TRANSCRIPTOME VARIATION IN BREAST CANCER

by

Michael Walker Seiler

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computational Biology and Molecular Biophysics

Written under the direction of

Prof. Gyan Bhanot

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2012

ABSTRACT OF THE DISSERTATION

TRANSCRIPTOME VARIATION IN BREAST CANCER

By MICHAEL WALKER SEILER

Dissertation Director:

Gyan Bhanot

Successful cancer treatment is based on our understanding of a number of biological considerations such as its mechanisms for survival, evasion of tumor suppressor programs, and proliferation. Unfortunately, cancer evolution is often chaotic and a single tumor may exhibit many different methods for achieving its goals, such as direct mutation of tumor suppressor genes, over-expression of genes which target tumor suppressors, or both. With that in mind, it is crucial for clinicians and researchers to be able to distinguish the properties of each tumor and identify similarities between them, so that broad-impact treatments can be

devised. Recently, a number of advances have been made which allow researchers to gather more detailed information in a high-throughput manner on the behavior of individual tumors. Where once only gross gene expression information could be gleaned using a microarray chip, now sequencing technology enables us to understand what individual isoforms of genes are being expressed, and in what abundance. Sequencing technology advances have also enabled us to find novel sites of expression on the genome which do not correspond to known proteins, and in fact provide evidence of a new class of large non-coding RNA molecules with functional consequences for cancer tumors. In this thesis, we present novel methodologies for the identification of alternative transcript as well as non-coding RNA usage in subgroups of breast cancer tumors using data from next generation transcriptome sequencing. Using these methods, we have identified genes which are differentially spliced between breast cancer tumors belonging to estrogen positive (ER+) and negative (ER-) sets, as well as in novel subgroups, and validated the existence of these transcripts in tumor tissue RNA using RT-PCR. Additionally, we present evidence of non-coding RNA transcripts which are aberrantly expressed based on estrogen status, and validate these in a similar way. These discoveries and new methodologies will help elucidate the biological differences between these subgroups of breast cancer, and will assist ongoing research into transcriptome abnormalities in other cancers as sequencing data become available.

Acknowledgments and Dedication

Thanks to

...Gyan Bhanot for always driving to produce, and to focus on what's important

...Shridar Ganesan for his infectious enthusiasm for the science

...Ming Yao for patience with yet another student with no laboratory experience

...Vessela Kristensen and Anne-Lise Børresen-Dale, for sharing their data and supporting my projects

...Sunniva Bjørklund and Doug Robinson, for helping me with validation and being there so I can bounce my ideas off of you

...Christine Cambrook, for putting up with me

...My sister Sarah, for being a friendly rival during this quest. Good luck, you will finish soon.

...My mother for always supporting me

...And my dad, who never got to see me grow up to follow in his footsteps

This thesis is dedicated to the memory of my father, Steven Seiler (1953-2002)

Table of Contents

List of Tables

List of Figures

Chapter 1: Background of High-Throughput Analysis of Breast Cancer

1.1 Origins of breast cancer

Cancer is the result of a series of “hallmark” cellular changes that result in a proliferative phenotype which can evade signaling mechanisms that would normally result in cell cycle arrest and/or programmed cell suicide (apoptosis) [1, 2]. As the tumor cell population grows, additional driver mutations emerge which promote the ability to invade other tissue (metastasis) which eventually allows the tumor to form masses of cells in distant locations in a patient's body. Primary tumors rarely kill patients; instead, most patients die from tumor metastases. Tumors are a tremendous drain on biological resources necessary for normal function, because of their incessant demand for biomaterials to sustain their growth. This demand, coupled with the compromised function of organs harboring metastatic lesions, causes multiple organ failure and eventually, death.

Most breast cancers arise in cells that make up the epithelial lining of the milk ducts, branches in breast tissue which carry milk from milk glands (lobules) to the

nipple. Breast cancer which has not acquired the invasive phenotype may be confined to the ducts or the lobules, referred to as ductal carcinoma in situ (DCIS) or lobular carcinoma in situ (LCIS), respectively. Once the tumor invades outside the basement membrane, the tumor is metastatic and is referred to as infiltrating ductal carcinoma (IDC). Often, progression of DCIS to IDC is accompanied by metastatic invasion to bone and brain. Whereas in the DCIS stage tumors are often curable, once the tumor is classifiable as IDC, it becomes difficult to treat with prognosis dependent on a variety of factors, not all of which are completely understood.

The estrogen receptor (ER) protein is found to be overexpressed in ~70% of all breast cancer tumors [3], which are then referred to as “estrogen postive” (ER+). By itself, the estrogen receptor is a transcription factor which becomes activated in the presence of the hormone estradiol [4]. Upon ligand binding, ER dimerizes and is relocalized to the nucleus, where it acts as a transcription factor that binds to a 13bp palindromic sequence in promoter regions of DNA called the estrogen response element (ERE). This binding induces transcription of downstream genomic targets, some of which include genes involved in cell cycle progression, making it an attractive target for over-expression in cancer. It has been shown that the serine/threonine kinase CDK1 can cause the estrogen receptor to promote transcription even in the absence of estradiol [4, 5]. As a result, CDK1 is

often found to be upregulated in breast cancers. Estrogen receptor positive tumors are generally treated with the drug Tamoxifen, an ER antagonist, to which roughly half of ER+ tumors respond [3]. Tumors that do not express ER are referred to as estrogen negative (ER-).

The oncogene ERBB2 (also referred to as HER2/neu) is upregulated in 15-20% of breast tumors [6, 7], and these tumors are called HER2+. ERBB2 is a receptor tyrosine kinase which is responsible for both a proliferative signaling cascade as well as the promotion of anti-apoptotic factors in these tumors. Overexpression of ERBB2 is strongly associated with disease progression and recurrence after treatment. There is currently a monoclonal antibody treatment for HER2+ patients called Herceptin, which inhibits ERBB2-related signaling in cancer cells [8]. There is an overlap of ER+ and HER2+ tumors, however there is no clinical evidence to suggest that overall survival is different between ER+/HER2+ and ER-/HER2+ subsets [9].

Under current protocols for treatment of breast cancer patients, clinicians classify breast cancers into four distinct classes: ER+/HER2-, ER+/HER2+, ER-/HER2- and ER-/HER2+. This classification, combined with information on the stage and grade of tumor, patient's age and medical history, family history of cancer, Oncotype DX score [10] etc. are combined into a risk profile which is used to

determine therapy.

1.2 Breast cancer classification

The most significant technological advance in the analysis of cancer was the advent of the microarray, a technology that allows researchers to simultaneously interrogate the expression levels of thousands of genes. Microarrays are constructed by binding DNA probes to a substrate, each composed with a complementary sequence to the target gene of interest. Labeled cDNA from the sample of interest is then washed over the chip, causing cDNA with complementary sequences to bind to the appropriate probes. The labels are read and summarized into a raw score for each probe. The technology has now matured so that multiple probes are used to measure the expression of different parts of the same gene, the scores from which can be combined to estimate not only the the relative concentration of the gene in that sample [11] but also potential splice variants [12, 13].

Where previously researchers could only measure the expression of a few genes in parallel, microarray technology opened new doors to whole-genome measurement. For example, in knockdown experiments, a biologist would need

to anticipate the effects of their knockdowns and measure the results individually. With the microarray chip, the relative expression of every known gene could be measured before and after knockdown, allowing for the interrogation of the whole cascade of gene-level changes. Transcriptome changes over time, such as through growth and differentiation, could be measured by extracting RNA from a cell population in each stage of growth and performing a microarray experiment.

However, microarray technology did more than just make previous experiments easier and more wide-ranging. It also enabled *de novo* classification of a sample within a population based solely on the pattern of genome-wide expression [14]. This was made possible using “clustering,” a family of computational algorithms designed to find sets of samples (“clusters”) within a population that maximize intra-cluster similarity and minimize inter-cluster similarity, based on a sample-sample distance function specified by the user [15-18]. With a microarray chip, this is typically an appropriately chosen distance or correlation function across genes or samples. With sufficient sample sizes, it is possible to distinguish subsets of samples which have similar patterns of gene expression, allowing an unsupervised and unbiased analysis of molecularly distinct disease subtypes for tumors which look similar under histological and pathological analysis.

Such studies have shown that even within the clinically identifiable subtypes,

breast cancer is a heterogeneous disease, composed of a number of disease subtypes. These have been identified over the last ten years as a result of the analysis of many large gene expression datasets of RNA extracted from breast cancer tumor samples [9, 14, 19-22]. In general, these molecular subtypes expand upon known breast cancer biology by splitting ER+, ER-, and HER2+ tumors into molecular subclasses which cannot be identified by histopathology. In many cases, these subclasses correlate with overall prognosis and/or treatment efficacy. This is potentially of great benefit to patients and of value to clinicians. Tumor with a molecular signature associated with poor outcome can be identified and treated aggressively, while patients whose tumors are unlikely to progress to metastasis have the option of less aggressive treatment. Given the high toxicity of many chemotherapeutic agents, reducing a patient's exposure when the outcome will be favorable, regardless of aggressive treatment is extremely desirable.

A recent successful clustering method known as “consensus ensemble clustering” is widely used in the bioinformatics community because of the robustness of its classification. The method is based on the idea that combining clustering information from several different clustering methods and multiple bootstraps of of the data results in a more stable (robust) set of clusters [23, 24]. The rationale for using many clustering algorithms is that every clustering

algorithms has some built in assumptions which the data may not support. For instance, the well-known k-means algorithm expects to find spherical subsets, and hierarchical clustering [15] is easily misled by outliers [18]. Algorithms such as k-means and Self-Organizing Map [17] are also stochastic in that their results are partially dependent on random initial conditions. However, these problems can be reduced by running each algorithm multiple times on the same dataset and pooling the results (see Appendix: ConsensusCluster) [25]. Consensus clustering has been successfully applied to stratify data from breast cancer tumor microarray experiments [9, 25] as well as RNA sequencing (RNA-Seq). It has also found applications in identifying biologically-significant subsets of clear cell renal cell carcinoma [26].

1.3 Biomarker discovery

Although the potential clinical benefits of tumor classification are clear, it is not immediately apparent how to go about classifying new tumors into the previously discovered subtypes. The challenge, then, is to find a set of measurements which can uniquely indicate its subtype so that accurate treatment can be effected, which are collectively known as a “biomarkers.” Factors in the selection of effective biomarkers for clinical treatment include sensitivity and specificity,

ease of measurement, and the cost of measurement. Tests which are both sensitive (fewer false positives) and specific (fewer false negatives) are clearly better choices for biomarker selection. In the clinic, it is imperative that tests for clinical tumor type be accurate. Otherwise, there are legal and ethical issues to performing treatment for a disease the patient does not carry which may be at best ineffective and at worst hazardous. Ease of measurement is another strong factor in biomarker selection. A blood test is a simple procedure which can be performed in any general practitioner's examination room, while a test which requires the tumor to be surgically removed, followed by additional tissue processing, RNA extraction, and microarray measurement of a panel of target genes, requires hospitalization and extensive time and effort. The last, but not least important factor of biomarker selection is the cost. This can be related to ease of measurement in that a prohibitively expensive test can prevent patients from receiving accurate treatment, especially those from underprivileged countries. For example, the Oncotype DX platform, a popular diagnostic test for breast cancer which relies on a panel of 21 genes to determine the likelihood of recurrence within 10 years [10], is currently priced over $4000.

Apart from their role in the clinical management of patients, biomarkers are also useful in elucidating the unique biology of a tumor subtype, which can then be used to identify a potential therapeutic target gene and thus aid in the discovery

of novel drugs. Since the driving factors in tumorigenesis are largely unknown for many cancers, the identification of any unique aspects of a given subset of tumors, such as altered gene expression, altered DNA methylation, or altered transcription factor binding, can help to gain a clearer understanding of the dysregulated pathways which contribute to disease progression. For example, the oncogene GRB7 is a biomarker for HER2+ cancers. GRB7 lies proximally to ERBB2, an oncogene associated with poor prognosis that promotes proliferation and tumorigenesis. Given that genes in this region are always co-amplified in this subset of cancers, it is reasonable to suspect the entire genomic region may be copied multiple times, a chromosomal aberration known as copy number variation (CNV). Indeed, it has been shown that the copy number of this region, now known as the HER2 amplicon, is increased in HER2+ cancers [27].

As additional disease subtypes are discovered and our understanding of their underlying biology increases, we move closer to the ultimate goal of being able to target therapy to treating the precise disease present in each patient. Combined with the new movement towards collecting other nonclinical attributes that may alter treatment effectiveness (race, genomic traits, etc), we can envision a future of “personalized medicine.” In this future, treatment is based on a holistic knowledge of disease and the genomic attributes of each patient, leading towards more accurate treatment with fewer side effects. Towards this future, in

this thesis, we will present results from the analysis of breast cancer tumor RNA sequencing (RNA-Seq) data which helps to elucidate novel biological underpinnings of breast cancer.

1.4 RNA sequencing

As evidenced by the existence of copy number variation in HER2+ cancers, the data that can be gleaned from DNA microarrays paints an incomplete picture of the biological interactions that give rise to the driving systems promoting tumorigenesis. A novel technology called RNA-Seq [28, 29] has recently become available to interrogate other genomic features of tumors, such as altered gene splicing patterns or alternative promoter usage. In contrast to interrogating gene association through probe binding, RNA-Seq gives a complete snapshot of the transcriptome by directly measuring the amount (expression level) and the sequence of fragments of mRNA in a sample [28, 30]. Briefly, the Illumina RNA-Seq protocol is as follows. Messenger RNA is first isolated from total RNA using a poly-T bead purification step, where strings of Thymine bases are bound to a substrate and the total RNA solution washed over it, causing mature mRNA, which possesses a poly-A tail, to bind to it. Next, the mRNA is sheared to ~150bp for sequencing, and a cDNA library is created for each sample by adding random

6-nucleotide primers (“hexamers”) to the sample, followed by polymerase chain reaction (PCR). During this step, hexamers anneal to the nucleotide fragments created in the previous step, which causes DNA polymerase to elongate them into double-stranded DNA (see PCR in Appendix: Methods). Adapter labels are then ligated to the ends of these fragments, and the nucleotide bases of each fragment are read by the sequencer in parallel.

Once these fragments of the original mRNA sequence are obtained, there remains the computational task of reconstructing the original. This can be done in two ways: with, and without, a genomic DNA reference sequence. Without a reference, sequence fragments are aligned to one another to form sequence “contigs” in a process referred to as *de novo* sequence alignment [31]. In order to make an accurate prediction of the original mRNA sequence, a sufficient coverage of sequence fragments are required for species with complex genomes, such as the human genome. Especially in cases where the original gene is not heavily expressed, a lesser fraction of the available sequence pool will be available to represent the transcript, and it is much less likely that sufficient sequence coverage will be available to predict the entire transcript after alignment. The total coverage necessary to accurately reflect genomic diversity at a given locus varies, though Robertson et al [31] reported success in loci with 20x sequence coverage. With a reference genome, however, sequence

fragments can be aligned to the reference instead, negating the need for high coverage. This success of this method depends largely on the accuracy of the reference genome, as well as the concordance between it and the sample's genomic DNA sequence. Most software utilities designed to align sequence fragments to a reference genome, such as the popular package Bowtie [32], allow fragments to contain errors, whether from genomic DNA mutation or from the sequencer itself.

The primary advantage of RNA-Seq over DNA microarray technology is that the number of genomic loci measurable using RNA-Seq is not limited to the number of probes bound to a substrate. All aspects of the transcriptome are interrogated at once, providing a high-resolution picture of expression activity. In particular, whereas microarrays are typically designed to measure the expression of an entire gene, RNA-Seq can produce fragments from all parts of the gene individually. This illuminates more esoteric features of the genome, such as retained introns, alternative splicing, and the use of alternative promoters [30]. It is impractical to design microarray probes which can test these features of the transcriptome. Currently, the largest commercially available microarray is the HuEx exon array from Affymetrix [12, 13], which can interrogate ~550000 genomic features, using 4 probes each. However, the small number of probes per feature greatly increases noise [33], and this is still a relatively small fraction

of the human genome which necessarily excludes novel features unsupported by reference annotations. RNA sequencing is the first genome-wide high-throughput technology to measure all transcriptomic features. However, given the relative novelty of the technology and the difficulties of analyzing the vast quantities of data it generates, many of these informative features of tumors are not yet available in current annotations of the human genome.

1.5 Alternative splicing

Alternative splicing is a biological process which generates transcript diversity at a given genomic locus. When genomic DNA is transcribed to become pre-mRNA, some parts of the sequence, called introns, are excised from the final product. Segments of mRNA (called exons) which form the sequences between one intron and the next, are retained and ligated together to create the final mRNA transcript, which is either used in its mRNA form directly or translated into protein. The set of proteins and small ribonucleoproteins (snRNPs) that performs the task of splicing out the introns, ligating the exons, adding a poly-A tail to form a mature mRNA is called the spliceosome [34]. Components of the spliceosome bind exons at their boundaries and perform a transesterification reaction to ensure that the two sequence fragments remain together. The intron is then

excised, released and its components are recycled.

Spliceosome assembly at an exon junction begins with a snRNP called U1 which binds to the sequence GU on the mRNA transcript at the 5' splice site [35]. The snRNP U2 binds the “branch point,” an adenosine base located within the intron, assisted by the U2 auxiliary factor (U2AF) snRNP, which binds to the 3' splice acceptor site, AG. The GT-AG dinucleotide pair is sometimes called the canonical splice junction, since 98.71% of known mammalian splice junctions follow this pattern [36]. Following this, additional snRNPs U4, U5, and U6 are recruited, and U1 and U4 dissociate before the complex becomes catalytically active. Spliceosome formation, and subsequent intron excision, are primarily controlled by splicing regulatory proteins which bind to specific exonic and intronic splicing enhancers (ESEs and ISEs) as well as exonic and intronic splicing silencers (ESSs and ISSs, respectively) [35].

Alternative splicing is the process where exons are joined in multiple arrangements, leading to different genetic transcripts (“isoforms”) from the same genomic sequence of DNA. Alternative splicing can have far-reaching effects on the transcripts generated. For example, if the transcript codes for a protein, whole strings of amino acids may be gained or lost in the splice variant upon translation. These segments may form active domains, or structures which

obstruct active sites. Exon splicing may also cause a frame-shift, which can result in a premature stop codon in the open reading frame and early protein termination. In non-coding areas of the transcript, in particular the 5' or 3' untranslated region (UTR), post-transcriptional regulatory domains are often found. Splicing in these domains can affect mRNA stability, or restrict translation. This sort of transcript diversity has been found to be pervasive in the human genome [37]. In a recent study, it was found that 92-94% of polled transcripts exhibited alternative splicing over 15 human cell lines, and that approximately 86% of those transcripts also showed minor isoform frequency of at least 15% [30].

A specific example of the kind of combinatorial diversity possible at a given locus is the gene EPB41, which codes for an array of cytoskeletal proteins. EPB41 displays an impressive array of tissue-specific alternative start sites and alternative exon use, including 10 exons differentially spliced in various tissues, which could combine to form over 1000 transcript combinations [38]. Additionally, EPB41 has three related proteins in the human genome (EPB41L1, EPB41L2, EPB41L3) which have similar properties. This shows the enormous potential diversity of human transcripts and implies that perhaps the importance of the “gene” is overstated, and that the true unit of genomic information is the transcript.

Given the profound functional diversity of spliced transcript usage in cells, it is unsurprising that many such events are implicated both in tumorigenesis and the suppression of cancer progression. The gene survivin, commonly over-expressed in cancer for its anti-apoptotic properties, has three known variants [39, 40], two of which (survivin-dEx3 and survivin-3B) contribute to poor prognosis in both breast and prostate cancers [39]. However, the third variant, survivin-2B, is actually pro-apoptotic and may in fact be a naturally occurring antagonist [35, 40]. Other important variants in cancer include those in the Caspase family, many of which are known for their promotional roles in apoptotic pathways, which make them important targets for regulation in cancer progression. A shorter variant of Caspase-2, created when the inclusion of an extra exon causes a frameshift and subsequent inclusion of an early stop codon, was shown by Jiang et al [41] to inhibit cell death, whereas the proapoptotic long form has been shown to function as a tumor suppressor [42, 43]. Importantly, despite the fact that gene-level measurement (such as a DNA microarray) is broadly assumed to correlate with “gene activity,” such analysis would not be able to discern relative levels of the two isoforms, which presents misleading results.

Studies of alternative splicing in breast cancer have, in general, focused on

tumor samples versus normal tissue [44, 45]. Venables et al [44] presented a parallel RT-PCR approach to identifying variants, running an average of 30 experiments to poll 600 genes previously associated with cancer. They identified a panel of 41 genes (only 12 of which were required) to classify tumors from normal tissue, indicating there are specific splicing changes which either effect or are the direct result of tumorigenesis. Another study used the exon array ExonHit to identify putative variants [45] differentially expressed over 120 malignant and 45 benign breast tumors with the goal of developing a clinical assay to determine whether new breast lesions were cancerous. An astonishing 37,858 exonic probe sets were found differentially expressed, of which 1228 made up a molecular classifier which might be used clinically. Among the genes surveyed, ACOX2 was identified as having a strong splice index (higher exonic variability) in malignant tumors, which parallels our own results showing that the full-length transcript of ACOX2 is heavily expressed in normal breast tissue from reduction surgery, however a shorter intronic-start version is present in all tumor samples and heavily upregulated in ER+ tumors.

Two studies of breast cancer alternative splicing between tumor subtypes have recently been published, forgoing comparison with non-malignant tissue and focusing on inter-subset variability [46, 47]. The primary benefit to this type of study is that breast tissue is highly heterogeneous, containing lobules, ductal

tissue, connective tissue, and fatty tissue [48]. As such, RNA obtained from most normal breast tissue necessarily originates from a variety of sources, which may have highly variable transcriptome profiles. Wang et al [47] showed that when counting exon skipping events in tumor samples, a panel of 4 ER+ and 4 ER- tumors and cell lines could be separated via PCA and clustering. Another study, run on a custom Affymetrix exon array, compared exon expression levels on a wide panel of breast cancer tumor cell lines, separated by tumor subtype. Intriguingly, they reported that alternative variants associated with a rare ER- subclass referred to as Claudin-low [22] were significantly associated with the Fox2 splicing factor, and that upon Fox2 knockdown these variants were reduced.

In order to poll alternative splicing using RNA-Seq, splicing-aware alignment software must be used, such as the TopHat [49] software package. TopHat works by saving all fragments which do not align end-to-end to the reference genome, and splitting them into smaller pieces. If these smaller pieces can be aligned to the genome near to each other, TopHat records a putative exon-exon splice junction at that locus. Finally, using the novel junction database generated in the previous step, TopHat once again aligns the leftover fragments to the genome. Alternative splicing appears as differential junction usage at the same locus, such as a splice junction skipping an exon present in another transcript. Once noted,

assembly can be accomplished through manual annotation or using a referenced-based whole transcript assembler such as Cufflinks [37]. Also visible using this method are retained introns. Intronic regions are sometimes retained post-splicing, but these regions usually contain stop codons, and thus result in truncated proteins during translation. Some intronic regions contain microRNAs (miRNAs), small RNA (~70bp) which can promote gene silencing by inhibiting translation or causing direct mRNA degradation, depending on sequence complementarity. Specific proteins bind to these introns in pre-mRNA, preventing immediate splicing so that the miRNA can be processed separately. RNA-Seq can provide clues to the existence of this process and other examples of retained introns through sequence fragments which align to intronic regions otherwise spliced in alternate transcripts.

As previously mentioned, another prominent method of promoting transcript diversity is alternative start exons. This has also been shown in the EPB41 family of genes, which each have ~3 potential start sites. Another example is the gene lactotransferrin (LTF), which has two alternative start sites. One start site encodes a 44 amino acid sequence that marks the resultant protein for extracellular secretion, where it has antiseptic properties. Without this first exon, the protein (referred to as dLTF) localizes to the nucleus, where it becomes a transcription factor [50-53]. It was recently shown by Pal et al that alternative

start sites tend to result in larger changes in expression, i.e., the bulk of differential transcript usage in tissues is due to start exon variation [54]. Indeed, dLTF has been shown to be almost entirely absent in breast cancer [50, 51]. These changes most likely result from the use of an alternative promoter, which could be due to many factors, including occlusion of the original promoter through methylation or protein binding, activation of transcription factors that bind to the new promoter, and chromatin remodeling that results in exposing different parts of the genome to transcription. Taken together, alternative splicing and promoter usage combine to effect vast genomic diversity. For example, although the human genome has approximately 23,000 genes, in the most current human genome annotation in ENSEMBL, there are ~196,000 known transcripts, of which ~56,000 are known to code for proteins [55].

A significant fraction of the transcriptome, however, does not code for protein. This class, collectively referred to as non-coding RNA (ncRNA), includes, but is not limited to, microRNA, short-interfering RNA (siRNA), transfer RNA (tRNA), ribosomal RNA, and long, intergenic non-coding RNA (lincRNA). While tRNA and ribosomal RNA have been well known to biologists for their activity during the process of translation, knowledge of lincRNAs as a whole is in its infancy. A prominent example in this class is XIST, a large (~19kb) stretch of alternatively spliced RNA which coats the unused female X chromosome to prevent

transcription. XIST expression is absent in some breast, ovarian, and cervical cancer cell lines, indicating it may play some role in tumorigenesis [56]. The lincRNA HOTAIR was recently discovered to be responsible for the repression of hundreds of genes by associating with the Polycomb Repressive Complex 2 (PRC2) [57]. PRC2 alters histone methylation in such a way that transcription is reduced or impossible. The occupation pattern of PRC2 genome-wide was found to resemble that of embryonic fibroblasts in tumor cell lines with upregulated HOTAIR, and these cell lines grew aggressively with enhanced metastatic capability [57]. HOTAIR has been found to be a driving force to malignancy in many human cancers including breast cancer [57], gastrointestinal tumors [58], and hepatocellular carcinoma [59]. Estimates of the total number of ncRNA in the human genome (ignoring alternative transcripts) put the number at approximately ~6700 [60], though it remains unclear how many are also conserved in mammalian genomes. RNA-Seq has played a pivotal role in ncRNA discovery, used previously in concert with immunoprecipitation to identify the sequences of ncRNA bound to PRC2 [61]. We demonstrate a method to identify novel ncRNA using RNA sequencing data, and have discovered a number of ncRNA which are associated with ER status in breast cancer tumor tissue.

RNA sequencing can also be used to identify mutations in transcriptome coding sequences, such as single nucleotide polymorphisms (SNPs) and small

insertions and deletions (indels). Mutations of this nature in protein-coding sequences can cause missense errors which change the amino acid being produced, or replace an amino acid with a stop codon, referred to as a nonsense mutation. The latter causes the production of a truncated protein which may not retain normal function. Both of these types of mutations are of particular interest in cancer, as tumor suppressor genes are frequently the target of errors resulting in a loss of function. A prominent example is the well-known tumor suppressor gene p53, which is found to be mutated in roughly 50% of human cancers and for which now more than 35,000 mutations contributing to tumorigenesis have been found and cataloged [62]. Unfortunately, many important types of genomic mutation, such as errors in promoter sequences or splice sites, are invisible to RNA sequencing. These types of errors are only apparent following genomic DNA sequencing, though changes in RNA expression in certain samples may provide clues as to the existence of such errors in nearby DNA, reducing the search space considerably.

RNA sequencing does have limitations, mutations in DNA promoting transcriptome changes being just one example. Other epigenetic changes which require specialized experiments to discern include DNA methylation, which is used in mammalian genomes to silence transcription. Histone methylation, another epigenetic modification, alters the accessibility of chromatin to

transcription factors. Methods to determine these changes have been used previously to identify novel regions of transcription in the mouse genome, leading to the discovery of thousands of highly-conserved ncRNA, many of which were also found to have potential regulatory roles due to their association with PRC2 [63, 64]. There are bioinformatic limitations as well, especially if a reference genome is not available. In one study, more than 7.4 GB of high-quality sequence data was obtained from mouse liver RNA, and using a state-of-the-art *de novo* alignment software package the authors were only able to match the transcriptome assembly results of the reference-based transcript assembler utility Cufflinks [31]. Such a study may prove cost-prohibitive for many researchers at present, as current pricing for high quality RNA-Seq data is roughly 10x that of a microarray experiment per sample, which makes studies with very large numbers of biological replicates intractable.

To date, no large-scale studies of human breast tumors using RNA sequencing have been published, though RNA-Seq data from some breast cancer tumor cell lines have appeared in a study of transcriptome changes by Wang et al [30]. The splicing-aware alignment software package MapSplice was validated on breast cancer tissue RNA-Seq data [47], however these data were not made publicly-available. In this thesis, we present analysis of RNA-Seq data from 53 primary breast tumors. Novel transcriptome changes among established and non-

canonical subsets of breast cancer are studied, as well as the identification of novel non-coding RNA. We present methods both to classify breast cancer RNA-Seq data based on consensus clustering and identify these transcriptome changes.

Chapter 2: Thesis Goals

2.1 Show RNA-Seq is an effective way to study breast cancer biology

RNA sequencing has previously been shown to be an effective method for measuring the expression of transcripts in total RNA samples [28, 37], as well as the differential expression of these transcripts between two or more samples [37, 65]. It has also been used to identify novel alternative splicing patterns in tumor and normal tissue [30, 37]. One would therefore expect that RNA-seq data should be able to classify breast cancers into clinical subclasses (such as by ER and HER2 status). Furthermore, because RNA-seq presents a more complete picture of the transcriptome, more subtle differences between subclasses, such as alternative splicing and alternative promoter usage, should also be accessible using this technology. It may also be possible to identify novel coding and non-coding transcripts which are differentially expressed in breast cancer subtypes.

2.2 Develop methods to identify new breast cancer biology

In order to separate breast cancer tumor RNA-Seq samples into clinically-relevant subtypes, we will describe a suite of clustering software called ConsensusCluster [25] which we have developed and made freely available. ConsensusCluster uses the consensus ensemble clustering method [23, 24] to identify robust, reproducible subclasses in data which are resistant both to sample and feature perturbations and to biases present in the clustering algorithms used. This method combines results from many iterations of well-established clustering algorithms such as k-means and self-organizing map [17], reducing the effect of bias from each algorithm used, such as spherical clusters in the case of k-means [18], as well as stochastic variance from random initial conditions. Each iteration involves a fraction of the total sample and feature set, mitigating the effects of outliers and ensuring the resulting clusters are robust. We also developed and released CudaConsensusCluster [66], a powerful extension to ConsensusCluster which enables offloading of computationally-intensive clustering algorithms to graphics processing units (GPUs), using NVIDIA's C for CUDA architecture [67, 68]. Due to the inherently parallel nature of graphics processing, modern GPUs have a massively parallel architecture. This is particularly advantageous to consensus clustering, given that several steps, namely subsampling sample and feature data, the clustering iterations,

and several steps within the algorithms themselves, are all “embarrassingly” parallel operations, resulting in significant speedup in overall runtime (see Appendix: ConsensusCluster).

Identifying alternative splicing and alternative start site usage in breast cancer subtypes is a challenging task, requiring the identification of differential transcript usage both in known transcript isoforms and novel ones. To accomplish this, we have developed a method for identifying regions of putative differential splicing in known transcripts, which can then be rigorously annotated using manual and computational methods to identify the underlying transcript changes. This method identifies changes in the ratios of exon expression between two sets of samples within a known transcript. In other words, it can determine whether there are two exons in a given transcript which have vastly different expression ratios between the two subsets (details in Chapter 3.10). Significantly large differences indicate potential splicing, which can then be identified by assembling sequence fragments at that locus, followed by in vitro validation. We have discovered and validated transcripts which are alternatively spliced, or the result of alternative promoters, in known clinical breast cancer subtypes. We examine the results of this study in Chapter 4.

We also developed a second method to identify alternative splicing in sets of

samples unassociated with clinical subtype. This allows us to render a more complete picture of variation within the samples, uncovering new subclasses with alternatively spliced biomarkers. We first scale the total expression of all exons in a transcript to be the same over all samples, i.e., a scaling factor is determined by dividing *whole transcript* expression by the mean over all samples. The expression of *each exon* is then multiplied by this scaling factor (See Chapter 3.11), which allows for direct comparison of exonic expression between samples, irrespective of overall differences in transcript abundance. The exome is then screened for highly variant expression in re-scaled exons, which indicates putative transcript variation at that locus. As before, this is followed with an assembly step using manual and computational methods to identify novel transcripts responsible for the observed variation. We identify a number of these alternative variants and examine the results in Chapter 4.

Finally, we demonstrate a novel method to discover long, intergenic non-coding RNA (lincRNA) by comparing expression data in intergenic regions between breast cancer subtypes. Total expression in small, tiled “windows” across all unannotated regions of the genome is tabulated for each subtype, and compared directly using a signal-to-noise ratio, a strict method for identifying changes in mean intensity while being robust to sample variation [69]. These regions are then annotated and compared to known ncRNA collections [55, 60, 70, 71]. While

previous studies have successfully identified ncRNA in mammalian genomes [63, 64], the challenge to identify functional transcripts remains difficult. We propose that long non-coding RNA robustly expressed in a set of breast cancer tumors that share clinical characteristics, and absent or nearly absent in others, are very likely to be relevant, though further experimental validation is necessary to confirm this assertion. Additionally, we show in Chapter 5 that

2.3 Identify novel transcriptome variation in breast cancer tumor subsets

We will demonstrate that ConsensusCluster is a powerful tool to distinguish breast cancer subtypes based on ER and HER2 status using RNA-seq data, with excellent concordance to labeling using standard clinical protocols (See Chapter 4.1 for method description and figures). The clusters identified in this way form the foundation on which the other discoveries of splice variants and non-coding RNA is based. Using methods outlined above, we observed differential splicing in subclasses of our tumor sample sets. Specifically, 7 alternative splicing and alternative start site variants were observed to be differentially expressed between ER+/HER2- and ER+/HER2- samples, 2 of which were entirely novel, and a total of 4 had not been previously associated with breast cancer. We also

observed an additional 6 highly differentially expressed variants in subsets of breast cancer unassociated with ER or HER2 status. These may be biomarkers for as yet unknown breast cancer subsets, though further study of their association to clinical parameters is needed. Lastly, we identify a total of 14 ncRNA variously expressed in ER+/HER2- and ER-/HER2- breast cancer tumors, 4 of which are entirely novel, and 2 of which have transcript assemblies vastly different from previous annotations (See Chapters 5 and 6 for detailed results).

2.4 Extensions of the presented work

We have developed and applied methods to identify transcriptome changes in breast cancer tumor subsets, however, it is unclear what the significance of these changes may be. What is clear is that these changes are biomarkers, strong, unique indicators of the presence of representative subsets of the breast cancer population, which may be treated in a specialized manner. There are two clear follow up directions in which we can proceed: 1) Determine the biological basis of these alternative splicings and ncRNA within and between breast cancer subtypes. 2) Determine whether their inhibition or promotion has a measurable phenotype and whether this may suggest novel therapeutic targets.

In order to determine the functional annotations for discovered transcriptome variation, the integration of additional sources of information is crucial. The web of causal interactions that give rise to observed changes can be explored in a number of different ways. For instance, changes in splicing are likely either the result of a specific splicing factor acting in concert with enzyme activity associated with ER status, or epigenetic changes which enhance or silence splicing factors at that locus. The former might be accomplished by knockdowns of known splice factor genes and observing whether alternative splicing continues, however the latter requires a detailed investigation of the epitome at that locus, including methylation and splice promoter sequence mutations. Alternative start sites are generally the result of variation in promoter usage, which implies that in differentially expressed samples there exists epigenetic silencing of one promoter and exposure of another. These changes may be a result of histone methylation changes [54], and as such, genome-wide chromatin state maps would be extremely useful in this study. Alternative promoters may also be the result of transcription factor binding, which can be elucidated using a ChIP-Seq approach [72]. Similarly, many long non-coding RNA have been recently implicated in association with the polycomb repressive complex (PRC2), which causes targeted downregulation [57, 61, 63, 64]. In [61], RNA-PRC2 complexes were precipitated, followed by sequencing of the complete mRNA transcript bound to PRC2. Their method resulted in the identification of over 9000

bound RNA transcripts in mouse cells, and can be used to determine which of the ER status-dependent ncRNA are actively involved in downregulation.

The possibility of using the splice variants for therapeutics is always a primary consideration when novel biomarkers are discovered. Though ER+ tumors are treated with Tamoxifen, only ~40-60% respond [3]. Similarly, ER- tumors have no known therapeutic target. Therefore there is still a pressing need for additional targets for cancer therapy. Through knockdown experiments on tumor cell lines, each of these variants and ncRNA can be assessed for their ability to repress growth and metastasis of tumor cells. In addition, other clinical factors such as abnormal cell morphology can give clues as to the potency of these transcriptome changes as targets. By enumerating the transcriptional variability that distinguishes breast cancer tumor subpopulations, we hope to have identified key features which either form the underpinnings of known disease pathways, or represent the lynchpins of disease phenotypes, which can be targeted in the clinic to reduce recurrence and improve survival.

Chapter 3: Data Description, Methods, and the State of the Art

3.1 Sample data

Total RNA from primary breast tumors, which was collected from patients treated at the Cancer Institute of New Jersey (CINJ) and the Norwegian Radium Hospital in Oslo, Norway, was used in our sequencing study. These RNA samples were analyzed in two separate sets which differed in the style and quality of sequencing. RNA for the first sequencing set, referred to as sample set A, came from a cohort of 13 patients from CINJ and 16 from Radium Hospital. RNA for the second set, sample set B, came from a cohort of 24 patients from CINJ and an additional 6 “normal RNA samples” from reduction mammoplasty patients from Radium Hospital. RNA was extracted using the Trizol reagent [73], which is used to capture total RNA and DNA from cell lysate. Trizol extraction is more difficult than column-based methods, however many column methods filter sequences smaller than 200bp, resulting in an incomplete RNA pool. Total DNA was also extracted from each sample for follow-up research purposes.

The standard RNA-Seq protocol recommended by Illumina was used to sequence the total RNA from each sample. First, mRNA was isolated from total RNA using poly-T bead purification, which is used to purify polyadenylated RNA fragments in a sample, removing ribosomal RNA as well as immature pre-mRNA. Purified mRNA fragments were sonicated to shear into fragments of ~150bp. A cDNA library for the sheared fragments in each sample was created using PCR and random hexamer primers. Hexamers are 6 nucleotide fragments which anneal to the RNA in a sample and provide a basis for elongation using DNA polymerase. All possible combinations of 6 nucleotide fragments are used to provide a consistent amplification. Following this, 50bp adapter fragments were ligated to cDNA in the sample in order to label the DNA as belonging to each sample, and all cDNA plus adapter sequences were run through agarose gel. The gel was cut at approximately 200bp to ensure that the captured fragments were the appropriate size for sequencing. A second amplification step was then performed using adapter-specific primers, amplifying fragments that underwent successful ligation. A third and final amplification step was performed on an Illumina cluster station, to hybridize target DNA fragments to a “flow cell,” on which they become amplified to form surface-bound “clusters.” Clusters on the flow cell surface were then read by the sequencer. Samples from sample set A were then sequenced using an Illumina Genome Analyzer IIx at a read length of 29bp, whereas sample set B fragments were sequenced on an Illumina HiSeq

2000 at 100bp. All samples were sequenced at the Mount Sinai School of Medicine (MSSM), in Dr. Ravi Sachidanandam's laboratory. Samples from both datasets were sequenced using a single-end, unstranded protocol.

Because of the short read length limitations of current RNA sequencing technology, certain protocols have previously employed a “paired-end” sequencing method [31, 37]. The method itself is identical to a single-end sequencing protocol, however following initial sequencing the fragments are then reversed and sequenced from the opposite end of the fragment, leading to sequence reads from both ends of the same RNA. Paired-end sequencing provides an abundance of useful assembly information, including not only the additional sequence configuration but also physical location as well. Since the RNA was sheared after purification, mate pairs must come from loci ~200bp apart, falling into a Gaussian distribution of distances with mean 200. This information can dramatically reduce ambiguity regarding complex transcriptome structure resulting from sequence inversions, alternative splicing, and chromosome fusions among others.

Another class of sequencing protocols which are increasing in popularity are strand-specific protocols, reviewed in [74]. In contrast to unstranded sequencing, where the 5' and 3' end orientation is lost during cDNA library generation, in

these experiments strand information is preserved before sequencing, such as through 5'/3'-specific adapter ligation to the mRNA transcript. In many complex genomic loci, such as the GNAS locus in the mouse and human genomes [75], there exists multiple transcripts which have alternate orientations with respect to the genome, i.e., a “sense” transcript and an “anti-sense” transcript. A strand-specific protocol would assist post-sequencing alignment in resolving the assignment of fragments to transcripts even if these sense and anti-sense transcripts directly overlapped. During novel transcript annotation, unstranded protocols can lead to ambiguity as to the orientation of the transcript. Stranded protocols also substantially reduce the work involved in *de novo* sequence assembly, since the assembler only need consider a single orientation when attempting to align each fragment to assembled contigs [76].

3.2 Sequence assembly

Once sequence fragments are obtained, the task becomes to assemble these fragments into a cohesive picture of the transcriptome. As previously discussed, there are two primary ways to perform sequence assembly: with, and without a reference. Transcriptome assembly without a reference genome sequence requires significant resources to ensure that there is sufficient coverage over all

relevant loci, that is, enough fragments overlap a region to call the sequence with some measure of confidence. This is especially important where sequences diverge, such as at exon-exon splice junctions, where the sequence following some common exon might be selected from a large pool of follow-up exons. Two popular options for *de novo* sequence alignment are Abyss [31] and Trinity [76]. Both methods assemble fragment reads into unique contigs of size *k* (*k*-mers), and then compile a de Bruijn graph of the resulting dataset. Following this, the graph is trimmed to remove errors stemming from imperfect sequence information by filtering short contigs that end prematurely, and complete sequences are assembled from connections along the graph. Using a de Bruijn graph to store sequence information has the advantage that computational memory usage is proportional to the size of the genome, rather than the sample libraries themselves, which are in general substantially larger. Both utilities can also take advantage of paired-end and strand-specific protocols, which provide useful additional information to guide assembly.

Reference-based transcript assembly makes use of a reference genomic DNA sequence common to the species from which the target sample was obtained. Assembly begins with an alignment step to map sequenced transcripts to their locations on the reference genome. This is done using a splicing-aware utility such as TopHat [49] or MapSplice [47]. In general, these utilities operate by

dividing sequence fragments into smaller pieces, which are aligned individually to the reference genome using an end-to-end alignment utility such as Bowtie [32]. Fragments which are separated by smaller physical genomic distances (typically less than 500kb) become candidates for putative spliced alignment. From here, alignment software utilities may either attempt to discover the splice junction via sequential search over the intervening region, or instead, restrict the search to known (canonical) junctions. A splice site usually begins with bases GT and ends with AG with respect to genomic DNA (5' to 3'). In mammals, 98.71% of known splice junctions are GT-AG [36]. Both guanine bases are part of the coding sequence, transcribed at the edges of the two exons of the splice junction. The thymine and adenosine bases of the GT-AG junction form the 5' and 3' boundaries of the intron, respectively. Non-canonical junctions include GC-AC (0.56%) and other junctions that are found even more rarely. By restricting search to canonical junctions, spliced alignment can proceed rapidly.

Once all junctions are defined, transcript assembly can be carried out manually or using a dedicated transcript assembler, the most popular of which is Cufflinks [37]. Cufflinks follows contiguous sequences as long as there is sufficient coverage along every base of the transcript, and where gaps in the transcript sequence are present, the program will leverage paired-end fragment information to guide assembly. Since paired-end protocols involve sequencing at both ends

of a longer cDNA fragment, fragment pairs are guaranteed to have originated in the same mRNA transcript, and thus the assembler can safely join contigs to which each fragment belongs. By the same token, single-end reads provide no additional location information, and any gaps in coverage will cause Cufflinks to assume the transcript does not cover these portions. This can pose serious challenges during assembly, since a novel transcript with low coverage might appear to be “broken up” into separate transcripts. This necessitates manual annotation, i.e., joining nearby transcripts based on assumptions of missing coverage, or prior knowledge, such as the observation that the region is ungapped in other samples. Cufflinks can also take advantage of strand-specific sequence information to resolve loci with overlapping transcripts of different orientations.

3.3 Abundance estimation

When the completed transcriptome is assembled, the total abundance of each isoform, or any transcriptome element, can be estimated on a per-sample basis. This process typically involves first counting the number of sequence fragments in a given sample which align to the target features, followed by normalization. A number of accepted methods for fragment counting exist, the simplest of which

was introduced by Cloonan et al [29]. This methodology involved simply counting all fragments which aligned end-to-end entirely within an exon as a hit, and ignoring all others. It can be argued that counts generated using this method are incomplete, since fragments that align across splice junctions are not considered. A more complex counting method, employed by both HTSeq [77] and BEDTools [78], requires spliced alignment data. This method counts all features overlapped by any portion of a sequence fragment, which improves counting accuracy over the previous method. Unfortunately, neither counting method attempts to identify the transcript of origin for each fragment. The software utility Cufflinks [37, 79] attempts to estimate the abundance of a transcript by calculating the likelihood of an abundance value given the fragments in the sample and the known assembly information at that locus. This method is much more effective with higher quality sequences, especially paired-end sequence information [79], since fragments can then be assigned more accurately to transcripts, which increases the robustness of the results.

3.4 Normalization

Normalization of count data is a necessary step in order to accurately gauge the relative abundance of transcripts among two or more samples. Recent RNA-Seq

analysis protocols begin this step by scaling the counts in each sample by the total number of fragments in a sample, referred to as the sequence depth [28, 29, 79]. It has been argued that this parameter is more accurately defined by the number of fragments that align anywhere on the genome [37, 65], as this will reduce error due to systematic sequencing errors or sample contamination. However, Robinson and Oshlack [80] noted that this value is artificially inflated by heavily overexpressed transcripts. For example, HER2+ cancers have extremely high expression of genes in the HER2 amplicon due to high copy number at this locus, such as ERBB2. In these samples, ERBB2 occupies a large proportion of the total sequencing “real-estate,” which artificially depresses the proportion of transcripts in those samples which belong to other genes. The solution is to utilize an “effective” sequence depth based on the total sum of all fragments that align to transcripts that occupy the middle 75% of the expression curve, chosen using a trimmed mean approach. Similar solutions are employed by both DESeq [65] and Cufflinks [37] in their normalization protocols.

Additional confounding factors which require normalization are feature length and biases introduced by the sequencing protocol itself, namely the use of random hexamer priming. As expected from experiments in which fragments are selected randomly from the transcriptome, the expected count of any feature is proportional to the length of the feature in question. Thus, it is natural to scale the

count data for each feature based on the relative length [28, 29, 37]. However, it has been shown by Oshlack and Wakefield [81] that this introduces a bias to subsequent data modeling in that the variance of the gene is no longer equivalent to the mean, as expected from a Poisson process. As a result, genes which are longer are more likely to be found to be differentially expressed regardless of the statistical method or cutoff chosen. Bullard et al [82] pointed out by scaling the test statistic by the square root of the length this effect disappears, though Type I error is increased in the process.

Random hexamer priming is used in many RNA-Seq protocols to convert the fragmented RNA library into cDNA. Because the cDNA library must provide an accurate representation of the transcript abundances present in the original sample, it is imperative that cDNA library generation be an unbiased sampling of the total mRNA pool. Hansen et al [83] showed that strong biases exist in the likelihood of observing a given base in a given position at the start of each sequence fragment which were significantly greater than the background distribution. Intriguingly, the bias fell off as a function of the distance from the start of each fragment. The authors also showed that this effect was reproducible across several studies, among which the most common factor was the use of random hexamer priming. It was concluded that a preference for elongation exists for hexamers of a certain configuration, and that this effect could be

accounted for using a simple linear weighting of each fragment towards the total count based on the relative proportion of its starting 6-mer against the background. This bias can be optionally accounted for in the Cufflinks utility during transcript abundance estimation [37].

3.5 Differential expression

In order to extract meaningful conclusions from feature abundance comparisons between samples, an appropriate statistical model of the data is necessary. RNA-Seq fragments are discreet counts resulting from a sampling of the total mRNA pool, and as such are frequently modeled as a Poisson process [84, 85]. Researchers Marioni et al [84] showed that the count distribution was effectively modeled using a Poisson distribution, and used this assumption to identify variability across technical replicates. However, over multiple biological replicates gene-level count data often shows high variance compared to the mean, which is not what is expected for a Poisson process [37, 65]. This indicated that additional parameters are needed to account for this over-dispersion. To date, methods have been suggested to model the count data include using a generalized Poisson model [86] or a negative binomial distribution [37, 65, 87], which results in greater statistical power to detect differential expression over the naïve

Poisson model. Software utilities which employ these methods include DESeq [65] and EdgeR [87].

These models are appropriate for gene-level count data, however, they cannot be used for isoform differential expression. The reason for this is that fragments which align to a region which is overlapped by multiple transcripts at a given locus cannot be assigned to the correct transcript, and thus isoform-level count data at each locus is inaccurate. As previously discussed, paired-end and strand-specific sequencing protocols are both RNA-Seq methodologies which provide additional transcript-specific information, both of which can be utilized by the software utility Cufflinks [37]. Cufflinks attempts to assign relative abundance estimates (fragment counts) to known transcripts using assembly information by calculating the maximum likelihood of these estimates given observed fragment count data at that locus. It can therefore perform differential expression analysis of individual transcripts using RNA-Seq data, which is done by modeling estimated counts as a negative binomial distribution and calculating a test statistic similarly to DESeq [65]. Though this method is appropriate for low numbers of samples, the uncertainty of isoform abundance estimation through this method is large, especially for data from single-ended protocols. The result is that the total variance of a comparison between two large sample pools must be the sum of both biological variation and the uncertainty of isoform abundance

measurements. This variance becomes intractable at high numbers of samples, which prevents accurate assessment of differential expression. We present a method in Chapter 3.10 which addresses these concerns.

Given the limitations inherent in fragment assignment to specific transcripts especially at complex loci, alternative experiments to quantify transcript abundances should be considered. The most popular method is quantitative real-time PCR (qRT-PCR) [88-90]. Quantitative PCR is an experiment wherein a specific transcript is amplified using a PCR reaction in solution with a double-stranded DNA-binding dye, which fluoresces upon binding. This fluorescence is measured over a number of PCR cycles which form an exponential curve based on both the abundance of the target transcript in the sample and how efficiently the primer is elongated [90]. After properly normalizing for the total cDNA in the sample and the primer efficiency (assessed on a per-plate basis using a standard dilution curve [88]), the number of cycles required before the rate begins to decay (called Ct), is proportional to the relative abundance of the target transcript. In order to control for experimental error, Gutierrez et al [89] recommend at least 3 technical replicates of each sample be performed and their results averaged. Finally, differences in RNA/cDNA quality may cause large changes in expression from sample to sample, and so when assessing differential expression, stable, highly-expressed “housekeeping” genes must be included as a control in each

experiment [91].

3.6 Identifying alternative splicing in RNA-Seq data

Currently, several methods exist for identification of alternative transcript variants in RNA Sequencing experiments. The basic steps include obtaining a reference annotation of transcript variants in a sample, followed by estimation of abundance and testing for differential expression. Compiled annotations exist for several genomes, including the human genome, provided by ENSEMBL [55] and RefSeq [70]. ENSEMBL combines annotation information from a wide variety of sources, including the Vertibrate Genome Annotation (VEGA) project [71], which provides manual annotation of novel transcripts over various species, and the ENCODE project [92], which aims to enumerate all functional aspects of the human genome. RefSeq is a heavily curated database which also combines transcript information from these and other sources, however, transcripts are only included in the RefSeq annotation if there is substantial evidence to validate their existence. Thus, RefSeq provides fewer transcripts than ENSEMBL, though the annotation is probably more robust [70]. Reference annotations can also be generated on a per-sample basis using the transcript assembly methods described above, which may reveal novel transcripts not present in currently

available databases. Alternatively, the RABT assembler, provided with Cufflinks [93] can improve the accuracy of assembly by combining novel transcript information with a supplied reference annotation from other sources.

Once annotation is complete, abundance is calculated and normalized as previously described. If subsets are predefined in a given sample pool, and sufficient biological replicates are included which represent these subsets, a differential expression calculation is all that is required to identify differential alternative splicing between these subsets. Unfortunately, the problems in abundance estimation described above are still present in that fragments cannot be easily assigned to their transcripts of origin. In single-ended, strand non-specific RNA-Seq protocols, as is the case in the present study, this problem is even more pronounced. While these concerns might be alleviated using a paired-end protocol and higher read depth, these procedures are considerably more expensive at present and may make studies with large numbers of biological replicates cost-prohibitive. The only currently-available method for identifying isoform-level differential expression across subsets is the software package Cufflinks [37]. Due to the large number of samples in our dataset and the uncertainty inherent in Cufflinks' method of isoform abundance, Cufflinks failed to identify any differentially expressed alternative transcripts across ER+/HER2- and ER-/HER2- samples in the present study. By contrast, our method (Chapter

3.10) was able to identify a number of putative variants alternatively expressed in these subtypes, two of which were chosen for and subsequently validated with RT-PCR (Chapter 4).

Another method previously used in RNA-Seq experiments to identify variants is alternative expression analysis by sequencing (ALEXA-Seq) [134]. ALEXA-Seq is a method that works by first creating a database of possible transcriptome expression variations by starting with ENSEMBL [55] and annotating and enumerating all possible features using known exon information, e.g., all introns, exons, alternative exon boundaries, and all possible exon-exon junctions including those which are unsupported in current annotations. Variant features are then determined by measuring the expression of individual features in each transcript from the assembly, and then identifying potential variants by attempting to find outlying feature expression levels in each transcript. When comparing transcripts between two sample libraries, feature expression is normalized to the total transcript expression level, which allows direct comparison of feature expression to determine changes. Large changes in exon measurements indicate, for example, the presence of exon-skipping events between two libraries. This process can be repeated for intronic features as well as splice site usage to enumerate all notable transcriptomic changes. ALEXA-Seq does not allow for multi-sample comparisons, and is thus inappropriate for identifying

variants across large studies with multiple biological replicates. We will show in Chapter 4, however, that a similar method can be used to identify variants that are differentially expressed in novel subsets of breast cancer, by normalizing exonic expression in all samples to a common total transcript expression level and then searching for outliers in individual exons.

3.7 Identifying ncRNA in RNA-Seq data

A completed transcriptome assembly, generated using one or more of the methods described above, will necessarily contain transcripts that map to non-coding regions of the genome. In an intensive study of the mouse transcriptome, Carninci et al [94] showed that over half of the transcriptional units (TUs) surveyed mapped to non-coding regions. In order to distinguish between RNA from coding and non-coding regions, sequence classification methods must be employed. In the most complete annotation of non-coding sources to date, Jia et al [60] employed an open reading frame (ORF) and BLASTP predictor to filter non-coding transcripts from expressed sequence tags (ESTs), short (500-800bp) fragments sequenced from tissue cDNA libraries, obtained from an earlier study of human transcriptome features [95]. Transcripts were screened for ORFs that did not contain stop codons, indicating they could be protein coding. Additionally,

the putative amino acid sequences of all ORFs found in the data were compared against known protein databases using the BLASTP tool [96], which searches for sequence conservation over various species. Amino acid conservation in an open reading frame indicates coded protein functionality, which were removed from consideration in their ncRNA database. In total, 5446 RNAs from ESTs were predicted to be non-coding, of which only a small number overlapped with ncRNA from previous annotations [60]. An alternate method was suggested by Kong et al [97], which utilizes a support vector machine (SVM) trained on sequence features in both coding and non-coding sequences, such as the number of open reading frames, the length of each frame, the presence of a start codon, and the presence of conserved protein domains. This SVM can be used to classify new sequences into coding/non-coding with high computational efficiency.

RNA sequencing studies which identify putative non-coding RNA transcripts can utilize these pipelines to verify that their transcripts do not code for protein. However, while much of the transcriptome is non-coding, early reports showed that many ESTs mapped to the human transcriptome showed weak sequence conservation over multiple species, leaving the functionality of these transcripts in doubt [63]. In two studies, specific histone modifications were shown to correlate with conserved regions of the genome [63, 64], indicating they served functional purposes. In the present study (Chapter 5), we will show that some

non-coding transcripts differentially expressed between subtypes of breast cancer are also heavily conserved over mammalian genomes, implying they are functional.

3.8 Problems with state of the art methods for identifying alternative splicing

The most pressing challenge for current research into transcriptome variation is the dependence of both reference-based and de novo assemblers on very high quality sequence information. For example, Robertson et al [31] used 174 million 50bp paired-end sequencing reads in order to assemble a mammalian transcriptome using the ABySS [98] software utility to the level of reference-based assembler Cufflinks [37]. The previously-unprecedented sequence depth used has a current list price of ~$5400 a sample [99]. The cost of such a study with multiple biological replicates, such as would be necessary to quantify splice variants in cancer samples (which have high heterogeneity), would be prohibitive. Cufflinks itself has only been shown to provide cohesive annotation using samples of ~60 million 75bp paired-end reads [37], which would cost in excess of $2000 a sample (price based from a listing at Otogenetics [99] for 40 million 50bp paired-end sequence reads). In experiments which utilize lower quality

sequencing information for larger sample pools, such as single-ended read protocols, there is a need to develop novel methods identify potential variants.

The other current state of the art method for identifying novel potential transcript variants, ALEXA-Seq [134], suffers from the same stringent sequence quality requirements as Cufflinks and *de novo* assemblers. In this method, a library of putative transcriptome features is first created from known annotations. Following this, a list of all features that are supported by the sample library and their expression is created, after which the expression of each feature can be pairwise compared between samples. ALEXA-Seq has a number of important drawbacks, most notable of which is that differential expression is only possible between two samples, which precludes the use of biological replicates. The software as currently available also does not support data from single-end read RNA-Seq experiments. However, the methodology is simply implemented, and we discuss a method in Chapter 3.12 which extends their algorithm to multi-sample comparisons.

3.9 Gaps in our knowledge of differential breast cancer transcriptome changes across subtypes

While many studies have presented splicing changes in breast cancer tumor samples [44-47, 100, 101], fundamental questions still remain. The most significant advances in our understanding of alternative splicing in breast cancer have come from studies of changes from normal tissue or benign lesions to malignancy [44, 45] and from tumors with poor metastatic capability to more aggressive phenotypes [100]. However, recent studies have shown that “breast cancer” is not simply one disease but a highly heterogeneous population, composed of the clinical subdivisions ER+/HER2-, ER-/HER2-, and HER2+ tumors, as well as further subsets of these with distinct molecular and phenotypic signatures [9, 14, 19, 20, 22]. These subclasses also have very different survival rates and response to treatment [9], suggesting that the biological changes underpinning these subclasses represent important targets for future therapeutic intervention. At the same time, “normal” breast tissue is generally highly heterogeneous as well [48], and is composed of milk ducts, lobules, connective tissue, and fat. It is therefore prudent when conducting a study of molecular changes, including alternative splicing, to reduce heterogeneity among sample populations to obtain a clearer picture of the biological changes taking place. We propose to do this by focusing on the differences between breast cancer disease subpopulations, which are less heterogeneous overall.

Two studies analyzed alternative splicing changes between breast cancer tumor

subtypes [46, 47]. Wang et al [47] counted exon skipping events on four ER+ and four ER- samples, composed of two tissue and two cell lines, each. The authors only showed that the quantity and location of some exon skipping events varied between the two subtypes, and did not report genes differentially regulated. Furthermore, these data were not made available to the public. Lapuk et al [46] is the only comprehensive study using multiple biological replicates which attempts to elucidate splicing changes between breast cancer subtypes, which was accomplished using an Affymetrix Human Junction Array, a custom, non-commercial array which was developed specially to tile across known exon-exon junctions in the human genome. However, this study has several shortcomings. First, while human breast cancer cell lines are a popular surrogate for cancer cells *in vivo*, there are marked differences in gene-level expression as well as copy number variation between cell lines and tissue samples. Using DNA microarray data Ertel et al [102] showed significant differences in genes relating to a number of pathways, including energy production, cell communication, cell cycle, and metabolism of nucleotide and cell signaling molecules. To date, no studies have been published comparing splicing changes between cell lines and tissue samples, which may impact results. Additionally, many genomic copy number changes are associated with cell lines that occur rarely in tissue samples [103]. Given these differences, a study of alternative splicing changes using breast tumor tissue samples (such as presented here) seems both warranted

and relevant. Another challenge to interpretation of results from Lapuk et al [46] is in regard to the use of a microarray for analysis. Microarrays are limited to predefined probesets, and thus cannot capture all aspects of the transcriptome, especially those resulting from novel changes. Microarray analyses are also hampered by cross-hybridization between probes [33], which contribute to difficulties reconstructing isoform models from data [104]. A previous study has also directly compared the Affymetrix HuEx exon array [12] with RNA sequencing data using proteomic results as a standard, and reported that RNA sequencing results better estimated absolute isoform abundances in their samples [105].

3.10 A novel method to identify differential alternative splicing in sample subsets

We present novel algorithms for locating putative differential splicing changes across and within subtypes in a sample population. Each method begins with raw fragment sequence information from both sample set A (29bp reads) and sample set B (100 bp reads - see Chapter 3.1). These data originated from Illumina sequencing of total RNA using single-end, strand non-specific sequence fragments. These fragments were aligned to the human reference genome, revision hg19 [106], using the TopHat spliced sequence alignment software

package [49]. TopHat aligns sequence fragments to reference exon junctions. When fragments fail to align contiguously and are not in the RefSeq annotation, they are identified as putative novel splice junctions. Reference splice junctions were provided to TopHat from the most current UCSC RefSeq human genome annotation [70], which is a curated database of known transcriptome annotations. For each sample, the coverageBed utility [78] is used to count all reads which overlap exonic features from the RefSeq annotation. Reads which span exon-exon junctions are counted as a hit for both features, ensuring count information most closely represents actual coverage.

Normalization of raw count information is an important step to being able to compare feature abundance levels between multiple sample libraries. First, counts are scaled linearly to account for preferential elongation of certain 6-mers during cDNA library generation, as suggested by [83]. All count data is then scaled by the total number of fragments in each sample which align to the human genome (sequence depth). The depth parameter in this calculation is adjusted using a trimming method suggested by Robinson and Oshlack [80] to remove biases due to heavily overexpressed transcripts. Each exon is then normalized by its length due to the increased likelihood for a random sampling of sequence fragments to originate in longer transcript features, proportional to their length [82]. Finally, the counts are log2 re-expressed [29, 31, 84]. It should be noted that

each of these normalization methods were also used to calculate isoform abundance in results which utilize Cufflinks' [37] estimation method, though these results were not used to calculate differentially expressed isoforms except to compare the success of these methods.

To identify potential differential alternative splicing between known subclasses, we developed a test statistic based on the observation that transcript exonic expression is correlated in these subsets if the ratio of composite transcripts for that gene is the same in both sets. This can be intuitively described with the following Gedanken experiment: A genetic locus with uniform fragment coverage from RNA sequencing experiments contains two transcripts, Alpha and Beta. Alpha and Beta exhibit alternative splicing, which indicates each contain an exon or exons which overlap similar exon(s) on the other transcript on the same strand. These transcripts are present in a sample in a certain relative, nonzero proportion. Suppose Alpha and Beta each contain one exon which is identical in the other, that is, they occupy the same genomic location and are the same length. After normalization by exon length, the abundance of each exon in Alpha is proportional to the total abundance of Alpha, except for the exon which is shared by Beta, which is increased in RNA-Seq expression due to fragments originating from the Beta transcript in that sample. Thus, given the relative proportion of Alpha and Beta and total gene-level expression at this locus (the

sum of fragment coverage for both transcripts), this sample would have an exonic pattern of expression for each transcript as described above.

Now, suppose the same conditions exist for a second sample, however, the total expression at this locus, the gene-level expression, is significantly larger. In this case, the exonic expression of both Alpha and Beta would be perfectly correlated between samples, having the same pattern produced by the identical relative proportion of transcripts. This indicates that no differential splicing has taken place between the two samples; instead, the locus itself is differentially expressed. It follows that the ratio of exon abundance between the two samples will be identical for all exons. To test for differential splicing, then, all exon abundance ratios in a transcript are compared between two samples, or sets of samples, and the existence of a high difference in ratios for any pair of exons in a transcript indicates that it is a potential candidate for splice variation. For sets of samples, the mean difference is normalized by the square root of the total variance for all exon measurements (eq 1), which reduces the influence of intra-subtype variability on the splicing score, *S*.

Eq 1

$$d_{i,m} = \mu_{s,i} - \mu_{t,i}$$

$$ER_{i,j} = \frac{d_i - d_j}{\sqrt{Var(s_i) + Var(t_i) + Var(s_j) + Var(t_j) + 1}}$$

$$S_m = Max_{(i,j)\in m} ER_{i,j}, i \neq j$$

Where *s* and *t* denote subsets of samples in a population, *m* is an isoform containing exons *i* and *j*, and *μ* refers to the mean expression values for each exon *i and j*, respectively. The exon ratio (ER) is the difference in the difference of means for any pair of exons in a transcript, normalized by the variance of all four exon measurements. This value is conceptually related to the signal-to-noise ratio (SNR) [69], which has been used in microarray experiments as a strict measurement of the difference in means between two sets [9]. Since the data are log2 re-expressed, this translates to the difference of exon expression ratios between the two sets. The splicing score, *S*, is the maximum exon ratio for all pairs of exons in the transcript. Higher values indicate higher divergence from perfectly correlated exonic expression behavior, and may indicate changes in the relative proportion of transcripts present in these two sets.

Two potential drawbacks of this method which result in higher scores for false positive transcripts can be reduced or eliminated using a number of data filtering procedures as discussed below. The method assumes that gene-level expression at a given genetic locus is non-zero for a given subset. If this assumption is violated, the exon ratio will be equal to the largest difference in

exonic expression within the expressed subset alone. To alleviate this, transcripts for which the average exon expression for either subset does not exceed a certain threshold (at least 1.0) are removed from further analysis. Another special case which will result in a false positive occurs in transcripts which contain exons that are entirely unexpressed in both subsets while the locus as a whole (gene-level expression) is increased in one subset versus the other. This can occur in many cases, such as if transcript diversity at a given locus is due to an alternative starting exon and only one transcript is in fact expressed in either subset. For the absent transcripts at this locus, a false positive will be generated and the exon ratio will be equal to the difference in gene-level expression between the two subsets. Other instances where a false positive will occur include exon-skipping events wherein the annotated transcript contains an extra exon, and this exon is absent in all samples. This can be due to incorrect annotation, universal exon skipping in breast cancer samples, or if transcript diversity at that locus is entirely the result of exon-skipping events and only one transcript is expressed in either subset. To eliminate this, we use a simple filter for the total average expression over both subsets in a single exon. Any exons which are effectively unexpressed for both subsets (below 1.0 after normalization) are not considered in the exon ratio.

To assess the significance of splicing scores obtained using this method, a

permutation test is employed. The null hypothesis for each score is that it happened by chance given two random sets of breast cancer tumor samples of size equal to the sizes of the subsets being considered. The following procedure is used to calculate this probability, given a set of transcripts *T* of total number *M*:
All breast tumor RNA-Seq samples are combined into a single set *N*

- 100 bins *B* representing discrete increments in splicing score space from a score *S* of 0.0 through 5.0 are created, each with an initial count *C* of 0
- 1000 datasets are created by randomly partitioning *N* into new subsets of size equal to that of the clinical subsets for which differential splicing is being testing
- For each transcript *t* in each dataset, if the score $S_t > B_i$ for any *i* in *B*, C_i is incremented
- The probability of seeing a score *S* by chance is given by C_i / (1000 * *M*)

  where $i = max_i\ S > B_i$

We account for multiple hypothesis testing to control the Type I error rate using the false discovery rate (FDR) correction suggested by Benjamini and Hochberg [107]. All p-values which correspond to a false discovery rate of less than 0.10 after FDR correction are considered potentially significant and worth validating experimentally.

Transcripts which display high putative splicing variation between breast cancer subsets are subjected to an independent review of their transcript annotations. Frequently, high splicing scores are due to the expression of novel transcripts, e.g., an alternative start beginning in intron 9 for an isoform of ACOX2 (see chapter 4) is differentially expressed in ER+ tumor cells, which results in increased expression of exons 10-15 while exons 1-9 are not differentially expressed between subsets. The intronic start is manually annotated based on coverage information from sequence fragment alignments provided by TopHat [49]. Where possible, a literature search is performed for each putatively spliced gene to identify independently validated assemblies. In transcripts where an exon is reported low in one subtype versus the other, implying a skipped cassette exon, spliced fragments from the surrounding exons are checked to ensure the change is due to an alternative exon junction event.

3.11 A novel approach to identifying potential variants in unknown subsets

There are many situations where sub-populations are not known in a sample set, such as if the set is extremely heterogeneous or if clustering fails to identify

robust subgroups due to noisy, uninformative features. In these cases, a methodology to identify differential alternative splicing without having prior knowledge of these classes becomes necessary. We present such a method which bears similarity to ALEXA-Seq [134], however the method can be used with any number of samples, rather than comparing two samples pairwise. As with our method to identify differential splicing between known subgroups, the data are first aligned to the reference genome using TopHat [49], and alignments within known RefSeq annotations [70] are counted using coverageBed [78]. Raw exon counts are then normalized to an effective sequence depth as suggested by Robinson and Oshlack [80]. These counts are then log2 re-expressed.

Both ALEXA-Seq [134] and our proposed method proceed by normalizing the exon count data by the total expression of the transcript in that sample. This is done to make exon expression directly comparable between two samples. We extend this over all samples, and perform a screen for highly variant exons over the entire dataset rather than comparing samples pairwise. The maximum exonic variance for a transcript defines the putative splicing score for that transcript in a subset of tumor samples. To reduce false positives, samples which have very low expression of the transcript (less than 1.0, average) are not considered when calculating this score. Transcripts with high putative splicing scores are manually annotated based on available splicing data provided by TopHat and compared to

other assemblies of that locus in the literature.

3.12 A method for discovering functional non-coding RNA differentially expressed between breast cancer subsets

It is estimated that more than half of the transcriptome consists of non-coding RNA (ncRNA) [94]. However, these ncRNA are not well annotated [60]. Furthermore, many of the non-coding transcripts discovered to date have very little conservation across species [64]. As a result, their functional relevance has been called into question. To ensure that ncRNA we discover are functional, we impose a strict measurement of differential expression between known clinical breast cancer tumor sub-populations, ensuring that the resulting transcripts are directly or indirectly the result of some systemic biological difference underlying these subtypes. Of these, roughly half also show sequence conservation over all mammals (Chapter 5). We have developed a novel method to identify these differentially expressed ncRNA which we now describe.

We define intergenic regions of the genome as those regions spanning genomic DNA which were not occupied by either exons or introns contained in the latest RefSeq [70] annotation of the human genome. These regions are split into 300bp

“windows” using the complement utility found in the BEDTools software package [78]. We then align RNA sequencing fragments from each sample to these intergenic regions, and the number of aligned fragments is counted for each window using the coverageBed utility [78]. These raw counts are normalized by the effective sequence depth [80] (based on alignments to the complete annotation, as calculated in Chapter 3.10) and log2 re-expressed.

To identify windows which are differentially expressed between subsets, we employ a signal-to-noise ratio (SNR) test [69], which has been shown to be a conservative measurement for selecting differentially expressed features between sample populations in microarray datasets [9]. Windows which represent an SNR ratio of expression between the two sets of at least 0.8 are considered to be putative locations for novel transcripts. To facilitate assembly of all non-coding transcripts that could be associated with differentially expressed windows, all fragments which align to the entire intergenic region which contains the window or windows are assembled separately using three complementary methods. First, the Cufflinks reference-based assembler [37] is used to assemble the combined pool of RNA-Seq fragments from all samples. Next, the Trinity *de novo* sequence assembler [76] is used to form complete sequence contigs from fragments in the same pool, and these contigs are mapped to the reference genome using both Bowtie [32] and TopHat [49]. Manual annotation is then

performed to disambiguate transcripts from these regions using information from raw fragment alignment, the Cufflinks assembly, and the Trinity assembly. Where possible, novel transcripts are compared to existing manual annotations of expressed sequence tags [55, 71] in order to provide evidence for exon junctions the data does not perfectly support, e.g., if fragments align to two separate novel exonic regions and database annotations provides a splice junction between them. However, novel features supported by sequence fragment information (exon junctions, exons) are considered canonical for assembly purposes and subsequent validation.

Recently, Jia et al [60] proposed a pipeline to discern non-coding RNA from translated sequences. We utilize this method to validate novel transcripts as non-coding RNA, which begins by enumerating all possible open reading frames (ORFs) in each isoform of the gene assembly. These potential amino acid sequences—all sequences in each reading frame which end in a stop codon—are used as input in the BLAST [96] alignment tool to identify protein sequence conservation over a wide range of species. Sequences which lack significant homology to any known conserved protein domains are considered non-coding. We also compared these results to those obtained from the Coding Potential Calculator [97], which is a support vector machine (SVM) capable of identifying non-coding sequences based on sequence features such as the

number of open reading frames, the length of each frame, the presence of a start codon, and the presence of conserved protein domains. Each transcript sequence is then assigned a score based on this classification.

The completed assembly of transcript isoforms at each non-coding RNA gene can then be tested for differential expression between breast cancer tumor subtypes using the DESeq software package [65]. DESeq attempts to model gene-level count data as a negative binomial distribution, and tests for differential expression using overdispersed estimates of variance fitted to this distribution. To obtain count data for a completed genetic locus assembly, the program HTSeq-count [77] is used. HTSeq-count assigns fragments to genes if their spliced alignments fall unambiguously within the gene annotation. These counts are then normalized by the sequence depth in each sample, which is adjusted using a methodology similar to [80]. After testing, the p-values are corrected for multiple testing using the false discovery rate ($p < 0.10$ after correction) [107]. The results of this analysis are given in Chapter 5.

Chapter 4: Identification and Validation of Splice Variants in Breast Cancer Subsets

4.1 Classification of breast cancer subtypes using RNA-Seq

Breast cancer is a heterogeneous disease, and can be classified into a number of clinical subtypes with variable patient outcomes and response to therapy. The primary clinical subtypes are ER+/HER2-, ER-/HER2-, and HER2+ breast cancer tumors [9, 14]. In order to divide the 29 tumor samples from sample set A and 24 tumor samples from sample set B into these subclasses, a multi-step method adhering to previous classification measures on DNA microarray analyses was performed [9]. First, HER2+ tumors were separated from other tumor samples based on overexpression of the ERBB2, GRB7, and PPARBP genes in the HER2 amplicon (Figure 1B), which are part of a genomic locus with elevated copy number in HER2+ breast cancers. Expression of these genes was determined by counting sequence fragments which aligned to isoforms of genes at these loci, and then normalizing count data by the effective sequence depth and length of each feature, as described in Chapter 3. Consensus ensemble clustering [25] (see Appendix: ConsensusCluster) of these features was used to separate

samples into “high HER2 amplicon” and a “low HER2 amplicon” groups, which correspond to HER2+ and a combination of ER-/HER2- and ER+/HER2- samples, respectively. Sample set A contained 6 HER2+ tumors, whereas just 2 HER2+ samples were found in sample set B, leaving 23 HER2- samples and 22 HER2- samples in set A and set B, respectively. The rationale for first stratifying samples based on HER2 status and then applying clustering analysis is discussed in detail in [9]. Briefly, the logic is that HER2 status is a driver of disease that supersedes ER status (HER2+ tumors are always more aggressive). Hence, to identify clinically relevant features, one should worry about ER status only after stratifying by HER2 status. The HER2- samples were clustered using all informative RefSeq [70] exons as features, determined using PCA selection [25, 108]. This separated sample set A into 15 ER+ samples and 7 ER- samples, and also separated sample set B into 9 ER+ and 10 ER- samples (Figure 1).

Validation of this classification from RNA sequencing was accomplished by comparing with immunohistochemistry (IHC) and fluorescent in situ hybridization (FISH) status for the samples, which is a routine classification used by pathologists and was available for all the samples. IHC is a tissue staining process which uses a protein antibody to target the estrogen (ERa) and HER2 receptors on breast cancer cells in separate experiments to identify ER+ and HER2+ samples, respectively. Secondary antibodies which bind to receptor-

targeted antibodies are tagged with the enzyme peroxidase which catalyzes the production of a colored precipitate when the appropriate substrate is added. When viewed under microscopy, the level of precipitate production indicates relative expression of the target receptor. Fluorescent in situ hybridization is a method in which a polynucleotide probe is bound to a fluorescent dye, which is visible under a fluorescent microscope. Like IHC, FISH is used both to indicate the presence of a specific cellular component as well as the physical location within the cell. FISH is used in breast cancer classification to indicate the presence of copy number changes, specifically the amplification of the HER2 genomic locus. Classification based on this protocol is necessarily subjective, since the decision is made partly by a trained (but not infallible) pathologist. Hence, one does not expect that RNA seq classification would agree 100% with IHC classification. We found that annotation using FISH and IHC was in 88% agreement with ER status and 80% agreement with HER2 status over the 53 samples when compared to classification based on RNA seq data. Where annotation disagreed between clustering and IHC/FISH, a separate clustering iteration was performed using all samples which were positively identified and each ambiguous sample independently. If the ambiguous sample did not associate with either ER+/HER2- or ER-/HER2- groups robustly, i.e., associated with either cluster with equal probability over multiple bootstraps of the data, it was removed from further analysis. In total, 1 sample was removed from sample

set A and 3 samples were removed from sample set B.

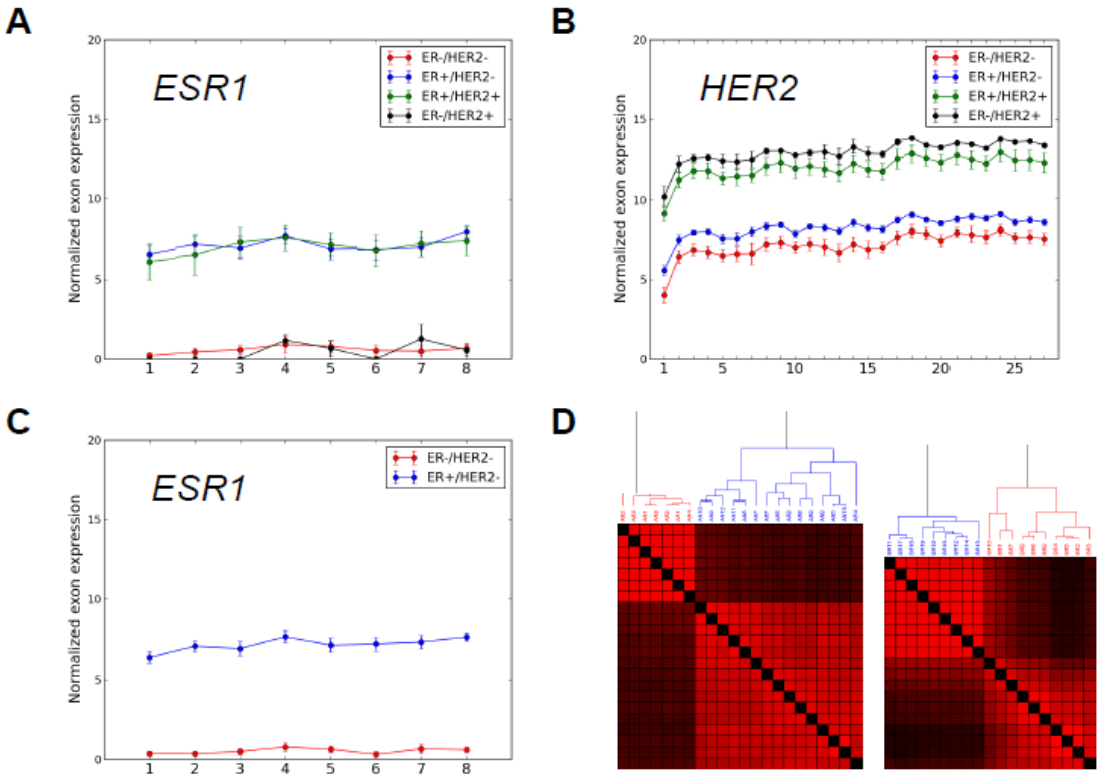

Figure 1

Figure 1. A-C) The exonic overexpression of ESR1 and HER2 (ERBB2) mRNA can be used to distinguish ER+ (A, C) and HER2+ (B) tumors, respectively. In figures A-C, the x-axis represents the canonical transcript exons in order, while the y-axis represents the normalized, log2 re-expressed expression of each feature averaged over all samples from sample set A, colored according to the assigned breast cancer tumor subtype. The error bars indicate standard error of sample expression. HER2+ samples are also separated by ER status for clarity, however, this classification was

not used for analysis. D) HER2- samples from set A (left) and set B (right) separate into ER+ and ER- subtypes when *de novo* consensus ensemble clustering is applied using all RefSeq exons as features. Figure 1D depicts a consensus matrix, which is a symmetric sample by sample representation of cluster association. Lighter squares represent a higher likelihood of row and column samples being clustered together over 1000 k-means iterations applied to bootstrapped sample and feature data. The final matrix is used as a distance matrix for single-link hierarchical clustering, which is used to produce the cluster tree. The cluster tree is colored according to final subtype association. Figure 1D was produced using the ConsensusCluster software utility [25].

4.2 Differentially expressed alternative transcripts in ER+/HER2- and ER-/HER2- breast cancers

We focused on the identification of splice variants differentially enriched in tumor samples based on their ER status, restricting ourselves only to the HER2- sample set. We applied a novel method based on the difference of normalized exon measurement ratios between ER+ and ER- subclasses, which is further normalized by the sample variance, as discussed in Chapter 3.10. This difference is used as a test statistic, the significance of which is measured using a permutation test. Application of this method to the combined HER2- sample pool

from both datasets (24 ER+, 17 ER-), resulted in 301 putative differentially expressed alternative transcripts that distinguished ER+ samples from ER- samples (after false discovery rate or FDR correction). These transcripts were then manually curated and isoform-level quantification was performed using the Cufflinks software package [37]. The top 7 biologically-motivated genes which we identified are presented in Table 1.

Gene Name	Variant Class	Log2 Fold Change ER+/ER- 29bp	Log2 Fold Change ER+/ER- 100bp	Adjusted p-value
TPD52	Alternative start	8.02	2.97	0.002659
NET1	Alternative start	-1.37	-1.97	0.006071
EPB41L1	Skipped exon	3.12	13.68	0.003422
IQCG	Alternative intronic start	-4.01	-1.98	0.003422
ACOX2	Alternative intronic start	8.28	2.95	0.076178
MYO6	Skipped exon	4.24	5.06	0.003385
KRT18	Alternative start	3.23	3.25	0.000967

Table 1

Table 1. The top biologically-motivated genes which contain differentially expressed transcripts between ER+/HER2- and ER-/HER2- breast cancer tumor samples. In column 2, the type of alternative splicing, identified during manual curation, is given. Log2 fold changes of the variant transcripts in set A and set B were estimated independently using Cufflinks [37]. The adjusted p-value for each transcript using a permutation test is also given. P-values corresponding to FDR less than 0.10 were considered significant.

One of the most promising transcripts identified in our study is an alternative start variant of Tumor Protein D52 (TPD52), which is a putative oncogene located on 8q21. This chromosomal region is often increased in copy number in breast tumors, and may in fact be the “driving” gene selected for in these chromosome changes [109]. Copy number variability in breast cancer has been previously implicated in aggressive phenotypes such as drug resistance [110]. We find that a specific alternative start isoform of TPD52 is upregulated in ER+ cancers, whereas a separate “canonical” transcript is expressed equally in both subtypes (Figure 2A-B). This raises the possibility of different phenotypic “goals” for increased copy number in this region depending on tumor subtype. Other researchers have shown that this variant is expressed in the ER+ breast cancer cell line MCF7 in response to estradiol [111], which implies that this product is the result of alternative promoter usage by the estrogen receptor. This alternative start variant has been previously described as an androgen-regulated prostate cancer gene [112], which protects cancer cells from apoptosis when overexpressed [113]. We validated the existence of this transcript in a subset of tissue samples from sample set A and a panel of cell lines (Figure 2). We also validated the significant association of ER status to its expression on a subset of clinical tissue samples from sample set B using quantitative RT-PCR (p-value 0.022, Figure 10), and measured the expression of this transcript on a panel of

breast cancer cell lines (Figure 11).

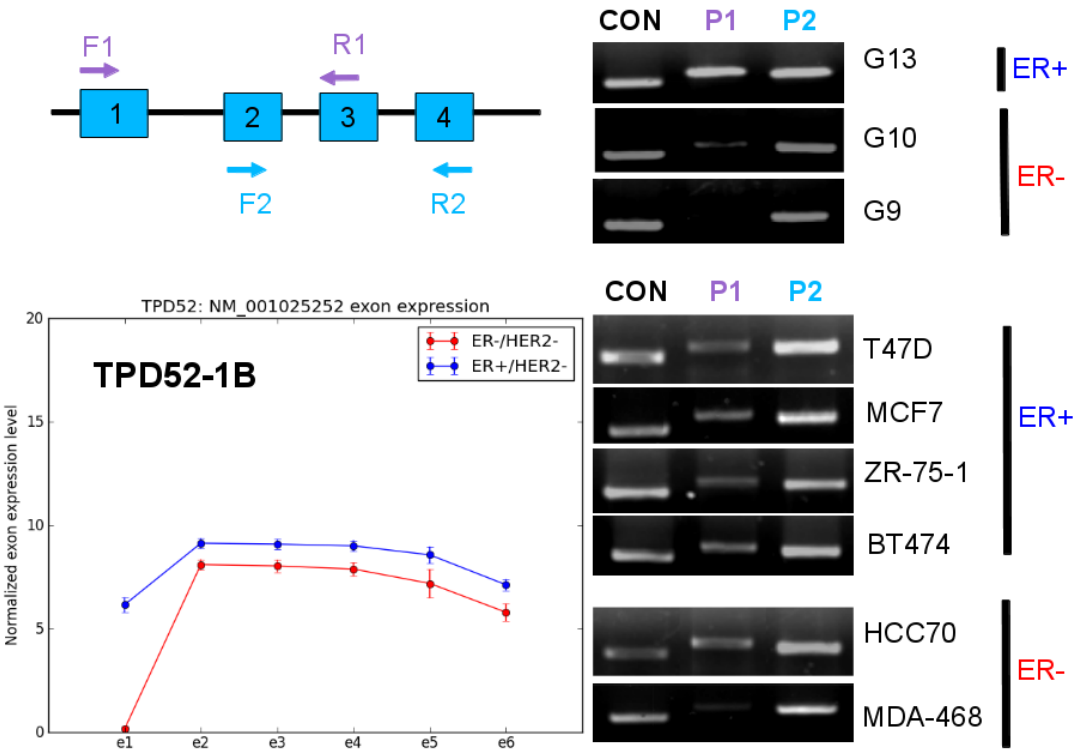

Figure 2

Figure 2. Validation of TPD52 isoforms via gel PCR. We designed primers to exons 1-3 (P1) of the variant to assay its presence in tissues (top) and cell lines (bottom). Both GAPDH and the presence of exons 2-4 were used as a control. The exon chart shows the normalized expression in both ER- (red) and ER+ (blue) samples over all exons.

Another transcript significantly differentially expressed in ER+/HER2- samples was a variant of the acyl-CoA oxidase 2 gene (ACOX2), which is involved in the metabolism of long, branched-chain fatty acids and the synthesis of bile acid precursor molecules in peroxisomes [114]. Peroxisomes, primarily responsible for metabolizing long chain fatty acids for transport into the mitochondria and eventual breakdown, have been previously shown to have reduced activity in breast tumor cells [115], though the mechanism for this phenotype is unknown. Our methods showed an enrichment of exons 10-15 in RNA-Seq data from ER+ breast cancers (Figure 3). By manually annotating this region, we found enhanced expression of intronic fragments in intron 9 which formed a putative contig with exon 10, suggesting that a novel alternative start isoform of the ACOX2 gene beginning in intron 9 and extending through exon 15 (validated on gel PCR in Figure 3) was present and significantly overexpressed in these samples (Figure 5C-D). Previously, Hodo et al [116] demonstrated the existence of an ACOX2 transcript starting in intron 9 which was significantly upregulated in hepatocellular carcinomas. Intriguingly, estrogen receptor overexpression has been shown to be associated with some high-grade hepatocellular carcinomas [117], and Joseph et al [118] reported the existence of an estrogen receptor binding site in the presence of estradiol near exon 10 in the ER+ T47D breast cancer tumor-derived cell line, suggesting the estrogen receptor itself plays a part in regulating the intronic-start ACOX2 transcript. The predicted protein

sequence of this variant lacks a fatty-acid binding domain yet retains the domain responsible for bile acid precursor synthesis [119], which may negatively impact the metabolism of branched-chain fatty acids in breast tumor cells and contribute to reduced peroxisomal activity. We also found that the full-length ACOX2 product was enriched in “normal” breast tissue from reduction mammoplasty with respect to ER+ tumor cells, and that these tissues also had on average a 3.4-fold reduction of intronic variant expression (Figure 6). We validated the association of the intronic variant of ACOX2 with ER+ samples using quantitative RT-PCR (p-value 0.023, Figure 10), and also quantified the abundance both the intronic and full length variants in a panel of breast cancer cell lines (Figure 11).

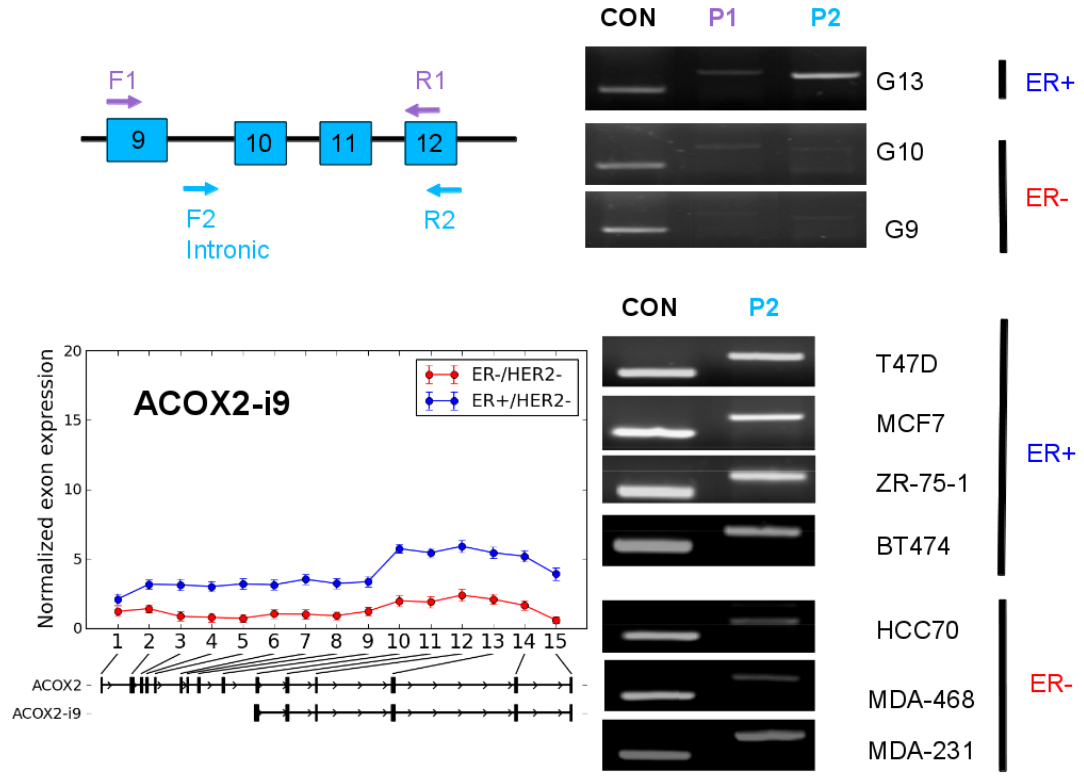

Figure 3

Figure 3. Validation of ACOX2 intronic variant on gel PCR. We assayed the intronic start i9 to exon 12 (P2) on a panel of tissues (top) and cell lines (bottom). GAPDH was used as a control (CON). The full product, represented by exons 9-12 (P1), was absent in all samples. The exon chart shows the normalized expression of each exon in ACOX2 in both ER- (red) and ER+ (red) samples.

Similarly to ACOX2, our analysis also found that exons 9-12 of the gene IQCG are overexpressed in ER- breast cancer tissue samples (Figure 4). Based on

amino acid sequence homology, IQCG is a putative calmodulin binding protein, though the true function is unknown. Manual annotation of IQCG in ER- breast cancer samples indicated high expression of intron 8 which formed a contig with exon 9, suggesting a novel potential alternative start intronic start, which was confirmed via gel PCR (Figure 4). Using Cufflinks' isoform-level quantitation [37], we found the full-length product to be weakly expressed in most tumor tissue samples, while the variant transcript was only heavily expressed in ER- samples (Figure 5F). Previously, Gorello et al [120] reported a fusion of exons 9-12 of IQCG with nuclear pore protein NUP98, which resulted in tumorigenesis in acute T-lymphoid/myeloid leukemia. This suggests that dysregulation of this region of IQCG may contribute to tumorigenesis in ER- breast cancer tumors. The predicted coding sequence of the intronic variant of IQCG retains the IQ coiled-coiled domain which mediates interaction with calmodulin in a Ca++-independent manner in homologous proteins, which may contribute to enhanced calmodulin signaling. A similar protein, EWS (Ewing's Sarcoma), also contains an IQ domain critical for oncogenesis [121].

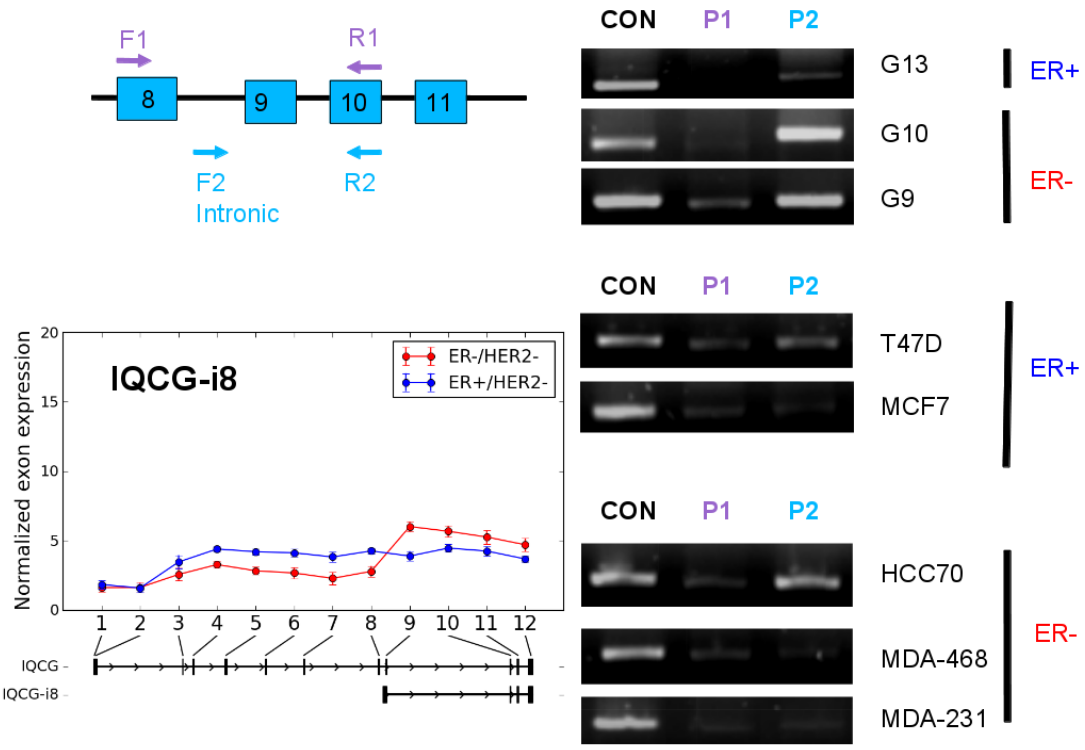

Figure 4

Figure 4. Validation of IQCG intronic variant on gel PCR. We assayed the intronic start i8 to exon 10 (P2) on a panel of tissues (top) and cell lines (bottom). GAPDH was used as a control (CON). The full product, represented by exons 8-10 (P1), was absent or nearly absent in all samples. The exon chart shows the normalized expression of each exon in ACOX2 in both ER- (red) and ER+ (red) samples.

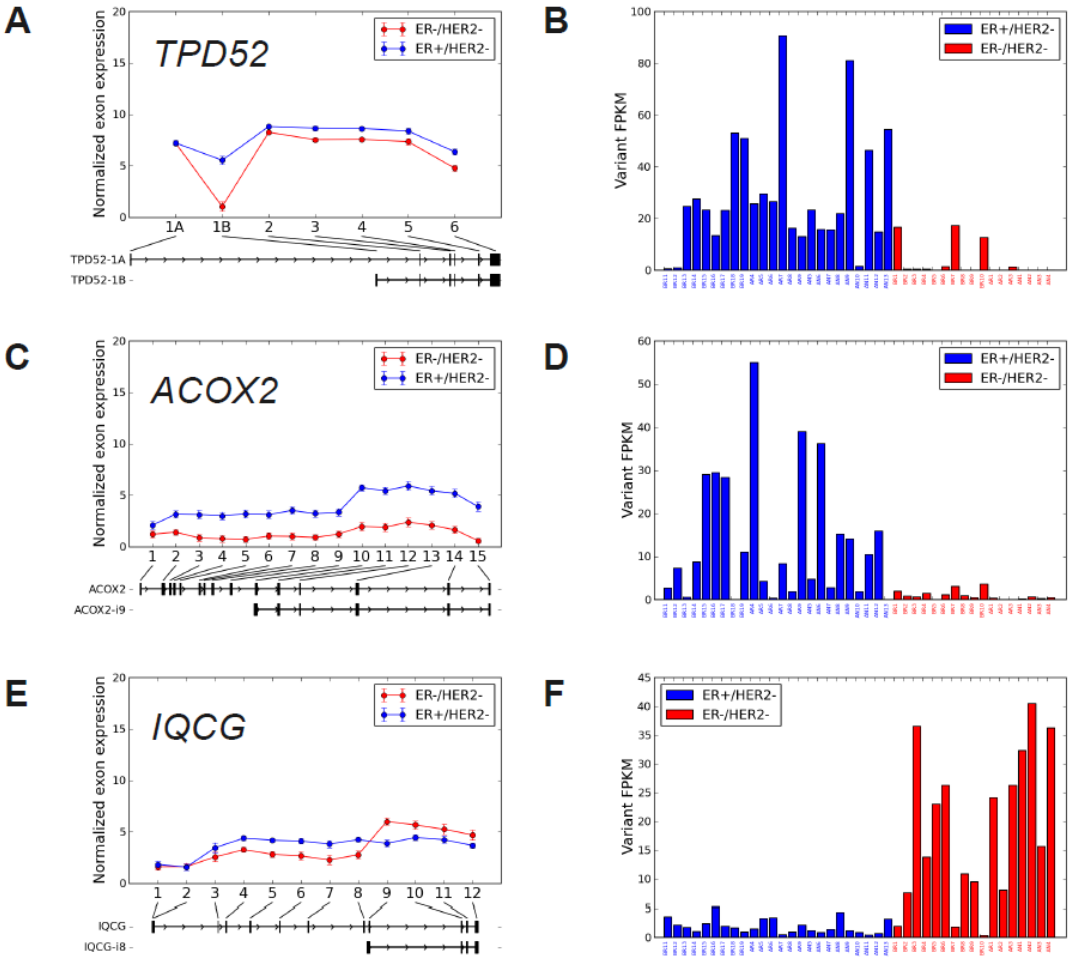

Figure 5

Figure 5. TPD52, ACOX2, and IQCG are among the top hits for genes with differential alternative expression between ER+/HER2- (blue) and ER-/HER2- (red) breast cancer tissues. Figures 5A, C, and E show the normalized exonic expression of the labeled gene transcripts along the y-axis, while the x-axis represents the exons for each transcript, in order. Here TPD52 is depicted with both alternative starting exons to show the relative expression of the

"canonical" transcript (1A and exons 2-6) and the variant (1B and exons 2-6) in both tumor subtypes. Figures 5B, D, and F show the expression of the variant transcript as measured by Cufflinks' isoform-level quantification [37].

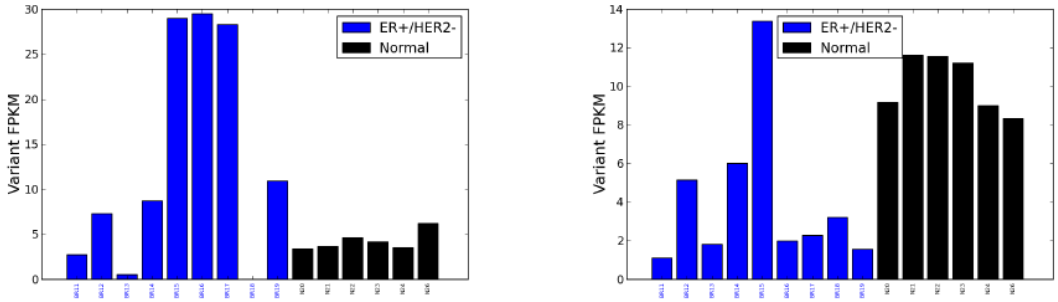

Figure 6

Figure 6. The intronic ACOX2 variant (left) is overexpressed in ER+/HER2- samples from set B, whereas normal breast tissues from reduction mammoplasty express this variant at an average 3.4 fold smaller level. The full-length ACOX2 product (right) is clearly expressed in all normal tissue samples, and relatively absent in tumors. Relative isoform-level quantification was measured using Cufflinks [37].

Two more alternative start variants were identified, both by starting at a canonical exon downstream from the original promoter. The locus KRT18, an epithelial filament protein, has two known variants in RefSeq, both of which have an identical coding sequence, which implies their function is also identical. One variant (RefSeq ID NM_199187) has an additional non-coding starting exon,

which we find absent in ER+/HER2- tissue samples (Figure 7). This UTR likely functions as a regulatory mechanism, which may alter the abundance of the KRT18 protein in ER-/HER2- samples due to mRNA destabilization. The other alternative start variant, NET1, has been shown by other investigators to have two variants in the ER+ MCF7 breast cancer cell line, which they called a “long form” and a “short form” [111]. Intriguingly, Dutertre et al were able to show that, in MCF7 cells, the NET1 long form was increased in response to estradiol, and that at the same time short form expression was silenced, forming a “switch” mechanism. However, our results show significant upregulation of the NET1 long form in *estrogen negative* tumors (Figure 7), in apparent direct contradiction. Further studies are warranted, though it is possible that MCF7 represents a different cell of origin than other ER+ tumors, or that changes in the MCF7 cell line have caused estrogen receptor-mediated transcription to behave differently than many other ER+ tissue samples, as we show is the case with ACOX2 (Figure 11).

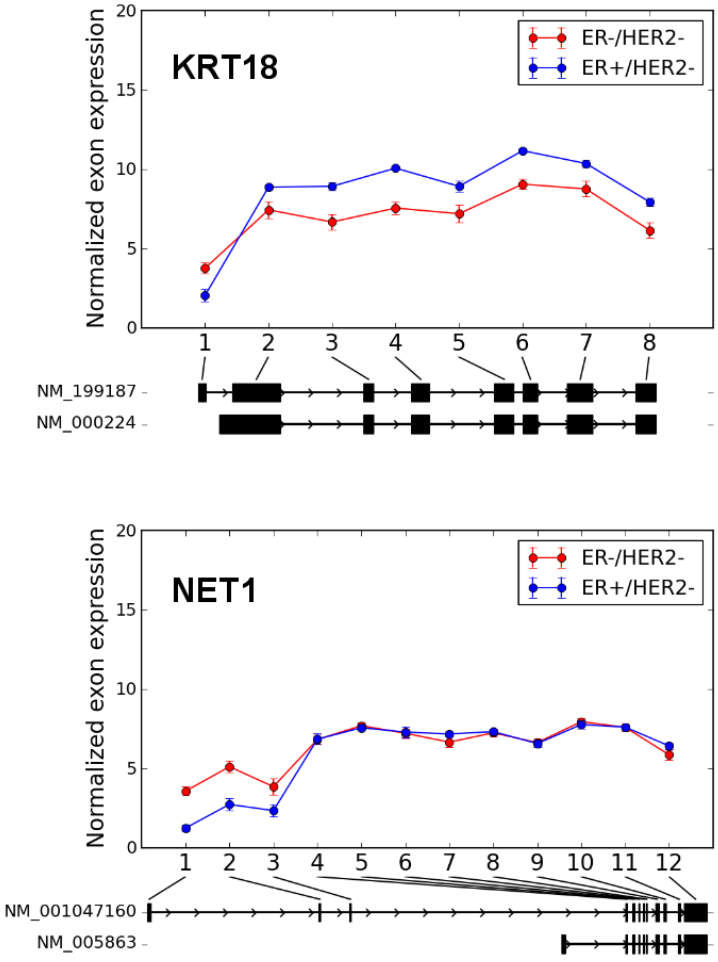

Figure 7

Figure 7. KRT81 and NET1 are examples of alternative starts differentially expressed in ER- and ER+ tumor tissue samples. Each chart shows the normalized exonic expression of the labeled gene transcripts along the y-axis, while the x-axis represents the exons for each transcript, in order. Below each x-axis is a scale representation of the full-length and variant transcripts, where

vertical lines denote exons and horizontal lines represent introns.

Finally, two transcripts were identified as having been the result of cassette exon skipping. At the MYO6 locus, a shorter transcript with exons 29 and 30 missing was found to be expressed in ER-/HER2- tumor samples (Figure 8). MYO6 has previously been associated with the tumor suppressor gene p53 [122], where researchers found the absence of MYO6 prevented DNA damage-induced apoptosis in the breast cancer cell line MFC7. As this variant has not been previously reported, though Lapuk et al [46] showed that MYO6 was differentially spliced in an ER-dependent manner on an exon microarray, Cho et al [122] did not test whether p53 interaction was modified in the presence of the shorter variant. The EPB41L1 locus encodes a ubiquitously-expressed cytoskeletal protein that can result from a prodigious number of combinatorial permutations of alternative starting exons and exon-skipping events [38]. It has also been shown to be involved in cell membrane stability, as well as to mediate cell-cell interactions. We report a variant skipping exon 13 in ER+ samples (Figure 8), which has been previously reported by Parra et al [38] as a variant expressed in all tissues *except* brain, which retained the exon as ER- samples did in our study, over a wide range of tissues tested. This lends credence to the idea that ER+ and ER- samples may have different cells of origin.

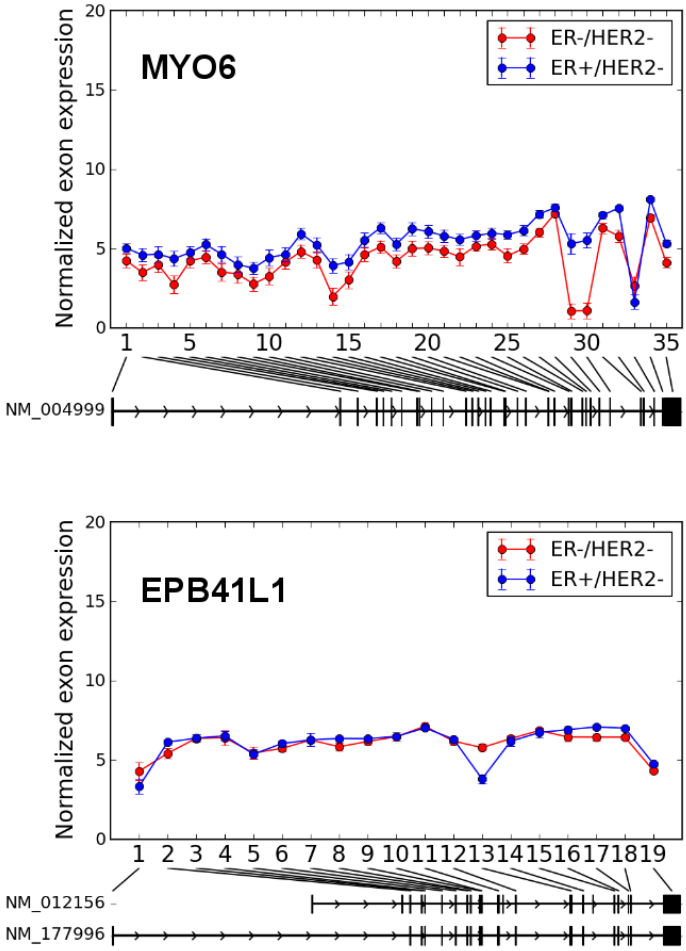

Figure 8

Figure 8. KRT81 and NET1 are examples of alternative starts differentially expressed in ER- and ER+ tumor tissue samples. Each chart shows the normalized exonic expression of the labeled gene transcripts along the y-axis, while the x-axis represents the exons for each transcript, in order. Below each x-axis is a scale representation of the full-length and variant transcripts, where

vertical lines denote exons and horizontal lines represent introns.

4.3 Differentially expressed alternative transcripts in novel subsets of breast cancer independent of ER or HER2 expression

Many predicted subtypes of breast cancer exist, based mainly on clinical annotations and patterns of gene expression from DNA microarrays [9, 14, 19-22]. To date, no large-scale study of breast tumor classification has been made to discover novel subsets of breast cancers denoted by transcriptome variability. With that goal in mind, we developed a method inspired by ALEXA-Seq [134] which performs a screen for high exon variability in known RefSeq transcripts [70] over all samples following normalization (see Chapter 3.11). Once transcripts with variant exons are identified, a manual annotation process which combines assembly information from Cufflinks [37] and raw sequence fragment alignment information is used to elucidate the transcript changes responsible for the observed phenotype. Where available, we also perform a literature search to validate our assembly information against previous research. We chose a conservative threshold of exon variance (3.0) which identified 187 transcripts putatively spliced in individual breast tumor samples. The top 7 transcripts by

total variance are presented in Table 2, below.

Gene Name	Variant Class	ER- Frequency (17)	ER+ Frequency (24)	HER2+ Frequency (14)
GNAS	Alternative start	0.47	0.17	0.50
KRT81	Alternative start	0.71	0.58	0.86
LTF	Alternative start	0.65	0.38	0.64
SCGB2A2	Early termination	0.24	0.13	0.07
TPD52	Alternative start	0.29	0.92	0.29
GSTM1	Multiple exon skipping	0.41	0.17	0.67
MIA	Early termination	0.29	0.54	0.50

Table 2

Table 2. The top genes identified in a screen for highly variant exons over all samples, normalized by transcript expression. The type of transcript variant was identified during manual annotation, and is listed in column 2. Columns 3-5 indicate the overall frequency each variant was found to be overexpressed (total fraction of all transcripts at that locus > 75%) in each sample clinical subtype. Among the genes identified using this method and the previous method for identifying differential alternative splicing between ER+ and ER- samples, only TPD52 was found to be strongly indicated by both methods.

One example of a highly-scoring variant which is not directly associated with a known clinical subtype is GNAS. In the most current RefSeq human annotation,

GNAS has 8 separate isoforms which are only variant in their first exons. We found that one of these variants, a transcript which encodes the protein NESP55, is heavily expressed in some breast cancer tumors and entirely absent in others (Figure 9B). NESP55 is a 55kd protein encoded entirely by a sequence present in its first exon, whereas the common downstream exons form the 3' UTR of this transcript [75]. By contrast, another alternative starting exon at this locus forms an in-frame transcript with these same downstream exons to encode G-protein alpha subunit 2 (GSa), which is a widely-expressed component of the cAMP response pathway and found in all breast tumor tissue samples surveyed (Figure 9A). GNAS is a complex locus which undergoes both maternal and paternal imprinting [75], and the NESP55 transcript is itself expressed exclusively from the maternal allele. The NESP55 transcript product encodes a neuroendocrine secretory protein, which raises the possibility that breast tumors which overexpress NESP55 may be derived from neuroendocrine cells of origin. We have validated the RNA Sequencing-derived quantification of the variant transcript using quantitative RT-PCR (Figure 10), with excellent correlation ($r=0.99$).

Another variant found in subsets of all clinical subtypes of breast cancer tumors surveyed is found at the KRT81 locus. We observed highly variant expression of exons 5-9 in these tumors (Figure 9C-D), which suggested the presence of an

alternative transcript formed from these exons alone. KRT81 is a hair keratin constitutively expressed in human hair shafts, and also found in nails. Other investigators have previously reported that a truncated form of KRT81 was present in some human breast tumor carcinomas [123], and Boulay et al [124] showed that this truncated protein was the result of a cryptic alternative promoter present in intron 4 of this gene. Using a GFP-reporter bound to this variant and transfected into HeLa cells, Boulay et al also found that truncated KRT81 was integrated into the cytoskeletal keratin network, and that expression of this protein may alter the adhesive properties of these tumor cells.

Figures 9E and 9F describe a novel variant of the Lactoferrin (LTF) gene which lacks the canonical first exon, which we find expressed in some ER+ and ER- breast cancers. Lactoferrin is a secreted, iron-binding protein with antiseptic properties which is heavily expressed in breast tissues during pregnancy. However, an alternative first exon has been previously reported for LTF [50-52] which lacks a coding sequence, causing translation to begin at an alternative downstream ATG and resulting in a separate protein, referred to as dLTF. The dLTF variant lacks 44 amino acids at the 5' end, which form a signal peptide sequence required for secretion. Rather than being secreted following translation, dLTF has been shown to localize to the nucleus, where it functions as a transcription factor involved in cell cycle control [51-53]. However, our analysis

did not indicate expression of the alternative non-coding first exon which is unique to dLTF in breast tumor samples (consistent with prior reports of its absence in breast cancer [51, 52]), rather, we observed the independent absence of canonical LTF exon 1 in a number of samples. This would result in an identical coding sequence to the previously reported dLTF, which is associated with good prognosis when highly expressed [53]. We observed both LTF and dLTF transcripts in normal breast tissue samples, indicating that the absence of dLTF exon 1 is not caused by errors in sequencing.

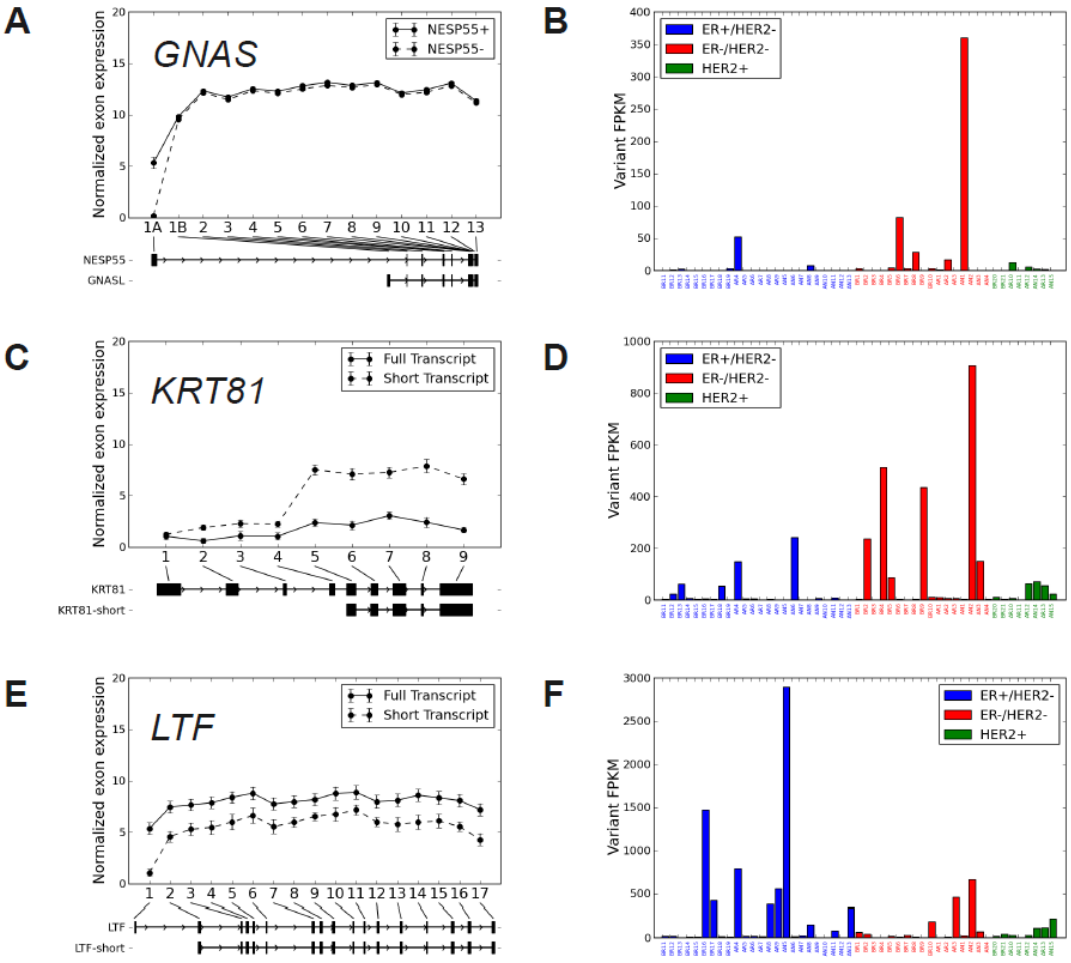

Figure 9

Figure 9. Variant isoforms of GNAS, KRT81, and LTF are heavily expressed in a subset of breast tumor samples and absent in others, regardless of clinical annotation. Figures 9A, C, and E depict the labeled gene's exonic expression in subsets which overexpress the variant versus those that do not. The y-axis represents mean normalized exon expression, while the x-axis represents the canonical transcript exons, in order. In each case,

the variant class is indicated by a dashed line. Figures 9B, D, and F show the relative expression of variant transcripts in each sample, labeled by clinical subtype. Isoform-level quantification for these images was estimated using Cufflinks [37].

4.4 Breast cancer cell lines express alternative splicing variants

While tumor-derived cell lines are not an ideal model of breast cancer *in vivo* [102, 103], they possess useful properties which encourage their use in research. Primarily, tumor cell lines are a renewable source of cancer cells which are completely homogeneous and can be manipulated freely, while being faster and easier to grow than animal models. Cell lines can be forced to overexpress genes of interest by transfecting them with plasmids containing a copy of the DNA transcript. This can be used to study the effects of alternative transcript expression on cell morphology, proliferative ability, and aggressiveness. Similarly, target transcripts can be silenced in cell lines using small interfering RNAs with complementary sequences to unique features on these transcripts, which promotes degradation via the RNA interference cellular pathway. The effectiveness of these studies is contingent on the similarity of basal transcriptomic and epigenetic factors to the tumor in question, and also on the cell's prior ability to produce the transcript. Therefore, it is beneficial to discover

whether cell lines express alternative transcript variants described in tumor tissue samples, so that the best model can be chosen for continued study.

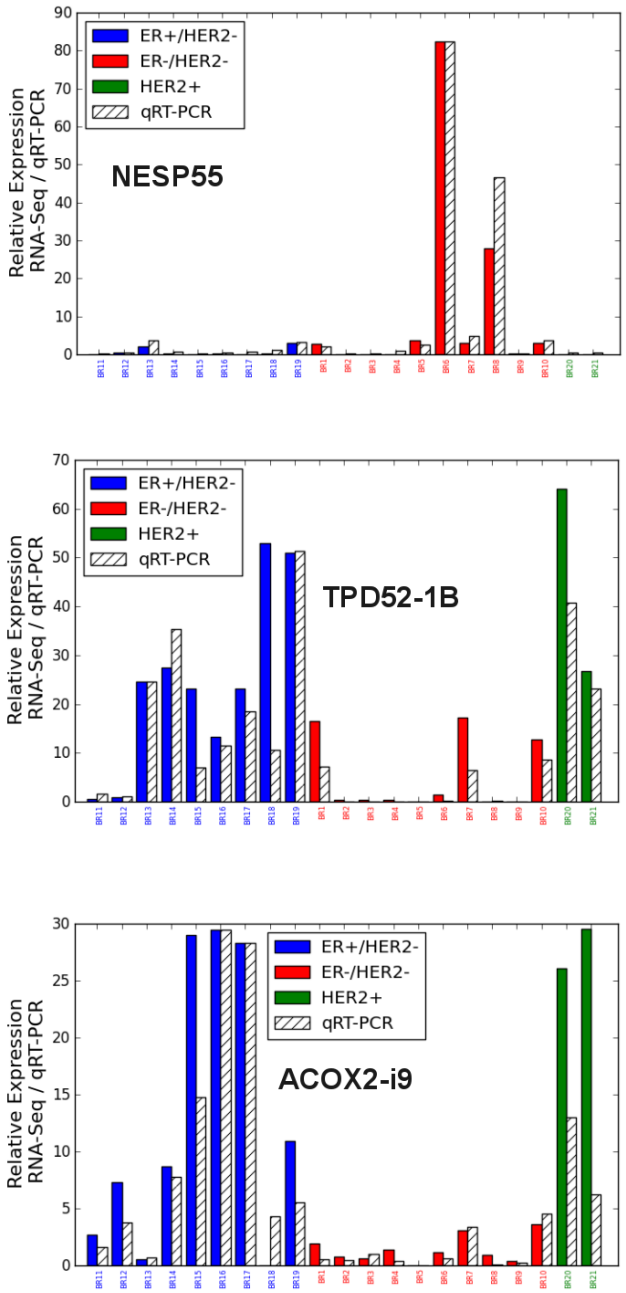

Figure 10

Figure 10. The expression of GNAS (NESP55), TPD52, and ACOX2 variants is consistent with RNA-Seq results as measured by quantitative RT-PCR. Each chart shows the expression of the labeled variant as quantified using Cufflinks' isoform-level measurement versus quantitative RT-PCR on a subset of tissue samples from set B. The correlation coefficients between the two measurements are 0.99, 0.85, and 0.85, for NESP55, TPD52-1B, and ACOX2-i9, respectively. As measured by qRT-PCR, the p-value of association between variants of TPD52 and ACOX2 and the ER+ tumor tissues is 0.022 and 0.023, respectively, while the p-value for NESP55 is 0.208, which is consistent with the results of our method.

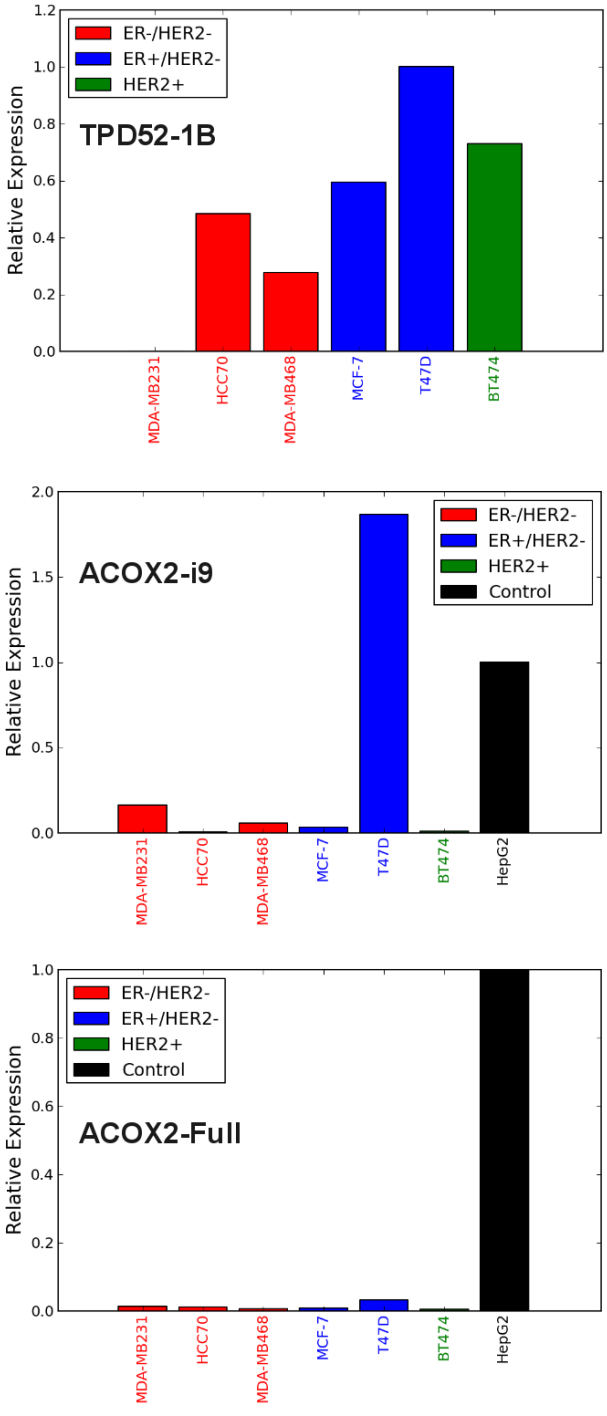

Figure 11

Figure 11. Variants of TPD52 and ACOX2 are expressed in cell

lines. Each chart depicts the qRT-PCR expression of the labeled TPD52 and ACOX2 transcripts over a number of breast cancer tumor-derived cell lines. TPD52 shows consistent association of ER+ cell lines with subtype, though the ACOX2 intronic variant is only found in the ER+ T47D cell line. The full-length transcript of ACOX2 was absent in all cell lines except the liver-derived control, HepG2.

We assessed a panel of 5 breast cancer cell lines using quantitative RT-PCR to determine whether the TPD52-1B (PrLZ) variant of TPD52 and both the full-length and intronic products of ACOX2 were present, and whether the associated ER/HER2 status of these cell lines affected expression (Figure 11). For the two ACOX2 transcripts, an additional transformed “normal” human liver cell line, HepG2, was also used as a control to provide a reference for full-length transcript expression. In general, the variant TPD52 transcript was heavily expressed in ER+ cell lines (T47D, MCF7) and the ER+/HER2+ cell line BT474, and expressed to a lesser degree in ER- cell lines (MDA-MB-468, HCC70), or entirely absent (MDA-MB-231). The ACOX2 intronic variant was only heavily expressed in one ER+ cell line, T47D, relative to the control line HepG2. This implies that MCF7 may have a different cell of origin to other ER+ tumors, or that it represents a subdivision of ER+ tumors which does not also express ACOX2-i9. This is consistent with a previous study that places an estrogen receptor binding site in the presence of estradiol near the start of the intronic transcript in T47D

cells [118], which may enhance expression in this cell line. The same report did not identify estrogen receptor binding in the same location in MCF7 cells, which may explain its diminished intronic transcript expression.

4.5 Normal breast tissues from reduction mammoplasty procedures express alternative splicing variants

Normal breast tissue is composed of a number of heterogeneous cell populations, including containing lobules, ductal tissue, connective tissue, and fatty tissue [48]. This presents a challenge to researchers attempting to locate cellular changes that promote tumorigenesis. We sequenced 6 additional breast tissue samples from Radium Hosptial in Norway which were obtained from reduction mammoplasty procedures using the same 100bp single-end read protocol used for sample set B. We found evidence of most tested variants, including TPD52, NESP55, IQCG, LTF, and KRT81, with no discernible pattern of expression that would indicate the presence of a specific “ER+” or “ER-” subpopulation of normal cells. This suggests many of these changes can also be the result of shifting regulatory programs that occur naturally in breast tissue, which become dysregulated in tumor cells. We found that, consistent with previous studies [51, 52], normal cells did exhibit the first non-coding exon of

dLTF, unlike any tumor samples surveyed. Also, normal samples overexpressed the full-length transcript of ACOX2 (Figure 3), whereas this transcript was heavily downregulated in tumor samples. Other investigators have shown that later exons of the ACOX2 transcript were differentially expressed between normal and tumor breast tissues on custom exon arrays [45], consistent with our results.

Chapter 5: Identification and Validation of Long Intergenic Non-Coding RNA in Breast Cancer Subsets

5.1 Identifying long non-coding RNA in intergenic regions

One of the primary benefits of RNA-Seq over traditional DNA microarrays is the ability to interrogate expression in unannotated areas of the transcriptome. At the present time, methods for identifying novel genes in genomic DNA sequence generally depend on features such as an open reading frame and conserved functional protein elements [60]. However, non-coding RNA structures are more difficult to detect, and most of our knowledge comes from expressed sequence tags (ESTs), which are large, sequenced contigs from transcriptome cDNA libraries. Carninci et al [94] showed that over half of these sequences appear to originate from non-coding sources, though very few are annotated. Moreover, many of these non-coding sources are poorly conserved over multiple species [63], indicating that these transcripts may not be functional, but instead, may be the result of “random” transcription. To avoid this problem, we took the approach of annotating non-coding transcripts, identified using RNA sequencing, which were differentially expressed between breast cancer tumor subtypes. We reason

that such transcription is more likely to be the result of selective dysregulation in tumors as they progress from normal tissue to distinct subtypes of disease. Hence, such non-coding RNA are more likely to be either functional themselves or the result of a functional change in some regulatory pathway (e.g., induction of a promoter suppressed in normal tissue), and therefore useful as a biomarker. Our results show that unlike the ncRNA identified previously [63] some of the non-coding RNA we identified are evolutionarily conserved, suggesting that they may indeed be functional.

In order to systematically assemble differentially expressed non-coding transcripts, we first identified all regions of transcriptome expression where the number of sequence fragments in a window passed a signal-to-noise ratio test [69] between two breast cancer subtypes (see Chapter 3.12 for detailed methodology) in intergenic regions. Briefly, we compared the mean and variance of fragment expression in intergenic windows of 300 base pairs over 24 ER+/HER2- and 17 ER-/HER2- samples from set A and set B. If the signal-to-noise ratio exceeded 0.8, the region was considered a putative locus for a non-coding RNA. These transcripts were assembled using reference-based assembler Cufflinks [37], the de novo transcript assembler Trinity [76], and manual annotation. This procedure identified 59 intergenic regions representing a total of 132 300 bp which were differentially expressed between ER+/HER2- and

ER-/HER2- breast cancers. Of these, 19 transcripts were spurious fragment alignment caused by sequence homology to other annotated genes (pseudogenes), 6 were unannotated exons associated with known transcripts, 24 were unannotated longer UTRs associated with known transcripts, and 10 contained putative RNA transcripts. One region was removed because of low fragment coverage, which produced an inconsistent annotation. Of the remaining 9 locations, one was found to contain both sense and anti-sense transcripts, which we considered to be separate genes for the purposes of annotation. This resulted in 4 potential non-coding genes associated with ER-/HER2- tumors, and 6 potential non-coding genes associated with ER+/HER2- tumors, containing a total of 30 transcript isoforms. The results of this survey are given in Table 3, below.

Location	ER Status	Conserved	DESeq Adjusted p-value
chr1:27384973-27391665	Basal	NO	0.00063
chr10:37524732-37721347	ER	YES	0.00089
chr11:69291900-69294708	ER	NO	0.00249
chr15:71371962-71387290	ER	LOW	0.00041
chr15:71371962-71387290*	ER	LOW	0.00001
chr22:42760405-42765242	Basal	YES	0.00249
chr4:159091903-159124029	ER	YES	0.17946
chr5:170171879-170174363	Basal	NO	0.00365
chr6:128900610-128901089	Basal	NO	0.00081
chr8:87345502-87355543	ER	YES	0.10164

Table 3

Table 3. Non-coding RNA found to be differentially expressed

between ER+/HER2- and ER-/HER2- breast tumor tissue samples. Conservation was determined using the program phastCons [125] on a multiple alignment of 46 vertebrate species to the most current human annotation, which uses a phylo-HMM to estimate regional selection given a neutral model of evolution. The FDR-adjusted p-value was given by DESeq [65]. We considered p-values less than 0.05 to be significant, and these are indicated in red. The non-coding RNA found on chromosome 15 was found to have an anti-sense transcript, denoted by an asterisk, and this gene was analyzed separately.

5.2 Putative intergenic genes do not code for protein

To assert the intergenic transcripts identified were untranslated, we utilized two separate methods. The first, used by Jia et al [60] to identify non-coding ESTs in the human genome, can be described as a ORF predictor/BLASTP pipeline to decide the coding status of transcripts. In this method, all six possible protein coding sequences were enumerated for each transcript sequence from three possible reading frames in two directions (sense and anti-sense with respect to genomic DNA). These amino acid sequences were aligned to conserved functional domain and full protein sequences from known species using the protein BLAST tool [96]. The mean maximum ORF length over all 30 transcripts was 84.4 amino acids, with a maximum of 202 (both transcripts from chr5). None

of the transcript ORFs contained significant homology to known functional domains, and only one sequence contained significant homology to a known gene. Approximately 19% of this transcript, part the conserved estrogen positive associated RNA found on chromosome 10 (Figure 12), contained a small region of human aminophospholipid transporter, class I, member 2 (ATP8A2) with 89% sequence identity. This suggests this ncRNA originated as part of a recent gene duplication event and subsequently evolved separate functionality. Based on the lack of homology to known functional protein sequence domains, all transcripts were classified as non-coding using this method.

The second method, based on work by Kong et al [97], utilizes a support vector machine (SVM) trained on features from coding and non-coding sequences to make a classification decision. This method, called the coding potential calculator (CPC), attempts to predict a likely open reading frame while taking into account possible sequencing errors. The length of the open reading frame, the "correctness" of the ORF (reduced by point mutations or small indels), and whether or not the ORF begins with a start codon form the "frame score," three points of data used to train the SVM. The "hit score" is calculated from a BLAST search for homology to the predicted ORF, and takes into account the number of protein sequence hits, the quality (E score) [96], and the enrichment of his in the predicted ORF versus other available open reading frames. The CPC classified

all putative transcripts as non-coding except for the putative gene on chr5, which was classified as “weakly coding” (within SVM margin) due to having a long ORF (202 aa) and a present start codon. However, there were no significant BLAST hits which indicated homology to known proteins, which reduced its coding score. Given these results, we concluded 9 of the 10 potential genes contained non-coding sequences, and that 1 gene remained ambiguous.

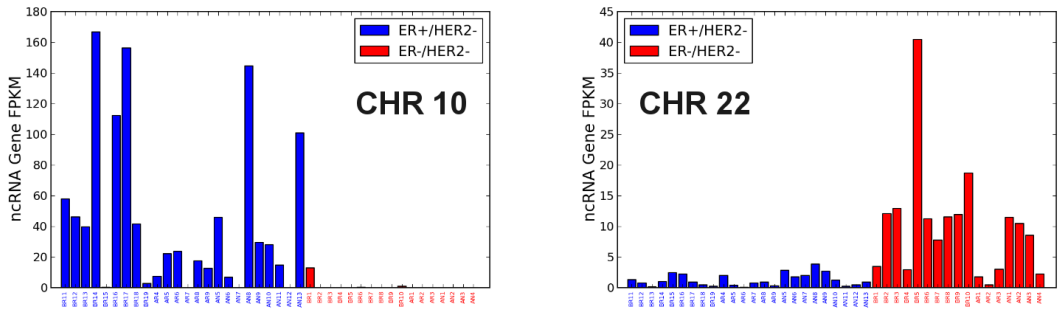

Figure 12

Figure 12. Gene-level expression of putative non-coding RNA on chromosome 10 and chromosome 22 are significantly associated with breast cancer tumor subtype. The y-axis represents normalized gene-level fragment counts, while the x-axis represents 24 ER+/HER2- (blue) and 17 ER-/HER2- (red) breast tumor tissue samples. Putative non-coding genes sequences on both chr10 and chr22 were found to be conserved over multiple species.

5.3 Transcripts associated with differentially expressed genomic windows are themselves differentially expressed

We used the DESeq software package [65] to examine whether count levels in our putative non-coding RNA gene assemblies were significantly different between breast tumor subtypes. DESeq models fragment count data as a negative binomial distribution to accurately estimate the individual variance of each gene in each sample, then performs a significance test based on the difference of means between samples in each subtype given the total biological and fragment count variances. These p-values are then FDR adjusted to give a final result. Using this method, we found that 8 of the 10 putative ncRNA were significantly differentially expressed ($p < 0.05$ after adjustment, Table 3), demonstrating that it is possible to identify differentially expressed transcripts from examining count data in windows as we have suggested (Figures 12 and 13).

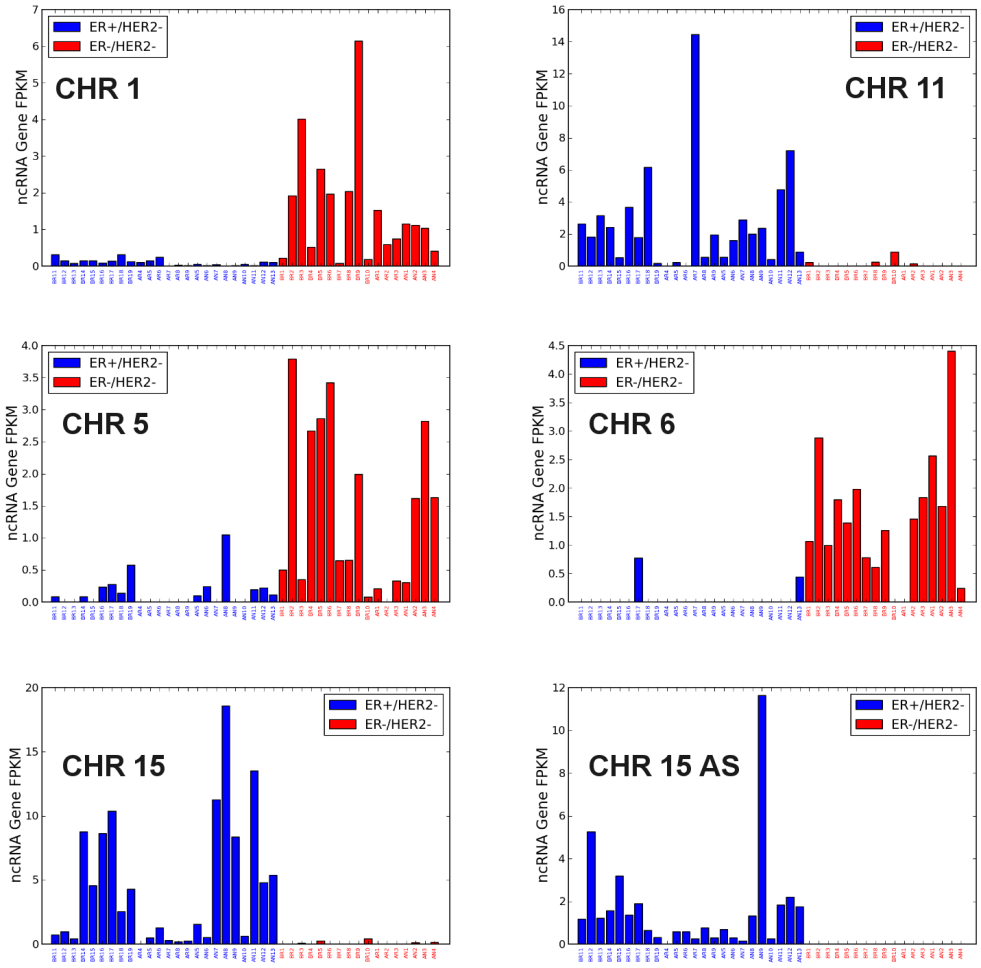

Figure 13

Figure 13. Gene-level count data from 6 non-conserved gene sources are differentially expressed between ER+/HER2- and ER-/HER2- breast tumor samples. The expression level of each gene, given on the y-axis, was calculated by normalizing the number of fragments which aligned to any exonic region with the putative genes by the sequence depth as well as the total length of

the gene.

5.4 Non-coding RNA gene sequences on chromosomes 4, 8, 10, and 22 contain elements conserved over multiple species

Though many non-coding transcriptome elements have been discovered, many are not evolutionarily conserved, which casts doubt on their functional significance [63]. Transcript sequence conservation implies that selective pressures are at play to ensure the survival of specific sequence elements, implying that sequence confers an evolutionary advantage. Although lack of conservation does not necessarily mean the transcript is non-functional (e.g., the well-known X-chromosome silencing regulator, XIST, is poorly conserved [126]), additional study is warranted. We examined conservation in putative non-coding gene regions over 46 species using phastCons [125], using data made available from the UCSC genome browser. PhastCons uses a phylo-HMM algorithm to decide whether a genomic region of arbitrary size evolved less rapidly than would be expected assuming a neutral model of evolution, given a multiple alignment of whole species genomes and the estimated phylogenetic branch lengths between them. We found evidence of conserved elements in multiple genes, two of which (chr10 and chr22) were also differentially expressed between ER+/HER2- and ER-/HER2- breast tumor samples (Figure 14). Evidence of sequence

conservation was weak for non-coding genes on chromosomes 1, 5, 6, 11, and 15, though significant differential expression with low variability indicates transcriptional programs actively dysregulated by these cancers are promoting expression in these regions either directly or indirectly, and that even if these transcripts prove non-functional they may be used as a biomarker for these disease populations.

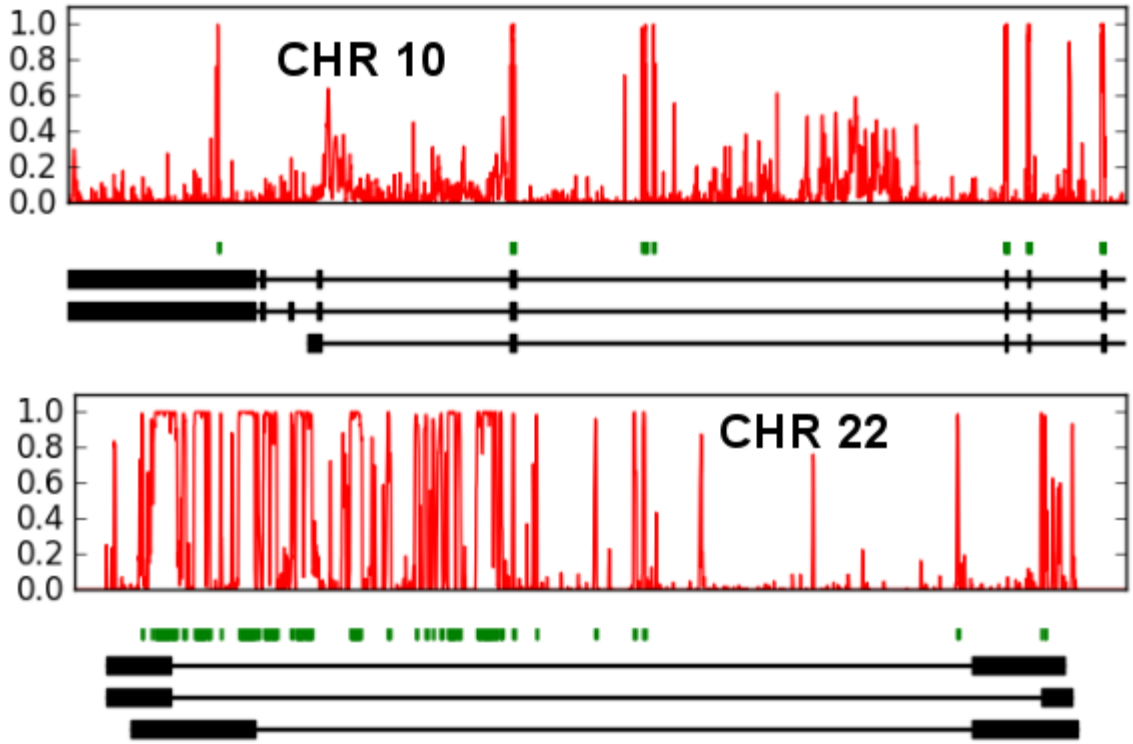

Figure 14

Figure 14. Putative non-coding regions on chromosomes 10 and 22 show regions of conservation over 46 diverse species genomes, as measured by phastCons [125]. Conservation (in red) is

measured as 1 – p, where p is the probability of conservation based on a neutral model of evolution. Regions corresponding to $p < 0.05$ are shown in green. Below the conservation graph for each gene is a scale representation of multiple transcript assemblies at the given locus. Transcripts belonging to the non-coding locus on chromosome 10 are presented up to the first 30,000 bases for clarity.

5.5 Validation of putative non-coding RNA on chromosome 10

We chose to validate our discovery of novel non-coding RNA on chromosome 10 using a gel PCR assay. This gene made an ideal candidate for validation due to heavy, consistent expression in ER+/HER2- tissues (Figure 12), which were also conserved (Figure 14), and therefore likely to be functional. With respect to the primary transcript (arbitrarily chosen), we designed primers to exons 1-2, 1-3, and 6-9, which are listed below in Table 4.

	Forward	**Reverse**
1-2	CTCGAAGCCATCAATGACAA	GATCCTAGAGGAGCCAGTTTCA
1-3	TGTTTGTCACGTGGTTGTTG	CTTTGGCATTCTGGGTGATT
6-9	TCCGCTGTGGAAGACTTTTT	TGAGAATGGTGGACCAGATG

Table 4

Table 2. Forward and reverse PCR primers for validation of putative

non-coding RNA located on chromosome 10.

Primers were chosen to encompass multiple exons to prevent contamination from genomic DNA from being amplified in the PCR process, which would result in an inaccurate gel band. We ran two separate gel PCR experiments on a single ER+/HER2- tissue sample from set A, and used GAPDH as a control. We first verified transcript regions from exons 1-2 and 6-9 (Figure 15), which produced bands close to 310bp and 286 bp as expected, respectively. Figure 15 provides a scaled representation of the genome track containing 3 example transcript assemblies of the putative non-coding RNA located on chromosome 10 (of a total of 9 isoforms), and shows the forward and reverse primer locations for both regions and their gel products. The second validation experiment focused on exons 1-3 (Figure 16), which produced a band close to 343 (exons 1-3, with an additional exon present in isoform 2) and 245 bp (exons 1-3, inclusive) as predicted by our assemblies. The existence of a longer product band was unexpected and indicates the presence of an additional splice variant at that locus, likely due to intron retention, which was observed at this locus in RNA-Seq data from tissue samples.

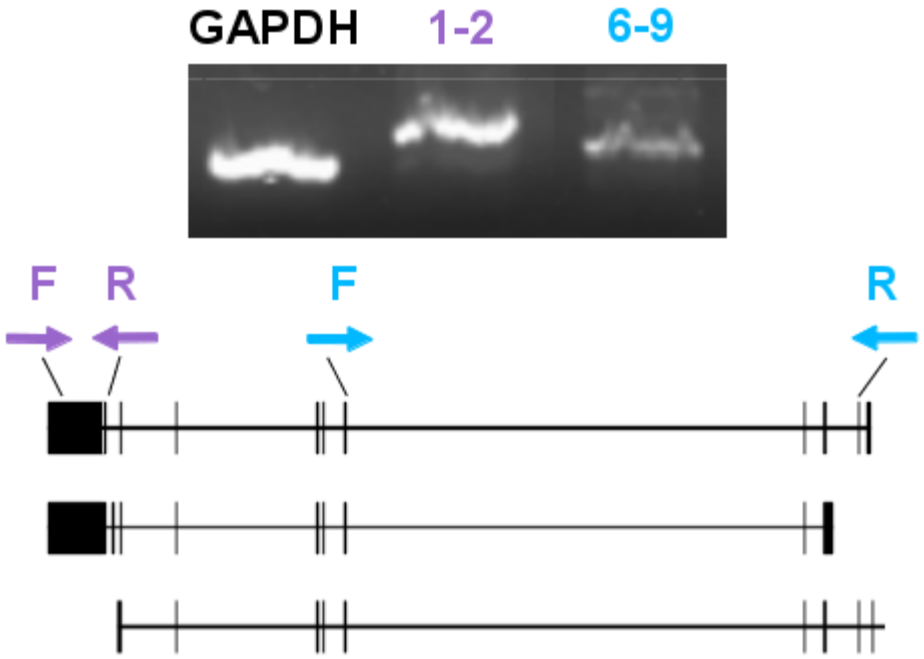

Figure 15

Figure 15. Gel PCR validation of putative ncRNA transcript assemblies at the chr10 locus on ER+ tissue sample RNA. Both bands at exons 1-2 and 6-9 were close to their expected sizes of 310 and 286 bp, respectively.

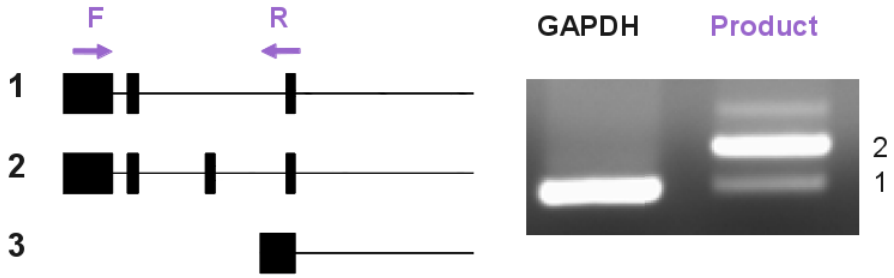

Figure 16

Figure 16. Gel PCR validation of exons 1-3 reveals multiple splicing isoforms, one of which is predicted by our assemblies. An unexpected longer product is likely the result of intron retention.

Ongoing studies to ascertain the function aspects of putative non-coding sources, especially those with conserved elements (Figure 14), should continue by investigating whether ER+/HER2- or ER-/HER2- tumor-derived cell lines express the variants we have discovered. Short hairpin RNA (shRNA) with sequence complementarity to common exons in ncRNA gene transcripts can be used to silence expression, and tumor cells can be interrogated both for genome-wide expression changes (by performing DNA microarray or RNA-Seq before and after knockdown) and changes in morphology, proliferative capability, and metastatic aggressiveness. Recent publications [61, 63, 64] have shown that

more than half of identified long, intergenic non-coding RNA transcripts are directly responsible for genome-wide transcriptional repression programs initiated by interaction with the polycomb repressive complex (PRC2). To identify whether identified non-coding RNA are functionally involved in this manner, ChIP-Seq experiments [61] can identify RNA bound to PRC2 and other repressive protein complexes. Another recent study [57] studied the effects of the non-coding RNA HOTAIR by examining histone modification maps before and after expression. This experiment can be adapted to the present work by transfecting plasmids which contain our ncRNA into cell lines which do not express them natively, and observing the effects of this transfection on chromatin state.

Chapter 6: Discussion and Future Directions

6.1 Benefits of transcriptome research to our understanding of breast cancer biology

Therapeutic treatment of breast cancer is currently based on our limited understanding of the foundations of breast cancer biology. Namely, that “estrogen positive” tumors over-express the estrogen receptor, and that estrogen receptor antagonism, in the form of Tamoxifen or other treatments, results in effective treatment response (lower recurrence, higher five-year survival) in ~50% of these patients [3]. Also, “HER2 positive” tumors over-express HER2/Neu, and that treatment via a monoclonal antibody which causes immuno-targeting of this receptor results in better prognosis [8]. At present, studies of the underlying biological pathways activated in these cancers are based mainly on gene-level dysregulation of expression [14]. However, as we have shown in this thesis, the “gene” is in many cases made up of diverse transcriptional elements and regulation. Through careful, systematic study of the disease transcriptome, we can increase our understanding of the specific changes which result in tumorigenesis, and use this information to assist in designing additional

therapeutic measures that combat these changes.

Towards this future, the present study presents the transcriptional alternative splicing changes which distinguish clinical subtypes of breast cancer. Our analysis identified several genes previously shown to be involved in tumorigenesis, such as TPD52 [109, 112, 113], LTF [50, 52, 53], and ACOX2 [116] which were found to be differentially spliced across tumor subclasses, implying that the resulting proteins may have altered activity or function in different disease profiles. Some of the alternative transcripts identified here are known to have very specific functions. For instance the GNAS transcripts [75] can alternately produce a secreted neuroendocrine protein or a G-coupled protein receptor, and the LTF transcripts [53] can alternately produce a secreted iron-binding protein with antiseptic ability or a nuclear transcription factor involved in cell cycle control.

Alternative transcripts have also previously shown to have directly opposing functions, as in the case of the anti-apoptotic gene survivin, which has a transcript variant that promotes apoptosis, and may be a naturally-occurring antagonist [35, 40]. Alternative splicing variants identified in this study may yet prove to be important contributors to disease phenotypes such as proliferative potential or aggressiveness, whereas other, ubiquitously-expressed transcripts at

the same locus do not. To determine this, a rigorous functional annotation must be undertaken, which is discussed below (Section 6.3).

We have also made a systematic study of potential non-coding RNAs which are differentially expressed in breast cancer tumor subclasses. Non-coding RNA has been previously implicated in various functional roles in cancers. For example, the large, non-coding RNA XIST is used to silence the extra X chromosome in females. However it is notably absent in various breast, ovarian and cervical cancer cell lines [56], suggesting that it has a potential role in tumorigenesis. Recently, the non-coding RNA HOTAIR has been shown to “reprogram” hundreds of genes in epithelial cancer cells, causing them to adopt a polycomb repressive complex 2 (PRC2) occupation pattern resembling embryonic fibroblast cells [57]. HOTAIR has since been shown to increase metastasis and promote growth in many cancers, including breast [57], gastrointestinal [58], and hepatocellular carcinoma [59]. Human non-coding RNA are poorly annotated [60], despite the fact that the majority of transcription originates in non-coding sources [94]. Like HOTAIR, many of these non-coding RNA have been shown to associate with the PRC2 complex to silence gene expression in mouse cells [61, 63, 64], implying active functional roles in gene regulation. We have discovered 9 non-coding RNA differentially expressed between ER+ and ER- breast cancers, none of which had been previously published (though 5 had been predicted from sequence

composition and EST support) of which 4 show clear sequence conservation over mammalian genomes, which is direct evidence for a functional role. In continuing experiments, we will elucidate the roles these non-coding RNA play in their respective disease classes.

6.2 Continuing studies

As sequencing technology continues to improve and sequencing costs decrease, we expect that our understanding of cancer will improve. The methods and research we have presented here were borne of necessity; the most current annotations of the human transcriptome are incomplete in terms of combinatorial splicing of known genes, as well as the relatively unknown gulf of non-coding genes and their variants. Thus, our methods are designed to incorporate the identification of novel variants and non-coding RNA. With powerful, cheaper RNA sequencing it will be possible to make complete tumor transcriptome annotations using deep, paired-end sequencing methods, an eventuality which some researchers have already begun to explore [31, 92]. With complete annotations, it will be easier to judge the origination of sequence fragments, or perhaps even develop a microarray chip (or chips) which can interrogate all known variations to gain a clear image of the isoform-level dysregulation present in individual tumors

[127]. With the knowledge gained in the present study and future work based on better transcriptome annotation, we will have gained a complete picture of isoform-level changes in breast cancer tumors.

At that point, continuing research should focus on two major paths: function, and epigenetic interaction. To determine the function of each ncRNA and alternative splicing variant, a systematic study of the effects of knock-down and knock-in experiments on changes in cell morphology, proliferative capability, and aggressiveness would be beneficial. In a cell line model, for instance, it is possible to knock down the splice variants we have discovered to identify functional consequences. For example, the expression of ACOX2 in the ER+ cell line T47D (Figure 6) can be accomplished with an shRNA designed to form a hairpin structure which binds with sequence complementarity to the target mRNA. This facilitates its degradation through the RNA interference cellular pathway. To upregulate a transcript in a cell line which otherwise does not express the variant, the transcript could be encoded in a plasmid, which is then transfected into the cell line. If changes in expression of these transcripts has an interesting phenotype, such as for instance cell death, change in drug response, reduced proliferative ability, or reduced metastatic potential, the transcript might be a biomarker or a target of therapeutic treatment.

Functional roles might be ascertained for non-coding RNA by studying the effects these transcripts have on isoform expression levels elsewhere. Many ncRNA in mammalian cells are predicted to interact with PRC2 or other repressive complexes [64], and their global expression changes could be monitored before and after knock-in and knock-down experiments to identify their functional role, if any. Additionally, ChIP experiments could be performed to identify protein complexes to which each ncRNA is bound, which may identify those involved in regulatory functions from others which are expressed only incidentally (eg. from chance promoter up-regulation or histone modification of a regulatory locus inherited by humans but no longer functional). An RNA-FISH experiment could also be performed to examine the cellular localization of these ncRNA. A nuclear localization would suggest a regulatory role, while localization in other parts of the cell might suggest more exotic functionality.

The functional import of each transcriptome variant is a component of the whole interaction network which forms the backbone of the epitome. Differential alternative transcript variants are generally the result of alternative promoter usage [54], which implies changes in transcription factors, transcription co-factors, or chromatin-level changes took place. Chromatin methylation markers have been previously shown to influence transcription, such as the H3K4Me marker which has been shown to be a marker for alternative promoter usage [54]

and also a marker for non-coding RNA transcriptional start sites [63, 64]. Similarly, transcriptional silencing due to the H3K27Me marker, maintained by the PRC2 complex [128], may have resulted in the absence of transcription in breast cancer subtypes which did not express the variants we observed. Thus, the integration of information from histone methylation maps may prove useful in determining the epigenetic changes which led to the observed phenotype. Transcription factor occupation maps, which can be determined through ChIP-Seq experiments [118, 129], show where transcription factors and their co-factors bind to genomic DNA and initiate transcription. The transcription factor(s) responsible for alternative transcriptome variation, such as the estrogen receptor in the case of the intronic ACOX2 transcript in T47D cells, is a powerful determinate of the cellular changes which gave rise to these changes.

6.3 Conclusion

We have performed a study of the breast cancer tumor transcriptome by sequencing and analyzing RNA from 53 tumors. We developed two software utilities, ConsensusCluster and CUDAConsensusCluster, to differentiate between clinical disease subtypes. We found these tools to be accurate determinants of prior clinical subpopulations, and have used them to discover novel disease

classes in other tumors (see Appendix A: ConsensusCluster). We developed new methodology to identify alternative splicing variants in breast cancer subsets, both clinical and novel, and we identified novel intergenic non-coding RNA which are both evolutionarily conserved, and differentially expressed in these subsets. These we have shown to be effective and accurate biomarkers for the represented breast cancer disease subtypes. We have laid out a course of study which would elucidate the functionality of these transcriptome variants, and divine their place in the epigenetic landscape of cellular changes which give rise to these variants. It is our hope that we have contributed to our understanding of the underlying biology of breast cancer, and that our work will be used to develop future treatment for this disease.

APPENDIX A: ConsensusCluster

A.1 Consensus ensemble clustering

Clustering is the process of finding related groups in an unlabeled sample pool, using some numeric metric of relatedness between samples. In general, clustering algorithms focus on maximizing intra-cluster similarity, and minimizing inter-cluster similarity. These methods have been used in diverse fields such as disease classification and mtDNA phylogeny [9, 19, 130]. In many popular algorithms, however, inherent biases or stochastic variance in the methods limit the robustness of cluster analysis. For example, the well known k-means algorithm relies on centroids to label clusters, and is thus assumes clusters in the input data are spherical [18]. The k-means and self-organizing map [17] methods both rely on random initial conditions, which can cause variable cluster output. To mitigate these concerns, the consensus clustering method [23-25] calculates the likelihood of two samples being clustered together over many bootstraps of both sample and feature data and multiple clustering methods, then uses this information as a robust distance metric. The resulting clusters are robust against outliers and data perturbations, and variance/bias in individual clustering

algorithms is reduced [9, 25].

To facilitate the use of consensus clustering, we developed and released a software utility titled ConsensusCluster [25], which is freely available from http://code.google.com/p/consensus-cluster/. ConsensusCluster is a multi-platform package written in Python and C, distributed in both source and binary form. There is a simple Python API for developing data input parsers, and a parser for the simplest form (tab-delimited text file with headers in the first row and column) is provided. The data are stored internally as a 32-bit floating point matrix, along with any available sample metadata. ConsensusCluster also includes a number of methods for combining and filtering datasets from a Python console interface, making it a convenient tool for data manipulation.

A.2 Clustering steps

Each step in the ConsensusCluster process is documented both in the main window of ConsensusCluster and also in detailed log files. For each clustering iteration, performed on a given *k*-cluster value, each sample is assigned to one of *k* clusters and a log file is generated which presents the top genes separating each pair of clusters, as measured by signal-to-noise ratio (SNR) [69]. When all

clustering iterations have completed for each *k*, the consensus matrix is created by calculating the fraction of clustering iterations each sample appeared in the same cluster as each other sample. This matrix serves as a distance matrix for a final clustering attempt using single-link hierarchical clustering [15], which is used to calculate a cluster dendrogram. An image containing the dendrogram and consensus matrix is output to provide a production-quality visual representation of clustering information (e.g., Figure 1).

1. PCA Feature Selection

Principal Components Analysis (PCA) [108] is a data analysis and visualization method which decomposes the input data into orthogonal eigenvectors (principal components) of the covariance matrix. Sorted by largest eigenvalue, these eigenvectors can be thought of as independent vectors representing variance in the input data, i.e., the component with the largest eigenvalue is the unit vector in feature space which has the highest variance. ConsensusCluster selects features by first enumerating principal components which make up a user-specified fraction of the total variance. Then, features corresponding to the largest values in the selected eigenvectors (again, greater than a user-specified threshold) by absolute value are used in clustering. A PCA plot is produced and labeled based on sample metadata (Figure 17).

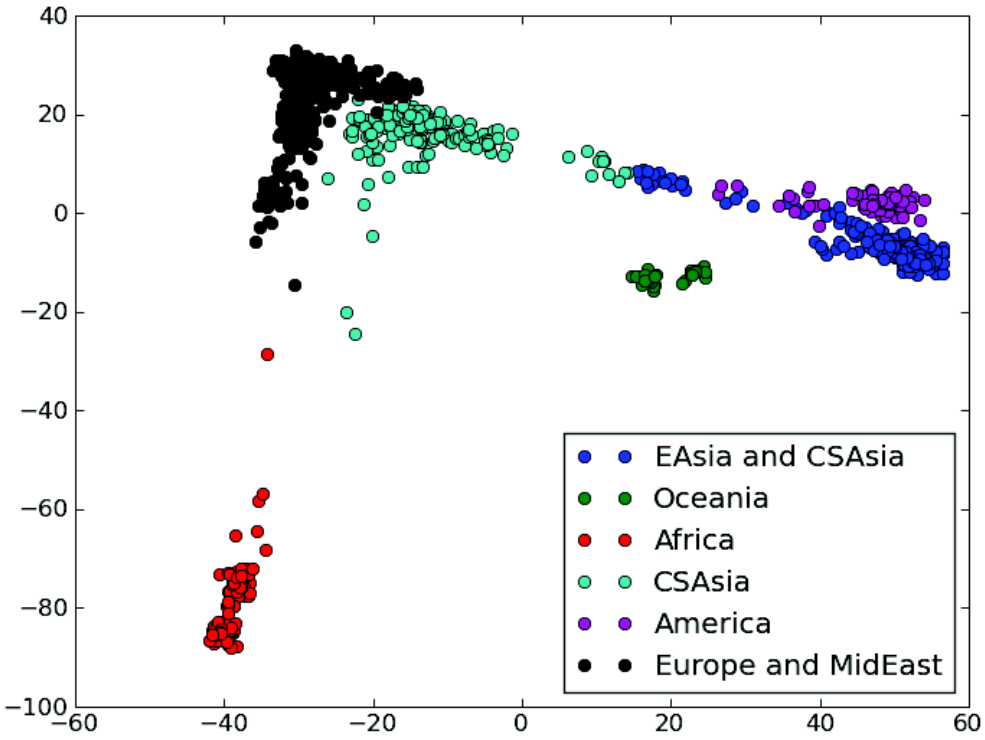

Figure 17

Figure 17. Samples from the Human Genome Diversity Project [131] are clearly separated by PCA, as performed by ConsensusCluster. The data are composed of 938 autosomal SNP samples, each containing 650,000 features. Clusters are determined based on 300 k-means clustering iterations, $k = 6$, then labeled based on geographical location.

2. Sample and Feature Bootstrap

In order to reduce influence of outlier samples and features, we perform

bootstrap aggregation ('bagging') of both sample and feature input, which has been shown to reduce variability in classification algorithms [18]. At each cluster iteration, a new dataset composed of a random subset (with replacement) of both samples and features is created. This new dataset is then used for clustering. When the consensus matrix is created, only the number of times each pair of samples appeared in the same bootstrap dataset is used towards the sample-sample cluster likelihood distance calculation. Bootstrapping also reduces the effects of sample perturbation, resulting in more robust clustering [9, 130].

3. Clustering Methods

ConsensusCluster implements four separate clustering algorithms, which can be selected through the Settings tab (Figure 18), or via the command line interface. The algorithms implemented are as follows:

- K-Means – *k* random initial centroids in feature space are iteratively moved towards local minima in sample density space. By running this algorithm many times (and many different initial conditions) and averaging the results, the expected value of the k-means algorithm can be found [18].
- Self-Organizing Map (SOM) – The SOM algorithm [17] is a simple neural network which proceeds by iteratively adjusting centroid “nodes” in feature space. These nodes are laid out in a multi-dimensional grid, and training

each node also adjusts other nodes based on physical grid distance, which results in a visual representation of the distance and relatedness between clusters. Once training is complete, the SOM network is used to classify each sample into a cluster assignment.

- Partition Around Medoids (PAM) – Rather than using centroids in feature space as in k-means, PAM [16] iteratively selects *k* samples as medoids, samples which represent the mean or median of each cluster. PAM requires only a distance matrix to function, which qualifies it to be used to cluster the consensus matrix after all iterations have completed.
- Hierarchical Clustering – Samples are iteratively assigned to clusters, and a new distance matrix is calculated at each iteration which includes the new “cluster” as a sample data point in place of the sample and sample(s) to which it is now joined [15]. This proceeds until all samples are assigned, and a min-cut to *k* clusters is performed. ConsensusCluster implements the “average,” “single,” and “complete” linkage options, which indicate how distance should be calculated from a sample to a cluster, based on the average of all samples in that cluster, the closest sample, and the farthest sample, respectively.

4. Building a Consensus

ConsensusCluster creates a new symmetric *n* x *n* matrix *M*, where each value

$M_{i,j}$ is the fraction of times samples *i* and *j* were clustered together, divided by the number of times *i* and *j* appeared in a bootstrap. This measurement of similarity forms the likelihood that this pair of samples are clustered together. By reversing this value ($1 - M_{i,j}$), the similarity matrix becomes a distance matrix, and can be used to perform the final clustering. Whereas hierarchical clustering is used by default, PAM is also an option.

5. Reorder the Consensus Matrix

Simulated annealing, a local optimization method, is used to reorder the consensus matrix so that more similar samples are grouped near to each other [130]. This has the effect that clusters are then represented on the consensus matrix heatmap as “boxes” along the diagonal (Figure 19). Both the consensus matrix heatmap and dendrogram, if available, are output to the user.

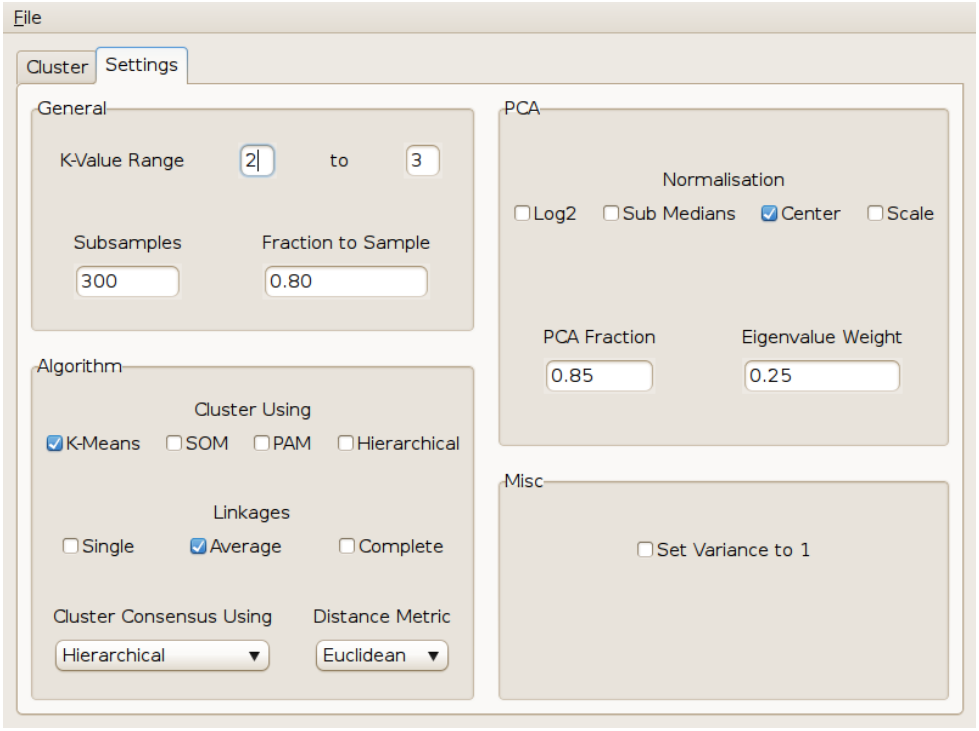

Figure 18

Figure 18. The settings available to the user in ConsensusCluster via the graphical user interface. A variety of simple input data normalization options are available, such as log2 re-expression, mean centering, and median subtraction. To change feature selection parameters, the total variance from selected principal components can be adjusted, as well as the absolute value of the weight of selected features from those eigenvectors. Four clustering algorithms are available to the user, and the distance metric for those methods can be changed from euclidean distance to sample correlation. ConsensusCluster will perform a complete run of the specified number of subsampled clustering iterations for each *k* value in the range entered here. All configuration options are

available from the command line interface.

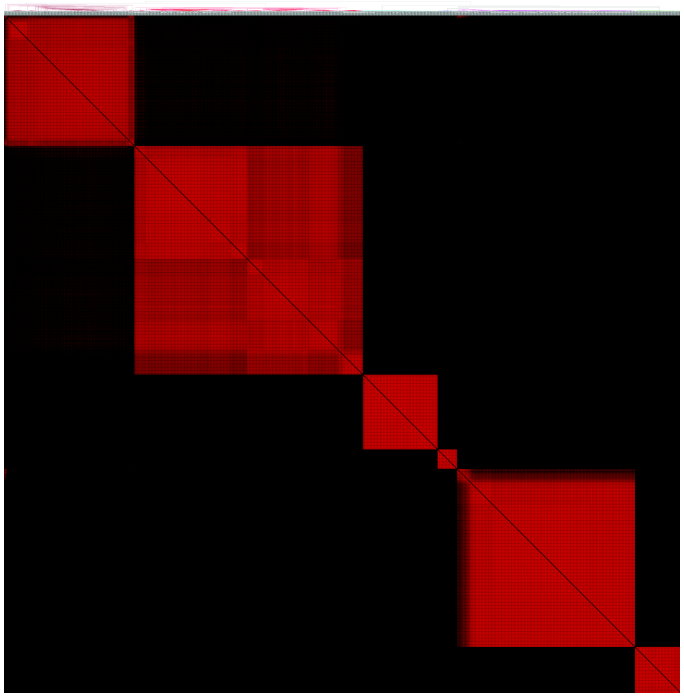

Figure 19

Figure 19. The consensus matrix generated after 300 k-means clustering iterations (k = 6) on 938 autosomal SNP samples from various geographical locations [131], with 650,000 measured SNPs in each sample. The consensus matrix is an n x n symmetric sample matrix, where each row and column represent the total fraction of times those samples were clustered together over all iterations. Brightness indicates the likelihood of cluster association, where lighter squares represent high association and darker squares less likely association.

A.3 ConsensusCluster in current research

ConsensusCluster [25] has recently found applications in both disease classification and phylogenetic tree identification. The first published usage of ConsensusCluster was in a phylogenetic application, where it was shown that without consensus clustering it is not possible to identify robust signatures for phylogenetic tree branch polymorphisms [130]. Since then, it has found applications in cancer biology, as a method for finding unknown subsets of tumor samples with correlated gene expression. Most prominently, ConsensusCluster was used to identify robust subtypes of clear cell renal cell carcinomas [26]. In this analysis, clustering revealed two distinct subpopulations, termed ccA and ccB, which were validated in a separate cohort. Later meta-analysis of this work found that both subtypes were robust, but that with additional information ccA could be again split into gender-related phenotypes [132], a result which may affect therapeutic measures taken against the disease in the future. As described in the present thesis, ConsensusCluster was also used to robustly identify breast cancer tumor subtypes in RNA-Seq data.

A.4 CUDAConsensusCluster

Consensus ensemble clustering is a method to utilize the information from many clustering methods and bootstraps of the data to form a more robust estimation of cluster association. The most significant limitation of this method is the time needed to run many clustering iterations, which increases linearly with the number of iterations run. To mitigate this, a separate version of ConsensusCluster, termed CUDAConsensusCluster, was created. This software utilizes the inherent massive parallelism of commodity computer graphics hardware to perform many clustering iterations simultaneously, resulting in high speed gains, especially over many clustering iterations. CUDAConsensusCluster is freely available from http://code.google.com/p/cuda-consensus-cluster/.

The Compute Unified Device Architecture (CUDA), developed by NVIDIA, is a parallel computing architecture which enables general purpose computation on graphics processing units (GPUs) [67]. Programming is generally accomplished using the “C for CUDA” language, though bindings to other languages, including Python, exist [68]. At present, the most powerful consumer GPU is the NVIDIA GTX 690, which contains 3072 processor cores [133]. An indefinite number of threads can be created for each task, and a hardware thread scheduler is responsible for ensuring processor cores are constantly in use if worker threads are available. Threads are split into a virtual “grid” to allow the user to specify which section of the grid is responsible for a given task. To that end, CUDA

requires that programs use the Single Input, Multiple Data (SIMD) model, where a single program is sent to all threads and work can be divided based on thread location on the grid. Each GPU is provided with a large global memory bank which is accessible through random access by computation threads, however it has very high latency. To ameliorate this, threads must access memory sequentially and in parallel, and a large amount of computation should be performed to use the GPU effectively. Each section of the thread grid, called a warp, also has its own shared memory, which is two orders of magnitude faster on average. Thus, commonly software which makes use of GPU programming will perform a parallel read of a sequential memory space into shared memory, followed by computation.

CUDAConsensusCluster uses an identical overall approach to ConsensusCluster, however, it provides accelerated parallel implementations of both k-means and self-organizing map [17] in order to provide substantial gains in speed. The parallel k-means implementation proceeds in the following steps.

1. A parallel Mersenne Twister implementation is used to generate random numbers for both bootstrapping of samples and features, and also to provide the initial *k* centroids.
2. The samples are each compared to *k* centroids, also in parallel. During this step, each thread warp reads the centroid information into shared

memory, up to 6 centroids. The same threads then read each sample datapoint and calculate correlation or euclidean distance. This calculation is “embarrassingly parallel” and scales nearly linearly with the number of processors. This calculation is serial in k, however, $k << n$ in most cases.

3. The centroids are updated in parallel, however, k is generally much smaller than the number of processors, and this step is minimally accelerated.
4. Steps 1-3 repeat until the centroids are no longer changed in step 3.

Self-organizing map [17] is a simple neural network which is trained to learn unknown groups in a dataset, and this classifier is then used to assign the input samples to clusters. Similarly to k-means, comparing input samples to centroid nodes and updating centroids are inherently parallel steps. However, this step must be repeated iteratively many times to ensure that later classification is accurate, which is a serial step. The result is that a parallel implementation of the SOM algorithm itself provides very little gain over the conventional serial implementation, unless the number of nodes is extremely large (on the order of millions). To provide speedup in CUDAConsensusCluster, we run many separate SOM implementations, which is parallel in the number of subsamples specified by the user. In theory, this provides speed gains linearly according to the number of iterations divided by the number of processors.

To complete the clustering task, CUDAConsensusCluster then calculates the consensus matrix in parallel. Further tasks, such as the final hierarchical clustering and consensus matrix reordering process, are accomplished using the same means present in ConsensusCluster. In practice, using an NVIDIA GTX 260 GPU with 192 GPU cores on a dataset with 260 samples, we have observed speedups of upwards of three orders of magnitude over the single core implementation. As the number of cores in GPU processors continues to increase, CUDAConsensusCluster should continue to scale when using the k-means algorithm on datasets with more samples than cores, and using the SOM algorithm if the number of subsamples is larger than the number of cores.

APPENDIX B: Experimental Methods

B.1 qRT-PCR

Total RNA was first extracted from each sample using the Trizol [73] reagent according to the manufacturer's protocol. A complementary DNA (cDNA) library was created for each sample using the Transcriptor First Strand Synthesis kit from Roche. The real-time polymerase chain reaction (RT-PCR) was run on the Mx3005p QPCR system using the SYBR Green dye for fluorescent detection. Each experiment was run in triplicate, and a dilution curve was run on each plate for each primer pair to assess primer efficiency. All expression levels were calculated relative to GAPDH mRNA expression, using the method suggested in [88]. Primers were designed using the Primer3 software package.

B.2 Alternative splicing validation primers

Gene	RefSeq ID	Forward 5-->3	Reverse 5-->3
TPD52 E1-E3	NM_005079	ATGGACCGCGGCGAGCAA	GTTTCCGCTTGATCTCTGCT
TPD52 E2-E4	NM_001025252	CACAGAGACCCTCTCGGAAG	GAGCCAACAGACGAAAAAGC
ACOX2 int9/10-E11	NM_003500	ACAGGGTTGGTCCCTATGGT	AGGTCAGGTGCGGTGAGATA
ACOX2 E9-E11	NM_003500	GCAAAGGTCCTGGACTACCA	CCAGGGGACATCTGAGTCTG
NESP55	NM_016592	GAGGCAGACCTTGAGCTGTC	CAACTTGAGAGCGTGCAGAC

REFERENCES

1. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
2. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
3. Umar, A., et al., *Identification of a putative protein profile associated with tamoxifen therapy resistance in breast cancer.* Mol Cell Proteomics, 2009. **8**(6): p. 1278-94.
4. Zwijsen, R.M., et al., *CDK-independent activation of estrogen receptor by cyclin D1.* Cell, 1997. **88**(3): p. 405-15.
5. Bindels, E.M., et al., *Involvement of G1/S cyclins in estrogen-independent proliferation of estrogen receptor-positive breast cancer cells.* Oncogene, 2002. **21**(53): p. 8158-65.
6. Harari, D. and Y. Yarden, *Molecular mechanisms underlying ErbB2/HER2 action in breast cancer.* Oncogene, 2000. **19**(53): p. 6102-14.
7. Klapper, L.N., et al., *Biochemical and clinical implications of the ErbB/HER signaling network of growth factor receptors.* Adv Cancer Res, 2000. **77**: p. 25-79.
8. Le, X.F., F. Pruefer, and R.C. Bast, *HER2-targeting antibodies modulate the cyclin-dependent kinase inhibitor p27Kip1 via multiple signaling pathways.* Cell Cycle, 2005. **4**(1): p. 87-95.
9. Alexe, G., et al., *High expression of lymphocyte-associated genes in node-negative HER2+ breast cancers correlates with lower recurrence rates.* Cancer Res, 2007. **67**(22): p. 10669-76.
10. Cronin, M., et al., *Measurement of gene expression in archival paraffin-embedded tissues: development and performance of a 92-gene reverse transcriptase-polymerase chain reaction assay.* Am J Pathol, 2004. **164**(1): p. 35-42.
11. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
12. Okoniewski, M.J. and C.J. Miller, *Comprehensive analysis of affymetrix exon arrays using BioConductor.* PLoS Comput Biol, 2008. **4**(2): p. e6.
13. Affymetrix. Available from: http://www.affymetrix.com.
14. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.
15. Hartigan, J.A., *Clustering algorithms*. Wiley series in probability and mathematical statistics. 1975, New York,: Wiley. xiii, 351 p.
16. Kaufman, L. and P.J. Rousseeuw, *Finding groups in data : an introduction*

to cluster analysis. Wiley series in probability and mathematical statistics. Applied probability and statistics,. 1990, New York: Wiley. xiv, 342 p.
17. Kohonen, T., *Self-organizing maps*. 2nd ed. Springer series in information sciences,. 1997, Berlin ; New York: Springer. xvii, 426 p.
18. Alpaydin, E., *Introduction to machine learning*. Adaptive computation and machine learning. 2004, Cambridge, Mass.: MIT Press. xxx, 415 p.
19. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.
20. Sorlie, T., et al., *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8418-23.
21. Wang, Y., et al., *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* Lancet, 2005. **365**(9460): p. 671-9.
22. Hennessy, B.T., et al., *Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics.* Cancer Res, 2009. **69**(10): p. 4116-24.
23. Strehl, A. and J. Ghosh, *Cluster ensembles --- a knowledge reuse framework for combining multiple partitions.* J. Mach. Learn. Res., 2003. **3**: p. 583-617.
24. Monti, S., et al., *Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data.* Mach. Learn., 2003. **52**(1-2): p. 91-118.
25. Seiler, M., et al., *ConsensusCluster: a software tool for unsupervised cluster discovery in numerical data.* OMICS, 2010. **14**(1): p. 109-13.
26. Brannon, A.R., et al., *Molecular Stratification of Clear Cell Renal Cell Carcinoma by Consensus Clustering Reveals Distinct Subtypes and Survival Patterns.* Genes Cancer, 2010. **1**(2): p. 152-163.
27. Pauletti, G., et al., *Detection and quantitation of HER-2/neu gene amplification in human breast cancer archival material using fluorescence in situ hybridization.* Oncogene, 1996. **13**(1): p. 63-72.
28. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.
29. Cloonan, N., et al., *Stem cell transcriptome profiling via massive-scale mRNA sequencing.* Nat Methods, 2008. **5**(7): p. 613-9.
30. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.
31. Robertson, G., et al., *De novo assembly and analysis of RNA-seq data.* Nat Methods, 2010. **7**(11): p. 909-12.
32. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

33. Kapur, K., et al., *Cross-hybridization modeling on Affymetrix exon arrays.* Bioinformatics, 2008. **24**(24): p. 2887-93.
34. Rio, D.C., *RNA processing.* Curr Opin Cell Biol, 1992. **4**(3): p. 444-52.
35. Schwerk, C. and K. Schulze-Osthoff, *Regulation of apoptosis by alternative pre-mRNA splicing.* Mol Cell, 2005. **19**(1): p. 1-13.
36. Burset, M., I.A. Seledtsov, and V.V. Solovyev, *Analysis of canonical and non-canonical splice sites in mammalian genomes.* Nucleic Acids Res, 2000. **28**(21): p. 4364-75.
37. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.
38. Parra, M., et al., *Differential domain evolution and complex RNA processing in a family of paralogous EPB41 (protein 4.1) genes facilitate expression of diverse tissue-specific isoforms.* Genomics, 2004. **84**(4): p. 637-46.
39. Mahotka, C., et al., *Survivin-deltaEx3 and survivin-2B: two novel splice variants of the apoptosis inhibitor survivin with different antiapoptotic properties.* Cancer Res, 1999. **59**(24): p. 6097-102.
40. Li, F., *Role of survivin and its splice variants in tumorigenesis.* Br J Cancer, 2005. **92**(2): p. 212-6.
41. Jiang, Z.H., et al., *Regulation of Ich-1 pre-mRNA alternative splicing and apoptosis by mammalian splicing factors.* Proc Natl Acad Sci U S A, 1998. **95**(16): p. 9155-60.
42. Fushimi, K., et al., *Up-regulation of the proapoptotic caspase 2 splicing isoform by a candidate tumor suppressor, RBM5.* Proc Natl Acad Sci U S A, 2008. **105**(41): p. 15708-13.
43. Kumar, S., *Caspase 2 in apoptosis, the DNA damage response and tumour suppression: enigma no more?* Nat Rev Cancer, 2009. **9**(12): p. 897-903.
44. Venables, J.P., et al., *Identification of alternative splicing markers for breast cancer.* Cancer Res, 2008. **68**(22): p. 9525-31.
45. Andre, F., et al., *Exonic expression profiling of breast cancer and benign lesions: a retrospective analysis.* Lancet Oncol, 2009. **10**(4): p. 381-90.
46. Lapuk, A., et al., *Exon-level microarray analyses identify alternative splicing programs in breast cancer.* Mol Cancer Res, 2010. **8**(7): p. 961-74.
47. Wang, K., et al., *MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.* Nucleic Acids Res, 2010. **38**(18): p. e178.
48. Deng, G., et al., *Loss of heterozygosity in normal tissue adjacent to breast carcinomas.* Science, 1996. **274**(5295): p. 2057-9.
49. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

50. Siebert, P.D. and B.C. Huang, *Identification of an alternative form of human lactoferrin mRNA that is expressed differentially in normal tissues and tumor-derived cell lines.* Proc Natl Acad Sci U S A, 1997. **94**(6): p. 2198-203.
51. Benaissa, M., et al., *Expression and prognostic value of lactoferrin mRNA isoforms in human breast cancer.* Int J Cancer, 2005. **114**(2): p. 299-306.
52. Hoedt, E., et al., *Discrimination and evaluation of lactoferrin and delta-lactoferrin gene expression levels in cancer cells and under inflammatory stimuli using TaqMan real-time PCR.* Biometals, 2010. **23**(3): p. 441-52.
53. Mariller, C., et al., *Delta-lactoferrin, an intracellular lactoferrin isoform that acts as a transcription factor (1) (1) This article is part of a Special Issue entitled Lactoferrin and has undergone the Journal's usual peer review process.* Biochem Cell Biol, 2012.
54. Pal, S., et al., *Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development.* Genome Res, 2011. **21**(8): p. 1260-72.
55. Flicek, P., et al., *Ensembl 2012.* Nucleic Acids Res, 2012. **40**(Database issue): p. D84-90.
56. Weakley, S.M., et al., *Expression and function of a large non-coding RNA gene XIST in human cancer.* World J Surg, 2011. **35**(8): p. 1751-6.
57. Gupta, R.A., et al., *Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis.* Nature, 2010. **464**(7291): p. 1071-6.
58. Niinuma, T., et al., *Upregulation of miR-196a and HOTAIR drive malignant character in gastrointestinal stromal tumors.* Cancer Res, 2012. **72**(5): p. 1126-36.
59. Geng, Y.J., et al., *Large intervening non-coding RNA HOTAIR is associated with hepatocellular carcinoma progression.* J Int Med Res, 2011. **39**(6): p. 2119-28.
60. Jia, H., et al., *Genome-wide computational identification and manual annotation of human long noncoding RNA genes.* RNA, 2010. **16**(8): p. 1478-87.
61. Zhao, J., et al., *Genome-wide identification of polycomb-associated RNAs by RIP-seq.* Mol Cell, 2010. **40**(6): p. 939-53.
62. Edlund, K., et al., *Data-driven unbiased curation of the TP53 tumor suppressor gene mutation database and validation by ultradeep sequencing of human tumors.* Proc Natl Acad Sci U S A, 2012. **109**(24): p. 9551-6.
63. Guttman, M., et al., *Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.* Nature, 2009. **458**(7235): p. 223-7.
64. Khalil, A.M., et al., *Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene*

expression. Proc Natl Acad Sci U S A, 2009. **106**(28): p. 11667-72.
65. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.
66. Seiler, M. *CUDA Consensus Cluster*. 2010; Available from: http://code.google.com/p/cuda-consensus-cluster/.
67. Nickolls, J., et al., *Scalable Parallel Programming with CUDA.* Queue, 2008. **6**(2): p. 40-53.
68. Kl, A., et al., *PyCUDA and PyOpenCL: A scripting-based approach to GPU run-time code generation.* Parallel Comput., 2012. **38**(3): p. 157-174.
69. Hengpraprohm, S. and P. Chongstitvatana. *Selecting Informative Genes from Microarray Data for Cancer Classification with Genetic Programming Classifier Using K-Means Clustering and SNR Ranking*. in *Frontiers in the Convergence of Bioscience and Information Technologies, 2007. FBIT 2007*. 2007.
70. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2007. **35**(Database issue): p. D61-5.
71. Wilming, L.G., et al., *The vertebrate genome annotation (Vega) database.* Nucleic Acids Res, 2008. **36**(Database issue): p. D753-60.
72. Welboren, W.J., et al., *ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands.* EMBO J, 2009. **28**(10): p. 1418-28.
73. Chomczynski, P. and N. Sacchi, *Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction.* Anal Biochem, 1987. **162**(1): p. 156-9.
74. Levin, J.Z., et al., *Comprehensive comparative analysis of strand-specific RNA sequencing methods.* Nat Methods, 2010. **7**(9): p. 709-15.
75. Hayward, B.E., et al., *Bidirectional imprinting of a single gene: GNAS1 encodes maternally, paternally, and biallelically derived proteins.* Proc Natl Acad Sci U S A, 1998. **95**(26): p. 15475-80.
76. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotechnol, 2011. **29**(7): p. 644-52.
77. Anders, S. *HTSeq: Analysing high-throughput sequencing data with Python*. 2010; Available from: http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html.
78. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
79. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.
80. Robinson, M.D. and A. Oshlack, *A scaling normalization method for*

differential expression analysis of RNA-seq data. Genome Biol, 2010. **11**(3): p. R25.
81. Oshlack, A. and M.J. Wakefield, *Transcript length bias in RNA-seq data confounds systems biology.* Biol Direct, 2009. **4**: p. 14.
82. Bullard, J.H., et al., *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.* BMC Bioinformatics, 2010. **11**: p. 94.
83. Hansen, K.D., S.E. Brenner, and S. Dudoit, *Biases in Illumina transcriptome sequencing caused by random hexamer priming.* Nucleic Acids Res, 2010. **38**(12): p. e131.
84. Marioni, J.C., et al., *RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.* Genome Res, 2008. **18**(9): p. 1509-17.
85. Jiang, H. and W.H. Wong, *Statistical inferences for isoform expression in RNA-Seq.* Bioinformatics, 2009. **25**(8): p. 1026-32.
86. Srivastava, S. and L. Chen, *A two-parameter generalized Poisson model to improve the analysis of RNA-seq data.* Nucleic Acids Res, 2010. **38**(17): p. e170.
87. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
88. Pfaffl, M.W., *A new mathematical model for relative quantification in real-time RT-PCR.* Nucleic Acids Res, 2001. **29**(9): p. e45.
89. Gutierrez, L., et al., *Towards a systematic validation of references in real-time rt-PCR.* Plant Cell, 2008. **20**(7): p. 1734-5.
90. Rieu, I. and S.J. Powers, *Real-time quantitative RT-PCR: design, calculations, and statistics.* Plant Cell, 2009. **21**(4): p. 1031-3.
91. Szabo, A., et al., *Statistical modeling for selecting housekeeper genes.* Genome Biol, 2004. **5**(8): p. R59.
92. Rosenbloom, K.R., et al., *ENCODE whole-genome data in the UCSC Genome Browser: update 2012.* Nucleic Acids Res, 2012. **40**(Database issue): p. D912-7.
93. Roberts, A., et al., *Identification of novel transcripts in annotated genomes using RNA-Seq.* Bioinformatics, 2011. **27**(17): p. 2325-9.
94. Carninci, P., et al., *The transcriptional landscape of the mammalian genome.* Science, 2005. **309**(5740): p. 1559-63.
95. Engstrom, P.G., et al., *Complex Loci in human and mouse genomes.* PLoS Genet, 2006. **2**(4): p. e47.
96. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
97. Kong, L., et al., *CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine.* Nucleic Acids Res,

2007. **35**(Web Server issue): p. W345-9.
98. Birol, I., et al., *De novo transcriptome assembly with ABySS.* Bioinformatics, 2009. **25**(21): p. 2872-7.
99. Otogenetics, *Otogenetics RNA-Seq price quote*.
100. Dutertre, M., et al., *Exon-based clustering of murine breast tumor transcriptomes reveals alternative exons whose expression is associated with metastasis.* Cancer Res, 2010. **70**(3): p. 896-905.
101. Olsson, E., et al., *CD44 isoforms are heterogeneously expressed in breast cancer and correlate with tumor subtypes and cancer stem cell markers.* BMC Cancer, 2011. **11**: p. 418.
102. Ertel, A., et al., *Pathway-specific differences between tumor cell lines and normal and tumor tissue cells.* Mol Cancer, 2006. **5**(1): p. 55.
103. Tsuji, K., et al., *Breast cancer cell lines carry cell line-specific genomic alterations that are distinct from aberrations in breast cancer tissues: comparison of the CGH profiles between cancer cell lines and primary cancer tissues.* BMC Cancer, 2010. **10**: p. 15.
104. Hiller, D., et al., *Identifiability of isoform deconvolution from junction arrays and RNA-Seq.* Bioinformatics, 2009. **25**(23): p. 3056-9.
105. Fu, X., et al., *Estimating accuracy of RNA-Seq and microarrays with proteomics.* BMC Genomics, 2009. **10**: p. 161.
106. *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
107. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J. Roy. Statist. Soc. Ser. B, 1995. **57**(1): p. 289--300.
108. Jolliffe, I.T., *Principal component analysis*. 2nd ed. Springer series in statistics. 2002, New York: Springer. xxix, 487 p.
109. Balleine, R.L., et al., *The hD52 (TPD52) gene is a candidate target gene for events resulting in increased 8q21 copy number in human breast carcinoma.* Genes Chromosomes Cancer, 2000. **29**(1): p. 48-57.
110. Bilal, E., et al., *Amplified Loci on Chromosomes 8 and 17 Predict Early Relapse in ER-Positive Breast Cancers.* PLoS One, 2012. **7**(6): p. e38575.
111. Dutertre, M., et al., *Estrogen regulation and physiopathologic significance of alternative promoters in breast cancer.* Cancer Res, 2010. **70**(9): p. 3760-70.
112. Wang, R., et al., *PrLZ, a novel prostate-specific and androgen-responsive gene of the TPD52 family, amplified in chromosome 8q21.1 and overexpressed in human prostate cancer.* Cancer Res, 2004. **64**(5): p. 1589-94.
113. Zhang, D., et al., *PrLZ protects prostate cancer cells from apoptosis induced by androgen deprivation via the activation of Stat3/Bcl-2 pathway.* Cancer Res, 2011. **71**(6): p. 2193-202.

114. Baumgart, E., et al., *Molecular characterization of the human peroxisomal branched-chain acyl-CoA oxidase: cDNA cloning, chromosomal assignment, tissue distribution, and evidence for the absence of the protein in Zellweger syndrome.* Proc Natl Acad Sci U S A, 1996. **93**(24): p. 13748-53.
115. el Bouhtoury, F., et al., *Peroxisomal enzymes in normal and tumoral human breast.* J Pathol, 1992. **166**(1): p. 27-35.
116. Hodo, Y., et al., *Comprehensive gene expression analysis of 5'-end of mRNA identified novel intronic transcripts associated with hepatocellular carcinoma.* Genomics, 2010. **95**(4): p. 217-23.
117. Tsiambas, E., et al., *Significance of estrogen receptor 1 (ESR-1) gene imbalances in colon and hepatocellular carcinomas based on tissue microarrays analysis.* Med Oncol, 2011. **28**(4): p. 934-40.
118. Joseph, R., et al., *Integrative model of genomic factors for determining binding site selection by estrogen receptor-α.* Mol Syst Biol, 2010. **6**: p. 456.
119. Tokuoka, K., et al., *Three-dimensional structure of rat-liver acyl-CoA oxidase in complex with a fatty acid: insights into substrate-recognition and reactivity toward molecular oxygen.* J Biochem, 2006. **139**(4): p. 789-95.
120. Gorello, P., et al., *t(3;11)(q12;p15)/NUP98-LOC348801 fusion transcript in acute myeloid leukemia.* Haematologica, 2008. **93**(9): p. 1398-401.
121. Olsen, R.J. and S.H. Hinrichs, *Phosphorylation of the EWS IQ domain regulates transcriptional activity of the EWS/ATF1 and EWS/FLI1 fusion proteins.* Oncogene, 2001. **20**(14): p. 1756-64.
122. Cho, S.J. and X. Chen, *Myosin VI is differentially regulated by DNA damage in p53- and cell type-dependent manners.* J Biol Chem, 2010. **285**(35): p. 27159-66.
123. Regnier, C.H., et al., *Expression of a truncated form of hHb1 hair keratin in human breast carcinomas.* Br J Cancer, 1998. **78**(12): p. 1640-4.
124. Boulay, A., et al., *Transcription regulation and protein subcellular localization of the truncated basic hair keratin hHb1-DeltaN in human breast cancer cells.* J Biol Chem, 2001. **276**(25): p. 22954-64.
125. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.
126. Pang, K.C., M.C. Frith, and J.S. Mattick, *Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function.* Trends Genet, 2006. **22**(1): p. 1-5.
127. Xu, W., et al., *Human transcriptome array for high-throughput clinical studies.* Proc Natl Acad Sci U S A, 2011. **108**(9): p. 3707-12.
128. Young, M.D., et al., *ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity.* Nucleic Acids Res, 2011. **39**(17):

p. 7415-27.
129. Ross-Innes, C.S., et al., *Differential oestrogen receptor binding is associated with clinical outcome in breast cancer.* Nature, 2012. **481**(7381): p. 389-93.
130. Alexe, G., et al., *PCA and clustering reveal alternate mtDNA phylogeny of N and M clades.* J Mol Evol, 2008. **67**(5): p. 465-87.
131. Li, J.Z., et al., *Worldwide human relationships inferred from genome-wide patterns of variation.* Science, 2008. **319**(5866): p. 1100-4.
132. Brannon, A.R., et al., *Meta-analysis of clear cell renal cell carcinoma gene expression defines a variant subgroup and identifies gender influences on tumor biology.* Eur Urol, 2012. **61**(2): p. 258-68.
133. NVIDIA. *NVIDIA GTX 690 Specifications*. 2012; Available from: http://www.geforce.com/hardware/desktop-gpus/geforce-gtx-690/specifications.
134. Griffith, M., et al. *Alternative expression analysis by RNA sequencing.* Nat Methods, 2010 **7**(10): p. 843-847.