography/tandem mass spectrometry (LC-MS/MS) as described previously [23]. The concentration versus time profiles of paclitaxel in plasma and tumor are presented in Figure 3.1. The mean ± SD of paclitaxel C_{max} and $AUC_{0-\infty}$ in plasma following IP administration were 2.1 µg/mL ± 1.5 and 6.3 µg/mL•h respectively.

The mean ± SD of paclitaxel C_{max} and $AUC_{0-\infty}$ in tumor following IP administration were 3.7 µg/g ± 2.1 and 42.4 µg/g•h respectively.

Response Criteria. Tumor volume was calculated from two-dimensional measurements as (*Volume = [(width)^2 x length]/2)*. The percent change in volume at 21 days was used to quantify response, except in the case of the *T11/TP53*$^{-/-}$ model where its faster growth rate required a 14-day treatment response assessment. Twenty-one day response was chosen as our primary response endpoint based on the fact that most of the untreated animals do not survive much longer than 21 days when starting with a tumor of >0.5 cm. Survival was measured from the first day of drug treatment.

Microarray Analysis. DNA microarray analyses of murine tumors was performed as described in Herschkowitz et al. [12]. We used using Agilent 4x44,000 feature mouse DNA microarrays and a common reference strategy. For hierarchical clustering analyses, the genes/rows were median centered and clustering of arrays was performed using Cluster v3.0 [24] with correlation centered genes and arrays, and centroid-linkage. Array cluster viewing and display was performed using JavaTreeview v1.1.4 [25].

Statistical Analyses

(A) Identification of significant differential genes in response to treatments. We performed two unpaired two-class SAM [26] analyses to identify genes that showed differential expressions as following: (i) between carboplatin/paclitaxel treated *C3(1)-T-antigen* tumors that responded versus those that did not and (ii) between carboplatin/paclitaxel treated *C3(1)-T-antigen* tumors versus those untreated. The primary SAM analysis to identify tumor response related genes included three responding tumors (shrinkage >20%) versus nine non-responding tumors (growth >20%). The secondary SAM analysis to identify treatment up-regulated or down-regulated genes included seven untreated tumors versus the twelve treated tumors. Two gene lists were obtained with a FDR of 1%: 348 genes (428 probes) showing significantly high expression in the untreated samples (called UNTREATED) and 61 genes (74 probes) showing significantly high expression in the samples from responders (called RESP-HIGH); the identified genes are listed in Appendix 1. Using the Mouse Genome Database [27], these lists were converted to orthologous human genes. In order to refine the list of these candidate genes relevant to human tumors, a hierarchical clustering analysis of these orthologous human gene lists was performed using the 337 tumor samples from Prat et al. [1]. From these clusters, we chose a dendrogram node based on the criteria that it would include the largest number of highly expressed genes and have a node correlation of >0.4. Figure 3.2b illustrates the gene set called UNTREATED-HUM that includes 30 unique genes. Figure 3.2d illustrates the gene set called RESP-HUM that includes 12 unique genes.

In the UNC337 human tumors sets, these two gene lists showed “homogeneous” expression patterns, and thus we decided that taking the mean of the genes within each list/dendrogram node was the most appropriate method to assign the signature score for each tumor sample. In brief, an UNTREATED-HUM score was assigned to each test sample by taking the mean of the 26 genes in the list. A RESP-HUM score was assigned to each test sample by taking the mean of the 12 genes in the list. Since we also aimed to compare the performance of these two signatures as well as including published genomic signatures, we standardized the signature scores with a standard deviation equivalent to 1 to bring all the signature scores to the same scale. We applied this same methodology to two independent data sets of neoadjuvant human tumors described below.

(**B) Association of the identified signatures with tumor response for neoadjuvant anthracycline/taxane containing chemotherapy regimens**. The performance of UNTREATED-HUM and RESP-HUM signatures to predict pathological complete response (pCR) was first tested on 462 patients with HER2 normal tumors in MDACC data set (Hatzis et al. [28], GEO # GSE25066) and validated on 81 patients with HER2 normal tumors in JSE data set (Miyake et al. [29] GEO # GSE32646). Patients on both data sets were treated with neoadjuvant anthracycline-taxane containing regimens. Univariable logistic regression analysis was used to assess the odds ratio and significance of the two signatures to predict pCR. Multivariable logistic regression analysis was used to determine the adjusted odds ratio and significance taking into account for the standard clinical variables measured at baseline and other published genomic signatures as appropriate. The area under the curve (AUC) value was calculated from the Receiver Operating Characteristics analysis of the univariable and multivariable logistic model respectively. The published genomic signatures included the PAM50 intrinsic subtypes [2], Claudin-low predictor[1], and 11-gene proliferation signature [9]; we also included signatures developed by Hatzis and colleagues (including Hatzis Sensitivity to endocrine therapy (SET) index, Hatzis signature chemo sensitive RCB-I predict, and Hatzis signature chemo resistance (RCB-III predict)) that were available for the data set [28]. Finally, survival outcome data after neoadjuvant treatment was available for the Hatzis et al. data set and Kaplan Meier analysis and log-rank test were used to determine the differential survival estimates of the two signatures to distant relapse free survival.

RESULTS

Sensitivity of GEMMs to chemotherapeutic agents

Our ultimate goal was to use GEMMs to develop predictors of therapeutic response for humans. Details of the work flow are outlined in the study design Figure 3.3. As a first step, we tested three different mammary cancer GEMMs with multiple therapeutics to find a GEMM, and a drug regimen, which gave a range of responses; from this GEMM, we then profiled sensitive and resistant tumors in order to identify a signature associated with response**.** We first therefore, determined the sensitivity of three distinct GEMMs/OSTs models of human breast cancer subtypes versus two cytotoxic chemotherapeutics and two small molecule kinase inhibitors. The models used were *C3(1)-T-antigen, MMTV-Neu,* and *T11/TP53*$^{-/-}$, with these models chosen based on their similarity in gene expression to Basal-like, Luminal B and Claudin-low human tumor subtypes respectively [12, 18]. Tumor volume changes at 21 days (or 14 days in the *T11/TP53*$^{-/-}$ model), and long-term survival were the primary endpoints. Response at 21 days (or 14 days for *T11/TP53*$^{-/-}$) was measured for 304 treated and control mice (150 *C3(1)-T-antigen*, 97 *MMTV-Neu*, 57 *T11/TP53*$^{-/-}$*)* with the percent volume change of each model's non-treated controls (i.e. growth rate) shown in Figure 3.4 (bottom rows). Although there was overlap in the average growth rates of tumors from each GEMM, the untreated *T11/TP53*$^{-/-}$ tumors grew significantly faster than their *MMTV-Neu* counterparts ($p<0.01$, Student's t-test), with the *C3(1)-T-antigen* model exhibiting an intermediate growth rate (Figure 3.4).

With the growth kinetics of these models established, we next tested two chemotherapeutics that are widely used to treat many solid epithelial human cancers, namely paclitaxel and carboplatin. Although the standard of care for most breast cancer patients is doxorubicin/cyclophosphamide with or without a taxane (i.e. AC-T) [30], platinum agents (carboplatin/cisplatin) are also gaining in use [31], and thus are relevant to breast cancers, especially triple-negative breast cancers (TNBC). As a single agent, carboplatin elicited a modest but significant responses in all three models, while paclitaxel alone elicited no response; however, systemic and tumor drug delivery was confirmed for paclitaxel (Figure 3.1).

Next we tested the commonly used chemotherapy doublet of carboplatin/paclitaxel (CT). A varied response profile was seen for the CT combination where the combination demonstrated no activity in the *T11/TP53*$^{-/-}$ model, and only modest activity in the *MMTV-Neu* model. Importantly, in the *C3(1)-T-antigen* model, a clear bimodal response was observed to the CT combination: ~2/3 of the tumors showed little response and ~1/3 showed near complete regression (Figure 3.4a). This finding is in accord with the observation that human Basal-like tumors exhibit a ~30-40% complete pathological response rate (pCR) to taxane containing neoadjuvant regimens, while the other 60-70% show residual disease and a worse overall survival [1, 5, 10].

Sensitivity to targeted agents

Two classes of biologically targeted agents are used in patients with breast cancer: agents blocking estrogen and progesterone receptor (ER/PR) signaling (e.g. tamoxifen or aromatase inhibitors) and drugs targeting HER2 (e.g. trastuzumab and lapatinib). Given that none of our GEMMs were ER+ or PR+ [12], we chose to focus on the HER2/EGFR family of kinases by using the small molecule inhibitor lapatinib (which targets HER2/ERBB2 primarily [27]), and the EGFR inhibitor erlotinib [32]. In the *MMTV-Neu* model, erlotinib and lapatinib were both highly effective, with lapatinib causing near 100% regression in all *MMTV-Neu* tumors. Conversely, neither erlotinib nor lapatinib were effective at reducing the growth rate of the *T11/TP53-/-* tumors. Lapatinib was similarly ineffective in the *C3(1)-T-antigen* tumors, but as was the case for the CT doublet, erlotinib showed potent activity in a subset (~40%) of treated mice. These data show that HER2/EGFR inhibitors exhibit potent activity in the *Neu/ERBB2/HER2*-driven model as expected, and provide further evidence for at least two subtypes of *C3(1)-T-antigen* tumors with regard to therapeutic sensitivity.

We also assessed the effects of anti-cancer therapies on the overall survival of tumor-bearing mice. Baseline survival for the *MMTV-Neu* (29 days) and *C3(1)-T-antigen* models (33 days) was similar in the absence of therapy, while the *T11/TP53-/-* animals showed significantly shorter median survival (15 days) (Figure 3.5). In the *MMTV-Neu* model, single-agent lapatinib (and to some extent erlotinib) greatly extended lifespan from a median of 29 days to 154 days (Figure 3.5b). Conversely, no single or combination regimen was able to extend survival in the *C3(1)-T-antigen* or *T11/TP53-/-* models.

Development of murine chemotherapy response signatures

A heterogeneous response to CT was seen in the *C3(1)-T-antigen* tumors that ranged from progressive disease to complete response (Figure 3.4a). We sought to explore these findings and develop a genomic predictor of this response using this GEMM by performing RNA expression profiling of treated vs. untreated tumors. For these experiments, we treated *C3(1)-T-antigen* tumors with carboplatin/paclitaxel for two or three cycles and measured response (n=12), and then harvested the tumor for molecular analysis. In addition, an independent set of seven untreated tumors was used as the non-treated controls (Table 3.2).

Significance Analysis of Microarray (SAM) [26] was used to derive two sets of differentially expressed genes by (A) comparing those mice that responded to treatment (n=3) versus those that did not (n=9), and by (B) comparing the untreated (n=7) versus treated tumors (n=12) (Table 3.2 and Appendix 1). When testing untreated versus treated tumors at a FDR of 1%, this analysis identified 428 probes corresponding to 348 mouse genes that were more highly expressed in untreated tumors (called UNTREATED gene list, Appendix 1a); a Gene Ontology analysis of the UNTREATED list identified multiple significant terms including "cellular macromolecule metabolic process", "nucleic acid metabolic process", "regulation of macromolecule biosynthetic process", "chromosome organization", "DNA metabolic process" and "cell cycle'. We applied a modules/signatures analysis to the untreated versus treated tumors where we examined if 302 previously defined expression signatures [33] varied with treatment (Appendix 2). This modules/signatures analysis showed that multiple signatures of fibroblasts/extracellular matrix, and signatures of the Claudin-low phenotype [1, 18] were more highly expressed after treatment, with this last result recapitulating findings observed in post-chemotherapy treated human tumors [34]. Multiple signatures decreased after treatment including one of proliferation and one of HER1-RAS-pathway activation. These data show that CT treatment induced expression of genes associated with Claudin-low/mesenchymal phenotype, and reduced cellular proliferation.

When the cohort of treated tumors was subdivided into responders versus non-responders at a FDR of 1%, a list of 74 differentially expressed probes corresponding to 61 mouse genes was obtained (Appendix 1b). These genes were more highly expressed in the mice that responded to treatment and the list was named RESP-HIGH. A gene ontology analysis of the RESP-HIGH list revealed the presence of no significant GO terms after Bonferroni or Benjamini corrections. We also applied the 302 signatures analysis above on the responder versus non-responder sample set, and only a small number of proliferation signatures were more highly expressed in non-responders.

Human testing of the murine chemotherapy response signatures

Next, the murine 348 gene UNTREATED and 74 gene RESP-HIGH lists were converted into human lists using gene orthology, and both lists were then further refined using hierarchical cluster analyses of 337 human breast tumors from Prat et al. [1] (Figure 3.2). This mouse-to-human filtering was necessary because a homogenous gene list from a cell line, or murine experiment, when applied to human primary tumors, will typically fragment into multiple signatures/modules when using *in vivo* human data [35]. We observed this type of gene list heterogeneity here, and thus, from these cluster analyses we chose a single dendrogram node that contained the highest homogenously expressed gene set observed within this human primary tumor data set, and for each gene list separately. This gave a set of 30 genes from the UNTREATED list that we call UNTREATED-HUM, and 12 genes from the RESP-HIGH list that we call RESP-HUM (Figure 3.2b and d); it should be noted that we did not test all possible dendrogram nodes, but instead limited our analyses to a single node from each cluster analysis. These two refined gene lists were also analyzed for GO terms with the UNTREATED-HUM list enriched for the terms ‘cell cycle’, ‘M phase’, ‘nuclear division’ and ‘mitosis’, and we also noted that 12/30 entries were ATP-binding proteins. The RESP-HIGH was not enriched for any GO term.

We next tested both humanized gene lists for their ability to predict distant relapse-free survival (DRFS), and most importantly, pathological complete response (pCR) using a completely independent set of human breast cancer patients treated with neoadjuvant chemotherapy. For both clinical endpoints, we used the Hatzis et al. data set (See Figure 3.3), which is a combined data set of patients who were treated with a taxane and anthracycline-containing neoadjuvant chemotherapy regimen [28]. We first stratified patients into low-medium-high (tertiles) groups based upon their rank-ordered mean expression values for the RESP-HUM and UNTREATED-HUM signature and then tested these stratifications for their ability to predict DRFS. These analyses showed that the RESP-HUM ($p < 0.001$) and UNTREATED-HUM ($p = 0.003$) signatures were able to predict DRFS, as was pCR vs. not, intrinsic subtype, and an 11-gene proliferation signature (Figure 3.6). In multivariable analyses, however, neither of these murine signatures added prognostic information beyond that conveyed by the PAM50 11-gene proliferation signature [9] (data not shown).

We then tested the humanized gene lists for their ability to predict pathological complete response (pCR), which is the most relevant endpoint for these chemotherapy response-based signatures. Within this patient set, 462 patients had pathological response data; 91 patients achieved a pCR and 371 did not (20% overall pCR rate). The pCR rates varied according to intrinsic subtype as follows: Basal-like (n=129, 40% pCR), Claudin-low (n=70, 23% pCR), HER2-Enriched (n=27, 19% pCR), Luminal A (n=140, 3% pCR), Luminal B (n=68, 16% pCR), and Normal-like (n=28 total, 14% pCR). To determine the possible significance of our two response signatures on this test set of human patients, the mean expression values for each gene list was calculated and the distribution of values between pCR patients versus not pCR patients determined. As shown in Table 3.3, when all 473 patients were considered, the UNTREATED-HUM signature was significantly correlated with pCR ($p<0.001$) and the RESP-HUM signature was trending toward significance ($p=0.051$). As we further stratified patients into the five, and even six intrinsic subtypes the UNTREATED-HUM signature continued to maintain significance. Interestingly, the RESP-HUM signature predicted pCR more strongly in the Normal-like and Claudin-low subtypes while the UNTREATED-HUM signature better tracked response within the Basal-like subtype (Table 3.3). Lastly, the triple-negative breast cancer distinction is a highly clinically relevant group because these patients are not candidates for the current targeted therapies in the breast clinic [10, 11]; within this group, the UNTREATED-HUM signature was also a significant predictor (p=0.003).

To more rigorously test the predictive significance of these new expression signatures, multivariable analysis using logistic regression was performed that included the common clinical variables, the intrinsic subtypes, the RESP-HUM and UNTREATED-HUM signatures, and three predictive genomic signatures identified by Haztis et al. (Table 3.4). For these analyses, we used the subset of patients that had pCR/response data, survival data, and who were treated with an anthracycline and taxane chemotherapy regimen (n=441). As shown in Table 3.4, multiple biomarkers were predictive in univariate analyses, but only the UNTREATED-HUM, Basal-like, Normal-like, and one of the Haztis et al. chemotherapy predictor signatures (i.e. RCB-III/resistance) were found significant in both the univariate and multivariate tests. To further assess the strength of the predictive powers of these genomic signatures, each was used to calculate an Area Under the Curve (AUC) for pCR, both alone (univariate AUC) and in the multivariate model (Table 3.4). The UNTREAT-HUM signature provided a good univariate AUC, and the multivariate model provided improvement with a high AUC (0.879). When the three Hatzis et al. signatures were removed from the multivariate analysis, most of the variables that were significant in the initial MVA remained significant, and the overall model continued to show a high AUC (0.82) (data not shown). Lastly, an additional test data set of anthracycline and taxane treated human patients was tested, which represents 81 patients treated neoadjuvantly from Japan [29]; similar predictive results were seen for the UNTREAT-HUM signature, which was again a significant predictor in both the univariate and multivariate analyses (Table 3.5). These data show that the UNTREATED-HUM signature (and possibly the RESP-HUM) provided predictive information for pCR beyond 1) the commonly used clinical variables, 2) breast cancer subtype, and 3) other genomic signatures derived from one of the data sets tested here.

DISCUSSION

As new agents for breast cancers are developed, validated preclinical models for assessing these agents' activity alone and in combination with approved therapies are needed. In this study, we chose genomically credentialed GEMM representatives for three human breast tumor subtypes (Basal-like, Luminal B and Claudin-low) as our preclinical models. While using single representatives of different tumor subtypes does not allow for the identification of subtype-specific effects, we believe this approach does make future predictions of therapeutic efficacy more robust by including results from a biologically diverse group of tumor-bearing individuals.

For therapeutic efficacy, each GEMM was treated with identical regimens and for most drugs, variable responses were seen. Our findings show that the *MMTV-Neu* tumors were the most responsive in general, with multiple agents being able to achieve complete tumor regression, especially the HER2 targeted agent lapatinib. Next in sensitivities was the Basal-like *C3(1)-T-antigen* model, which was generally more resistant than the *MMTV-Neu* model, but in some cases complete responses were documented (CT and carboplatin/erlotinib); interestingly, a heterogeneity of responses was common in this GEMM (Figure 3.4a), suggesting that two or more sub-classes of tumors may be present. Importantly, a similar heterogeneous response pattern is seen within human Basal-like patients when treated with comparable agents where many patients achieve a pCR and have good overall survival, but the majority show residual disease and worse outcomes (Figure 3.6c and see [5, 36]). Lastly, the Claudin-low *T11/TP53-/-* model was the most resistant with only small responses seen in this model.

We ultimately chose to focus our analysis on expression-signatures associated with chemotherapy treatment of one of our GEMMs and response for two main reasons. First, we reasoned transcripts highly expressed in sensitive murine tumors (i.e. the RESP-HIGH list) might also be highly expressed in sensitive human tumors; although this list was predictive in human tumors, it was not obvious from gene ontology analysis what molecular characteristics drive this biology, and this list was not significant when accounting for other variables (MVA p-value = 0.058). Second, in a tumor treated *in vivo*, we reasoned chemotherapy might deplete the most sensitive cells and their characteristic transcripts. Therefore, the collection of transcripts that were highly expressed in untreated cells and depleted with treatment (i.e. the UNTREATED list) similarly seemed rational for testing in humans. Specifically, an analysis showed the 26-gene UNTREAT-HUM signature (Figure 3.6) was a significant predictor of response and may also provide mechanistic insight. This 26-gene list suggests that the cells actually undergoing DNA synthesis and mitosis (i.e. in S/G2/M-phase) are more sensitive to cytotoxic agents than cells in other parts of the cell cycle (G0 or G1), which is a concept dating back to the 1960's (reviewed in [37]). It is important to note that this list added independent information above and beyond strict assessments of proliferation (e.g. an 11-gene proliferation signature that contains Ki-67), suggesting this list may better capture specific features of the cell cycle (e.g. length of time spent in S/G2/M) associated with sensitivity to carboplatin/paclitaxel. The UNTREAT-HUM list is in fact a biologically rich list that contains at least two different sets of genes/proteins that physically form a multi-protein complex, namely SMC2 and SMC4, and MCM4 and MCM6. In addition, this list has two different E2F family members (E2F3 and E2F8), for which a poor prognostic signature has already been linked to E2F3 [38]. These data also suggest that no single gene/protein is likely to be a robust biomarker of chemosensitivity because a multitude of genes, each involved in different aspects of the cell cycle, were collectively identified as being predictive of response. These new expression signatures were derived from murine models that, despite their specific chemoresponses not being a mirror of their human counterparts (i.e. paclitaxel), added a significant predictive component to the multivariate model that at least equaled the ability of those tested signatures that were derived directly from this human tumor data set.

In terms of human biomarker advances, we made progress using the *C3(1)-Tag* GEMM. As shown in Tables 3.3 and 3.4, the UNTREATED-HUM signature was predictive of response to a multi-agent neoadjuvant chemotherapy regimen, not only across all HER2-normal human breast cancer patients but also within the clinically relevant triple-negative subset, as well as the more biologically relevant Basal-like subset. Interestingly, this UNTREATED-HUM signature was also able to predict pCR even when accounting for intrinsic subtype, the common clinical variables, and two other genomic signatures specifically designed to predict neoadjuvant response (Table 3.4). Although the murine treatment and human treatment involved the use of different chemotherapeutics, both species studies used paclitaxel and at least one DNA damaging agent (carboplatin in mice and doxorubicin/epirubicin in humans). Overall, a multivariate model that contained the UNTREAT-HUM, the intrinsic subtypes, and the common clinical variables showed an AUC of 0.82, which may be sufficiently predictive to be of value for routine clinical use.

We were surprised to find that the results from mice treated with single agent paclitaxel did not mimic the effectiveness of this drug in human breast cancer patients. Delivery of higher therapeutic doses of paclitaxel to the mice (i.e. doses closer to those received by human patients) may have proven more efficacious; however, our chosen formulation of paclitaxel contained chremaphor and ethanol in amounts that precluded higher dosing. Another caveat to our studies is that these two GEMM-derived signatures were both predictive and prognostic; however, it must be noted that it is often difficult, if not impossible, to disentangle these two features. For example, both ER and HER2 in breast cancer are prognostic (they predict outcomes in the absence of therapy) and they are predictive (ER predicts hormone therapy benefit and HER2 predicts trastuzumab benefit) and thus, our new signatures are showing dual properties similar to those seen for the existing breast cancer biomarkers. Much additional validation work is needed before these two murine-derived signatures could be used to guide patient treatment. However, this study has laid the groundwork of a general strategy for evaluating new drugs, combinations, and schedules using GEMMs and has shown it is possible to use mice as a tool to identify a biomarker that may be of predictive value for human cancer patients.

TABLES

Table 3.1 Summary of drugs used in this study, their doses, and utilized schedules of administration.

Chemical Agent	Activity	Dose (mpk)	Route	Schedule	Notes
Carboplatin	DNA cross links	50	Parenteral	weekly	Carboplatin was dosed in a range from 50-75 mpk as a single agent. A dose of 50mpk was determined to illicit a tumor response and was tolerable for overall survival.
Erlotinib	EGFR/HER1 inhibitor	25	PO	continuous in food	Doses were based upon literature review.
Lapatinib	HER2/ERBB2 inhibitor	220	PO	continuous in food	Drug was first dosed on the targeted model MMTV-Neu at 75 mpk PO in food. Stable disease was reached in 21 days but no toxicity was noted. The dose was escalated systematically to 220 mpk. This dose caused tumor complete regression in the MMTV-Neu model within 14 days and was tolerated well with mild toxicity only showing in some animals after 120 days of continuous treatment. Plasma was drawn to confirm drug presence.
Paclitaxel	stabilizes microtubules	10	Parenteral	weekly	Drug was dosed by tail IV at 10 and 20 mpk and IP at 10 mpk. IV 20 mpk caused moderate skin lesions on the tail around 28 days. IV and IP 10 mpk were well tolerated and no tumor response differences between IV and IP were noted. The IP route was pursued for all subsequent studies due to the significantly greater ease of repeated administration.

Abbreviations: mpk, mg per kilogram; PO, by mouth; IV, intravenous; IP, Intraperitoneal

Table 3.2 List of mouse *C3(1)-T-antigen* gene expression microarrays used to derive the murine gene lists.

Treatment	Experiment name	SlideName	GEO ID
Non-treated	FVB_C3(1)-Tag_116409C_untreated	Agilent-251486822731-2	GSM929889
Non-treated	FVB_C3(1)-Tag_116410B_untreated	Agilent-251486822731-1	GSM929888
Non-treated	FVB_C3(1)-Tag_117338-1_untreated	Agilent-251486822730-4	GSM929887
Non-treated	FVB_C3(1)-Tag_117517_untreated	Agilent-148681502-1	GSM929880
CT Treated	FVB_C3(1)-Tag_118657_three-week-treatment_Non-responder	Agilent-251486822800-2	GSM929890
Non-treated	FVB_C3(1)-Tag_120157_untreated	Agilent-148681503-3	GSM929881
CT Treated	FVB_C3(1)-Tag_120865_two-week-treatment_Non-responder	Agilent-251486822800-4	GSM929891
CT Treated	FVB_C3(1)-Tag_121491-three-week-treatment_Non-responder	Agilent-148682256-4	GSM929882
CT Treated	FVB_C3(1)-Tag_123051-three-week-treatment_Non-responder	Agilent-148682257-1	GSM929883
CT Treated	FVB_C3(1)-Tag_123240-two-week-treatment_Non-responder	Agilent-148682257-4	GSM929884
CT Treated	FVB_C3(1)-Tag_125653a_two-week-treatment_Non-responder	Agilent-1486822814-2	GSM929885
CT Treated	FVB_C3(1)-Tag_125905_two-week-treatment_Non-responder	Agilent-1486822814-3	GSM929886
Non-treated	FVB_C3(1)-Tag-120555_untreated	Mouse 4X44K-251486820747-120555	GSM929879
Non-treated	FVB_C3(1)-Tag-121415_untreated	Mouse 4X44K-251486819757-121415	GSM929873
CT Treated	FVB_C3(1)-Tag-121450T2-three-week-treatment_Non-responder	Mouse 4X44K-251486819760-121450-T2	GSM929878
CT Treated	FVB_C3(1)-Tag-122387_two week-treatment_Responder	Mouse 4X44K-251486819758-122387	GSM929874
CT Treated	FVB_C3(1)-Tag-122738-three week-treatment_Responder	Mouse 4X44K-251486819750-122738	GSM929875
CT Treated	FVB_C3(1)-Tag-124051T1-three-week-treatment_Non-responder	Mouse 4X44K-251486819750-124051-T1	GSM929876
CT Treated	FVB_C3(1)-Tag-124051T2-three-week-treatment_Responder	Mouse 4X44K-251486819751-124051-T2	GSM929877

Table 3.3 Pathological Complete Response (pCR) rates across different patient subsets for the RESP-HUM and UNTREATED-HUM signatures.

			RESP-HUM			**UNTREATED-HUM**		
	pCR	**Risidual disease**	**P-value**	**AUC**	**Odds ratio**	**P-value**	**AUC**	**Odds ratio**
All patients	91(19.7%)	371(80.3%)	*0.051*	*0.586*	*0.788 (0.61-0.99)*	*<0.001*	*0.752*	*2.72 (2.08-3.63)*
ER-negative only	62(33.3%)	124(66.7%)	0.821			*<0.001*	*0.683*	*2.09 (1.44-3.11)*
ER positive only	29(10.5%)	246(89.5%)	0.405			*<0.001*	*0.747*	*2.64 (1.67-4.31)*
PAM50 (5 intrinsic subtypes)								
Basal-like	62(35.8%)	111(64.2%)	0.707			*0.001*	*0.649*	*2.05 (1.33-3.24)*
HER2-enriched	5(17.9%)	23(82.1%)	0.43			0.076		
Luminal A	4(2.8%)	141(97.2%)	0.208			0.172		
Luminal B	12(16.7%)	60(83.3%)	0.638			0.079		
Normal-like	8(18.2%)	36(81.8%)	*0.009*	*0.837*	*4.47 (1.69-16.9)*	0.274		
PAM50 + Claudin-low (6 intrinsic subtypes)								
Basal-like	51(39.5%)	78(60.5%)	0.356			*0.001*	*0.68*	*2.33 (1.44-3.93)*
Claudin-low	16(22.9%)	54(77.1%)	*0.054*	*0.660*	*1.55 (1-2.47)*	0.474		
HER2-enriched	5(18.5%)	22(81.5%)	0.426			0.086		
Luminal A	4(2.9%)	136(97.1%)	0.223			0.169		
Luminal B	11(16.2%)	57(83.8%)	0.976			0.272		
Normal-like	4(14.3%)	24(85.7%)	0.052			0.137		
Triple-Negative only	56(33.5%)	111(66.5%)	0.651			*0.003*	*0.651*	*1.8 (1.24-2.68)*

NOTE: The P-value and Area-Under-the-Curve (AUC) columns indicate whether the RESP-HUM or UNTREATED-HUM signature (as a continuous variable from low to high expression) was associated with response (italics) within that patient set/subset.

Table 3.4 Univariate and Multivariate Analysis for pCR using clinical and genomic features including the RESP-HUM and UNTREATED-HUM signatures on the Hatzis et al. data set.

		Univariate			**Multivariate**		
	No. of pts*	**p-value**	**odds ratio**	**AUC**	**p-value**	**odds ratio**	**AUC**
UNTREATED-HUM	441	<0.001	2.57 (1.96-3.45)	0.740	0.013	2.3 (1.21-4.52)	0.879
RESP-HUM	441	0.073	0.796 (0.618-1.02)	0.583	0.058	1.45 (0.99-2.15)	
PAM50 proliferation	441	<0.001	2.57 (1.9-3.56)	0.730	0.917	0.96 (0.443-2.11)	
ER							
Negative	175(40%)		1	0.562		1	
Positive	266(60%)	<0.001	0.234 (0.139-0.385)		0.992	0.99 (0.391-2.52)	
PR							
Negative	227(51%)		1	0.467			
Positive	214(49%)	<0.001	0.30 (0.176-0.51)		0.683	0.83 (0.363-1.97)	
Clinical T Stage							
1	27(6%)		1	0.571		1	
2	226(51%)	0.364	0.652 (0.269-1.75)		0.902	0.92 (0.261-3.39)	
3	126(29%)	0.746	0.854 (0.34-2.36)		0.844	0.87 (0.236-3.38)	
4	62(14%)	0.054	0.306 (0.0885-1.02)		0.195	0.35 (0.071-1.69)	
Clinical Grade							
1	28(6%)		1	0.481		1	
2	170(39%)	0.498	2.05 (0.38-38.1)		0.967	1.05 (0.13-23.7)	
3	243(55%)	0.019	11.1 (2.3-201)		0.574	2.02 (0.235-46.6)	
PAM50							
LumA	141(32%)		1	0.633		1	
Basal	167(38%)	<0.001	18.2 (7.21-61.5)		0.026	5.76 (1.3-29.4)	
Her2	24(5%)	0.010	6.85 (1.51-31.1)		0.161	3.55 (0.59-21.7)	
LumB	67(15%)	0.003	6.01 (1.92-22.6)		0.400	1.92 (0.438-9.49)	
Normal	42(10%)	0.001	8.06 (2.39-31.7)		0.002	10 (2.34-47.6)	
Hatzis signature SET index							
1	386(88%)		1	0.136		1	
2	36(8%)	0.091	0.353 (0.0835-1.02)		0.729	1.34 (0.223-6.49)	
3	19(4%)	0.302	0.457 (0.0715-1.64)		0.953	0.94 (0.105-6.17)	
Hatzis signature chemo sensitive (RCB-I predict)							
1	296(67%)		1	0.553		1	
2	145(33%)	<0.001	2.97 (1.82-4.84)		0.127	1.76 (0.855-3.67)	
Hatzis signature chemo resistance (RCB-III predict or 3 year survival)							
1	197(45%)		1	0.603		1	
2	244(55%)	<0.001	0.089 (0.0449-0.166)		<0.001	0.129 (0.0565-0.277)	

NOTE: Univariate and Multivariate analyses were performed using all Hatzis et al. patients who received anthracycline and taxane chemotherapy only, and who had overall survival data (n=441).

* The number of patients with clinical ER status, PR status, T stage, grade and pCR status available.

Table 3.5 Univariate and Multivariate Analysis for pCR using clinical and genomic features including the UNTREATED-HUM and RESP-HUM signatures on the Miyake et al. data set.

		Univariate			**Multivariate**		
	No. of pts*	**p-value**	**odds ratio**	**AUC**	**p-value**	**odds ratio**	**AUC**
UNTREATED-HUM	81	0.019	2.38 (1.22-5.3)	0.712	0.038	45 (2.11-3290)	0.900
RESP-HUM	81	0.395	0.778 (0.424-1.37)	0.593	0.351	1.98 (0.488-9.23)	
PAM50 proliferation	81	0.097	1.95 (0.971-4.73)	0.672	0.152	0.109 (0.00383-1.92)	
ER							
Negative	26(32%)		1	0.631		1	
Positive	55(68%)	0.003	0.16 (0.044-0.518)		0.037	0.001 (7.29e-07-0.22)	
PR							
Negative	43(53%)		1	0.446			
Positive	38(47%)	0.091	0.34 (0.0877-1.11)		0.227	5.94 (0.408-188)	
Clinical T Stage							
1+2	66(81%)			0.205		1	
3+4	15(19%)	0.218	0.265 (0.014-1.5)		0.100	0.0668 (0.00126-1.02)	
Clinical Grade							
1	13(16%)		1	0.624		1	
2	54(67%)	0.818	0.819 (0.168-6)		0.790	0.722 (0.0688-10.5)	
3	14(17%)	0.131	4.12 (0.729-33.6)		0.303	6.41 (0.231-335)	
Clinical Nodal Status							
Negative	23(28%)		1	0.322		1	
Positive	58(72%)	0.068	7 (1.28-131)		0.055	13.4 (1.4-399)	
PAM50							
LumA	25(31%)		1	0.713		1	
Basal	15(19%)	0.057	5.75 (1.05-45.2)		0.118	0.00262 (3.92e-07-1.4)	
Her2	9(11%)	0.026	9.2 (1.41-81.8)		0.160	0.0135 (1.12e-05-2.08)	
LumB	19(23%)	0.428	2.16 (0.322-17.8)		0.880	1.26 (0.0627-29.5)	
Normal	13(16%)	0.973	0.958 (0.042-11)		0.102	0.00494 (3.86e-06-1.08)	

NOTE: Univariate and Multivariate analyses were performed using the clinically HER2 negative subset of patients sets taken from Miyake et al. 2012.

* The number of patients with clinical ER status, PR status, T stage, grade, nodal status and pCR status available.

FIGURES

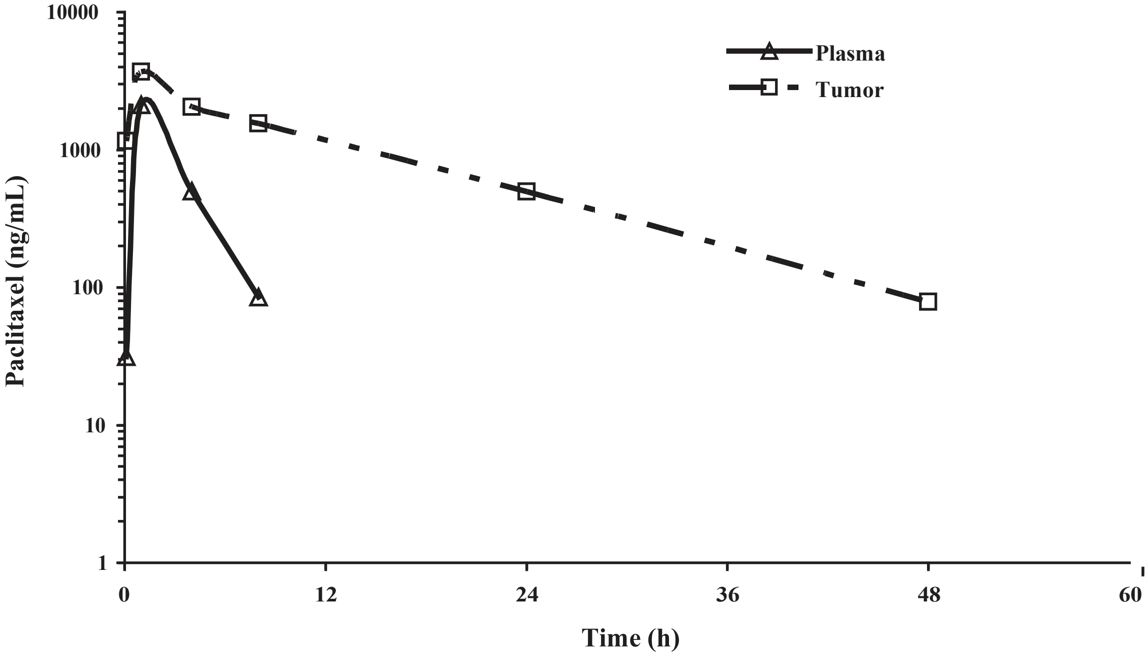


Figure 3.1 Pharmacokinetic evaluation of paclitaxel delivery. Paclitaxel drug concentrations were measured by mass spectroscopy and samplings of MMTV-Neu tumors and plasma. The results show significant systemic delivery of this drug when administered using intraperitoneal injections, both in the tumor and in the plasma.

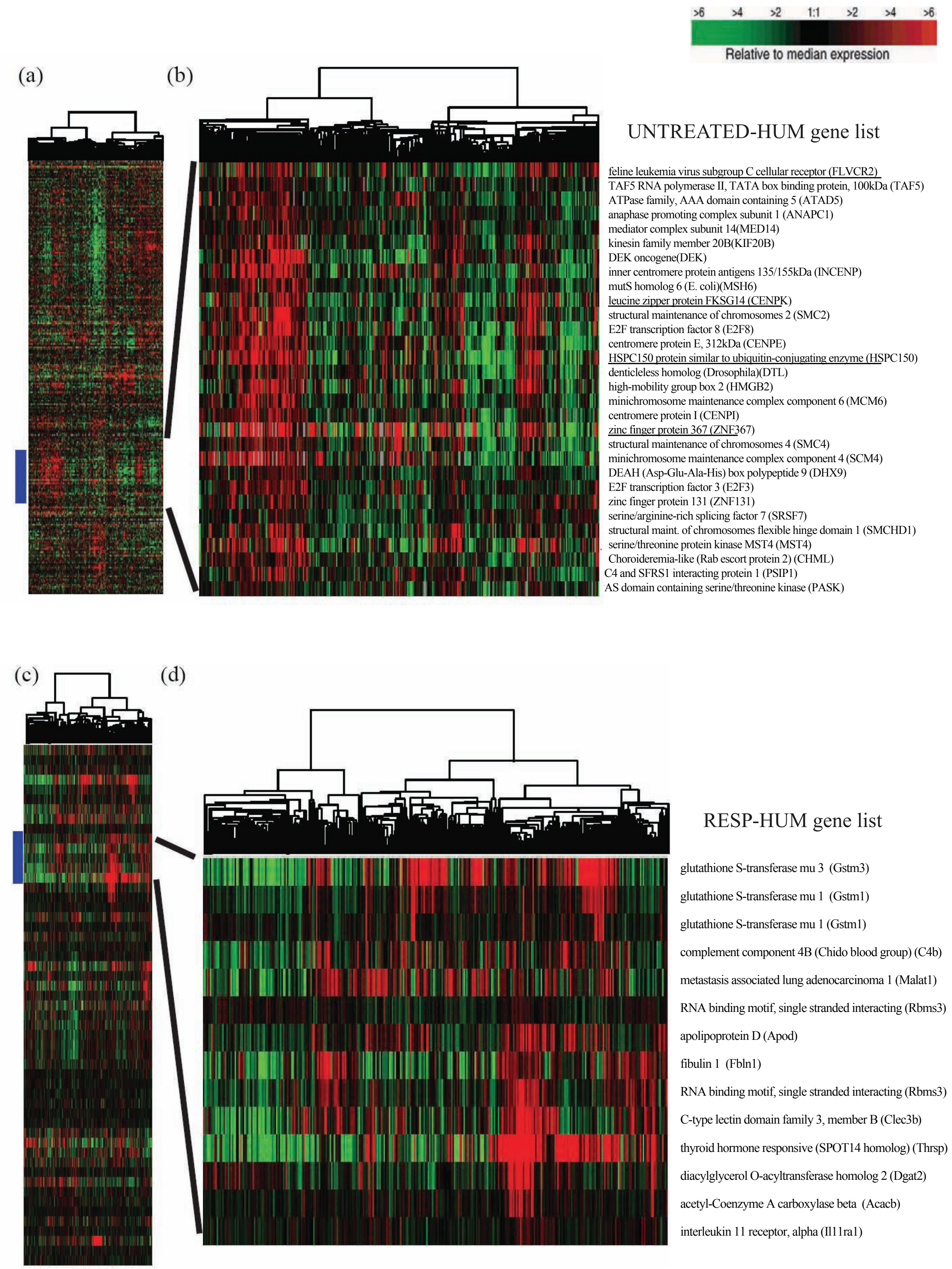
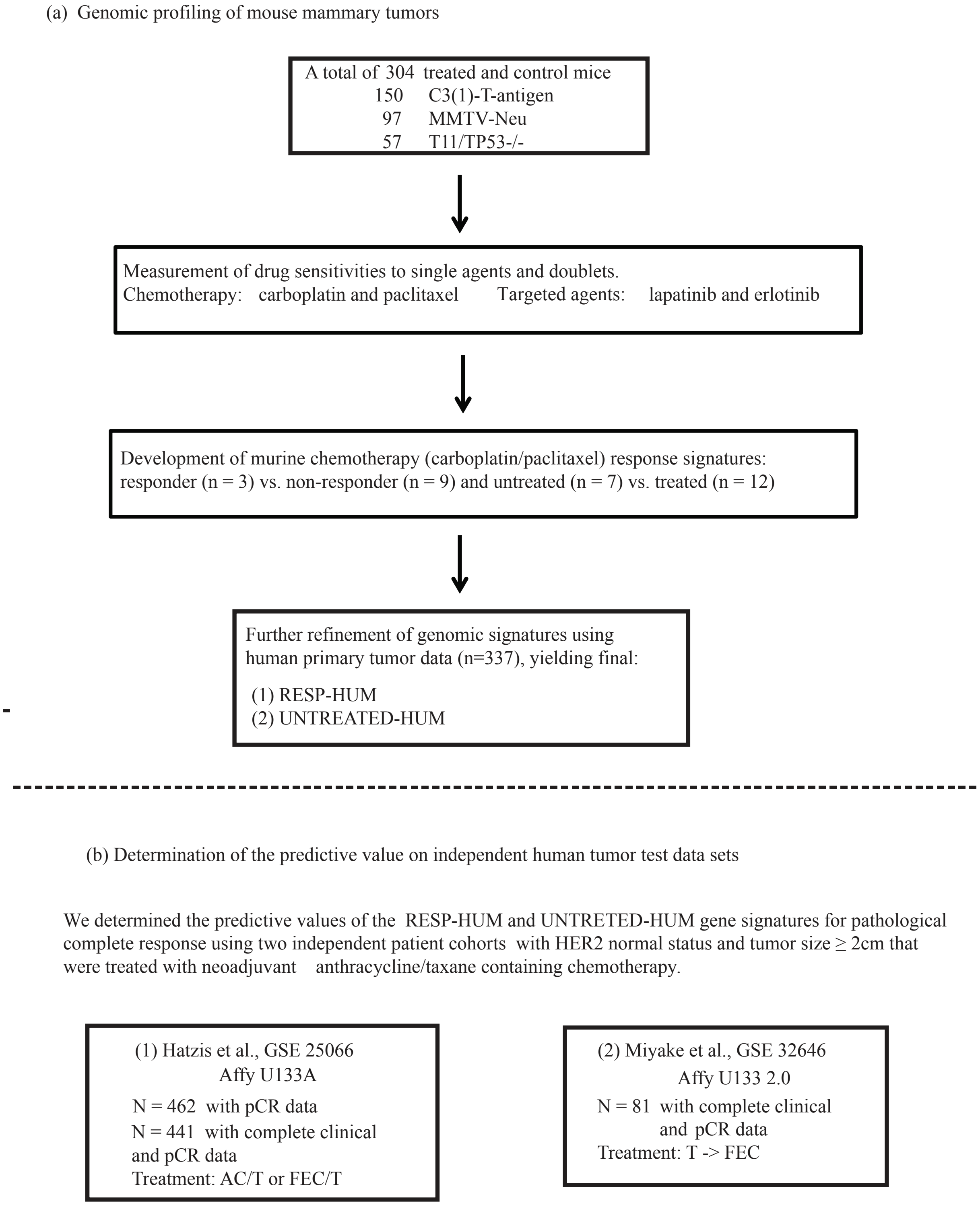

>6
>4
>2
1:1
>2
>4
>6
Relative to median expression
(a)
(b)
UNTREATED-HUM gene list
feline leukemia virus subgroup C cellular receptor (FLVCR2)
TAF5 RNA polymerase II, TATA box binding protein, 100kDa (TAF5)
ATPase family, AAA domain containing 5 (ATAD5)
anaphase promoting complex subunit 1 (ANAPC1)
mediator complex subunit 14(MED14)
kinesin family member 20B(KIF20B)
DEK oncogene(DEK)
inner centromere protein antigens 135/155kDa (INCENP)
mutS homolog 6 (E. coli)(MSH6)
leucine zipper protein FKSG14 (CENPK)
structural maintenance of chromosomes 2 (SMC2)
E2F transcription factor 8 (E2F8)
centromere protein E, 312kDa (CENPE)
HSPC150 protein similar to ubiquitin-conjugating enzyme (HSPC150)
denticleless homolog (Drosophila)(DTL)
high-mobility group box 2 (HMGB2)
minichromosome maintenance complex component 6 (MCM6)
centromere protein I (CENPI)
zinc finger protein 367 (ZNF367)
structural maintenance of chromosomes 4 (SMC4)
minichromosome maintenance complex component 4 (SCM4)
DEAH (Asp-Glu-Ala-His) box polypeptide 9 (DHX9)
E2F transcription factor 3 (E2F3)
zinc finger protein 131 (ZNF131)
serine/arginine-rich splicing factor 7 (SRSF7)
structural maint. of chromosomes flexible hinge domain 1 (SMCHD1)
serine/threonine protein kinase MST4 (MST4)
Choroideremia-like (Rab escort protein 2) (CHML)
C4 and SFRS1 interacting protein 1 (PSIP1)
AS domain containing serine/threonine kinase (PASK)
(c)
(d)
RESP-HUM gene list
glutathione S-transferase mu 3 (Gstm3)
glutathione S-transferase mu 1 (Gstm1)
glutathione S-transferase mu 1 (Gstm1)
complement component 4B (Chido blood group) (C4b)
metastasis associated lung adenocarcinoma 1 (Malat1)
RNA binding motif, single stranded interacting (Rbms3)
apolipoprotein D (Apod)
fibulin 1 (Fbln1)
RNA binding motif, single stranded interacting (Rbms3)
C-type lectin domain family 3, member B (Clec3b)
thyroid hormone responsive (SPOT14 homolog) (Thrsp)
diacylglycerol O-acyltransferase homolog 2 (Dgat2)
acetyl-Coenzyme A carboxylase beta (Acacb)
interleukin 11 receptor, alpha (Il11ra1)

Figure 3.2 Hierarchical clustering analysis of the untreated and responding murine chemotherapy signature using 337 human breast tumors. (a) The 348 genes highly expressed in untreated C3(1)-T-antigen tumors versus carboplatin/paclitaxel treated tumors was used to cluster the human breast tumor data set from Prat et al. 2010. (b) The highlighted dendrogram node identifies the 30 genes that were selected for additional analyses, for which 26 orthologs were found in the other human data sets (missing genes are identified by underlining) and used to evaluate correlations with pathological complete response. (c) The 74 genes highly expressed in those C3(1)-T-antigen tumors that responded to carboplatin/paclitaxel treatment versus those tumors that did not respond was used to cluster the human breast tumor data set from Prat et al. 2010. (d) The highlighted dendrogram node identifies the 12 genes that were selected for additional analyses and were used to evaluate correlations with pathological complete response on other data sets.



(a) Genomic profiling of mouse mammary tumors

A total of 304 treated and control mice
150 C3(1)-T-antigen
97 MMTV-Neu
57 T11/TP53-/-

Measurement of drug sensitivities to single agents and doublets.
Chemotherapy: carboplatin and paclitaxel Targeted agents: lapatinib and erlotinib

Development of murine chemotherapy (carboplatin/paclitaxel) response signatures:
responder (n = 3) vs. non-responder (n = 9) and untreated (n = 7) vs. treated (n = 12)

Further refinement of genomic signatures using human primary tumor data (n=337), yielding final:

(1) RESP-HUM
(2) UNTREATED-HUM

(b) Determination of the predictive value on independent human tumor test data sets

We determined the predictive values of the RESP-HUM and UNTRETED-HUM gene signatures for pathological complete response using two independent patient cohorts with HER2 normal status and tumor size ≥ 2cm that were treated with neoadjuvant anthracycline/taxane containing chemotherapy.

(1) Hatzis et al., GSE 25066
Affy U133A
N = 462 with pCR data
N = 441 with complete clinical and pCR data
Treatment: AC/T or FEC/T

(2) Miyake et al., GSE 32646
Affy U133 2.0
N = 81 with complete clinical and pCR data
Treatment: T -> FEC

Figure 3.3 Study design overview. (a) Drug treatment and genomic profiling of mouse mammary tumors for the development of chemotherapy response signatures. (b) Testing of genomic signatures on two human tumor neoadjuvant treatment data test data sets.

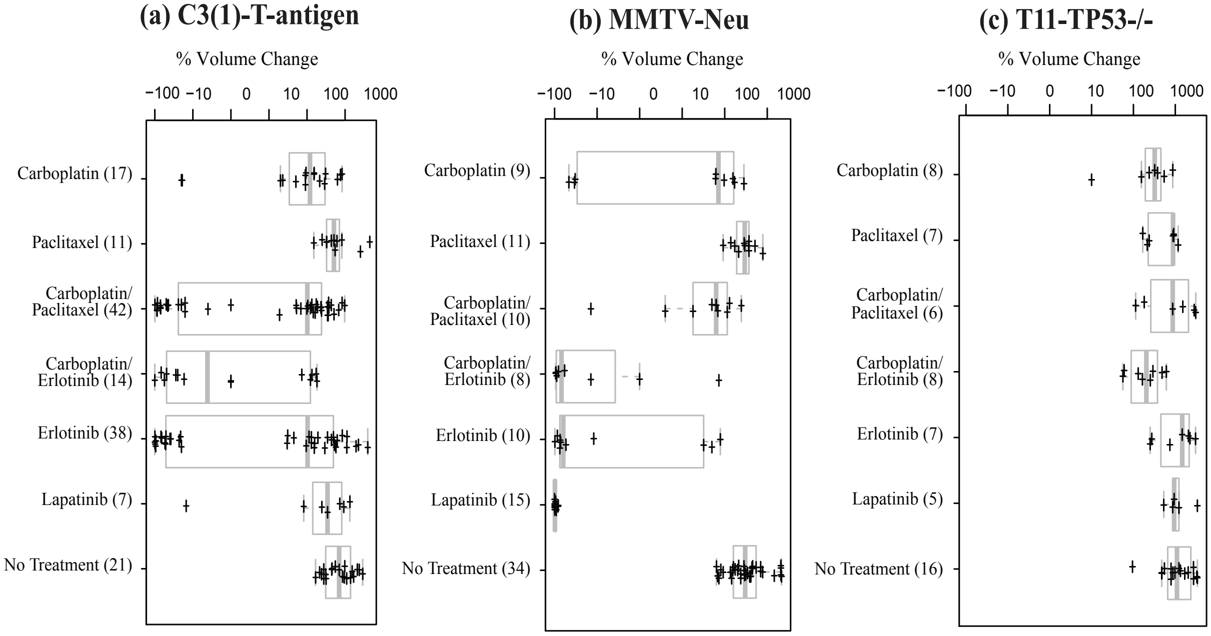


Figure 3.4 Short-term treatment responses for three mouse models of mammary cancer. Box and whisker plots are shown as measures of tumor responsiveness. In each case, 2-3 cycles of therapy was administered for all chemotherapeutics (1 dose/week), while in the case of erlotinib and lapatinib, the drug was continuously administered via the chow. Tumor size was measured at baseline and at weekly intervals thereafter. The change in tumor volume over a 21-day treatment period is plotted for (a) C3(1)-T-antigen model, (b) MMTV-Neu model, and (c) T11/TP53-/- model; note that the T11/TP53-/- model is based upon a 14-day treatment period due to its faster growth rate. Drugs that elicited a statistically significant response as assessed by a t-test when compared versus its matched untreated controls are identified by being underlined. The number of animals in each treatment group is indicated in parentheses.

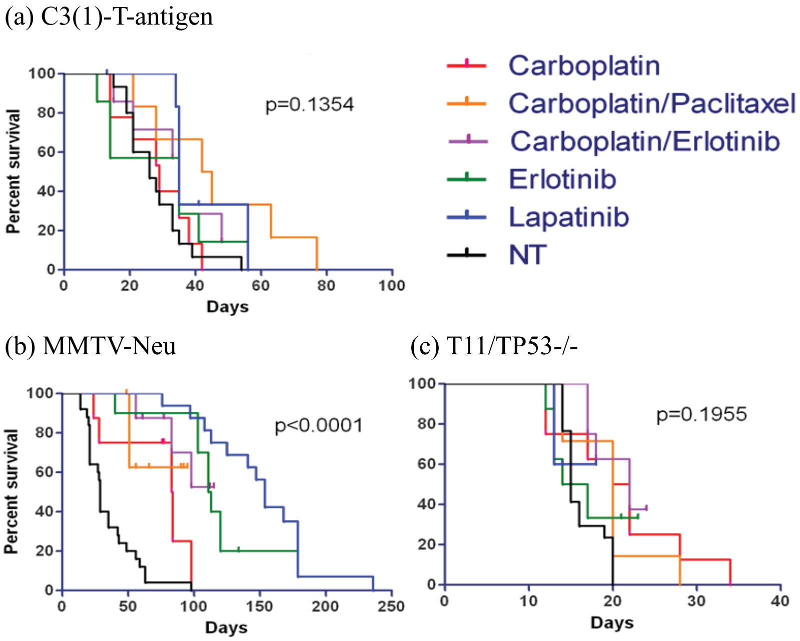


Figure 3.5 Long term survival results for three mouse models of mammary cancer. Kaplan-Meier analyses for overall survival of tumor bearing mice was performed. A) C3(1)-T-antigen, B) MMTV-Neu, and C) T11/TP53-/- results for chemotherapeutic treatments, targeted agents, and combinations. A log-rank test was performed to determine significance of all treatment groups and is shown.

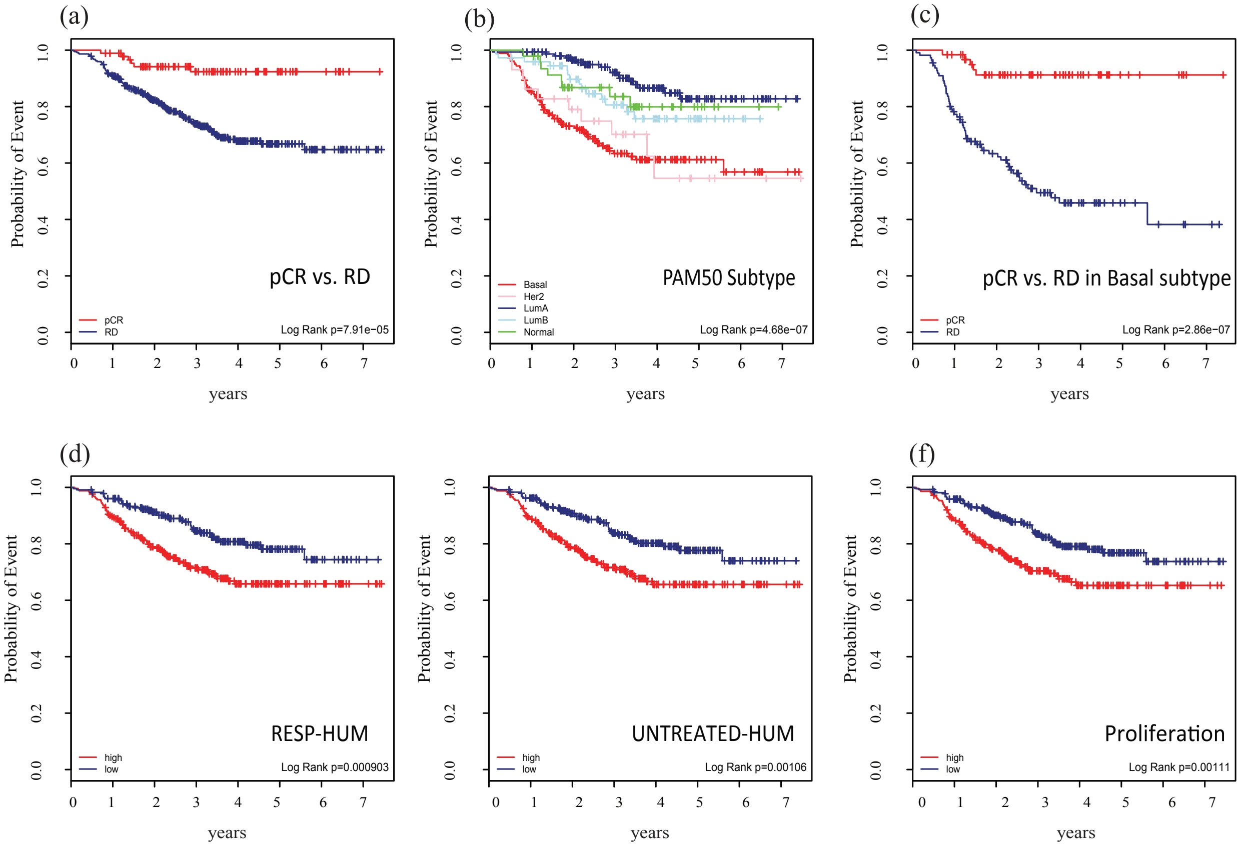


Figure 3.6 Kaplan-Meier analyses for the prediction of Distant Relapse Free Survival. Using the Hatzis et al. data set, Kaplan-Meier plots were performed for A) pCR vs. residual disease (RD), B) the five PAM50-defined intrinsic subtypes, C) pCR vs. RD within just Basal-like subtype patients, D) high versus low expression of the RESP-HUM 12-gene signature, E) high versus low expression of the UNTREATED-HUM 26-gene signature, and F) high versus low expression of an 11-gene proliferation signature taken from Nielsen et al. 2010.

REFERENCES

1. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, He X, Perou CM: **Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer.** *Breast Cancer Res* 2010, **12**:R68.

2. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS: **Supervised risk predictor of breast cancer based on intrinsic subtypes.** *J Clin Oncol* 2009, **27**:1160–1167.

3. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406**:747–752.

4. Hugh J, Hanson J, Cheang MCU, Nielsen TO, Perou CM, Dumontet C, Reed J, Krajewska M, Treilleux I, Rupin M, Magherini E, Mackey J, Martin M, Vogel C: **Breast cancer subtypes and response to docetaxel in node-positive breast cancer: use of an immunohistochemical definition in the BCIRG 001 trial.** *J Clin Oncol* 2009, **27**:1168–1176.

5. Carey LA, Dees EC, Sawyer L, Gatti L, Moore DT, Collichio F, Ollila DW, Sartor CI, Graham ML, Perou CM: **The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes.** *Clin Cancer Res* 2007, **13**:2329–2334.

6. Martin M, Romero A, Cheang MCU, López García-Asenjo JA, García-Saenz JA, Oliva B, Román JM, He X, Casado A, de la Torre J, Furio V, Puente J, Caldés T, Vidart JA, Lopez-Tarruella S, Diaz-Rubio E, Perou CM: **Genomic predictors of response to doxorubicin versus docetaxel in primary breast cancer.** *Breast Cancer Res Treat* 2011, **128**:127–136.

7. Glück S, Ross JS, Royce M, McKenna EF, Perou CM, Avisar E, Wu L: **TP53 genomics predict higher clinical and pathologic tumor response in operable early-stage breast cancer treated with docetaxel-capecitabine ± trastuzumab**. *Breast Cancer Res Treat* 2012, **132**:781–791.

8. Dunbier AK, Anderson H, Ghazoui Z, Salter J, Parker JS, Perou CM, Smith IE, Dowsett M: **Association between breast cancer subtypes and response to neoadjuvant anastrozole**. *Steroids* 2011, **76**:736–740.

9. Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T, Davies SR, Snider J, Stijleman IJ, Reed J, Cheang MCU, Mardis ER, Perou CM, Bernard PS, Ellis MJ: **A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer.** *Clin Cancer Res* 2010, **16**:5222–5232.

10. Perou CM: **Molecular stratification of triple-negative breast cancers.** *Oncologist* 2011, **16 Suppl 1**:61–70.

11. Prat A, Adamo B, Cheang MCU, Anders CK, Carey LA, Perou CM: **Molecular characterization of basal-like and non-basal-like triple-negative breast cancer.** *Oncologist* 2013, **18**:123–33.

12. Herschkowitz JI, Zhao W, Zhang M, Usary J, Murrow G, Edwards D, Knezevic J, Greene SB, Darr D, Troester MA, Hilsenbeck SG, Medina D, Perou CM, Rosen JM: **Comparative oncogenomics identifies breast tumors enriched in functional tumor-initiating cells**. *Proc Natl Acad Sci* 2012, **109**:2778–2783.

13. Van Dyke T, Jacks T: **Cancer modeling in the modern era: Progress and challenges**. *Cell* 2002, **108**:135–144.

14. Sharpless NE, Depinho RA: **The mighty mouse: genetically engineered mouse models in cancer drug development**. *Nat Rev Drug Discov* 2006, **5**:741–754.

15. Roberts PJ, Usary JE, Darr DB, Dillon PM, Pfefferle AD, Whittle MC, Duncan JS, Johnson SM, Combest AJ, Jin J, Zamboni WC, Johnson GL, Perou CM, Sharpless NE: **Combined PI3K/mTOR and MEK Inhibition Provides Broad Antitumor Activity in Faithful Murine Cancer Models**. *Clin Cancer Res* 2012, **18**:5290–5303.

16. Chen Z, Cheng K, Walton Z, Wang Y, Ebi H, Shimamura T, Liu Y, Tupper T, Ouyang J, Li J, Gao P, Woo MS, Xu C, Yanagita M, Altabef A, Wang S, Lee C, Nakada Y, Pe ña CG, Sun Y, Franchetti Y, Yao C, Saur A, Cameron MD, Nishino M, Hayes DN, Wilkerson MD, Roberts PJ, Lee CB, Bardeesy N, et al.: **A murine lung cancer co-clinical trial identifies genetic modifiers of therapeutic response**. *Nature* 2012, **483**:613–617.

17. Nardella C, Lunardi A, Patnaik A, Cantley LC, Pandolfi PP: **The APL Paradigm and the "Co-Clinical Trial" Project**. *Cancer Discov* 2011, **1**:108–116.

18. Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z, Rasmussen KE, Jones LP, Assefnia S, Chandrasekharan S, Backlund MG, Yin Y, Khramtsov AI, Bastein R, Quackenbush J, Glazer RI, Brown PH, Green JE, Kopelovich L, Furth PA, Palazzo JP, Olopade OI, Bernard PS, Churchill GA, Van Dyke T, Perou CM: **Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors.** *Genome Biol* 2007, **8**:R76.

19. Maroulakou IG, Anver M, Garrett L, Green JE: **Prostate and mammary adenocarcinoma in transgenic mice carrying a rat C3(1) simian virus 40 large tumor antigen fusion gene.** *Proc Natl Acad Sci U S A* 1994, **91**:11236–11240.

20. Guy CT, Webster MA, Schaller M, Parsons TJ, Cardiff RD, Muller WJ: **Expression of the neu protooncogene in the mammary epithelium of transgenic mice induces metastatic disease.** *Proc Natl Acad Sci U S A* 1992, **89**:10578–10582.

21. Green JE, Shibata MA, Yoshidome K, Liu ML, Jorcyk C, Anver MR, Wigginton J, Wiltrout R, Shibata E, Kaczmarczyk S, Wang W, Liu ZY, Calvo A, Couldrey C: **The C3(1)/SV40 T-antigen transgenic mouse model of mammary cancer: ductal epithelial cell targeting with multistage progression to carcinoma.** *Oncogene* 2000, **19**:1020–1027.

22. Jerry DJ, Kittrell FS, Kuperwasser C, Laucirica R, Dickinson ES, Bonilla PJ, Butel JS, Medina D: **A mammary-specific model demonstrates the role of the p53 tumor suppressor gene in tumor development.** *Oncogene* 2000, **19**:1052–1058.

23. Hou W, Watters JW, McLeod HL: **Simple and rapid docetaxel assay in plasma by protein precipitation and high-performance liquid chromatography-tandem mass spectrometry**. *J Chromatogr B Anal Technol Biomed Life Sci* 2004, **804**:263–267.

24. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95**:14863–14868.

25. Saldanha AJ: **Java Treeview--extensible visualization of microarray data.** *Bioinformatics* 2004, **20**:3246–3248.

26. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci U S A* 2001, **98**:5116–5121.

27. Spector NL, Xia W, Burris H, Hurwitz H, Dees EC, Dowlati A, O'Neil B, Overmoyer B, Marcom PK, Blackwell KL, Smith DA, Koch KM, Stead A, Mangum S, Ellis MJ, Liu L, Man AK, Bremer TM, Harris J, Bacus S: **Study of the biologic effects of lapatinib, a reversible inhibitor of ErbB1 and ErbB2 tyrosine kinases, on tumor growth and survival pathways in patients with advanced malignancies.** *J Clin Oncol* 2005, **23**:2502–12.

28. Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, Vidaurre T, Holmes F, Souchon E, Wang H, Martin M, Cotrina J, Gomez H, Hubbard R, Chacón JI, Ferrer-Lozano J, Dyer R, Buxton M, Gong Y, Wu Y, Ibrahim N, Andreopoulou E, Ueno NT, Hunt K, Yang W, Nazario A, DeMichele A, O'Shaughnessy J, Hortobagyi GN, Symmans WF: **A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer.** *JAMA* 2011, **305**:1873–1881.

29. Miyake T, Nakayama T, Naoi Y, Yamamoto N, Otani Y, Kim SJ, Shimazu K, Shimomura A, Maruyama N, Tamaki Y, Noguchi S: **GSTP1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in ER-negative breast cancer.** *Cancer Sci* 2012, **103**:913–20.

30. Harris L, Fritsche H, Mennel R, Norton L, Ravdin P, Taube S, Somerfield MR, Hayes DF, Bast RC: **American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer.** *J Clin Oncol* 2007, **25**:5287–5312.

31. Tutt ANJ, Lord CJ, McCabe N, Farmer H, Turner N, Martin NM, Jackson SP, Smith GCM, Ashworth A: **Exploiting the DNA repair defect in BRCA mutant cells in the design of new therapeutic strategies for cancer.** *Cold Spring Harb Symp Quant Biol* 2005, **70**:139–148.

32. Hidalgo M, Siu LL, Nemunaitis J, Rizzo J, Hammond LA, Takimoto C, Eckhardt SG, Tolcher A, Britten CD, Denis L, Ferrante K, Von Hoff DD, Silberman S, Rowinsky EK: **Phase I and pharmacologic study of OSI-774, an epidermal growth factor receptor tyrosine kinase inhibitor, in patients with advanced solid malignancies**. *J Clin Oncol* 2001, **19**:3267–3279.

33. Fan C, Prat A, Parker JS, Liu Y, Carey LA, Troester MA, Perou CM: **Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures.** *BMC Med Genomics* 2011, **4**:3.

34. Creighton CJ, Li X, Landis M, Dixon JM, Neumeister VM, Sjolund A, Rimm DL, Wong H, Rodriguez A, Herschkowitz JI, Fan C, Zhang X, He X, Pavlick A, Gutierrez MC, Renshaw L, Larionov AA, Faratian D, Hilsenbeck SG, Perou CM, Lewis MT, Rosen JM, Chang JC: **Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features.** *Proc Natl Acad Sci U S A* 2009, **106**:13820–13825.

35. Hoadley KA, Weigman VJ, Fan C, Sawyer LR, He X, Troester MA, Sartor CI, Rieger-House T, Bernard PS, Carey LA, Perou CM: **EGFR associated expression profiles vary with breast tumor subtype.** *BMC Genomics* 2007, **8**:258.

36. Liedtke C, Mazouni C, Hess KR, André F, Tordai A, Mejia JA, Symmans WF, Gonzalez-Angulo AM, Hennessy B, Green M, Cristofanilli M, Hortobagyi GN, Pusztai L: **Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer.** *J Clin Oncol* 2008, **26**:1275–1281.

37. Norton L: **Implications of kinetic heterogeneity in clinical oncology.** *Semin Oncol* 1985, **12**:231–49.

38. Huang E, Ishida S, Pittman J, Dressman H, Bild A, Kloos M, D'Amico M, Pestell RG, West M, Nevins JR: **Gene expression phenotypic models that predict the activity of oncogenic pathways.** *Nat Genet* 2003, **34**:226–230.

CHAPTER IV

COMPARISON OF RNA-SEQ BY POLY(A) CAPTURE, RIBOSOMAL RNA DEPLETION, AND DNA MICROARRAY FOR EXPRESSION PROFILING[3]

Background: RNA sequencing (RNA-Seq) is often used for transcriptome profiling as well as the identification of novel transcripts and alternative splicing events. Typically, RNA-Seq libraries are prepared from total RNA using poly(A) enrichment of the mRNA (mRNA-Seq) to remove ribosomal RNA (rRNA), however, this method fails to capture non-poly(A) transcripts or partially degraded mRNAs. Hence, a mRNA-Seq protocol will not be compatible for use with RNAs coming from Formalin-Fixed and Paraffin-Embedded (FFPE) samples.

Results: To address the desire to perform RNA-Seq on FFPE materials, we evaluated two different library preparation protocols that could be compatible for use with small RNA fragments. We obtained paired fresh-frozen (FF) and FFPE RNAs from multiple tumors and subjected these to different gene expression profiling methods. We tested 11 human breast tumor samples using: (a) FF RNAs by microarray, mRNA-Seq, Ribo-Zero-Seq and DSN-Seq (Duplex-Specific Nuclease) and (b) FFPE RNAs by Ribo-Zero-Seq and DSN-Seq. We also performed these different RNA-Seq protocols using 10 TCGA tumors as a validation set.

The data from paired RNA samples showed high concordance in transcript quantification across all protocols and between FF and FFPE RNAs. In both FF and FFPE, Ribo-Zero-Seq removed rRNA with comparable efficiency as mRNA-Seq, and it provided an equivalent or less biased coverage on gene 3’ ends. Compared to mRNA-Seq where 69% of bases were mapped to the

[3] Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM: **Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling**. *BMC Genomics* 2014, **15**:419.

transcriptome, DSN-Seq and Ribo-Zero-Seq contained significantly fewer reads mapping to the transcriptome (20-30%); in these RNA-Seq protocols, many if not most reads mapped to intronic regions. Approximately 14 million reads in mRNA-Seq and 45-65 million reads in Ribo-Zero-Seq or DSN-Seq were required to achieve the same gene detection levels as a standard Agilent DNA microarray.

Conclusions: Our results demonstrate that compared to mRNA-Seq and microarrays, Ribo-Zero-Seq provides equivalent rRNA removal efficiency, coverage uniformity, genome-based mapped reads, and consistently high quality quantification of transcripts. Moreover, Ribo-Zero-Seq and DSN-Seq have consistent transcript quantification using FFPE RNAs, suggesting that RNA-Seq can be used with FFPE-derived RNAs for gene expression profiling.

INTRODUCTION

The development of massively parallel sequencing for use in gene expression profiling is known as RNA-sequencing (RNA-Seq). RNA-Seq has had an enormous impact on gene expression studies. Compared to hybridization-based technologies like DNA microarrays, it provides consistent quantification and manifests its superiority in terms of the dynamic range, sampling depth, and has independence from pre-existing sequence information[1, 2]. RNA-Seq can be used for traditional transcriptome profiling[3, 4] , identification of novel transcripts[5], identification of expressed SNPs[6, 7], alternative splicing, and for the detection of gene fusion events[8–11].

To allow for mRNA/gene detection, highly abundant ribosomal RNAs (rRNAs) must be removed from total RNA before sequencing. One standard solution is to enrich for the polyadenylated (poly(A)) RNA transcripts (so called mRNA-Seq) with oligo (dT) primers, similar to how DNA microarrays are primed; however, this method eliminates all non-poly(A) RNAs in addition to rRNAs. Recent studies suggested that certain non-polyA RNAs, either non-coding or protein coding, are functionally important[12–15]. Moreover, mRNA-Seq poorly captures partially degraded mRNAs, hence it is not an optimal method to use when the starting materials are from Formalin-Fixed and Paraffin-Embedded (FFPE) samples, because the RNAs from FFPE are degraded to a small average size[16]. To overcome these challenges, several rRNA depletion protocols have been developed. The Ribo-Zero method removes rRNA through hybridization capture of rRNA followed by binding to magnetic beads for subtraction. Another method involves Duplex-Specific Nuclease (DSN) degradation by the C_0t-kinetics-based normalization method to deplete abundant sequences that reanneal quickly, such as those derived from the highly abundant rRNAs and tRNAs[17]. In this study, we examined rRNA-depleted libraries from total RNA of fresh-frozen (FF) and FFPE samples sequenced by mRNA-Seq, Ribo-Zero-Seq and DSN-Seq and compared these results across methods and with conventional DNA microarrays.

METHODS

RNA samples. We constructed RNA-Seq libraries using eleven UNC breast tumor samples using different sample preparation protocols including: (a) FF RNA samples by mRNA-Seq, Ribo-Zero-Seq and DSN-Seq and (b) FFPE samples by Ribo-Zero-Seq and DSN-Seq (Figure 4.1b). One of the FF-DSN samples, 3 of the FFPE-Ribo-Zero samples, and 7 of the FFPE-DSN samples failed sequencing QC (i.e. too few reads) and were not included in the study. To augment the UNC sample set, we also tested an additional sample set of FF and FFPE samples collected as part of the TCGA project, where total RNA of ten tumors, including 6 breast tumors and 4 prostate tumors, were prepared in three ways: (a) FF samples with mRNA-Seq, (b) FFPE with Ribo-Zero-Seq and 8 technical replicates, and (c) FFPE with DSN-Seq. In addition, we prepared FF samples for 6 of the 10 TCGA tumors with Ribo-Zero-Seq protocol (Figure 4.1b). All library construction and sequencing were performed at UNC for both the UNC and TCGA samples. For fresh-frozen tissues, we isolated total RNA with Qiagen RNeasy mini kit. For FFPE samples, total RNA was isolated using Roche High Pure RNA paraffin kit, Cat# 03270289001. The extent of RNA degradation was assessed using a BioAnalyzer (Agilent).

Library construction and sequencing. mRNA-Seq library: Illumina TruSeq RNA Sample Prep Kit (Cat# RS-122-2001) was used with 1ug of total RNA for the construction of libraries according to the manufacturer’s protocol. Ribo-Zero library: rRNA was removed from FF or FFPE total RNA using Epicentre's Ribo-Zero rRNA Removal kit (Cat# RZH11042). For FF samples, 30-100ng Ribo-Zero RNA was used for the construction of the library using the Illumina TruSeq RNA Sample Prep Kit and followed the manufacturer’s instruction, except for omitting the purification step before fragmentation. For FFPE samples, 30-100ng Ribo-Zero RNA was then incubated with Random Primers (Invitrogen, Cat# 48190011) at 65^0C for 5 minutes then Illumina TruSeq™ RNA Sample Prep Kit was used to construct the library according to the manufacturer’s protocol from the step of First Strand cDNA Synthesis. DSN library: Illumina TruSeq RNA Sample Prep Kit was used with 100ng of total RNA for the construction of libraries following the manufacturer's protocol, except for omitting the purification of mRNA step in FF samples, and the purification and fragmentation step in FFPE samples. The total RNA libraries went through DSN treatment and PCR enrichment according to Illumina DSN Normalization Sample Preparation Guide (http://supportres.illumina.com/documents/myillumina/7836bd3e-3358-4834-b2f7-80f80acb4e3f/dsn_normalization_sampleprep_application_note_15014673_c.pdf). Sequencing: All cDNA libraries were sequenced using an Illumina HiSeq2000, producing 48x7x48 bp paired-end reads with multiplexing.

Read processing and alignment. All samples were processed and filtered as described in The Cancer Genome Atlas[18]. Bases and QC assessment of sequencing were generated by CASAVA 1.8. QC-passed reads were aligned to the NCBI build 37 (hg19) human reference genome using MapSplice v12_07 [9]. The alignment profile was determined by Picard Tools v1.64 (http://picard.sourceforge.net/).The aligned reads were sorted and indexed using SAMtools, and then translated to transcriptome coordinates and filtered for indels, large inserts, and zero mapping quality using UBU v1.0 (https://github.com/mozack/ubu). For the reference transcriptome, UCSC hg19 GAF2.1 for KnownGenes[19] was used, with genes located on non-standard chromosomes removed. The abundance of transcripts was then estimated using an Expectation-Maximization algorithm implemented in the software package RSEM[20] v1.1.13. Estimated counts were transformed by upper quartile normalization prior to comparison of expression across protocols.

Identification of RNA-Seq library complexity and random sampling. The RNA-Seq data was filtered by requiring the gross RSEM count to be $\geq$3 for each gene. For each protocol, the detected gene sets were defined as genes that were reported in >70% tumor lanes and with 3 or more reads. To determine the amount of input reads needed for sufficient transcriptome coverage, a simulation test was performed on the UNC data. A series of fixed number of reads were randomly selected from each protocol in a drawing without replacement method. For all the resampling levels, the simulated data followed the same alignment and filtering pipeline as described above. Gene sets detected were then identified for all the various levels.

Gene expression comparison methods

For all the FF tumors and the Common Reference Sample, Agilent 244,000 feature whole genome microarrays were hybridized with tumor RNAs (Cy5) and a human common reference (Cy3) and lowess normalized as described in Herschkowitz et al.[21]. In the RNA-Seq data, the detected gene sets were identified as above (i.e. 3 or more reads in >70% of samples). The log2 ratio of RNA-Seq tumor samples to RNA-Seq human Common Reference Sample (which was the same RNA used for the 2-color microarrays) was determined. Pearson correlation was determined and a Student's t-test was applied to evaluate the difference of RNA-Seq protocols in their consistency to microarray.

The RNA-Seq gene quantification data was next filtered by gene counts as above. The log2 transformed abundance of tumor samples was reported and was used to derive the correlation between RNA-Seq protocol pairs. Using R package MethComp, Deming regression was applied to compare the sensitivity in detecting differentially expressed genes. An unpaired two-class SAM analysis was used to identify genes that have differential expression level in a) mRNA-Seq versus Ribo-Zero-Seq, and b) Ribo-Zero-Seq versus DSN-Seq.

Gene expression quantification by microarray and RNA-Seq for all samples new to this manuscript can be found in GEO database under accession GSE51783. Aligned BAM files are available at dbGaP under the series ID of phs000676.v1.p1. TCGA sample RNA-Seq data is available at cgHub (BAM files, https://cghub.ucsc.edu/) and DCC (expression level data, https://tcga-data.nci.nih.gov/tcga/).

RESULTS

To rigorously evaluate the feasibility of reproducible gene expression profiling using RNA from clinically relevant FFPE materials, we collected FFPE and fresh-frozen (FF) tumor RNAs for matched sets of tumors from two different sources (UNC and TCGA). Most tumors were subjected to gene expression profiling using six different methods that included: 1) Agilent DNA microarrays using FF RNA, 2) mRNA-Seq using FF RNA, 3) Ribo-Zero-Seq using FF RNA, 4) DSN-Seq using FF RNA, 5) Ribo-Zero-Seq using FFPE RNA, and 6) DSN-Seq using FFPE RNA; see Figure 4.1 for a comparison of each RNA-Seq protocol and the number of samples tested for each protocol. Analytical comparisons were focused on several features including rRNA depletion efficiency, genome alignment profile, transcriptome coverage, transcript quantification accuracy and reproducibility, gene expression patterns and differential gene expression, as well as coverage of annotated genes at different sequencing depths.

rRNA depletion efficiency

The efficiency of rRNA removal is a key factor to maximize reads mapping to transcripts, because if left alone, rRNAs make up >80-90% of the total RNA of an un-enriched sample[22]. Due the nature of rRNA sequences, many rRNA short reads will produce poor alignments; hence, the estimation of absolute abundance of rRNA based on whole genome alignment tends to underestimate rRNA amounts. Thus we evaluated the relative level of rRNA components across protocols by comparing the levels to those observed in mRNA-Seq. Ribo-Zero-Seq reduced rRNA levels to a similar order of magnitude as mRNA-Seq in both FF and FFPE RNA, while the rRNA fraction in DSN-Seq libraries were significantly higher ($p<0.001$) and with greater variation, particularly within the FFPE samples (Table 4.1). Consistent with the analysis of the UNC dataset, Ribo-Zero-Seq provided the same rRNA removal efficiency as mRNA-Seq in the TCGA samples; the level of rRNA reduction observed here for the Ribo-Zero-Seq protocol was similar to that reported by the company that makes the Ribo-Zero kit (data not shown).

Genome alignment profile

The precision of RNA-Seq gene quantification is directly dependent on the number of reads that are mapped to transcripts, thus we first assessed the fraction of reads aligning to the reference human genome UCSC hg19 (Table 4.1). In FF samples, mRNA-Seq and Ribo-Zero-Seq provided comparable percentage of nucleotide bases mapping to the genome (94.0%, 93.8%), while DSN-Seq aligned a smaller number (85.5%). In FFPE samples, Ribo-Zero-Seq and DSN-Seq both had good performance in alignment on average (81.5% in Ribo-Zero-Seq-FFPE, 93.5% in DSN-Seq-FFPE); TCGA samples had a similar result for both FF and FFPE (Table 4.1). Compared to FF, the FFPE samples tended to exhibit a greater variation in the % aligned, most likely related to more variable quality of FFPE RNAs.

Transcriptome coverage

The coverage of the transcriptome directly affects the accuracy of transcript abundance estimation and the sensitivity of transcript detection, which are two critical features of all gene expression studies. Therefore, we evaluated two features of the transcriptome coverage: (a) relative coverage of exons, introns, and intergenic regions, and (b) uniformity of transcript coverage.

(a) Relative coverage of exons, introns, and intergenic regions. In FF samples, bases mapping to transcripts (i.e. coding and UTR regions) constituted 62.3% total bases in mRNA-Seq, while a marked reduction was observed in the two rRNA-depletion protocols (31.5% in Ribo-Zero-Seq and 22.7% in DSN-Seq, Figure 4.2a). Conversely, bases mapping to intronic and intergenic regions increased from 31.6% in mRNA-Seq to 62.5% in DSN-Seq and Ribo-Zero-Seq. In FFPE samples, DSN-Seq and Ribo-Zero-Seq provided similar coverage profiles, where ~20% of bases were mapped to transcriptome and >60% to intronic or intergenic regions. These results were concordant with that observed in the TCGA sample set (Figure 4.2b).

We further investigated the coverage across individual genes (Figure 4.3a, GATA3 as an example). In mRNA-Seq, most reads mapped almost exclusively to exons, and the coverage of intronic regions was low and comparable to the intergenic background. In contrast, in Ribo-Zero-Seq and DSN-Seq there was a more continuous coverage of both exons and introns, although the coverage of intergenic regions was more similar to what was seen with mRNA-Seq. This unique profile suggests that the rRNA depletion protocol may capture pre-mRNAs in addition to mature mRNAs. To test this hypothesis, we examined the pile-up profile of a few individual genes and identified reads that spanned exon-intron boundaries in the Ribo-Zero-Seq and DSN-Seq protocols (Figure 4.3b, see red arrows for spanning reads).

(b) Uniformity of transcript coverage. We next determined the evenness of transcript coverage by comparing the median coefficient of variation (CV) for the read coverage of the 1000 most highly expressed transcripts (Table 4.1). In FF libraries, mRNA-Seq and Ribo-Zero-Seq had significantly lower CV than DSN-Seq (mRNA-Seq: p<0.001, Ribo-Zero-Seq: p=0.002), indicating a more uniform coverage across the full length of transcripts. In the FFPE libraries, there was an increase in CV in both protocols. Ribo-Zero-Seq-FFPE had slightly higher variation than the result reported in Adiconis et al.[23], while DSN-Seq-FFPE had the highest CV among all protocols.

Another measure of transcript coverage is the variation at 5' and 3' ends. We evaluated the ratio of coverage at the 5' end relative to the 3' end for the 1000 most highly expressed transcripts (Table 4.1). Previous studies have shown that the poly(A)-capture strategy shows substantially more reads from the 3' ends of transcripts. Our analysis revealed that on FF, Ribo-Zero-Seq provided less biased 5'-to-3' coverage ratio than mRNA-Seq (p<0.001), while DSN-Seq made no significant improvement. In FFPE samples, both protocols performed similar as mRNA-Seq with respect to 5'-to-3' bias.

Transcript quantification and reproducibility

RNA-Seq poly(A) enrichment strategies yield an accurate and reproducible measurement of transcript abundance with a wide dynamic range[1, 4, 24, 25]. Given the advantages of profiling multiple types of RNA species (i.e. mRNAs, lincRNAs, snoRNAs, etc.), it is critical to evaluate the performance of mRNA quantification in total RNA-Seq protocols. To determine the possible concordance of RNA-Seq with data generated by older genomic profiling platforms, we compared the gene expression levels of RNA-Seq data with that of Agilent DNA microarray data that were assayed using the same RNAs. With specific and standard gene filtering criteria[26], we detected 16,975 expressed Entrez genes by custom Agilent 244,000 feature microarrays, with 15,206 genes detected by both microarray and RNA-Seq across our paired samples. In FF samples, gene abundance measurements by all protocols of RNA-Seq were highly correlated with the microarray data (Pearson>0.8, Table 4.1). In FFPE samples, RNA-Seq measurements were lower but also significantly correlated with FF microarray (Pearson ~0.7, Table 4.1), which is at a level similar to that observed when comparing concordance between Agilent and Affymetrix microarrays[27].

We next examined the correlation of transcript abundance across the different RNA-Seq protocols. There was greater concordance and fewer outliers than when compared to the microarray data (Figure 4.4a and b). Among FF tissues, the correlation was >0.9 for all pair-wise, sample-matched comparisons. DSN-Seq and Ribo-Zero-Seq on FFPE were less correlated with FF mRNA-Seq (>0.8), but still higher than the correlation observed with microarrays. The two rRNA depletion protocols were the most highly correlated in both FF and FFPE samples (Pearson correlation 0.961 in FF and 0.934 in FFPE). The correlation plots for an individual sample (breast tumor 020678B) are shown in Figure 4.4c.

Additional quality assessments were made on the TCGA dataset, to account for the fact that a much smaller set of reads were mapped to transcriptome in RNA depletion protocols. We generated eight technical replicates with the Ribo-Zero-Seq-FFPE protocol to balance the total number of transcriptome reads for the comparison with FF mRNA-Seq. The assessment of technical reproducibility suggested that these FFPE replicates were indistinguishable (Pearson =0.991). The correlation between Ribo-Zero-Seq on FF and FFPE as well as between Ribo-Zero-Seq-FFPE replicate pairs has also been confirmed in Norton et al.[28].

Lastly, we applied Deming regression to estimate a statistically unbiased slope to determine the relative sensitivity of protocol pairs (Figure 4.4d). A slope of 1 indicates the equivalent sensitivity of the two libraries, whereas a smaller value is indicative of a higher sensitivity of the first protocol in the pair. mRNA-Seq exhibited its superiority over all the other protocols in terms of sensitivity, with a slope less than 1 in all the pair-wise comparison. In addition, DSN-Seq and Ribo-Zero-Seq both have higher sensitivity in FF samples than in FFPE.

Gene expression patterns and differential gene expression

Hierarchical clustering analysis provides a global examination whether biologically relevant expression signatures are consistently measured by distinct protocols. In this example, we tested whether the same sample assayed by different protocols “paired” or “partnered” together; if so, then this is a very high level of assay validation as not only are the overall subtype expression profiles maintained, but also the profiles that are unique to that sample are maintained. We performed hierarchical clustering analysis of the RNA-Seq data using a previously published ‘intrinsic gene list’[29] (Figure 4.5) and a set of 904 human breast tumor samples that consists of the 88 UNC and TCGA samples described here and 725 additional breast tumors and 91 normal breast tissues with mRNA-Seq from TCGA. 41/44 samples of the UNC tumor dataset were tightly co-clustered with their partner sample originating from the same tumor, and these clustered with other TCGA tumors based upon each tumor’s subtype profile. The 3/44 non-clustered samples were all prepared by Ribo-Zero-Seq on FFPE samples and their partner DSN-Seq samples on FFPE were not available. In the TCGA dataset, 40/44 samples were tightly co-clustered with their partners (i.e. libraries constructed from the same tumor using a different sequencing protocol); the four samples that were not clustered were on a separate branch, but were moderately correlated with their partner samples (correlation>0.6).

As another test of data quality, we determined the differentially expressed gene set in FF mRNA-Seq vs. Ribo-Zero-Seq and FF Ribo-Zero-Seq vs. DSN-Seq using Significance Analysis of Microarray (SAM). We identified 410 genes with a FDR of 0 that were differentially expressed between mRNA-Seq and Ribo-Zero-Seq (Appendicies 3a and b); this list was enriched with snoRNAs and histone RNAs that were more highly expressed in the Ribo-Zero-Seq samples. Many of these RNAs do not possess poly(A) tails, and therefore, are not targeted by poly(A) selection in mRNA-Seq. Conversely, 104 genes at a FDR of 0 were identified to be differentially expressed between Ribo-Zero-Seq and DSN-Seq libraries (Appendices 3c and d); among these, 38 genes were lowly quantified by DSN-Seq and most of these genes were snoRNAs and histone RNAs, which tend to exist at high abundance in total RNAs. Since DSN-Seq removes the most highly abundant components via CoT kinetics, these RNAs may also be subject to depletion in the DSN protocol relative to the Ribo-Zero, which uses beads to capture only the rRNAs.

Coverage of annotated genes at different sequencing depths

Compared to hybridization-based methods, the cost per sample by RNA-Seq is still higher. The utilization of multiplexing techniques provides a strategy to further lower the costs. However, too much multiplexing will inhibit the ability to detect lowly expressed genes; therefore, we   sought to determine the minimal number of reads required to provide the same transcriptome coverage as provided by an Agilent DNA microarray. The ENCODE Consortium guidelines and other studies have provided insights into the sufficient RNA-Seq coverage and depth for studies of various design goals[30], but these efforts were primarily focused on experiments with FF samples prepared by poly(A)-enrichment protocols. Here we extended the investigation to rRNA depletion approaches and FFPE samples.

We applied a simulation-based method on the pooled data of each protocol. The UCSC known gene reference database (GAF 2.1) includes 20,531 (non-ribosomal) genes. To reduce the noise, we only counted genes as present if there were 3 or greater read counts. Using the average number of genes detected on our Agilent microarrays as the baseline (n=16,975), 13.5 million reads from FF mRNA-Seq libraries would allow detection of the same number of genes (Figure 4.6), which is consistent with previous studies[30]. In the DSN-Seq and Ribo-Zero-Seq FF libraries, and Ribo-Zero-Seq-FFPE libraries, 35-65M reads were required to provide the same transcriptome coverage. Only the DSN-Seq-FFPE library required a much larger number of input reads (90M).

DISCUSSION

The growing popularity of RNA-Seq makes it one of the more desired methods to explore the transcriptome. Preparing RNA-Seq libraries with poly(A) enrichment provides an accurate method to characterize mRNAs, which is functionally equivalent to what DNA microarrays have been accomplishing for more than a decade. However, certain biologically relevant RNA species that do not possess poly(A) tails are largely undetected using a poly(A) selection protocol. In addition, FFPE samples, such as those collected as part of standard medical practice, also require library preparation methods that do not rely on the intact poly(A) structure due to the highly degraded nature of the FFPE RNA. In this study, we demonstrate that a Ribo-Zero-Seq protocol using either fresh-frozen (FF) or FFPE samples eliminates rRNA with good efficiency. In evaluation of a possible coverage bias, 5’-to- 3’ bias was reduced in FF Ribo-Zero-Seq libraries as it does not rely on poly(A) selection step.

One major distinction across these various protocols is the coverage of the transcriptome. To more directly investigate the relationship between sequencing depth and transcriptome coverage, we performed a simulation approach where mRNA-Seq was the most cost effective strategy to equal a microarray in terms of total genes detected with a minimum of ~13.5 million reads needed. For the same transcriptome coverage, the reads required for Ribo-Zero-Seq in FF and FFPE and DSN-Seq in FF were 35-65M reads. However, rRNA depletion protocols also appear to measure immature transcripts (pre-mRNAs) and therefore provide more information on splicing patterns and possible splice junctions. Thus to achieve the same level of exonic reads as FF mRNA-Seq, one needs to sequence 2-4 times the number of reads in rRNA-depletion on FFPE RNA libraries.

Despite fewer of the total reads mapping to exonic regions and a greater number of transcripts being detected, we did not observe a marked decrease in the correlation between microarray and RNA-Seq in rRNA-depleted libraries, where RNA-zero-Seq and DSN-Seq were found to be highly consistent in gene quantification. Our evaluation of the quantitative consistency of RNA-Seq on FFPE with microarray may be limited in two aspects: (a) the quality of a few UNC FFPE samples was less satisfactory, and (b) not all the tumors have RNA-Seq data on matched FFPE samples that passed our quality control available for this analysis. Yet we still observed very good correlations with microarray data for those samples with complete FFPE data, which gave correlation values nearly identical to those seen when comparing an Agilent microarray versus an Affymetrix microarray[27].

Given the consistent quantification, mRNA-Seq and rRNA depletion protocols exhibited their merits in different aspects. In the set of genes detected by all the protocols, mRNA-Seq provided the highest sensitivity in detecting differentially expressed genes, which was likely due to the greater fraction of reads mapping to the transcriptome. On the other hand, Ribo-Zero-Seq detected about 550 more annotated genes than mRNA-Seq (data not shown). With a much greater set of reads mapping to the intergenic and intronic regions in rRNA depletion protocols, the number of additional transcripts detected with the new protocols may be expected to be greater than our conservative estimation here. As shown in another recent studies[30], we also expect more novel transcripts to be identified from the rRNA depletion methods.

The very good quantification performance of the protocols on FFPE samples is of significant impact for researchers with clinical samples. Our results demonstrate that Ribo-Zero-Seq had high technical reproducibility on FFPE RNAs and high concordance with FF RNAs. Though the quantification of FFPE was less correlated to FF mRNA-Seq, the two rRNA depletion methods provided highly consistent gene profiles on FFPE. Thus, it is the quality of FFPE RNA samples, rather than the robustness of method, that likely contributes more to the variation of performance with respect to gene quantification. The hierarchical clustering analysis also validated that the biologically-based intrinsic gene profiles were present and highly correlated between FF and FFPE. Hence, we suggest that it is possible to apply the rRNA depletion protocols to FFPE samples and achieve quantitative accuracies comparable with standard genome profiling techniques that use FF tissues and RNAs.

CONCLUSIONS

In this study, we demonstrated that compared to mRNA-Seq, Ribo-Zero-Seq provides equivalent rRNA removal efficiency, coverage uniformity, genome-based mapped reads, and reduces 5’- to- 3’ bias. In addition, both Ribo-Zero-Seq and DSN-Seq provide highly consistent quantification of transcripts when compared to microarrays or mRNA-Seq, and substantially more information on non-poly(A) RNA. Moreover, the two rRNA depletion methods have consistent transcript quantification using FFPE RNAs and show high reproducibility.

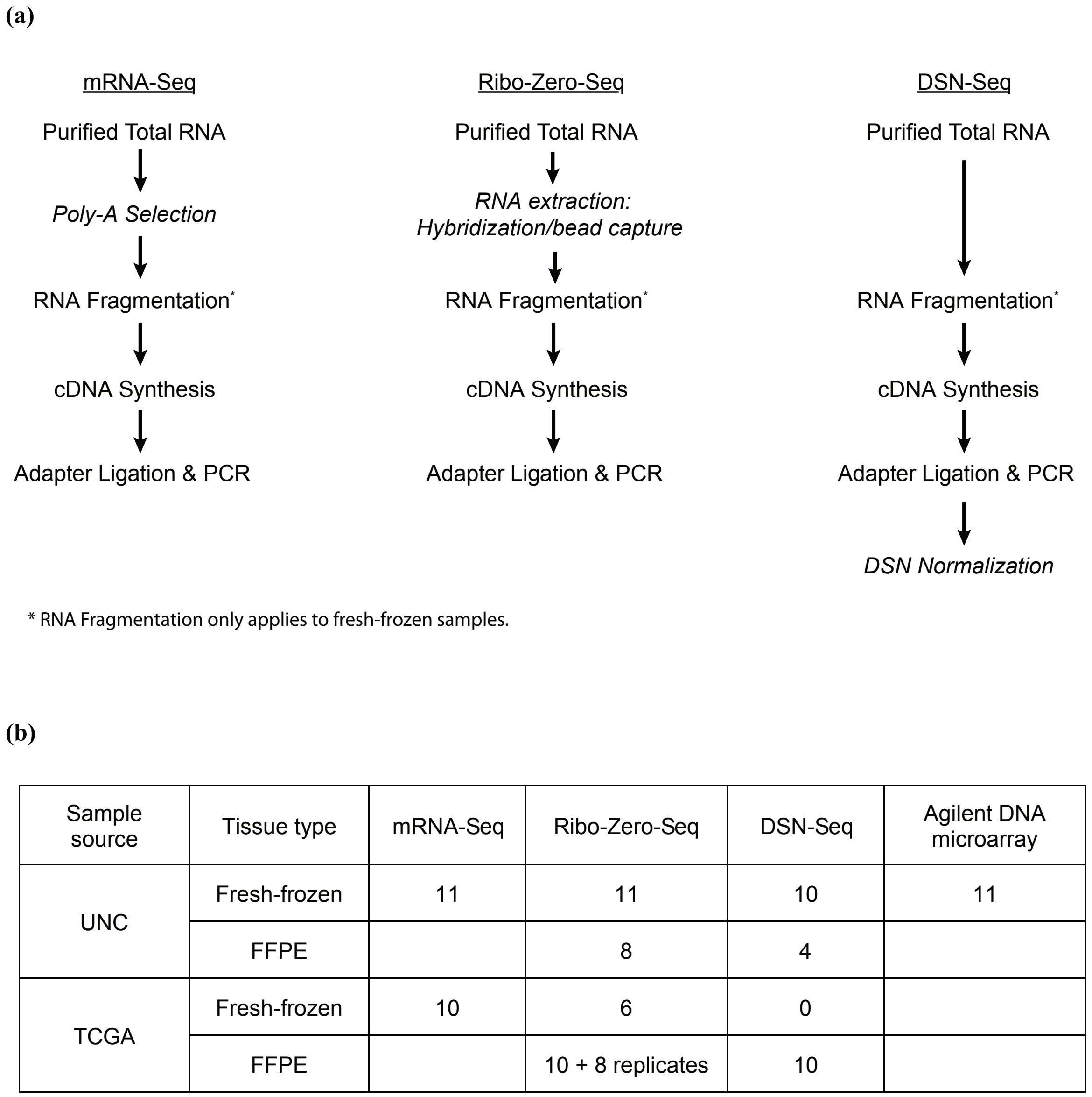
TABLES

Table 4.1 Analysis of performance for multiple RNA-Seq methods.

Five different analyses were performed in order to assess the capabilities of the different RNA-seq protocols. These included: 1) % rRNA relative to mRNA-Seq; 2) % Aligned bases; 3) Median CV coverage; 4) Median 5' to 3' bias; 5) The Pearson correlation coefficient between the RNA-Seq libraries methods and the same samples assayed by DNA microarray in UNC dataset.

	mRNA-Seq	**RiboZero-Seq**	**DSN-Seq**	**RiboZero-FFPE**	**DSN-FFPE**
UNC dataset					
Sample size	11	11	10	8	4
% rRNA relative to mRNA-seq	1 (1-1)	5.04 (1.42-8.66)	116 (78.9-154)	7.14 (3.48-10.8)	585 (-347-1,517)
% Aligned bases	94 (91.5-96.5)	93.8 (92-95.5)	85.5 (82.6-88.4)	81.5 (71-92)	93.5 (92.2-94.8)
Median CV coverage	0.533 (0.506-0.56)	0.525 (0.505-0.545)	0.56 (0.549-0.57)	0.744 (0.713-0.775)	0.929 (0.814-1.04)
Median 5' to 3' bias	0.27 (0.189-0.35)	0.64 (0.493-0.788)	0.209 (0.143-0.275)	0.356 (0.285-0.427)	0.242 (0.0329-0.451)
Pearson correlation to microarray	0.851 (0.825-0.878)	0.832 (0.809-0.854)	0.855 (0.84-0.871)	0.636 (0.601-0.671)	0.7 (0.628-0.771)
TCGA dataset					
Sample size	10	6	NA	18	10
% rRNA relative to mRNA-seq	1 (1-1)	11.2 (1.51-20.9)	NA	0.935 (0.631-1.24)	41.7 (22.1-61.3)
% Aligned bases	96.4 (95.4-97.5)	95.0 (93.9-96.2)	NA	93.4 (91.6-95.2)	93.2 (90.7-95.8)
Median CV coverage	0.534 (0.517-0.551)	0.478 (0.458-0.499)	NA	0.83 (0.791-0.869)	0.953 (0.896-1.01)
Median 5' to 3' bias	0.309 (0.244-0.374)	0.46 (0.37-0.551)	NA	0.417 (0.253-0.581)	0.157 (0.0856-0.229)

FIGURES



* RNA Fragmentation only applies to fresh-frozen samples.

Sample source	Tissue type	mRNA-Seq	Ribo-Zero-Seq	DSN-Seq	Agilent DNA microarray
UNC	Fresh-frozen	11	11	10	11
	FFPE		8	4	
TCGA	Fresh-frozen	10	6	0	
	FFPE		10 + 8 replicates	10	

Figure 4.1 Schematic overview of the rRNA removal protocols and list of samples tested. (A) mRNA-Seq, Ribo-Zero-Seq and DSN-Seq library preparation protocols are shown, with the key steps to remove the rRNA from the library show in italics. The full protocol was applied to the fresh-frozen (FF) samples, and a similar alternative protocol was applied to FFPE samples (omitting steps marked as *). (B) The list of samples tested by each RNA-Seq library protocol and their source.

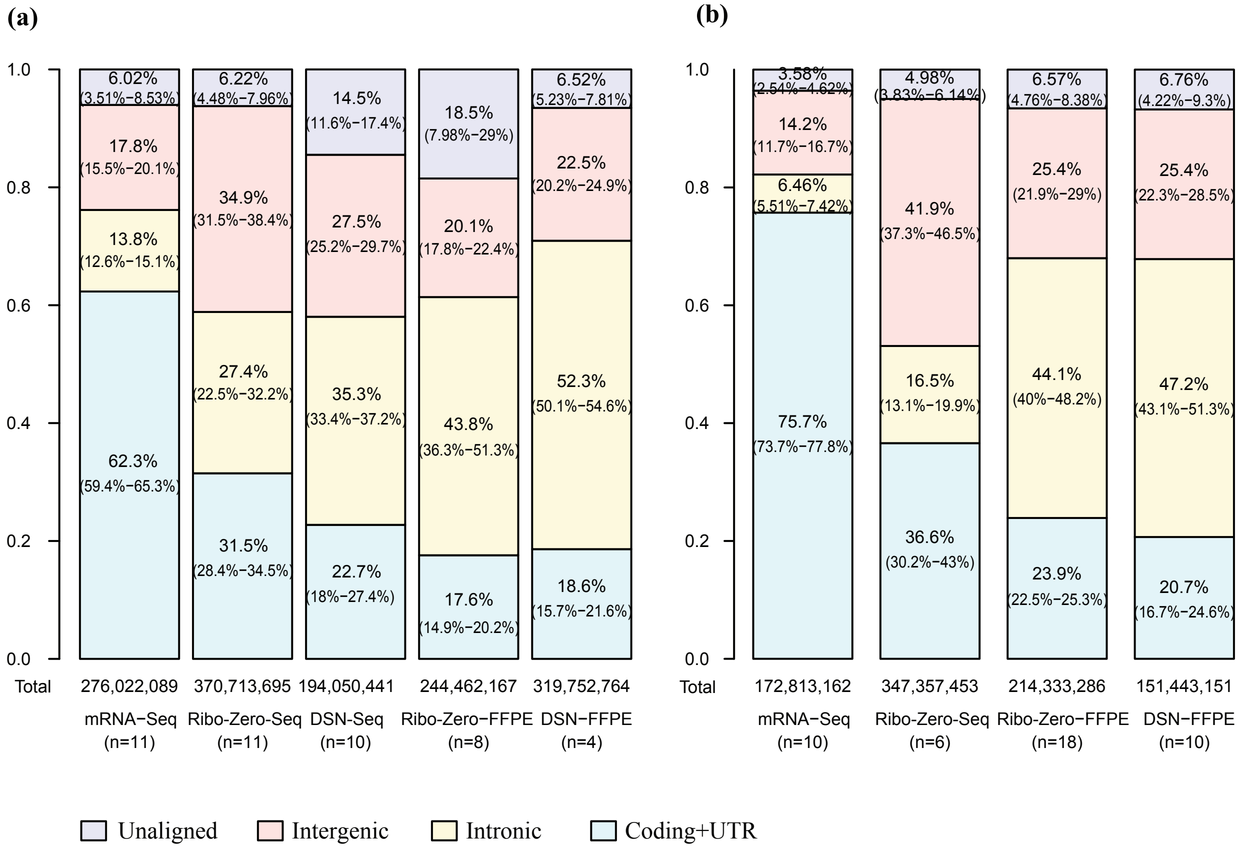


Figure 4.2 Genome alignment profile. The percentage of nucleotide bases mapping to three different regions of the genome: exonic/protein coding and UTR (green), intronic (yellow), intergenic (red), and the percentage of unmapped bases (purple). The data is shown separately for the UNC (a) and TCGA (b) datasets.

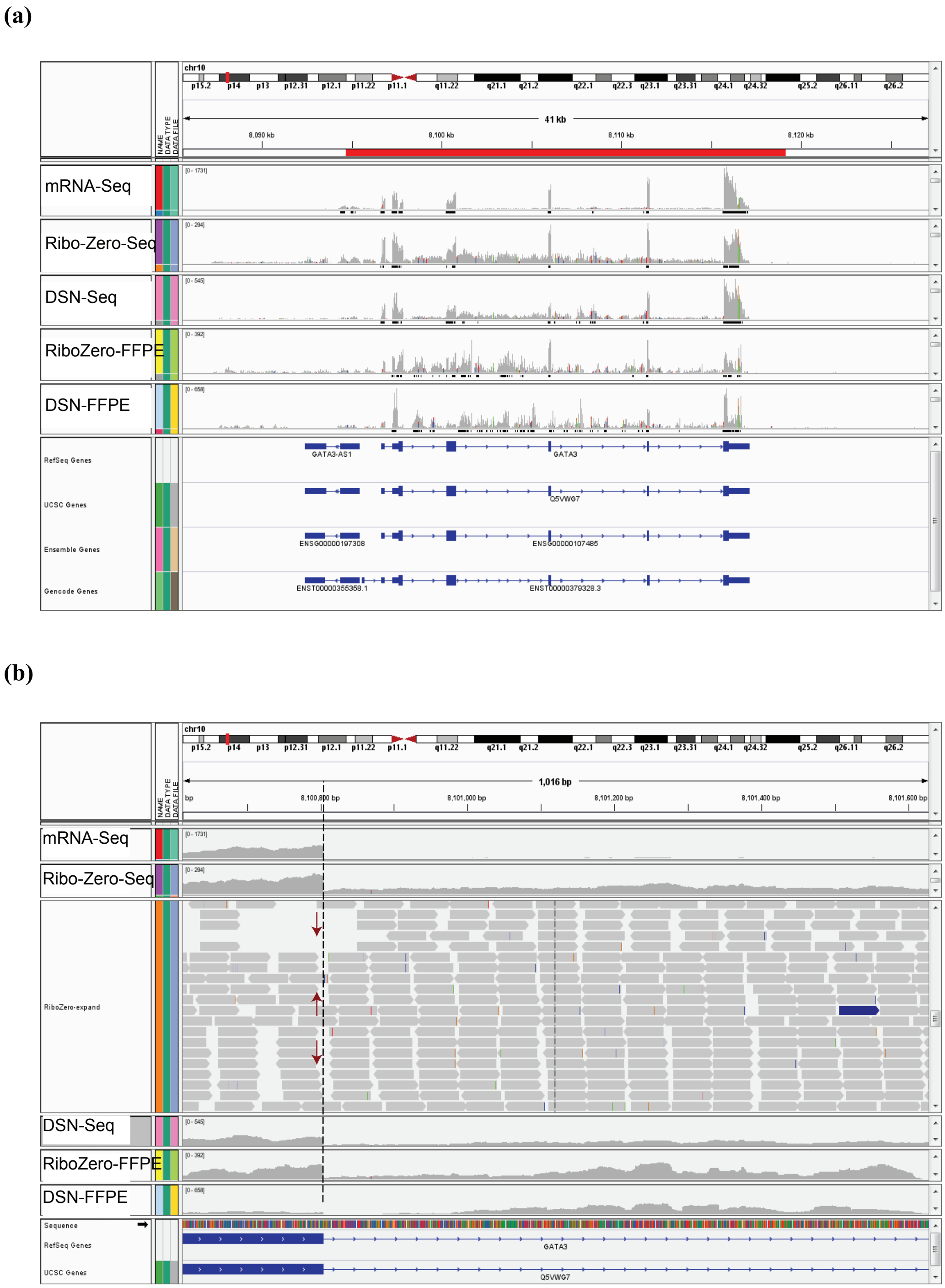


Figure 4.3 Visual display of reads aligning to GATA3. (A) Read pile-up plots of GATA3 in Sample 020578B showing data for five different RNA-Seq libraries. (B) Close-up of the read mapping identifying reads that span exon-intron boundaries, which identify unspliced mRNA species.

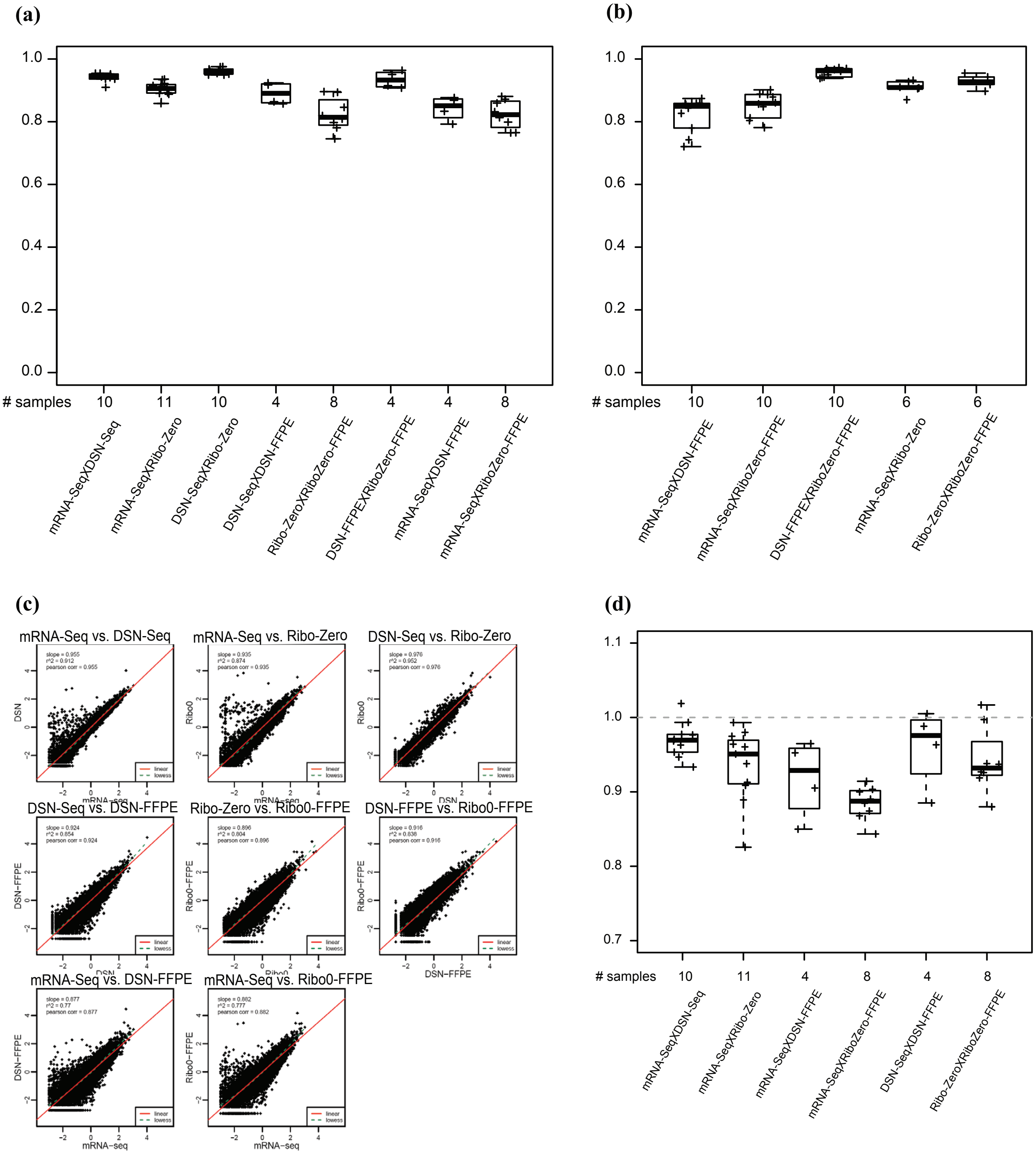


Figure 4.4 Comparison of gene quantification concordance across RNA-Seq library protocols. Pearson correlation coefficients of RNA-Seq libraries pairs in (A) UNC and (B) TCGA dataset. (C) Scatter plots of libraries of each pair of protocols for breast tumor sample 020578B. (D) Deming regression slope for pairs of RNA-Seq libraries in UNC dataset. A slope of 1 indicates the equivalent sensitivity of the two libraries, whereas a smaller value is indicative of a higher sensitivity of the first term/method in the pair.

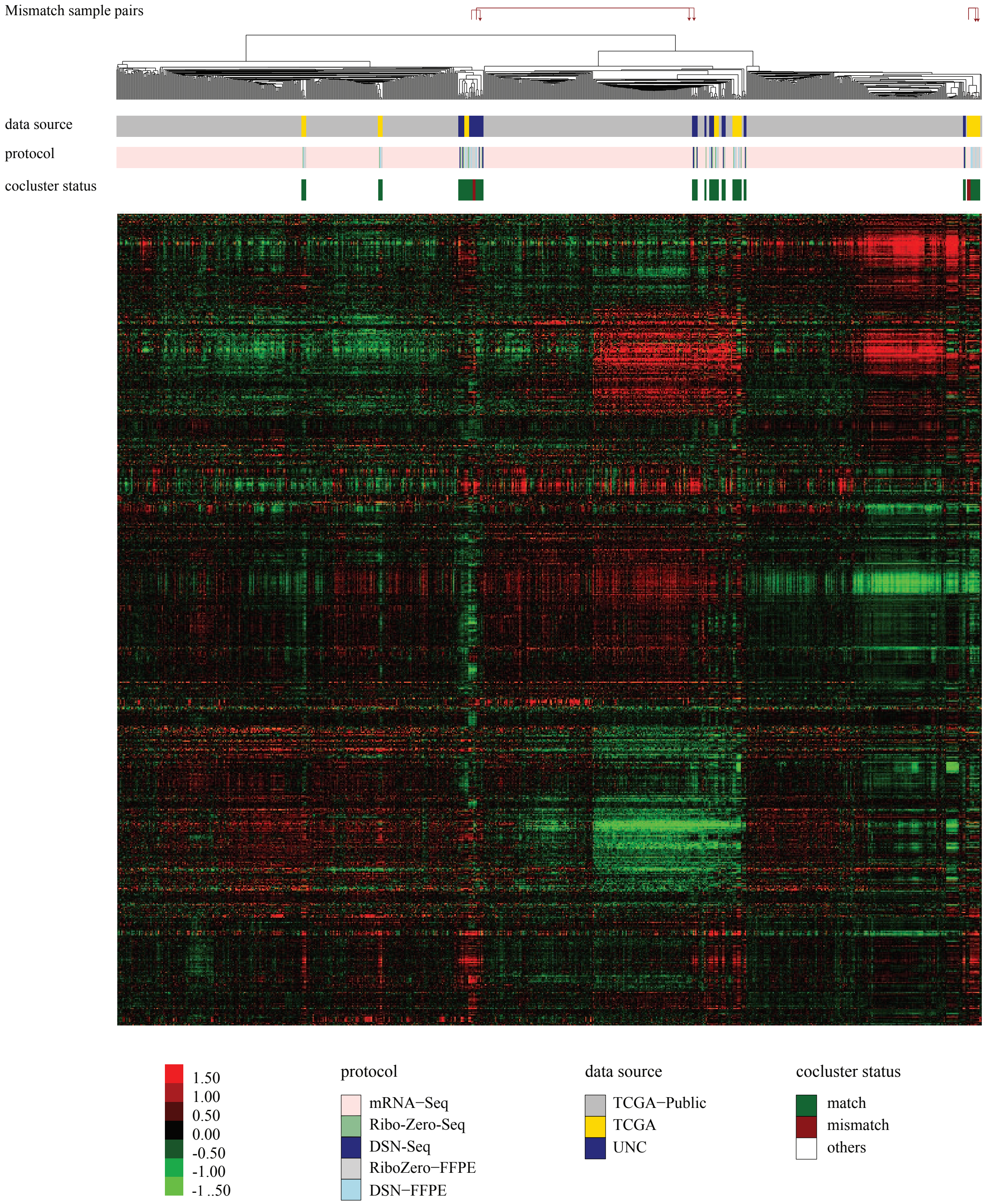


Figure 4.5 Intrinsic gene set clustering analysis. Hierarchical cluster using a breast cancer intrinsic gene set (~2000 genes) and 88 breast tumor samples prepared using the multiple protocols, with an additional 816 samples from the TCGA Breast Cancer Project (725 tumors and 91 normal tissues). The rows above the heat map identify the 88 samples from this study, their RNA-Seq protocol type, and the red arrows show the location of the few mismatched samples

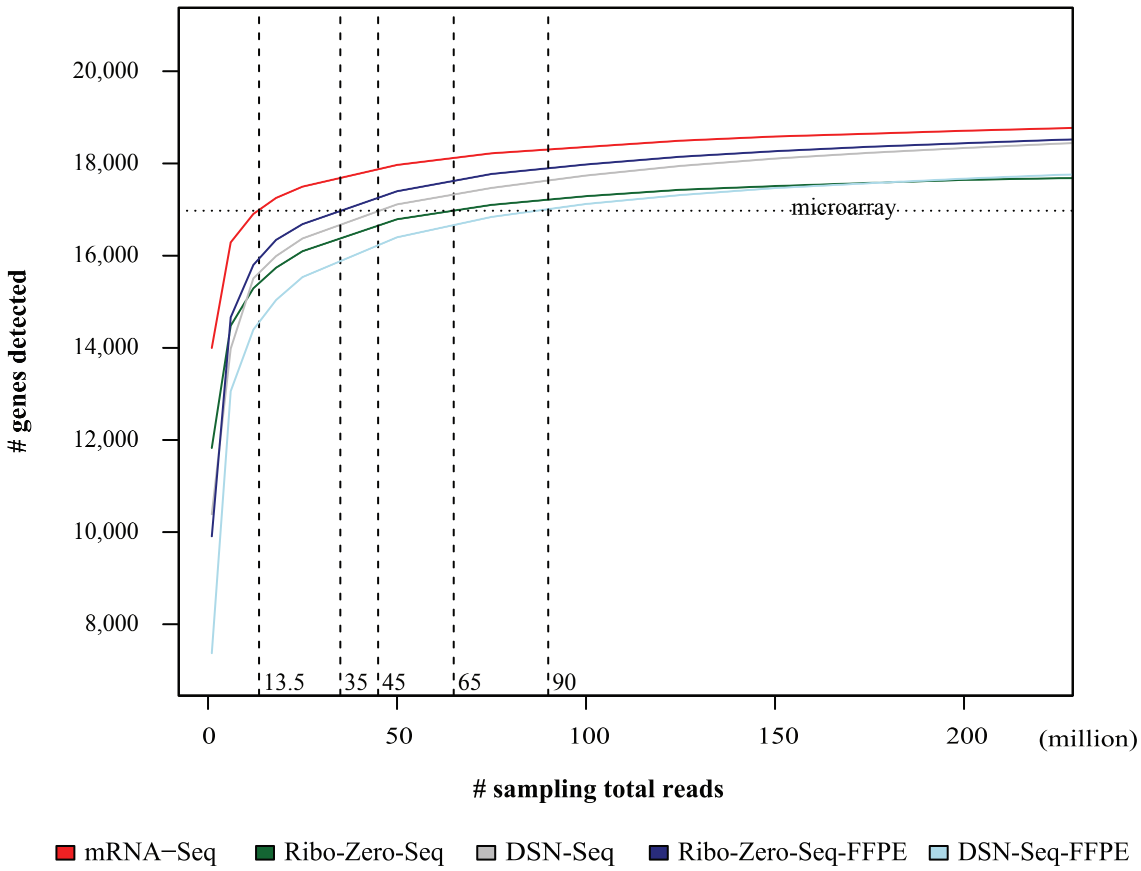


Figure 4.6 Determination of the number of reads needed for each RNA-Seq protocol to equal DNA microarray. The number of detected genes at different levels of sequencing depth is displayed relative to the number of genes detected via DNA microarray (dashed horizontal line).

REFERENCES

1. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res* 2008, **18**:1509–1517.

2. Guo Y, Sheng Q, Li J, Ye F, Samuels DC, Shyr Y: **Large Scale Comparison of Gene Expression Levels by Microarrays and RNAseq Using TCGA Data.** *PLoS One* 2013, **8**:e71462.

3. The Cancer Genome Atlas Network: **Integrated genomic characterization of endometrial carcinoma**. *Nature* 2013, **497**:67–73.

4. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**:621–628.

5. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, et al.: **Landscape of transcription in human cells**. *Nature* 2012, **489**:101–108.

6. Quinn EM, Cormican P, Kenny EM, Hill M, Anney R, Gill M, Corvin AP, Morris DW: **Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data.** *PLoS One* 2013, **8**:e58815.

7. Piskol R, Ramaswami G, Li JB: **Reliable Identification of Genomic Variants from RNA-Seq Data.** *Am J Hum Genet* 2013, **93**:641–51.

8. Chao H-H, He X, Parker JS, Zhao W, Perou CM: **Micro-scale genomic DNA copy number aberrations as another means of mutagenesis in breast cancer.** *PLoS One* 2012, **7**:e51719.

9. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J: **MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.** *Nucleic Acids Res* 2010, **38**:e178.

10. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, Keeffe SO, Haas S, Vingron M, Lehrach H, Yaspo M: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.** *Science* 2008, **321**:956–60.

11. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.** *Nat Genet* 2008, **40**:1413–1415.

12. Esteller M: **Non-coding RNAs in human disease**. *Nat Rev Genet* 2011, **12**:861–874.

13. Fatica A, Bozzoni I: **Long non-coding RNAs: new players in cell differentiation and development.** *Nat Rev Genet* 2013, **15**:7–21.

14. Du Z, Fei T, Verhaak RGW, Su Z, Zhang Y, Brown M, Chen Y, Liu XS: **Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer.** *Nat Struct Mol Biol* 2013, **20**:908–13.

15. Akrami R, Jacobsen A, Hoell J, Schultz N, Sander C, Larsson E: **Comprehensive Analysis of Long Non-Coding RNAs in Ovarian Cancer Reveals Global Patterns and Targeted DNA Amplification.** *PLoS One* 2013, **8**:e80306.

16. Mullins M, Perreard L, Quackenbush JF, Gauthier N, Bayer S, Ellis M, Parker J, Perou CM, Szabo A, Bernard PS: **Agreement in breast cancer classification between microarray and quantitative reverse transcription PCR from fresh-frozen and formalin-fixed, paraffin-embedded tissues.** *Clin Chem* 2007, **53**:1273–1279.

17. Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, Matz M V, Meleshkevitch E, Moroz LL, Lukyanov SA, Shagin DA: **Simple cDNA normalization using kamchatka crab duplex-specific nuclease.** *Nucleic Acids Res* 2004, **32**:e37.

18. The Cancer Genome Atlas Network: **Comprehensive genomic characterization of squamous cell lung cancers.** *Nature* 2012, **489**:519–25.

19. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hinrichs AS, Learned K, Lee BT, Li CH, Raney BJ, Rhead B, Rosenbloom KR, Sloan CA, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ: **The UCSC Genome Browser database: 2014 update.** *Nucleic Acids Res* 2014, **42**(Database issue):D764–70.

20. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics* 2011, **12**:323.

21. Herschkowitz JI, Zhao W, Zhang M, Usary J, Murrow G, Edwards D, Knezevic J, Greene SB, Darr D, Troester MA, Hilsenbeck SG, Medina D, Perou CM, Rosen JM: **Comparative oncogenomics identifies breast tumors enriched in functional tumor-initiating cells**. *Proc Natl Acad Sci* 2012, **109**:2778–2783.

22. O’Neil D, Glowatz H, Schlumpberger M: **Ribosomal RNA depletion for efficient use of RNA-seq capacity.** *Curr Protoc Mol Biol* 2013, **Chapter 4**:Unit 4.19.

23. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby M a, Berlin AM, Sivachenko A, Thompson DA, Wysoker A, Fennell T, Gnirke A, Pochet N, Regev A, Levin JZ: **Comparative analysis of RNA sequencing methods for degraded or low-input samples.** *Nat Methods* 2013, **10**:623–9.

24. ’t Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHAM, de Menezes RX, Boer JM, van Ommen G-JB, den Dunnen JT: **Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms.** *Nucleic Acids Res* 2008, **36**:e141.

25. Oshlack A, Robinson MD, Young MD: **From RNA-seq reads to differential expression results.** *Genome Biol* 2010, **11**:220.

26. The Cancer Genome Atlas Network: **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490**:61–70.

27. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, et al.: **Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.** *Cancer Cell* 2010, **17**:98–110.

28. Norton N, Sun Z, Asmann YW, Serie DJ, Necela BM, Bhagwate A, Jen J, Eckloff BW, Kalari KR, Thompson KJ, Carr JM, Kachergus JM, Geiger XJ, Perez EA, Thompson EA: **Gene expression, single nucleotide variant and fusion transcript discovery in archival material from breast tumors.** *PLoS One* 2013, **8**:e81925.

29. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS: **Supervised risk predictor of breast cancer based on intrinsic subtypes.** *J Clin Oncol* 2009, **27**:1160–1167.

30. Wang Y, Ghaffari N, Johnson CD, Braga-neto UM, Wang H, Chen R: **Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens**. *BMC Bioinformatics* 2011, **12**(Suppl 10):S5.

CHAPTER V

CONCLUSION

Studies in the past few decades have shown that cancer is a heterogeneous disease, with genetic diversity within and between tumors. This diversity highlights the necessity of tumor classification, using genomic and/or genetic biomarkers, which may ultimately lead to more personalized therapies. Advances in laboratory and computational techniques laid the foundation for comprehensive identification and measurement of various types of genetic aberrations and evaluation of their clinical significance. These techniques include the development of sophisticated model systems (i.e. Genetically Engineered Mouse Models/GEMM) and accurate and efficient high-throughput technologies (i.e. Massively Parallel Sequencing/MPS). My research focused on exploring the possibility of using mouse models and advanced RNA-Seq protocols to facilitate the translation of biological findings in models systems into clinically meaningful knowledge for human patients.

Our studies on the GEMM p53 null transplant model in **Chapter 2** showed an example that multiple disease types could present within a single murine model. Unsupervised clustering analysis of gene expression data revealed that these p53 null tumors fell into five groups, including two Basal-like, one Luminal, one Claudin-low and one subpopulation unique to this model, with each group recapitulates genetic signatures of their human cognate subtypes. Hence it provided a novel preclinical resource for investigating the human Claudin-low subtype, which was recently identified and only consists of 5-10% human breast tumors. Similar to human Claudin-low patients, p53 null Claudin-low tumors lack tight junction proteins, show high expression of EMT genes and features of normal mammary stem cells, and are enriched in TIC. Again, as expanded evidence demonstrated that this model recapitulates Claudin-low tumors faithfully, it could serve as an *in vivo* system to reveal the features of this subtype using various experimental approaches, which was previously restricted by the rarity and limitations in procurement of human Claudin-low tumors. For example, through the use of the murine Claudin-low tumors we were able to show a high Tumor Initiating Cell frequency in these tumors, and we were able to show enrichment for multiple stem cell associated features in both human and murine Claudin-low tumors

Interestingly, each p53 null tumor subgroup also displays distinct copy number alteration (CNA) landscape. By cross-species comparison, a few events are in common between human and mice. Some of these shared CNA regions contain cancer genes, such as INPP4B that is lost in both human and p53 null mouse basal-like tumors, while the driver genes in other subtypes are not as clear. Noteworthy, the copy number landscape of Claudin-low subtype has yet to be identified. As the results suggested that the p53 null Claudin-low mouse tumors showed a fair amount of genomic instability, it might provide some insights into putative subtype-specific CNA events, and even driver genes in tumor initiation and progression in Claudin-low patients. The appealing feature of high enrichment of TIC in this mouse model offers an opportunity to investigate important signaling pathways within the context of a model with a demonstrated enrichment of potential cancer stem cells. In addition, their transplantability into syngenic hosts allows for preclinical testing of novel therapeutics that target stem cells.

Another example of using GEMMs as a preclinical testing model was shown in **Chapter 3**. In this study, we extensively examined the efficacy of a set of chemotherapeutics using large cohorts of three genomically matched murine mammary tumor models. These genomically well-defined models represent three human breast tumor subtypes: Basal-like, Luminal B and Claudin-low, and were treated with identical regimens of commonly used chemotherapeutics taken from the human breast cancer clinic. With the exception of lapatinib in the *MMTV-Neu* model, single-agent regimens rarely elicited a strong response in all GEMMs. On the contrary, some combination regimens showed more potent effects. Intriguingly, in the Basal-like *C3(1)-T-antigen* model, heterogeneous responses to several therapeutics were observed. Of note, the same pattern has also been observed in human Basal-like patients that were treated with comparable agents, which suggested the existence of genetically distinct subgroups within this subtype[1, 2].

Therefore, we chose to focus on the expression patterns associated with Basal-like *C3(1)-T-antigen* mice treated with the chemotherapy doublet of carboplatin/paclitaxel (CT). About 2/3 of the tumors showed little response, while the remaining 1/3 showed a near complete regression. Two murine signatures were derived from the sets of differentially expressed genes obtained when comparing these tumor populations. The signatures were significantly associated with pCR and are predictive to pCR in two large, independent cohorts of human patients that were treated with similar chemotherapeutics (i.e. anthracycline/taxane). Importantly, even in multivariate analysis with a set of commonly used clinical and biological markers, these signatures were still significant predictors. Specifically, the UNTREATED-HUM signature was also predictive in clinically relevant triple-negative patients, which is of practical value because this subpopulation of patients are not candidates for the current targeted therapies [2]. Furthermore, gene ontology analysis revealed that the UNTREATED-HUM signature was enriched with genes involved in M phase, mitosis, and the cell cycle. This might provide some mechanistic insights in that cells undergoing DNA-synthesis and mitosis are more sensitive to cytotoxic agents. In sum, the efforts of testing therapeutics in faithful mouse models has laid the groundwork for expanded drug efficacy testing. The murine-derived biomarkers, though, might need further validation, but could potentially inform clinical practice of personalized medication.

On the other hand, the strategy by which researchers extract information of human genomics has been revolutionized by the development of RNA-Seq technology. Due to the many special characteristics of cancer studies, special considerations are needed in the application of RNA-Seq. In particular, tumor samples with low quantity and/or quality, as well as the large reservoir of FFPE tissues pose great challenges for the study design, as there is no clear standard for the choice of laboratory methods and optimized parameters. Our study in **Chapter 4** aims to address this need by evaluating two rRNA depletion protocols on paired Fresh Frozen (FF) and FFPE samples, and extensively comparing them with other transcript profiling techniques.

The initial evaluation of sequencing library quality demonstrated that the rRNA removal efficiency is equivalently high in Ribo-Zero-Seq and DSN-Seq in both FF and FFPE, as compared to the standard mRNA-Seq protocol. Especially, the Ribo-Zero-Seq has comparable or even better performance in two other types of biases: uniformity of transcript coverage and 5’-to-3’ bias. This feature was critical to derive information, particularly non-biased transcript abundance, from heavily fragmented FFPE materials, so that researchers could appreciate the untapped potential of FFPE archives that could be used in cancer transcriptome studies.

As expected, the relative coverage across genes exhibited remarkable distinct patterns in Ribo-Zero-Seq and DSN-Seq. Bases mapping to exons constitute only 20-30% total bases in both FF and FFPE, as compared to 60-70% in mRNA-Seq libraries of FF samples. Screening of individual genes identified reads that spanned exon-intron boundaries, providing direct evidence that unspliced pre-mRNAs are captured by these FFPE and whole transcriptome protocols. As this feature could affect experimental design, mostly influencing the number of read needed for complete coverage, we performed an objective analysis to determine the number of reads required by each protocol in FF and FFPE to match the performance of a DNA-microarray. The analysis indicated that approximately 14 million reads in mRNA-Seq and 45-65 million reads in Ribo-Zero-Seq or DSN-Seq are required for the same level of gene detection. This result, though might need adjustment based on the experimental setting, but should contribute to optimize study design and balance the cost versus detection sensitivity.

Another important feature of RNA-Seq technology is its accuracy and reproducibility to quantify transcript abundance. Our data showed that Ribo-Zero-Seq provide highly reproducible quantification even in FFPE samples. All RNA-Seq protocol pairs are highly consistent, with the two rRNA depletion protocols being most highly correlated in FF and FFPE. In addition, all the RNA-Seq data from FF libraries are highly correlated with DNA microarray, and results from FFPE were moderately correlated, but more than acceptable using current standards of concordance (i.e. a pearson correlation >0.7). As DNA microarray is considered as the old school gold standard method, and many previous studies were performed on the platform of microarray, these findings set the ground work for data integration and comparison across platforms. The analysis collectively suggested that using these new techniques, particularly Ribo-Zero-Seq, it is able to perform accurate, reproducible and comprehensive transcript profiling using FFPE-derived RNAs.

In closing, my dissertation work has demonstrated the potential clinical utility of two distinct experimental approaches, namely the use of GEMM for biomarker discovery and rRNA depletion methods of RNA-Seq, and then also takes these two a coordinated step farther thus showing how a mouse signature could be turned into a practically delivered human biomarker. With the expanding reservoir of GEMM available, their utility is not limited to basic biological discovery. As the restrictions involved human tumor samples are not applicable to GEMM, the advantages of mouse models in identifying drug target, testing drug efficacy or developing new biomarkers deserve more appreciation. The development of laboratory and computational techniques of RNA-Seq offers an unprecedented opportunity for cancer genomic studies with a wide range of interests. The possibility of performing RNA-Seq from FFPE samples is promising for retrospective studies and prospective clinical trials. On the other hand, understanding the impact of technical variations and choosing the optimal techniques and parameters for specific aims would be a new challenge for the cancer genomic studies today.

REFERENCES

1. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, He X, Perou CM: **Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer.** *Breast Cancer Res* 2010, **12**:R68.

2. Perou CM: **Molecular stratification of triple-negative breast cancers.** *Oncologist* 2011, **16 Suppl 1**:61–70.

APPENDIX 1

Gene List from SAM study

(a) UNTREATED list of 428 perturbed clones (348 mouse genes) from the untreated versus not SAM analysis. FDR = 1%

	Probe Name	Gene ID	Score (d)	Nume rator	Denom inator	Fold Change
1	AGI_MM_OLIGO_A_51_P408410	cathepsin A	4.004	0.897	0.224	1.833
2	AGI_MM_OLIGO_A_52_P253317	RIKEN cDNA 4930481A15 gene	3.950	1.603	0.406	3.077
3	AGI_MM_OLIGO_A_52_P217492	RIKEN cDNA 5031425E22 gene	3.875	1.018	0.263	2.015
4	AGI_MM_OLIGO_A_52_P346706	aldo-keto reductase family 1, member B3 (aldose reductase)	3.807	1.242	0.326	2.339
5	AGI_MM_OLIGO_A_51_P328613	Fc receptor, IgG, low affinity III	3.715	1.030	0.277	2.028
6	AGI_MM_OLIGO_A_52_P676510	T-cell specific GTPase 1	3.634	1.400	0.385	2.664
7	AGI_MM_OLIGO_A_51_P481494	roundabout homolog 3 (Drosophila)	3.564	1.282	0.360	2.540
8	AGI_MM_OLIGO_A_52_P653054	NA	3.550	1.017	0.286	1.958
9	AGI_MM_OLIGO_A_51_P164203	non-metastatic cells 4, protein expressed in	3.522	1.114	0.316	2.115
10	AGI_MM_OLIGO_A_52_P67088	non-metastatic cells 4, protein expressed in	3.503	0.959	0.274	1.961
11	AGI_MM_OLIGO_A_51_P184928	interferon induced transmembrane protein 7	3.503	0.943	0.269	1.878
12	AGI_MM_OLIGO_A_52_P582059	lysozyme 1	3.451	1.408	0.408	2.625
13	AGI_MM_OLIGO_A_51_P144438	zinc finger, NFX1-type containing 1	3.418	0.840	0.246	1.807
14	AGI_MM_OLIGO_A_52_P313856	solute carrier family 25 (mitochondrial carrier, phosphate carrier), member 26	3.399	0.752	0.221	1.687
15	AGI_MM_OLIGO_A_52_P464315	cathepsin A	3.376	0.858	0.254	1.762
16	AGI_MM_OLIGO_A_51_P301117	predicted gene 7582	3.371	1.294	0.384	2.396
17	AGI_MM_OLIGO_A_52_P627816	transglutaminase 1, K polypeptide	3.345	1.082	0.323	2.124
18	AGI_MM_OLIGO_A_51_P131358	selectin, platelet (p-selectin) ligand	3.319	1.068	0.322	2.155
19	AGI_MM_OLIGO_A_51_P255016	nuclear antigen Sp100	3.262	1.002	0.307	1.957
20	AGI_MM_OLIGO_A_51_P408471	CDP-diacylglycerol--inositol 3-phosphatidyltransferase (phosphatidylinositol synthase)	3.251	0.890	0.274	1.831
21	AGI_MM_OLIGO_A_51_P241457	leukocyte immunoglobulin-like receptor, subfamily B, member 4	3.238	1.037	0.320	2.074
22	AGI_MM_OLIGO_A_51_P197528	lymphocyte antigen 6 complex, locus C2	3.231	1.351	0.418	2.307
23	AGI_MM_OLIGO_A_51_P384629	cathepsin D	3.212	1.044	0.325	2.126
24	AGI_MM_OLIGO_A_51_P423578	schlafen 2	3.190	1.127	0.353	2.255
25	AGI_MM_OLIGO_A_51_P452227	poly (ADP-ribose) polymerase family, member 11	3.184	1.071	0.336	2.179
26	AGI_MM_OLIGO_A_52_P367791	methylthioribose-1-phosphate isomerase homolog (S. cerevisiae)	3.178	0.782	0.246	1.719
27	AGI_MM_OLIGO_A_51_P206518	N-acetylglucosamine kinase	3.170	0.799	0.252	1.748
28	AGI_MM_OLIGO_A_51_P405476	Fc receptor, IgE, high affinity I, gamma polypeptide	3.149	0.980	0.311	2.029
29	AGI_MM_OLIGO_A_51_P165342	annexin A2	3.137	1.057	0.337	2.004
30	AGI_MM_OLIGO_A_51_P463562	guanylate binding protein 4	3.136	1.236	0.394	2.360
31	AGI_MM_OLIGO_A_51_P386382	shisa homolog 5 (Xenopus laevis)	3.108	0.806	0.259	1.699
32	AGI_MM_OLIGO_A_52_P150565	ubiquitin specific peptidase 39	3.105	0.664	0.214	1.588

	Probe Name	Gene ID	Score (d)	Nume rator	Denom inator	Fold Change
33	AGI_MM_OLIGO_A_51_P183746	paired related homeobox 2	3.104	1.336	0.430	2.612
34	AGI_MM_OLIGO_A_51_P502132	matrix metallopeptidase 23	3.100	1.047	0.338	2.094
35	AGI_MM_OLIGO_A_51_P131800	cytochrome b-245, alpha polypeptide	3.094	0.915	0.296	1.869
36	AGI_MM_OLIGO_A_51_P303160	arginase, liver	3.085	1.492	0.484	2.727
37	AGI_MM_OLIGO_A_51_P354354	galactose-3-O-sulfotransferase 1	3.085	1.195	0.387	2.145
38	AGI_MM_OLIGO_A_51_P413866	complement factor B	3.075	0.910	0.296	1.884
39	AGI_MM_OLIGO_A_51_P493543	ferritin light chain 2	3.074	1.044	0.340	2.022
40	AGI_MM_OLIGO_A_52_P210164	RIKEN cDNA 4930471M23 gene	3.071	0.773	0.252	1.729
41	AGI_MM_OLIGO_A_51_P259726	differentially expressed in FDCP 8	3.060	0.805	0.263	1.765
42	AGI_MM_OLIGO_A_51_P321794	6-phosphogluconolactonase	3.057	0.788	0.258	1.704
43	AGI_MM_OLIGO_A_51_P179504	angiogenin, ribonuclease A family, member 6	3.056	0.962	0.315	1.970
44	AGI_MM_OLIGO_A_51_P169281	transmembrane protein 132A	3.050	0.836	0.274	1.774
45	AGI_MM_OLIGO_A_51_P321886	CKLF-like MARVEL transmembrane domain containing 3	3.041	0.966	0.318	1.989
46	AGI_MM_OLIGO_A_51_P474078	selenoprotein W, muscle 1	3.040	0.773	0.254	1.715
47	AGI_MM_OLIGO_A_52_P10041	aldo-keto reductase family 1, member B3 (aldose reductase)	3.035	1.076	0.355	2.107
48	AGI_MM_OLIGO_A_52_P185485	actin related protein 2/3 complex, subunit 4	3.031	0.824	0.272	1.759
49	AGI_MM_OLIGO_A_52_P463518	CD200 receptor 1	3.018	1.249	0.414	2.406
50	AGI_MM_OLIGO_A_51_P146149	napsin A aspartic peptidase	3.017	0.805	0.267	1.727
51	AGI_MM_OLIGO_A_51_P238734	major facilitator superfamily domain containing 11	3.008	0.765	0.254	1.661
52	AGI_MM_OLIGO_A_52_P607128	macrophage scavenger receptor 1	3.001	0.981	0.327	2.016
53	AGI_MM_OLIGO_A_52_P325477	tripartite motif-containing 16	2.995	0.985	0.329	1.870
54	AGI_MM_OLIGO_A_52_P179640	fucosidase, alpha-L- 1, tissue	2.992	0.809	0.270	1.729
55	AGI_MM_OLIGO_A_52_P523330	transmembrane protein 147	2.990	0.551	0.184	1.468
56	AGI_MM_OLIGO_A_51_P189208	RIKEN cDNA 4933417G07 gene	2.986	0.710	0.238	1.631
57	AGI_MM_OLIGO_A_52_P338180	translocase of inner mitochondrial membrane 13 homolog (yeast)	2.978	0.786	0.264	1.685
58	AGI_MM_OLIGO_A_51_P452876	adenylate kinase 1	2.977	1.067	0.359	2.093
59	AGI_MM_OLIGO_A_52_P35960	cathepsin D	2.975	0.891	0.300	1.884
60	AGI_MM_OLIGO_A_51_P338397	potassium channel tetramerisation domain containing 10	2.967	0.871	0.293	1.765
61	AGI_MM_OLIGO_A_51_P399106	RIKEN cDNA 9030617O03 gene	2.955	0.851	0.288	1.838
62	AGI_MM_OLIGO_A_52_P582394	mitochondrial ribosomal protein S11	2.953	0.685	0.232	1.597
63	AGI_MM_OLIGO_A_51_P237383	ribonuclease, RNase A family 4	2.942	1.007	0.342	1.999
64	AGI_MM_OLIGO_A_51_P327904	Yip1 domain family, member 1	2.941	0.752	0.256	1.651
65	AGI_MM_OLIGO_A_51_P117618	ethylmalonic encephalopathy 1	2.939	0.598	0.203	1.519
66	AGI_MM_OLIGO_A_51_P321150	lysozyme 2	2.936	0.891	0.303	1.815
67	AGI_MM_OLIGO_A_52_P262511	ribonuclease, RNase A family 4	2.932	0.989	0.337	2.008
68	AGI_MM_OLIGO_A_52_P62085	cathepsin Z	2.924	0.842	0.288	1.760
69	AGI_MM_OLIGO_A_52_P112110	transmembrane protein 82	2.910	0.854	0.294	1.838
70	AGI_MM_OLIGO_A_51_P460954	chemokine (C-C motif) ligand 6	2.908	1.085	0.373	2.041
71	AGI_MM_OLIGO_A_51_P169693	bone marrow stromal cell antigen 2	2.905	1.150	0.396	2.131

	Probe Name	Gene ID	Score (d)	Numerator	Denominator	Fold Change
72	AGI_MM_OLIGO_A_52_P422338	phosphatidylserine decarboxylase	2.904	0.766	0.264	1.675
73	AGI_MM_OLIGO_A_52_P474242	histocompatibility 2, K1, K region	2.889	0.982	0.340	1.984
74	AGI_MM_OLIGO_A_52_P136264	phosphatidylinositol glycan anchor biosynthesis, class S	2.879	0.748	0.260	1.628
75	AGI_MM_OLIGO_A_51_P165882	solute carrier family 22 (organic cation transporter), member 18	2.859	0.827	0.289	1.719
76	AGI_MM_OLIGO_A_51_P251325	translocase of inner mitochondrial membrane 13 homolog (yeast)	2.848	0.757	0.266	1.648
77	AGI_MM_OLIGO_A_51_P337412	epidermal growth factor-containing fibulin-like extracellular matrix protein 1	2.846	1.255	0.441	2.100
78	AGI_MM_OLIGO_A_51_P467889	N-acetyltransferase 9 (GCN5-related, putative)	2.842	0.714	0.251	1.643
79	AGI_MM_OLIGO_A_51_P314285	transmembrane protein 86A	2.842	0.831	0.292	1.798
80	AGI_MM_OLIGO_A_51_P150653	protein tyrosine phosphatase, receptor type, V	2.834	1.465	0.517	3.403
81	AGI_MM_OLIGO_A_51_P263471	C-type lectin domain family 4, member n	2.824	0.962	0.341	2.025
82	AGI_MM_OLIGO_A_51_P371051	GLI pathogenesis-related 1 (glioma)	2.824	1.004	0.356	2.088
83	AGI_MM_OLIGO_A_51_P208240	tumor necrosis factor (ligand) superfamily, member 14	2.823	0.879	0.311	1.886
84	AGI_MM_OLIGO_A_52_P465647	ubiquitin-conjugating enzyme E2D 2	2.823	0.666	0.236	1.589
85	AGI_MM_OLIGO_A_51_P310780	paired immunoglobin-like type 2 receptor alpha	2.813	1.030	0.366	2.056
86	AGI_MM_OLIGO_A_52_P657317	transmembrane protein 160	2.804	0.997	0.356	1.977
87	AGI_MM_OLIGO_A_52_P113190	myosin IF	2.799	0.796	0.285	1.773
88	AGI_MM_OLIGO_A_51_P181312	dicarbonyl L-xylulose reductase	2.799	0.819	0.293	1.800
89	AGI_MM_OLIGO_A_51_P408363	complement factor properdin	2.796	0.901	0.322	1.920
90	AGI_MM_OLIGO_A_52_P638457	RIKEN cDNA A430084P05 gene	2.793	1.371	0.491	2.970
91	AGI_MM_OLIGO_A_51_P347961	Niemann Pick type C2	2.786	0.897	0.322	1.756
92	AGI_MM_OLIGO_A_51_P359891	sialic acid binding Ig-like lectin 1, sialoadhesin	2.784	1.246	0.448	2.573
93	AGI_MM_OLIGO_A_52_P408757	Fc receptor, IgG, low affinity III	2.783	0.825	0.296	1.753
94	AGI_MM_OLIGO_A_51_P120239	polymerase I and transcript release factor	2.782	0.797	0.287	1.734
95	AGI_MM_OLIGO_A_52_P131466	RAS-related C3 botulinum substrate 2	2.781	0.756	0.272	1.713
96	AGI_MM_OLIGO_A_52_P1022311	NA	2.774	0.817	0.295	1.758
97	AGI_MM_OLIGO_A_51_P333349	transmembrane protein 120A	2.774	0.737	0.266	1.672
98	AGI_MM_OLIGO_A_52_P501733	ferritin light chain 1	2.768	0.924	0.334	1.888
99	AGI_MM_OLIGO_A_52_P48155	eukaryotic translation initiation factor 5A	2.758	0.832	0.302	1.817
100	AGI_MM_OLIGO_A_51_P337918	aldehyde dehydrogenase 4 family, member A1	2.758	0.837	0.304	1.811
101	AGI_MM_OLIGO_A_52_P197722	3-hydroxymethyl-3-methylglutaryl-Coenzyme A lyase-like 1	2.757	1.064	0.386	2.027
102	AGI_MM_OLIGO_A_51_P213706	stromal cell derived factor 4	2.756	0.707	0.256	1.583
103	AGI_MM_OLIGO_A_51_P295389	chondroitin polymerizing factor 2	2.755	0.972	0.353	1.915
104	AGI_MM_OLIGO_A_51_P212754	transforming growth factor, beta induced	2.748	1.246	0.453	2.372
105	AGI_MM_OLIGO_A_51_P212419	leucine rich repeat containing 41	2.740	0.713	0.260	1.623
106	AGI_MM_OLIGO_A_52_P555235	regulator of G-protein signaling 19	2.737	0.846	0.309	1.783
107	AGI_MM_OLIGO_A_52_P164017	secretory carrier membrane protein 4	2.734	0.628	0.230	1.545
108	AGI_MM_OLIGO_A_52_P684378	glutathione peroxidase 1	2.733	0.600	0.220	1.504
109	AGI_MM_OLIGO_A_51_P241715	RIKEN cDNA 4930579C12 gene	2.730	1.382	0.506	2.618

	Probe Name	Gene ID	Score (d)	Nume rator	Denom inator	Fold Change
110	AGI_MM_OLIGO_A_51_P391754	sterol O-acyltransferase 1	2.728	0.884	0.324	1.837
111	AGI_MM_OLIGO_A_51_P101146	serine (or cysteine) peptidase inhibitor, clade B, member 2	2.726	1.424	0.522	3.223
112	AGI_MM_OLIGO_A_52_P670026	radical S-adenosyl methionine domain containing 2	2.715	0.908	0.335	1.824
113	AGI_MM_OLIGO_A_51_P153053	sphingomyelin phosphodiesterase, acid-like 3A	2.713	0.974	0.359	1.986
114	AGI_MM_OLIGO_A_51_P485312	chemokine (C-C motif) ligand 5	2.712	0.780	0.288	1.683
115	AGI_MM_OLIGO_A_51_P144143	NA	2.710	0.802	0.296	1.750
116	AGI_MM_OLIGO_A_51_P181451	complement component 1, q subcomponent, alpha polypeptide	2.708	0.899	0.332	1.810
117	AGI_MM_OLIGO_A_52_P429749	transmembrane protein 115	2.707	0.653	0.241	1.563
118	AGI_MM_OLIGO_A_51_P432641	chemokine (C-X-C motif) ligand 10	2.703	1.139	0.421	2.068
119	AGI_MM_OLIGO_A_52_P51429	DENN/MADD domain containing 1C	2.703	0.792	0.293	1.763
120	AGI_MM_OLIGO_A_52_P57013	nucleoredoxin	2.700	1.019	0.377	2.072
121	AGI_MM_OLIGO_A_51_P495730	RIKEN cDNA 1700049L16 gene	2.698	0.963	0.357	1.869
122	AGI_MM_OLIGO_A_52_P356562	NA	2.698	0.668	0.247	1.596
123	AGI_MM_OLIGO_A_52_P481957	gremlin 1	2.691	1.148	0.427	2.423
124	AGI_MM_OLIGO_A_51_P138895	coiled-coil domain containing 102A	2.691	0.647	0.240	1.562
125	AGI_MM_OLIGO_A_51_P484329	T-cell receptor alpha chain	2.689	1.247	0.464	2.569
126	AGI_MM_OLIGO_A_52_P29879	differentially expressed in FDCP 8	2.683	0.748	0.279	1.645
127	AGI_MM_OLIGO_A_51_P232355	protease (prosome, macropain) 26S subunit, ATPase 5	2.682	0.829	0.309	1.780
128	AGI_MM_OLIGO_A_52_P110812	gamma-glutamyl carboxylase	2.676	0.637	0.238	1.562
129	AGI_MM_OLIGO_A_52_P223618	reticulocalbin 3, EF-hand calcium binding domain	2.675	0.878	0.328	1.883
130	AGI_MM_OLIGO_A_51_P511315	proline-serine-threonine phosphatase-interacting protein 1	2.674	0.878	0.328	1.858
131	AGI_MM_OLIGO_A_51_P205943	huntingtin	2.673	0.717	0.268	1.640
132	AGI_MM_OLIGO_A_52_P511269	malectin	2.672	0.875	0.327	1.814
133	AGI_MM_OLIGO_A_51_P211491	glucuronidase, beta	2.669	0.823	0.308	1.774
134	AGI_MM_OLIGO_A_51_P306047	SEC13 homolog (S. cerevisiae)	2.669	0.737	0.276	1.653
135	AGI_MM_OLIGO_A_52_P241732	peroxisomal biogenesis factor 16	2.657	0.616	0.232	1.536
136	AGI_MM_OLIGO_A_52_P409760	islet cell autoantigen 1	2.656	0.904	0.340	1.708
137	AGI_MM_OLIGO_A_51_P359636	lectin, galactoside-binding, soluble, 3 binding protein	2.655	0.890	0.335	1.781
138	AGI_MM_OLIGO_A_51_P297679	hematopoietic cell specific Lyn substrate 1	2.655	0.927	0.349	1.983
139	AGI_MM_OLIGO_A_51_P183685	RAB34, member of RAS oncogene family	2.655	0.813	0.306	1.732
140	AGI_MM_OLIGO_A_52_P393755	LIM motif-containing protein kinase 2	2.653	0.730	0.275	1.677
141	AGI_MM_OLIGO_A_52_P162298	YdjC homolog (bacterial)	2.652	0.584	0.220	1.498
142	AGI_MM_OLIGO_A_52_P479001	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58	2.649	1.022	0.386	2.111
143	AGI_MM_OLIGO_A_51_P414412	TRAF type zinc finger domain containing 1	2.648	0.762	0.288	1.681
144	AGI_MM_OLIGO_A_51_P154780	vav 1 oncogene	2.647	0.840	0.317	1.827
145	AGI_MM_OLIGO_A_52_P428354	histocompatibility 2, Q region locus 10	2.646	0.883	0.334	1.857
146	AGI_MM_OLIGO_A_51_P159042	NA	2.645	0.990	0.374	2.147
147	AGI_MM_OLIGO_A_51_P276235	patatin-like phospholipase domain containing 7	2.643	0.664	0.251	1.609

	Probe Name	Gene ID	Score (d)	Nume rator	Denom inator	Fold Change
148	AGI_MM_OLIGO_A_51_P327825	peptidylprolyl isomerase B	2.640	0.642	0.243	1.558
149	AGI_MM_OLIGO_A_51_P178063	RAS p21 protein activator 3	2.635	0.940	0.357	1.943
150	AGI_MM_OLIGO_A_51_P371942	procollagen C-endopeptidase enhancer protein	2.634	1.038	0.394	1.978
151	AGI_MM_OLIGO_A_51_P500550	major facilitator superfamily domain containing 1	2.633	1.073	0.408	2.029
152	AGI_MM_OLIGO_A_52_P537566	centromere protein T	2.629	0.529	0.201	1.448
153	AGI_MM_OLIGO_A_51_P377833	CKLF-like MARVEL transmembrane domain containing 7	2.629	0.687	0.261	1.630
154	AGI_MM_OLIGO_A_51_P119544	aspartylglucosaminidase	2.628	0.789	0.300	1.694
155	AGI_MM_OLIGO_A_52_P52618	colony stimulating factor 2 receptor, beta, low-affinity (granulocyte-macrophage)	2.627	0.841	0.320	1.799
156	AGI_MM_OLIGO_A_52_P231714	cAMP responsive element binding protein 3	2.627	0.703	0.268	1.593
157	AGI_MM_OLIGO_A_51_P506045	plasma glutamate carboxypeptidase	2.626	0.990	0.377	1.990
158	AGI_MM_OLIGO_A_52_P367792	methylthioribose-1-phosphate isomerase homolog (S. cerevisiae)	2.624	0.560	0.214	1.481
159	AGI_MM_OLIGO_A_51_P481930	cadherin 15	2.623	1.000	0.381	2.104
160	AGI_MM_OLIGO_A_51_P311785	mannosidase 2, alpha B2	2.621	0.831	0.317	1.820
161	AGI_MM_OLIGO_A_51_P515883	placenta-specific 8	2.620	1.189	0.454	2.398
162	AGI_MM_OLIGO_A_51_P341736	matrix metallopeptidase 2	2.620	1.260	0.481	2.492
163	AGI_MM_OLIGO_A_51_P151628	lactamase, beta	2.616	0.989	0.378	2.195
164	AGI_MM_OLIGO_A_52_P583973	low density lipoprotein receptor-related protein associated protein 1	2.614	0.800	0.306	1.691
165	AGI_MM_OLIGO_A_51_P148314	family with sequence similarity 117, member A	2.612	0.802	0.307	1.679
166	AGI_MM_OLIGO_A_51_P507787	secretion regulating guanine nucleotide exchange factor	2.608	0.802	0.307	1.708
167	AGI_MM_OLIGO_A_52_P400436	sphingosine kinase 1	2.605	0.838	0.322	1.775
168	AGI_MM_OLIGO_A_52_P96151	WD repeat domain containing 82	2.601	0.689	0.265	1.640
169	AGI_MM_OLIGO_A_51_P133381	dysbindin (dystrobrevin binding protein 1) domain containing 2	2.600	0.550	0.211	1.454
170	AGI_MM_OLIGO_A_51_P498267	ankyrin repeat domain 29	2.598	1.023	0.394	1.983
171	AGI_MM_OLIGO_A_51_P442097	solute carrier family 41, member 3	2.596	0.896	0.345	1.892
172	AGI_MM_OLIGO_A_51_P235835	NA	2.595	0.786	0.303	1.676
173	AGI_MM_OLIGO_A_51_P370825	coiled-coil domain containing 124	2.593	0.583	0.225	1.497
174	AGI_MM_OLIGO_A_51_P156857	RIKEN cDNA 2010002N04 gene	2.592	1.058	0.408	2.076
175	AGI_MM_OLIGO_A_52_P354844	ectodysplasin A2 receptor	2.591	1.130	0.436	2.232
176	AGI_MM_OLIGO_A_52_P516091	phospholipase A2, group XV	2.591	0.807	0.312	1.776
177	AGI_MM_OLIGO_A_51_P368571	D-tyrosyl-tRNA deacylase 1 homolog (S. cerevisiae)	2.586	0.697	0.270	1.631
178	AGI_MM_OLIGO_A_51_P130101	Fc receptor, IgG, low affinity IIb	2.579	0.806	0.312	1.771
179	AGI_MM_OLIGO_A_52_P128095	adaptor-related protein complex 2, sigma 1 subunit	2.578	0.621	0.241	1.513
180	AGI_MM_OLIGO_A_52_P490910	fibronectin type III domain containing 4	2.572	0.910	0.354	1.833
181	AGI_MM_OLIGO_A_52_P600318	secretory carrier membrane protein 2	2.570	0.675	0.263	1.569
182	AGI_MM_OLIGO_A_52_P582374	epithelial stromal interaction 1 (breast)	2.569	1.195	0.465	2.500
183	AGI_MM_OLIGO_A_51_P112223	glutathione S-transferase, alpha 4	2.565	0.914	0.356	1.924
184	AGI_MM_OLIGO_A_52_P604629	cysteine-serine-rich nuclear protein 1	2.563	0.694	0.271	1.606
185	AGI_MM_OLIGO_A_52_P613953	RIKEN cDNA 0610037L13 gene	2.561	0.554	0.216	1.472

	Probe Name	Gene ID	Score (d)	Numerator	Denominator	Fold Change
186	AGI_MM_OLIGO_A_51_P517843	GLI pathogenesis-related 2	2.561	1.010	0.394	2.057
187	AGI_MM_OLIGO_A_51_P291460	predicted gene 6498	2.560	0.686	0.268	1.586
188	AGI_MM_OLIGO_A_52_P62037	annexin A2	2.553	0.756	0.296	1.600
189	AGI_MM_OLIGO_A_51_P256202	cathepsin Z	2.553	0.726	0.284	1.616
190	AGI_MM_OLIGO_A_52_P196077	transcription factor 25 (basic helix-loop-helix)	2.551	0.573	0.225	1.484
191	AGI_MM_OLIGO_A_51_P162671	folate receptor 2 (fetal)	2.548	1.003	0.394	2.062
192	AGI_MM_OLIGO_A_52_P357133	selenoprotein M	2.546	1.015	0.399	2.107
193	AGI_MM_OLIGO_A_51_P517138	mitochondrial ribosomal protein L2	2.546	0.652	0.256	1.536
194	AGI_MM_OLIGO_A_51_P172231	gasdermin D	2.542	0.755	0.297	1.709
195	AGI_MM_OLIGO_A_52_P127069	serum/glucocorticoid regulated kinase 3	2.539	0.782	0.308	1.716
196	AGI_MM_OLIGO_A_51_P103586	RIKEN cDNA A730054J21 gene	2.537	1.023	0.403	2.055
197	AGI_MM_OLIGO_A_51_P254541	BCL2-associated X protein	2.536	0.725	0.286	1.670
198	AGI_MM_OLIGO_A_51_P155073	nudix (nucleoside diphosphate linked moiety X)-type motif 14	2.534	0.582	0.230	1.485
199	AGI_MM_OLIGO_A_51_P307644	general transcription factor III A	2.532	0.614	0.243	1.515
200	AGI_MM_OLIGO_A_51_P400166	three prime repair exonuclease 1	2.530	0.695	0.275	1.576
201	AGI_MM_OLIGO_A_52_P108845	CAP-GLY domain containing linker protein 3	2.530	0.840	0.332	1.791
202	AGI_MM_OLIGO_A_51_P328300	protein disulfide isomerase associated 4	2.529	0.764	0.302	1.678
203	AGI_MM_OLIGO_A_52_P560728	serine hydrolase-like	2.528	0.607	0.240	1.501
204	AGI_MM_OLIGO_A_51_P231820	RIKEN cDNA C130026I21 gene	2.528	1.246	0.493	2.750
205	AGI_MM_OLIGO_A_51_P347713	TAF10 RNA polymerase II, TATA box binding protein (TBP)-associated factor	2.520	0.639	0.253	1.580
206	AGI_MM_OLIGO_A_51_P402378	ADP-ribosylation factor-like 4D	2.520	0.804	0.319	1.827
207	AGI_MM_OLIGO_A_51_P149699	leprecan-like 2	2.518	0.948	0.377	1.921
208	AGI_MM_OLIGO_A_51_P265495	lymphocyte antigen 6 complex, locus A	2.517	1.077	0.428	1.883
209	AGI_MM_OLIGO_A_51_P237752	polymerase I and transcript release factor	2.513	0.908	0.361	1.854
210	AGI_MM_OLIGO_A_51_P134972	Sh3kbp1 binding protein 1	2.512	0.600	0.239	1.513
211	AGI_MM_OLIGO_A_52_P350519	histocompatibility 2, blastocyst	2.504	1.036	0.414	1.870
212	AGI_MM_OLIGO_A_51_P400366	Rhesus blood group-associated B glycoprotein	2.502	0.944	0.377	1.884
213	AGI_MM_OLIGO_A_51_P146168	collagen, type XII, alpha 1	2.497	1.199	0.480	2.481
214	AGI_MM_OLIGO_A_51_P268831	solute carrier family 2 (facilitated glucose transporter), member 6	2.494	0.975	0.391	1.969
215	AGI_MM_OLIGO_A_51_P146560	mesothelin	2.493	1.658	0.665	2.598
216	AGI_MM_OLIGO_A_51_P255699	matrix metallopeptidase 3	2.493	1.471	0.590	2.708
217	AGI_MM_OLIGO_A_52_P661982	obscurin-like 1	2.491	0.748	0.300	1.641
218	AGI_MM_OLIGO_A_51_P380178	inhibitor of DNA binding 3	2.490	1.052	0.422	2.012
219	AGI_MM_OLIGO_A_51_P236829	zinc finger, SWIM-type containing 7	2.490	0.620	0.249	1.559
220	AGI_MM_OLIGO_A_51_P282667	hexosaminidase A	2.483	0.865	0.348	1.803
221	AGI_MM_OLIGO_A_51_P160567	endothelial differentiation-related factor 1	2.477	0.520	0.210	1.431
222	AGI_MM_OLIGO_A_51_P241861	T-cell, immune regulator 1, ATPase, H+ transporting, lysosomal V0 protein A3	2.477	0.731	0.295	1.653
223	AGI_MM_OLIGO_A_51_P377452	neutrophil cytosolic factor 4	2.476	0.802	0.324	1.772
224	AGI_MM_OLIGO_A_52_P550491	tubulin, gamma complex associated protein 4	2.469	0.648	0.262	1.565

	Probe Name	Gene ID	Score (d)	Nume rator	Denom inator	Fold Change
225	AGI_MM_OLIGO_A_51_P446510	epithelial membrane protein 3	2.466	0.811	0.329	1.640
226	AGI_MM_OLIGO_A_51_P433824	triosephosphate isomerase 1	2.466	0.713	0.289	1.617
227	AGI_MM_OLIGO_A_51_P373226	ADP-ribosylation factor-like 2	2.466	0.614	0.249	1.527
228	AGI_MM_OLIGO_A_51_P465281	lectin, galactose binding, soluble 1	2.464	0.868	0.352	1.785
229	AGI_MM_OLIGO_A_51_P139108	carboxypeptidase X 1 (M14 family)	2.463	1.494	0.607	3.189
230	AGI_MM_OLIGO_A_51_P416419	calreticulin	2.460	0.743	0.302	1.682
231	AGI_MM_OLIGO_A_51_P359570	interferon-induced protein with tetratricopeptide repeats 3	2.455	0.645	0.263	1.538
232	AGI_MM_OLIGO_A_51_P438165	Rho GTPase activating protein 22	2.452	0.967	0.394	1.910
233	AGI_MM_OLIGO_A_52_P679273	RIKEN cDNA 2310001H17 gene	2.448	0.849	0.347	1.825
234	AGI_MM_OLIGO_A_51_P261517	TYRO protein tyrosine kinase binding protein	2.446	0.855	0.350	1.896
235	AGI_MM_OLIGO_A_51_P408346	interferon activated gene 204	2.445	0.989	0.404	1.858
236	AGI_MM_OLIGO_A_52_P548121	tubulin tyrosine ligase-like 1	2.445	0.939	0.384	2.012
237	AGI_MM_OLIGO_A_52_P533825	translocase of outer mitochondrial membrane 6 homolog (yeast)	2.443	0.705	0.288	1.609
238	AGI_MM_OLIGO_A_51_P118300	synuclein, gamma	2.442	1.210	0.495	2.058
239	AGI_MM_OLIGO_A_52_P230938	lymphocyte antigen 6 complex, locus C1	2.441	0.937	0.384	1.871
240	AGI_MM_OLIGO_A_52_P231079	synaptosomal-associated protein, 47	2.441	0.601	0.246	1.494
241	AGI_MM_OLIGO_A_52_P3029	1-acylglycerol-3-phosphate O-acyltransferase 4 (lysophosphatidic acid acyltransferase, delta)	2.440	0.920	0.377	1.892
242	AGI_MM_OLIGO_A_51_P433450	keratin 17	2.436	0.758	0.311	1.624
243	AGI_MM_OLIGO_A_51_P448664	thromboxane A synthase 1, platelet	2.436	0.805	0.330	1.795
244	AGI_MM_OLIGO_A_51_P398683	reticulocalbin 3, EF-hand calcium binding domain	2.432	1.027	0.422	2.158
245	AGI_MM_OLIGO_A_51_P405397	extracellular matrix protein 1	2.431	1.063	0.437	1.979
246	AGI_MM_OLIGO_A_51_P454943	G protein-coupled receptor 19	2.430	0.860	0.354	1.778
247	AGI_MM_OLIGO_A_51_P319917	a disintegrin and metallopeptidase domain 8	2.427	1.001	0.412	2.035
248	AGI_MM_OLIGO_A_51_P454002	RIKEN cDNA 2310035K24 gene	2.423	0.564	0.233	1.484
249	AGI_MM_OLIGO_A_51_P184484	matrix metallopeptidase 13	2.421	1.564	0.646	4.127
250	AGI_MM_OLIGO_A_51_P344399	Rab interacting lysosomal protein-like 2	2.420	0.569	0.235	1.466
251	AGI_MM_OLIGO_A_51_P381060	paired immunoglobin-like type 2 receptor beta 1	2.419	0.679	0.281	1.582
252	AGI_MM_OLIGO_A_52_P422557	zinc finger protein 362	2.417	0.776	0.321	1.622
253	AGI_MM_OLIGO_A_52_P612803	cyclin G1	2.417	0.915	0.379	1.921
254	AGI_MM_OLIGO_A_52_P521054	serine (or cysteine) peptidase inhibitor, clade B, member 8	2.416	0.968	0.401	1.887
255	AGI_MM_OLIGO_A_52_P676403	chemokine (C-X-C motif) ligand 11	2.416	1.292	0.535	2.320
256	AGI_MM_OLIGO_A_52_P654130	ornithine decarboxylase antizyme 2, pseudogene	2.415	0.781	0.323	1.678
257	AGI_MM_OLIGO_A_52_P215539	RIKEN cDNA C030006K11 gene	2.413	0.657	0.272	1.571
258	AGI_MM_OLIGO_A_51_P502152	solute carrier family 19 (sodium/hydrogen exchanger), member 1	2.412	0.620	0.257	1.544
259	AGI_MM_OLIGO_A_52_P304128	matrix metallopeptidase 14 (membrane-inserted)	2.410	0.729	0.302	1.601
260	AGI_MM_OLIGO_A_52_P641013	ankyrin repeat domain 29	2.410	0.893	0.371	1.870
261	AGI_MM_OLIGO_A_52_P650279	Sec61 alpha 1 subunit (S. cerevisiae)	2.410	0.631	0.262	1.535
262	AGI_MM_OLIGO_A_52_P86965	expressed sequence AI607873	2.409	0.997	0.414	1.792
263	AGI_MM_OLIGO_A_52_P44949	tropomyosin 1, alpha	2.408	0.812	0.337	1.708

	Probe Name	Gene ID	Score (d)	Nume rator	Denom inator	Fold Change
264	AGI_MM_OLIGO_A_51_P314153	nuclear receptor 2C2-associated protein	2.407	0.597	0.248	1.525
265	AGI_MM_OLIGO_A_51_P165870	nucleoredoxin	2.405	0.668	0.278	1.603
266	AGI_MM_OLIGO_A_52_P264368	RIKEN cDNA 2410001C21 gene	2.405	0.589	0.245	1.482
267	AGI_MM_OLIGO_A_51_P289223	LIM motif-containing protein kinase 2	2.401	0.763	0.318	1.739
268	AGI_MM_OLIGO_A_51_P509573	chemokine (C-C motif) ligand 4	2.400	1.260	0.525	2.193
269	AGI_MM_OLIGO_A_51_P175681	predicted gene 8995	2.398	1.135	0.473	2.314
270	AGI_MM_OLIGO_A_52_P569067	mevalonate kinase	2.398	0.711	0.296	1.604
271	AGI_MM_OLIGO_A_52_P451775	OTU domain, ubiquitin aldehyde binding 2	2.396	0.712	0.297	1.657
272	AGI_MM_OLIGO_A_51_P100327	transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)	2.395	0.864	0.361	1.826
273	AGI_MM_OLIGO_A_51_P449325	histocompatibility 2, O region alpha locus	2.395	1.000	0.418	2.051
274	AGI_MM_OLIGO_A_51_P439092	ribosomal protein SA	2.395	0.778	0.325	1.763
275	AGI_MM_OLIGO_A_51_P444669	transmembrane protein 106A	2.394	0.734	0.307	1.704
276	AGI_MM_OLIGO_A_51_P502203	ArfGAP with SH3 domain, ankyrin repeat and PH domain 3	2.392	0.671	0.281	1.567
277	AGI_MM_OLIGO_A_51_P108757	fucosidase, alpha-L- 1, tissue	2.389	0.640	0.268	1.521
278	AGI_MM_OLIGO_A_52_P84447	DDRGK domain containing 1	2.388	0.534	0.224	1.428
279	AGI_MM_OLIGO_A_51_P381506	family with sequence similarity 176, member B	2.387	0.667	0.279	1.550
280	AGI_MM_OLIGO_A_51_P326854	ubiquitin-conjugating enzyme E2C binding protein	2.387	0.603	0.253	1.516
281	AGI_MM_OLIGO_A_52_P545613	Fc receptor, IgG, low affinity IIb	2.385	0.777	0.326	1.693
282	AGI_MM_OLIGO_A_52_P321733	macrophage migration inhibitory factor	2.385	0.626	0.262	1.516
283	AGI_MM_OLIGO_A_51_P121891	RAS-related C3 botulinum substrate 2	2.385	0.828	0.347	1.876
284	AGI_MM_OLIGO_A_52_P196732	NIMA (never in mitosis gene a)-related expressed kinase 6	2.383	0.664	0.279	1.602
285	AGI_MM_OLIGO_A_52_P463936	ISG15 ubiquitin-like modifier	2.383	0.697	0.292	1.583
286	AGI_MM_OLIGO_A_51_P149714	membrane-spanning 4-domains, subfamily A, member 6D	2.381	0.876	0.368	1.952
287	AGI_MM_OLIGO_A_51_P174005	zinc finger, C3HC type 1	2.378	0.486	0.204	1.403
288	AGI_MM_OLIGO_A_52_P438957	ribosome binding protein 1	2.378	0.606	0.255	1.502
289	AGI_MM_OLIGO_A_51_P120093	sorting nexin 10	2.375	0.849	0.357	1.792
290	AGI_MM_OLIGO_A_51_P351860	complement component 1, q subcomponent, beta polypeptide	2.375	0.723	0.304	1.661
291	AGI_MM_OLIGO_A_51_P327261	TNFAIP3 interacting protein 1	2.372	0.631	0.266	1.519
292	AGI_MM_OLIGO_A_51_P449777	prostate transmembrane protein, androgen induced 1	2.372	0.750	0.316	1.579
293	AGI_MM_OLIGO_A_52_P111715	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 6	2.372	0.975	0.411	1.859
294	AGI_MM_OLIGO_A_51_P158161	MRT4, mRNA turnover 4, homolog (S. cerevisiae)	2.371	0.507	0.214	1.423
295	AGI_MM_OLIGO_A_51_P185660	chemokine (C-C motif) ligand 9	2.368	1.068	0.451	1.997
296	AGI_MM_OLIGO_A_52_P261846	collagen, type XVIII, alpha 1	2.368	0.912	0.385	1.791
297	AGI_MM_OLIGO_A_52_P326548	formin homology 2 domain containing 3	2.366	0.805	0.340	1.803
298	AGI_MM_OLIGO_A_51_P395164	peroxisomal biogenesis factor 16	2.365	0.536	0.226	1.451
299	AGI_MM_OLIGO_A_51_P496569	slit homolog 2 (Drosophila)	2.365	1.165	0.493	2.561
300	AGI_MM_OLIGO_A_52_P639522	serine (or cysteine) peptidase inhibitor, clade B, member 2	2.363	1.358	0.574	3.055

	Probe Name	Gene ID	Score (d)	Numerator	Denominator	Fold Change
301	AGI_MM_OLIGO_A_51_P327295	aldo-keto reductase family 1, member A4 (aldehyde reductase)	2.363	0.655	0.277	1.532
302	AGI_MM_OLIGO_A_51_P405478	Fc receptor, IgE, high affinity I, gamma polypeptide	2.363	0.710	0.301	1.679
303	AGI_MM_OLIGO_A_51_P175454	calcium and integrin binding 1 (calmyrin)	2.362	0.522	0.221	1.432
304	AGI_MM_OLIGO_A_51_P307964	keratin 13	2.362	0.840	0.356	1.800
305	AGI_MM_OLIGO_A_51_P286946	phospholysine phosphohistidine inorganic pyrophosphate phosphatase	2.361	0.790	0.334	1.723
306	AGI_MM_OLIGO_A_52_P157408	ornithine decarboxylase antizyme 2, pseudogene	2.360	0.735	0.311	1.638
307	AGI_MM_OLIGO_A_51_P419935	adaptor-related protein complex 2, sigma 1 subunit	2.359	0.608	0.258	1.499
308	AGI_MM_OLIGO_A_51_P453715	family with sequence similarity 54, member B	2.358	0.611	0.259	1.528
309	AGI_MM_OLIGO_A_52_P585104	FK506 binding protein 4	2.355	0.723	0.307	1.693
310	AGI_MM_OLIGO_A_52_P507802	lysophosphatidic acid receptor 1	2.351	0.830	0.353	1.798
311	AGI_MM_OLIGO_A_52_P683580	TBC1 domain family, member 9	2.350	0.658	0.280	1.587
312	AGI_MM_OLIGO_A_51_P449048	CNDP dipeptidase 2 (metallopeptidase M20 family)	2.349	0.632	0.269	1.575
313	AGI_MM_OLIGO_A_51_P336161	mannosidase 2, alpha B1	2.347	0.661	0.282	1.580
314	AGI_MM_OLIGO_A_52_P416879	branched chain ketoacid dehydrogenase E1, beta polypeptide	2.346	0.618	0.264	1.505
315	AGI_MM_OLIGO_A_52_P811150	Cdk5 and Abl enzyme substrate 1	2.341	0.734	0.313	1.615
316	AGI_MM_OLIGO_A_51_P358152	selenoprotein N, 1	2.341	0.657	0.281	1.570
317	AGI_MM_OLIGO_A_52_P660945	cathepsin F	2.340	0.681	0.291	1.637
318	AGI_MM_OLIGO_A_51_P429366	hairy and enhancer of split 6 (Drosophila)	2.337	0.657	0.281	1.562
319	AGI_MM_OLIGO_A_51_P235687	arachidonate 5-lipoxygenase activating protein	2.336	0.797	0.341	1.755
320	AGI_MM_OLIGO_A_51_P237754	histocompatibility 2, T region locus 23	2.335	0.779	0.333	1.595
321	AGI_MM_OLIGO_A_51_P154417	fibulin 1	2.334	1.250	0.536	2.609
322	AGI_MM_OLIGO_A_52_P515094	transmembrane protein 176A	2.333	0.698	0.299	1.584
323	AGI_MM_OLIGO_A_51_P320593	differentially expressed in FDCP 6	2.333	0.868	0.372	1.875
324	AGI_MM_OLIGO_A_51_P509746	porcupine homolog (Drosophila)	2.332	0.985	0.422	1.974
325	AGI_MM_OLIGO_A_51_P411389	RIKEN cDNA 5730437N04 gene	2.329	0.666	0.286	1.518
326	AGI_MM_OLIGO_A_52_P405177	C1q and tumor necrosis factor related protein 6	2.329	0.932	0.400	1.960
327	AGI_MM_OLIGO_A_52_P162775	spastic paraplegia 7 homolog (human)	2.329	0.519	0.223	1.429
328	AGI_MM_OLIGO_A_52_P383653	copine VIII	2.328	1.064	0.457	1.838
329	AGI_MM_OLIGO_A_52_P558368	zinc finger, DHHC domain containing 1	2.328	0.968	0.416	2.099
330	AGI_MM_OLIGO_A_52_P337357	2-5 oligoadenylate synthetase 1A	2.328	0.606	0.260	1.506
331	AGI_MM_OLIGO_A_52_P91658	RIKEN cDNA A930001C03 gene	2.328	0.908	0.390	2.042
332	AGI_MM_OLIGO_A_52_P117966	family with sequence similarity 173, member B	2.328	0.444	0.191	1.364
333	AGI_MM_OLIGO_A_52_P431483	IKAROS family zinc finger 2	2.327	0.695	0.299	1.678
334	AGI_MM_OLIGO_A_51_P302458	phospholipid scramblase 3	2.323	0.586	0.252	1.490
335	AGI_MM_OLIGO_A_51_P461578	ferredoxin reductase	2.321	0.568	0.245	1.465
336	AGI_MM_OLIGO_A_51_P175580	transformation related protein 53 inducible nuclear protein 1	2.321	0.773	0.333	1.770
337	AGI_MM_OLIGO_A_51_P350817	calponin 1	2.318	1.167	0.503	2.513
338	AGI_MM_OLIGO_A_51_P246924	tubulin polymerization-promoting protein family member 3	2.317	0.894	0.386	1.812

	Probe Name	Gene ID	Score (d)	Numerator	Denominator	Fold Change
339	AGI_MM_OLIGO_A_51_P332136	zinc finger, matrin type 5	2.313	0.543	0.235	1.450
340	AGI_MM_OLIGO_A_51_P400752	histocompatibility 2, Q region locus 5	2.311	0.754	0.326	1.597
341	AGI_MM_OLIGO_A_51_P284486	glutathione S-transferase, mu 2	2.308	1.082	0.469	2.381
342	AGI_MM_OLIGO_A_51_P173678	solute carrier family 10 (sodium/bile acid cotransporter family), member 6	2.308	0.671	0.291	1.607
343	AGI_MM_OLIGO_A_51_P506417	keratin 14	2.307	0.791	0.343	1.620
344	AGI_MM_OLIGO_A_51_P419147	leupaxin	2.305	0.617	0.268	1.556
345	AGI_MM_OLIGO_A_51_P169495	Moloney leukemia virus 10	2.303	0.516	0.224	1.439
346	AGI_MM_OLIGO_A_51_P126275	keratin 6B	2.302	0.953	0.414	2.087
347	AGI_MM_OLIGO_A_52_P582112	host cell factor C1 regulator 1 (XPO1-dependent)	2.302	0.598	0.260	1.472
348	AGI_MM_OLIGO_A_52_P487615	family with sequence similarity 105, member A	2.300	0.714	0.310	1.675
349	AGI_MM_OLIGO_A_51_P376347	heme binding protein 1	2.300	0.646	0.281	1.568
350	AGI_MM_OLIGO_A_52_P244723	tubulin folding cofactor E-like	2.300	0.762	0.331	1.601
351	AGI_MM_OLIGO_A_52_P148514	heparanase	2.300	0.930	0.404	2.050
352	AGI_MM_OLIGO_A_52_P592101	DNA segment, Chr 6, Wayne State University 116, expressed	2.300	0.543	0.236	1.449
353	AGI_MM_OLIGO_A_52_P132612	predicted gene, EG433923	2.299	0.900	0.391	1.814
354	AGI_MM_OLIGO_A_52_P28651	poliovirus receptor-related 1	2.298	0.638	0.277	1.560
355	AGI_MM_OLIGO_A_51_P407832	NA	2.297	1.034	0.450	1.892
356	AGI_MM_OLIGO_A_51_P505538	hemochromatosis	2.296	0.713	0.310	1.613
357	AGI_MM_OLIGO_A_51_P274259	adenylate kinase 5	2.296	0.755	0.329	1.713
358	AGI_MM_OLIGO_A_52_P5891	family with sequence similarity 101, member B	2.295	0.775	0.338	1.666
359	AGI_MM_OLIGO_A_52_P323111	LAG1 homolog, ceramide synthase 6	2.295	0.754	0.329	1.736
360	AGI_MM_OLIGO_A_52_P612518	solute carrier family 44, member 2	2.292	0.600	0.262	1.492
361	AGI_MM_OLIGO_A_51_P273170	nucleolar protein 3 (apoptosis repressor with CARD domain)	2.292	0.859	0.375	1.945
362	AGI_MM_OLIGO_A_51_P495492	syntaxin 4A (placental)	2.288	0.643	0.281	1.525
363	AGI_MM_OLIGO_A_51_P329928	pleckstrin homology-like domain, family A, member 3	2.288	0.740	0.323	1.735
364	AGI_MM_OLIGO_A_51_P512085	collectin sub-family member 12	2.288	0.995	0.435	2.066
365	AGI_MM_OLIGO_A_52_P570717	histocompatibility 2, class II antigen A, beta 1	2.288	1.025	0.448	1.862
366	AGI_MM_OLIGO_A_52_P1076740	NA	2.287	0.748	0.327	1.758
367	AGI_MM_OLIGO_A_51_P484869	guanidinoacetate methyltransferase	2.286	0.498	0.218	1.406
368	AGI_MM_OLIGO_A_51_P211436	G protein-coupled receptor 83	2.285	0.920	0.403	1.785
369	AGI_MM_OLIGO_A_52_P2659	NA	2.284	0.805	0.352	1.754
370	AGI_MM_OLIGO_A_51_P171382	CD163 molecule-like 1	2.283	0.902	0.395	1.807
371	AGI_MM_OLIGO_A_51_P244492	neuroblastoma, suppression of tumorigenicity 1	2.282	1.044	0.457	2.108
372	AGI_MM_OLIGO_A_51_P300277	coronin, actin binding protein 1A	2.282	0.775	0.340	1.723
373	AGI_MM_OLIGO_A_51_P376238	serine (or cysteine) peptidase inhibitor, clade G, member 1	2.282	0.954	0.418	1.861
374	AGI_MM_OLIGO_A_52_P5454	CD248 antigen, endosialin	2.280	0.888	0.389	1.941
375	AGI_MM_OLIGO_A_51_P242201	N-acylethanolamine acid amidase	2.280	0.641	0.281	1.544
376	AGI_MM_OLIGO_A_51_P437426	leucine rich repeat containing 33	2.280	0.809	0.355	1.785

	Probe Name	Gene ID	Score (d)	Numerator	Denominator	Fold Change
377	AGI_MM_OLIGO_A_51_P375267	drebrin-like	2.277	0.509	0.223	1.401
378	AGI_MM_OLIGO_A_51_P165182	basic leucine zipper transcription factor, ATF-like 2	2.276	0.849	0.373	1.823
379	AGI_MM_OLIGO_A_52_P83479	interferon activated gene 204	2.275	0.864	0.380	1.739
380	AGI_MM_OLIGO_A_52_P76196	NA	2.273	0.546	0.240	1.461
381	AGI_MM_OLIGO_A_51_P466490	armadillo repeat containing 5	2.273	0.585	0.257	1.483
382	AGI_MM_OLIGO_A_52_P371237	neuropilin 1	2.272	0.658	0.290	1.548
383	AGI_MM_OLIGO_A_51_P456870	forkhead box J1	2.271	1.152	0.507	2.093
384	AGI_MM_OLIGO_A_51_P160544	epidermal growth factor-containing fibulin-like extracellular matrix protein 2	2.270	0.719	0.317	1.680
385	AGI_MM_OLIGO_A_52_P50496	histocompatibility 2, K1, K region	2.270	0.651	0.287	1.528
386	AGI_MM_OLIGO_A_51_P496562	slit homolog 2 (Drosophila)	2.269	0.965	0.425	2.114
387	AGI_MM_OLIGO_A_51_P146753	colony stimulating factor 2 receptor, beta 2, low-affinity (granulocyte-macrophage)	2.267	0.693	0.306	1.632
388	AGI_MM_OLIGO_A_52_P296632	NA	2.266	0.631	0.278	1.561
389	AGI_MM_OLIGO_A_51_P369550	CD84 antigen	2.266	0.803	0.355	1.683
390	AGI_MM_OLIGO_A_52_P446363	mesoderm development candidate 2	2.265	0.989	0.437	2.158
391	AGI_MM_OLIGO_A_51_P433733	nucleobindin 1	2.264	0.600	0.265	1.519
392	AGI_MM_OLIGO_A_52_P21595	RAB38, member of RAS oncogene family	2.264	0.856	0.378	1.793
393	AGI_MM_OLIGO_A_51_P246653	C-type lectin domain family 7, member a	2.263	1.027	0.454	2.137
394	AGI_MM_OLIGO_A_51_P105068	LY6/PLAUR domain containing 6B	2.263	1.070	0.473	1.937
395	AGI_MM_OLIGO_A_51_P441426	platelet factor 4	2.262	1.076	0.476	2.314
396	AGI_MM_OLIGO_A_51_P154222	lysyl-tRNA synthetase	2.261	0.571	0.253	1.497
397	AGI_MM_OLIGO_A_52_P466993	ADP-ribosylation factor GTPase activating protein 2	2.261	0.554	0.245	1.447
398	AGI_MM_OLIGO_A_51_P417612	histocompatibility 2, D region locus 4	2.261	0.640	0.283	1.536
399	AGI_MM_OLIGO_A_51_P338443	angiopoietin-like 4	2.260	0.981	0.434	1.913
400	AGI_MM_OLIGO_A_51_P331831	hydrogen voltage-gated channel 1	2.260	0.813	0.360	1.806
401	AGI_MM_OLIGO_A_51_P188271	CD248 antigen, endosialin	2.259	1.042	0.461	2.170
402	AGI_MM_OLIGO_A_51_P436928	ribosomal protein S9	2.258	0.613	0.271	1.546
403	AGI_MM_OLIGO_A_52_P131548	ajuba	2.256	0.745	0.330	1.716
404	AGI_MM_OLIGO_A_52_P262914	zinc finger protein 593	2.255	0.610	0.271	1.508
405	AGI_MM_OLIGO_A_51_P295037	proteasome (prosome, macropain) subunit, beta type 5	2.255	0.576	0.255	1.458
406	AGI_MM_OLIGO_A_51_P345366	proteasome (prosome, macropain) subunit, beta type 8 (large multifunctional peptidase 7)	2.255	0.684	0.303	1.578
407	AGI_MM_OLIGO_A_51_P124748	transforming growth factor, beta 3	2.253	0.604	0.268	1.541
408	AGI_MM_OLIGO_A_51_P295237	low density lipoprotein receptor-related protein 11	2.252	0.787	0.350	1.711
409	AGI_MM_OLIGO_A_52_P195839	cathepsin C	2.252	0.742	0.330	1.676
410	AGI_MM_OLIGO_A_51_P278519	quiescin Q6 sulfhydryl oxidase 1	2.252	0.703	0.312	1.638
411	AGI_MM_OLIGO_A_51_P483311	MPV17 mitochondrial membrane protein-like 2	2.252	0.552	0.245	1.476
412	AGI_MM_OLIGO_A_51_P484111	matrilin 2	2.250	0.895	0.398	1.921
413	AGI_MM_OLIGO_A_51_P120738	purinergic receptor P2Y, G-protein coupled, 14	2.249	0.745	0.331	1.627
414	AGI_MM_OLIGO_A_52_P436238	ornithine decarboxylase, structural 1	2.248	0.763	0.339	1.751

	Probe Name	Gene ID	Score (d)	Nume rator	Denom inator	Fold Change
415	AGI_MM_OLIGO_A_51_P431649	ATPase type 13A1	2.247	0.498	0.222	1.413
416	AGI_MM_OLIGO_A_51_P404077	frizzled homolog 2 (Drosophila)	2.246	0.583	0.260	1.470
417	AGI_MM_OLIGO_A_51_P101474	transferrin receptor 2	2.245	0.817	0.364	1.734
418	AGI_MM_OLIGO_A_51_P150480	RIKEN cDNA 2900073G15 gene	2.244	0.490	0.218	1.402
419	AGI_MM_OLIGO_A_52_P30632	RIKEN cDNA 1700007K13 gene	2.243	1.423	0.634	4.308
420	AGI_MM_OLIGO_A_52_P451073	tumor necrosis factor receptor superfamily, member 21	2.242	0.656	0.293	1.539
421	AGI_MM_OLIGO_A_52_P470316	lysosomal-associated protein transmembrane 4B	2.241	0.698	0.311	1.542
422	AGI_MM_OLIGO_A_51_P432544	histocompatibility 2, T region locus 22	2.240	0.759	0.339	1.767
423	AGI_MM_OLIGO_A_51_P465292	histamine N-methyltransferase	2.240	0.858	0.383	1.789
424	AGI_MM_OLIGO_A_51_P300281	coronin, actin binding protein 1A	2.240	0.646	0.288	1.558
425	AGI_MM_OLIGO_A_51_P207031	neutrophil cytosolic factor 1	2.239	0.842	0.376	1.900
426	AGI_MM_OLIGO_A_51_P153765	carbonic anhydrase 13	2.238	0.730	0.326	1.576
427	AGI_MM_OLIGO_A_51_P408343	interferon activated gene 204	2.238	0.813	0.363	1.634
428	AGI_MM_OLIGO_A_51_P238523	shisa homolog 4 (Xenopus laevis)	2.237	0.505	0.226	1.429

(b) RESP-HIGH list of 74 perturbed probes (61 mouse genes) from the responding versus not SAM analysis. FDR = 1%

	Probe Name	Gene ID	Score (d)	Nume rator	Denom inator	Fold Change
1	AGI_MM_OLIGO_A_51_P500156	parvalbumin	-5.628	-3.216	0.571	0.116
2	AGI_MM_OLIGO_A_51_P220262	ring finger protein 165	-5.327	-1.707	0.321	0.309
3	AGI_MM_OLIGO_A_51_P222467	ATP-binding cassette, sub-family G (WHITE), member 1	-5.286	-1.327	0.251	0.396
4	AGI_MM_OLIGO_A_52_P670612	Meis homeobox 1	-5.254	-4.002	0.762	0.053
5	AGI_MM_OLIGO_A_52_P676518	NA	-4.849	-1.403	0.289	0.383
6	AGI_MM_OLIGO_A_51_P338072	myosin, heavy polypeptide 4, skeletal muscle	-4.583	-4.952	1.080	0.058
7	AGI_MM_OLIGO_A_52_P367294	fibronectin type III and SPRY domain containing 1-like	-4.398	-1.282	0.291	0.405
8	AGI_MM_OLIGO_A_52_P677036	RIKEN cDNA D430007A19 gene	-4.318	-1.157	0.268	0.454
9	AGI_MM_OLIGO_A_52_P843919	NA	-4.304	-1.850	0.430	0.273
10	AGI_MM_OLIGO_A_51_P217498	solute carrier family 2 (facilitated glucose transporter), member 4	-4.216	-1.375	0.326	0.390
11	AGI_MM_OLIGO_A_51_P172155	histidine ammonia lyase	-4.177	-1.421	0.340	0.375
12	AGI_MM_OLIGO_A_52_P484519	septin 11	-4.169	-1.498	0.359	0.348
13	AGI_MM_OLIGO_A_51_P479311	glutathione S-transferase, mu 1	-4.150	-1.330	0.320	0.391
14	AGI_MM_OLIGO_A_51_P353946	interleukin 11 receptor, alpha chain 1	-4.143	-1.262	0.305	0.423
15	AGI_MM_OLIGO_A_51_P496245	homeobox C6	-4.142	-1.688	0.408	0.311
16	AGI_MM_OLIGO_A_51_P279841	B-cell linker	-4.064	-1.241	0.305	0.423
17	AGI_MM_OLIGO_A_51_P305019	RIKEN cDNA 9530049O05 gene	-4.056	-1.732	0.427	0.289
18	AGI_MM_OLIGO_A_52_P602147	myosin, heavy polypeptide 4, skeletal muscle	-4.015	-3.253	0.810	0.072
19	AGI_MM_OLIGO_A_51_P112932	ectonucleoside triphosphate diphosphohydrolase 2	-4.014	-1.443	0.359	0.380

	Probe Name	Gene ID	Score (d)	Numerator	Denominator	Fold Change
20	AGI_MM_OLIGO_A_51_P486239	C-type lectin domain family 3, member b	-4.006	-1.691	0.422	0.321
21	AGI_MM_OLIGO_A_52_P1187697	NA	-4.004	-1.219	0.305	0.418
22	AGI_MM_OLIGO_A_52_P27576	syntaxin binding protein 3A	-3.952	-1.316	0.333	0.384
23	AGI_MM_OLIGO_A_52_P381311	amyloid beta (A4) precursor protein	-3.933	-1.482	0.377	0.372
24	AGI_MM_OLIGO_A_51_P438967	glycoprotein (transmembrane) nmb	-3.911	-1.938	0.495	0.266
25	AGI_MM_OLIGO_A_51_P380750	core-binding factor, runt domain, alpha subunit 2, translocated to, 3 (human)	-3.900	-1.622	0.416	0.325
26	AGI_MM_OLIGO_A_51_P384901	myostatin	-3.837	-1.687	0.440	0.324
27	AGI_MM_OLIGO_A_51_P141413	C-type lectin domain family 10, member A	-3.750	-1.969	0.525	0.257
28	AGI_MM_OLIGO_A_52_P254817	resistin like alpha	-3.724	-1.725	0.463	0.300
29	AGI_MM_OLIGO_A_52_P411315	NA	-3.717	-1.417	0.381	0.381
30	AGI_MM_OLIGO_A_51_P366811	apolipoprotein D	-3.712	-1.462	0.394	0.370
31	AGI_MM_OLIGO_A_52_P1064707	NA	-3.710	-1.804	0.486	0.241
32	AGI_MM_OLIGO_A_51_P234113	nucleotide-binding oligomerization domain containing 1	-3.682	-1.184	0.322	0.442
33	AGI_MM_OLIGO_A_52_P87843	aldehyde dehydrogenase family 1, subfamily A3	-3.675	-1.282	0.349	0.415
34	AGI_MM_OLIGO_A_51_P306229	importin 7	-3.666	-0.999	0.273	0.491
35	AGI_MM_OLIGO_A_52_P467096	golgi autoantigen, golgin subfamily a, 4	-3.664	-0.811	0.221	0.573
36	AGI_MM_OLIGO_A_52_P592909	diacylglycerol O-acyltransferase 2	-3.663	-1.415	0.386	0.378
37	AGI_MM_OLIGO_A_51_P454949	glutathione S-transferase, mu 3	-3.661	-1.064	0.291	0.472
38	AGI_MM_OLIGO_A_51_P428708	complement component 4B (Childo blood group)	-3.658	-1.334	0.365	0.406
39	AGI_MM_OLIGO_A_51_P210310	NA	-3.653	-1.354	0.371	0.403
40	AGI_MM_OLIGO_A_51_P204350	keratin 33A	-3.648	-1.826	0.500	0.298
41	AGI_MM_OLIGO_A_52_P18775	cytochrome b5 reductase-like	-3.645	-0.914	0.251	0.523
42	AGI_MM_OLIGO_A_52_P416123	metastasis associated lung adenocarcinoma transcript 1 (non-coding RNA)	-3.611	-1.027	0.284	0.487
43	AGI_MM_OLIGO_A_52_P142496	hect (homologous to the E6-AP (UBE3A) carboxyl terminus) domain and RCC1 (CHC1)-like domain (RLD) 1	-3.611	-1.164	0.322	0.442
44	AGI_MM_OLIGO_A_52_P779578	NA	-3.601	-1.160	0.322	0.458
45	AGI_MM_OLIGO_A_52_P534250	RNA binding motif, single stranded interacting protein	-3.586	-1.307	0.365	0.407
46	AGI_MM_OLIGO_A_51_P401792	titin	-3.568	-1.312	0.368	0.401
47	AGI_MM_OLIGO_A_51_P226453	acyl-CoA thioesterase 11	-3.559	-0.912	0.256	0.528
48	AGI_MM_OLIGO_A_51_P497451	RNA binding motif, single stranded interacting protein	-3.557	-1.332	0.374	0.403
49	AGI_MM_OLIGO_A_51_P154417	fibulin 1	-3.555	-1.685	0.474	0.321
50	AGI_MM_OLIGO_A_52_P348709	cellular repressor of E1A-stimulated genes 1	-3.532	-1.437	0.407	0.343
51	AGI_MM_OLIGO_A_52_P381665	AF4/FMR2 family, member 1	-3.525	-0.772	0.219	0.587
52	AGI_MM_OLIGO_A_51_P257951	resistin like alpha	-3.512	-1.736	0.494	0.299
53	AGI_MM_OLIGO_A_52_P475033	ubiquitin specific peptidase 15	-3.506	-1.264	0.361	0.398
54	AGI_MM_OLIGO_A_51_P316523	interferon regulatory factor 2	-3.502	-1.033	0.295	0.490
55	AGI_MM_OLIGO_A_52_P588483	fibulin 1	-3.497	-1.769	0.506	0.303
56	AGI_MM_OLIGO_A_51_P160439	beta-gamma crystallin domain containing 3	-3.497	-1.406	0.402	0.369
57	AGI_MM_OLIGO_A_51_P111762	cellular repressor of E1A-stimulated genes 1	-3.496	-1.425	0.408	0.346

	Probe Name	Gene ID	Score (d)	Nume rator	Denom inator	Fold Change
58	AGI_MM_OLIGO_A_51_P513311	retinoid X receptor gamma	-3.487	-0.879	0.252	0.542
59	AGI_MM_OLIGO_A_51_P239236	acetyl-Coenzyme A carboxylase beta	-3.467	-1.127	0.325	0.444
60	AGI_MM_OLIGO_A_52_P102846	similar to T-cell receptor beta-2 chain C region	-3.460	-1.656	0.479	0.342
61	AGI_MM_OLIGO_A_52_P153939	NA	-3.455	-0.888	0.257	0.543
62	AGI_MM_OLIGO_A_52_P338816	zinc finger with UFM1-specific peptidase domain	-3.443	-1.205	0.350	0.431
63	AGI_MM_OLIGO_A_51_P194099	thyroid hormone responsive SPOT14 homolog (Rattus)	-3.430	-1.213	0.354	0.415
64	AGI_MM_OLIGO_A_52_P566867	B-cell leukemia/lymphoma 10	-3.428	-1.075	0.314	0.462
65	AGI_MM_OLIGO_A_52_P358360	immunoglobulin heavy constant mu	-3.419	-0.987	0.289	0.505
66	AGI_MM_OLIGO_A_51_P516012	neurotrophic tyrosine kinase, receptor, type 2	-3.415	-1.358	0.398	0.397
67	AGI_MM_OLIGO_A_51_P370600	Friend leukemia integration 1	-3.378	-1.258	0.372	0.423
68	AGI_MM_OLIGO_A_52_P49136	signal-induced proliferation-associated 1 like 1	-3.360	-0.927	0.276	0.516
69	AGI_MM_OLIGO_A_52_P251672	RIKEN cDNA 6430548M08 gene	-3.353	-1.069	0.319	0.479
70	AGI_MM_OLIGO_A_51_P447976	family with sequence similarity 46, member C	-3.351	-1.443	0.431	0.380
71	AGI_MM_OLIGO_A_52_P922893	RIKEN cDNA 9530013L04 gene	-3.350	-1.419	0.424	0.356
72	AGI_MM_OLIGO_A_51_P196844	oxysterol binding protein-like 3	-3.349	-1.117	0.334	0.461
73	AGI_MM_OLIGO_A_52_P177661	CWF19-like 2, cell cycle control (S. pombe)	-3.339	-0.851	0.255	0.551
74	AGI_MM_OLIGO_A_52_P312084	zinc finger protein 266	-3.338	-0.805	0.241	0.572

APPENDIX 2

Modules/expression signatures analysis of carboplatin/paclitaxel treated *C3(1)-T-antigen* mouse tumors.

Module/Signature Name	mentioned in text	Anova	Anova adjusted	Reference
mouseCT_untreated_high		0.000	0.000	this paper
mouseCT_treated_high		0.000	0.000	this paper
19p13_Amplicon		0.000	0.000	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
ESC_CORE		0.000	0.000	Cell Stem Cell. 2008 Apr 10;2(4):333-44
ESC_MOUSE		0.000	0.000	Cell Stem Cell. 2008 Apr 10;2(4):333-44
Myc_targets1		0.000	0.000	Nature Genetics 2008 May;40(5):499-507
Myc_targets2		0.000	0.000	Nature Genetics 2008 May;40(5):499-507
Unknown_4		0.000	0.000	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
Unknown_6		0.000	0.001	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
ESC_HUMAN		0.000	0.001	Cell Stem Cell. 2008 Apr 10;2(4):333-44
Interferon_Response		0.000	0.001	Genome Biology 2007, 8:R191doi:10.1186/gb-2007-8-9-r191
ES_exp1		0.000	0.001	Nature Genetics 2008 May;40(5):499-507
HS_Green16		0.000	0.001	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
HS_Red11		0.000	0.001	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
BRCA1		0.000	0.001	Nature 2002;415:530-6.
Caldas_immune		0.000	0.001	Genome Biology 2007, 8:R157 (doi:10.1186/gb-2007-8-8-r157)
IGF		0.000	0.001	J Clin Oncol. 2008 Sep 1;26(25):4078-85.
MM_Red8		0.000	0.001	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
Nanog_targets		0.000	0.001	Nature Genetics 2008 May;40(5):499-507
Sox2_targets		0.000	0.001	Nature Genetics 2008 May;40(5):499-507
MM_p53null		0.000	0.001	Genome Biology 2007, 8:R76 (doi:10.1186/gb-2007-8-5-r76)
Estrogen_Reg		0.000	0.001	J Clin Oncol. 2006 Apr 10;24(11):1656-64. Epub 2006 Feb 27.
Oncogenic_SRC		0.000	0.001	Nature 2006;439:353-7.
GRANS		0.000	0.001	BMC Genomics 2006, 7:115 doi:10.1186/1471-2164-7-115
MCF7_Baylor_2		0.000	0.001	Cancer Res. 2008 Sep 15;68(18):7493-501.
Oncogenic_E2F3		0.000	0.001	Nature 2006;439:353-7.
Polyak_A		0.000	0.001	Cancer Cell. 2007 Mar;11(3):259-73.
Sotiriou_PNAS_485_Survival		0.000	0.001	Proc Natl Acad Sci U S A. 2003 Sep 2;100(18):10393-8. Epub 2003 Aug 13.
MCF7_Baylor_1		0.000	0.001	Cancer Res. 2008 Sep 15;68(18):7493-501.
MM_NeuPyMT		0.000	0.001	Genome Biology 2007, 8:R76 (doi:10.1186/gb-2007-8-5-r76)
SMC_Serum_Response		0.000	0.001	PLoS Genet. 2007 Sep;3(9):1770-84.
VEGF_Hypoxia		0.000	0.001	Cancer Research 67, 3441-3449, April 1, 2007. doi: 10.1158/0008-5472.CAN-06-3322
Oncogenic_BCAT		0.000	0.001	Nature 2006;439:353-7.
HER2_Immune		0.000	0.001	Cancer Research 67, 10669-10676, November 15, 2007. doi: 10.1158/0008-5472.CAN-07-0539

Module/Signature Name	mentioned in text	Anova	Anova adjusted	Reference
MM_BRCAwnt		0.000	0.001	Genome Biology 2007, 8:R76 (doi:10.1186/gb-2007-8-5-r76)
MM_C3Tag		0.000	0.002	Genome Biology 2007, 8:R76 (doi:10.1186/gb-2007-8-5-r76)
HER1_Cluster1	yes	0.000	0.002	BMC Genomics. 2007 Jul 31;8:258.
Oct4_targets		0.000	0.002	Nature Genetics 2008 May;40(5):499-507
KRAS2		0.000	0.002	Nat Genet. 2005 Jan;37(1):7-8.
MCF7_Baylor_5		0.000	0.002	Cancer Res. 2008 Sep 15;68(18):7493-501.
MM_WapINT3		0.000	0.002	Genome Biology 2007, 8:R76 (doi:10.1186/gb-2007-8-5-r76)
MCF7_Baylor_3		0.000	0.003	Cancer Res. 2008 Sep 15;68(18):7493-501.
Genomic_Grade		0.000	0.003	JNCI Journal of the National Cancer Institute 2006 98(4):262-272; doi:10.1093/jnci/djj052
MM_Green19		0.000	0.003	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
B_Cell		0.000	0.003	BMC Genomics 2006, 7:115 doi:10.1186/1471-2164-7-115
RDAM		0.000	0.003	Journal of Clinical Oncology, Vol 23, No 4 (February 1), 2005: pp. 732-740
Sotiriou_PNAS_706		0.000	0.003	Proc Natl Acad Sci U S A. 2003 Sep 2;100(18):10393-8. Epub 2003 Aug 13.
MM_Myc		0.000	0.003	Genome Biology 2007, 8:R76 (doi:10.1186/gb-2007-8-5-r76)
Oncogenic_RAS		0.001	0.003	Nature 2006;439:353-7.
Stromal_NatMed		0.001	0.003	Nat Med. 2008 May;14(5):518-27. Epub 2008 Apr 27.
MM_WAPTag		0.001	0.003	Genome Biology 2007, 8:R76 (doi:10.1186/gb-2007-8-5-r76)
MCF7_Baylor_4		0.001	0.003	Cancer Res. 2008 Sep 15;68(18):7493-501.
CD8		0.001	0.003	BMC Genomics 2006, 7:115 doi:10.1186/1471-2164-7-115
HS_Red4		0.001	0.004	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
HS_Green20		0.001	0.004	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MM_Potluck		0.001	0.004	Genome Biology 2007, 8:R76 (doi:10.1186/gb-2007-8-5-r76)
NOS_targets		0.001	0.004	Nature Genetics 2008 May;40(5):499-507
MUnknown_33		0.001	0.004	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MM_Red16		0.001	0.004	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MInterferon_Cluster		0.001	0.004	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
3p21Amplicon		0.001	0.004	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
HS_Green6		0.001	0.004	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
LYMPHS		0.001	0.005	BMC Genomics 2006, 7:115 doi:10.1186/1471-2164-7-115
ESC_MOUSE_Adult		0.001	0.005	Cell Stem Cell. 2008 Apr 10;2(4):333-44
HS_Red16		0.001	0.005	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MM_Green25		0.001	0.005	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MM_Red25		0.001	0.005	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MM_Green17		0.001	0.006	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
CLEVERS_TCF		0.002	0.007	Gastroenterology. 2007 Feb;132(2):628-32. Epub 2006 Aug 18.
MM_Green12		0.002	0.007	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse

Module/Signature Name	mentioned in text	Anova	Anova adjusted	Reference
MM_Green24		0.002	0.007	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MUnknown_8		0.002	0.007	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MM_Normal		0.002	0.007	Genome Biology 2007, 8:R76 (doi:10.1186/gb-2007-8-5-r76)
MM_Red22		0.002	0.008	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MUnknown_7		0.002	0.008	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
HS_Green22		0.002	0.009	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MM_Green11		0.002	0.009	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MProliferation	yes	0.002	0.009	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MUnknown_23		0.002	0.009	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
Immune_cell_Cluster		0.002	0.009	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MM_Green20		0.002	0.009	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MM_Green21		0.002	0.009	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
NKI_TAM		0.002	0.009	Cancer Res. 2005 May 15;65(10):4059-66.
WNT_Fibroblast_Brown		0.002	0.009	PLoS ONE. 2007 Sep 26;2(9):e945.
HS_Green12		0.003	0.009	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MM_Red9		0.003	0.009	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
Bone_Metastasis_Down		0.003	0.010	Cancer Cell. 2003 Jun;3(6):537-49.
MM_DMBAwnt		0.003	0.010	Genome Biology 2007, 8:R76 (doi:10.1186/gb-2007-8-5-r76)
WNT		0.003	0.010	BMC Developmental Biology 2002, 2:8doi:10.1186/1471-213X-2-8
Interferon_Cluster		0.003	0.010	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
ECM		0.003	0.010	J Pathol. 2007 Nov 29; : 18044827
MM_Red6		0.003	0.010	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
HS_Green7		0.003	0.010	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MM_Red10		0.003	0.010	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MM_Red17		0.003	0.010	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
Eed_targets		0.003	0.010	Nature Genetics 2008 May;40(5):499-507
SDDP		0.003	0.011	Nat Med. 2008 May;14(5):518-27. Epub 2008 Apr 27.
MM_Green8		0.004	0.011	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
Unknown_1		0.004	0.011	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
Claudin_High	yes	0.004	0.011	Genome Biology 2007, 8:R76 (doi:10.1186/gb-2007-8-5-r76)
Stromal_PLOS		0.004	0.011	PLoS Biol. 2005 Jun;3(6):e187. Epub 2005 May 10.
Suz12_targets		0.004	0.012	Nature Genetics 2008 May;40(5):499-507
Unknown_10		0.004	0.012	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
H3K27_bound		0.004	0.012	Nature Genetics 2008 May;40(5):499-507

Module/Signature Name	mentioned in text	Anova	Anova adjusted	Reference
RB_LOH	yes	0.004	0.012	Breast Cancer Res. 2008 Sep 9;10(5):R75
E2F1_Repressed_by_Serum		0.004	0.012	Cancer Cell 13, 11–22, January 2008
HS_Red20		0.004	0.012	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MBASAL		0.004	0.012	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MUnknown_10		0.004	0.012	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MM_Green1		0.005	0.013	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MM_Red1		0.005	0.013	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
T_Cell		0.005	0.013	BMC Genomics 2006, 7:115 doi:10.1186/1471-2164-7-115
Unknown_11		0.005	0.013	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
HS_Green9		0.005	0.015	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MUnknown_27		0.006	0.015	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MITO2		0.006	0.016	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
PRC2_targets		0.007	0.017	Nature Genetics 2008 May;40(5):499-507
HS_Green5		0.007	0.017	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MUnknown_4		0.007	0.017	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MCD3_CD8		0.007	0.018	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
Ramaswamy		0.007	0.018	Nat Genet. 2003 Jan;33(1):49-54. Epub 2002 Dec 9.
Unknown_16		0.008	0.020	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MECM	yes	0.008	0.020	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MM_Green10		0.008	0.020	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MUnknown_15		0.009	0.021	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
HS_Red7		0.010	0.023	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
Bone_Metastasis_Up		0.010	0.024	Cancer Cell. 2003 Jun;3(6):537-49.
NOS_TFs		0.011	0.025	Nature Genetics 2008 May;40(5):499-507
LUMINAL_Cluster		0.011	0.025	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MUnknown_26		0.011	0.026	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MDACC		0.012	0.028	J Clin Oncol. 2006 Sep 10;24(26):4236-44. Epub 2006 Aug 8.
MM_Green22		0.013	0.029	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
HS_Green4		0.014	0.031	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
RB_LOSS	yes	0.016	0.036	J Clin Invest. 2007 Jan;117(1):218-28. Epub 2006 Dec 7.
15q25_Amplicon		0.016	0.036	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
Fibromatosis	yes	0.016	0.036	Lab Invest. 2008 Jun;88(6):591-601. Epub 2008 Apr 14.
Sample_Handling		0.016	0.036	Journal of Clinical Oncology, Vol 24, No 23 (August 10), 2006: pp. 3763-3770
HS_Red21		0.017	0.037	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human

Module/Signature Name	mentioned in text	Anova	Anova adjusted	Reference
HS_Red23		0.018	0.040	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MUnknown_19		0.019	0.040	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
Polyak_TGFB		0.020	0.042	Cancer Cell. 2007 Mar;11(3):259-73.
Unknown_5		0.020	0.042	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
Fibroblast_Cluster	yes	0.022	0.046	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
HS_Green13		0.022	0.047	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
Autopsy_AJPath		0.025	0.052	American Journal of Pathology, Vol. 161, No. 5, November 2002
MUnknown_35		0.025	0.052	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MUnknown_34		0.026	0.054	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
HS_Green24		0.027	0.054	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
HS_Red3		0.028	0.056	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
StemCell_11genes		0.028	0.056	J Clin Invest. 2005 Jun;115(6):1503-21.
GATA3		0.028	0.057	J Clin Oncol. 2006 Apr 10;24(11):1656-64. Epub 2006 Feb 27.
HS_Green15		0.030	0.060	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MUnknown_3		0.033	0.064	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
PAM50_proliferation	yes	0.033	0.064	Journal of Clinical Oncology, 10.1200/JCO.2008.18.1370
Unknown_8		0.033	0.064	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MUnknown_24		0.034	0.065	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
Unknown_14		0.034	0.065	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MProtocadherin		0.036	0.068	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
HS_Red9		0.037	0.071	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
ES_exp2		0.038	0.071	Nature Genetics 2008 May;40(5):499-507
Proliferation_Cluster	yes	0.038	0.071	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
HS_Red24		0.040	0.075	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MUnknown_2		0.041	0.076	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
LKB1		0.044	0.081	Nature. 2007 Aug 16;448(7155):807-10. Epub 2007 Aug 5.
Unknown_7		0.044	0.081	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MM_Green4		0.045	0.081	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
HS_Green17		0.047	0.085	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
Claudin_9CELL_LINE	yes	0.048	0.086	Breast Cancer Res. 2010 Sep 2;12(5):R68
MUnknown_31		0.048	0.086	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
11q13_Amplicon		0.052	0.092	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MUnknown_30		0.052	0.092	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MUnknown_5		0.053	0.094	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse

Module/Signature Name	mentioned in text	Anova	Anova adjusted	Reference
MM_Green9		0.057	0.099	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MUnknown_1		0.065	0.113	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
CIN70		0.068	0.118	Nat Genet. 2006 Sep;38(9):1043-8. Epub 2006 Aug 20.
HS_Green14		0.069	0.119	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
HS_Red15		0.070	0.119	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MM_Red23		0.069	0.119	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
8p22_Amplicon		0.071	0.120	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
HS_Red22		0.072	0.122	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MM_Green15		0.073	0.122	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MM_Red20		0.074	0.124	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MM_Green6		0.083	0.138	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
HS_Red19		0.085	0.141	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MM_Green3		0.092	0.151	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MM_Green14		0.093	0.151	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MM_Green5		0.093	0.151	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MClaudin_Cluster		0.100	0.161	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MM_Red21		0.112	0.180	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
17PP13_Amplicon		0.114	0.182	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MITO1		0.119	0.189	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
16q23_Amplicon		0.122	0.193	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
4p16_Amplicon		0.124	0.195	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
Unknown_13		0.133	0.207	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
HS_Red2		0.136	0.211	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
IGG_Cluster		0.140	0.217	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MK14_K17		0.152	0.233	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
16.13_Amplicon		0.154	0.235	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
HER1_Cluster2		0.155	0.236	BMC Genomics. 2007 Jul 31;8:258.
MM_Red12		0.156	0.236	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
HS_Green10		0.162	0.245	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MM_Green2		0.169	0.254	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MNOtch4		0.173	0.259	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
E2F1_NOT_Repressed_by_Serum		0.175	0.260	Cancer Cell 13, 11–22, January 2008
CD34_CD36_Cluster		0.176	0.260	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human

Module/Signature Name	mentioned in text	Anova	Anova adjusted	Reference
MM_Green18		0.178	0.262	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MUnknown_16		0.183	0.268	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MM_Red7		0.194	0.282	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MUnknown_22		0.194	0.282	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
HS_Red10		0.198	0.286	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
ADM_S100A10_A110NDGR1_Cluster		0.201	0.288	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MNADH_CYTochrome		0.203	0.290	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
Claudin_Low		0.205	0.291	Genome Biology 2007, 8:R76 (doi:10.1186/gb-2007-8-5-r76)
MM_Red2		0.214	0.304	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MVEGFC		0.216	0.304	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
HS_Green3		0.218	0.306	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
Young_13genes		0.219	0.306	BMC Medicine 2009, 7:9 doi:10.1186/1741-7015-7-9
HS_Green8		0.222	0.308	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
HER2_Amplicon		0.224	0.310	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
Ribosomal_Cluster		0.226	0.312	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
Polyak_B		0.230	0.315	Cancer Cell. 2007 Mar;11(3):259-73.
1p36_Amplicon		0.236	0.322	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MKRAS_amplicon		0.237	0.322	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MM_Red3		0.244	0.329	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
16q24x		0.250	0.336	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MM_Red13		0.265	0.355	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MM_Green7		0.277	0.370	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
HER1_Cluster3		0.283	0.375	BMC Genomics. 2007 Jul 31;8:258.
MUnknown_14		0.284	0.375	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
HS_Red6		0.288	0.379	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
HS_Red25		0.290	0.380	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
HS_Red1		0.302	0.394	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MM_Green23		0.314	0.408	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MUnknown_17		0.328	0.425	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
Oncogenic_MYC		0.335	0.432	Nature 2006;439:353-7.
MUnknown_32		0.340	0.436	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
HS_Red8		0.348	0.443	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MUnknown_28		0.348	0.443	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse

Module/Signature Name	mentioned in text	Anova	Anova adjusted	Reference
MPYMT_NEU_Cluster		0.352	0.446	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MSquamous		0.366	0.462	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
17q25x		0.372	0.467	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
Claudin29		0.376	0.469	Genome Biology 2007, 8:R76 doi:10.1186/gb-2007-8-5-r76
MM_Green16		0.377	0.469	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MM_Red19		0.377	0.469	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
BASAL_Cluster		0.380	0.470	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
HS_Red14		0.384	0.471	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MNB1		0.384	0.471	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
HS_Green21		0.398	0.486	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
Unknown_15		0.400	0.487	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MUnknown_18		0.403	0.489	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
Glycolysis_Signature		0.408	0.491	BMC Medicine 2009, 7:9 doi:10.1186/1741-7015-7-9
MM_Red5		0.408	0.491	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MUnknown_6		0.411	0.493	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
8p_Amplicon		0.415	0.495	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
HS_Green1		0.421	0.497	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
HS_Green11		0.420	0.497	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
Unknown_3		0.422	0.497	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MNB3		0.430	0.503	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MRibosomal		0.429	0.503	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MM_Red18		0.474	0.552	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
KRAS_amplicon		0.482	0.560	Genome Biology 2007, 8:R76 (doi:10.1186/gb-2007-8-5-r76)
12qMDM4		0.500	0.577	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MUnknown_20		0.499	0.577	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
HISTONE		0.508	0.583	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
Fibrinogen_Cluster		0.523	0.596	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
HS_Red13		0.522	0.596	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
HS_Red12		0.543	0.616	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
Secretoglobin		0.549	0.620	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
HS_Red18		0.553	0.623	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
Mmyosin		0.556	0.624	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
mouseCT_resp_high		0.572	0.639	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse

Module/Signature Name	mentioned in text	Anova	Anova adjusted	Reference
MUnknown_21		0.579	0.645	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
13q14_Amplicon		0.581	0.645	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MM_Green13		0.589	0.652	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
HS_Green18		0.592	0.653	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
AMPH_EPIREGULIN_ Cluster		0.611	0.671	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
FOS_JUN_Cluster		0.618	0.676	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
DiffScore		0.625	0.679	Breast Cancer Res. 2010 Sep 2;12(5):R68
MUnknown_25		0.625	0.679	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MUnknown_11		0.653	0.707	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
Chromogramin		0.669	0.722	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
HS_Green25		0.679	0.730	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MM_Red15		0.704	0.754	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
HS_Green2		0.724	0.772	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MM_Red14		0.736	0.783	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
HS_Green19		0.791	0.835	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MUnknown_9		0.788	0.835	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MUnknown_29		0.794	0.835	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MM_Red11		0.803	0.842	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
HS_Red17		0.817	0.854	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MHistone		0.820	0.854	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
HS_Green23		0.823	0.854	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human
MM_Red24		0.835	0.860	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
S100A9_A8		0.834	0.860	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MM_Red4		0.916	0.939	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Mouse
MUnknown_13		0.917	0.939	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MUnknown_12		0.924	0.943	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
MFGFR2		0.942	0.958	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
Unknown_12		0.956	0.969	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
MNB2		0.963	0.970	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Mouse
Unknown_2		0.964	0.970	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
Unknown_9		0.974	0.978	Fan et al. BMC Medical Genomics 2011, 4:3, Unsupervised Cluster from Human
HS_Red5		0.986	0.986	Fan et al. BMC Medical Genomics 2011, 4:3, Bi-Cluster identified from Human