DECODING THE TRANSCRIPTIONAL LANDSCAPE OF TRIPLE-NEGATIVE BREAST CANCER USING NEXT GENERATION WHOLE TRANSCRIPTOME SEQUENCING

Milan Radovich

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Medical and Molecular Genetics
Indiana University

August 2011

Accepted by the Faculty of Indiana University, in partial
fulfillment of the requirements for the degree of Doctor of Philosophy.

Bryan P. Schneider, MD, Chair

David A. Flockhart, MD, PhD

Doctoral Committee

Mircea Ivan, MD, PhD

June 15, 2011

Brittney-Shea Herbert, PhD

Brenda R. Grimes, PhD

Harikrishna Nakshatri, PhD

ACKNOWLEDGEMENTS

As always, the success of our endeavors are built upon the support and love of so many people who help make life possible. Through my graduate career, I have undeservedly received an abundance of encouragement, support, and resources from so many which has allowed me to be where I am today.

To my mentor Dr. Bryan Schneider, we have been together since the lab first started, and what an amazing time it has been! Through victories and failures, and time of joy and sorrow, you have always been there. I truly appreciate your friendship and mentorship through the years. Your persistence and support is what has allowed me to be the translational scientist I am today. Thank you for believing in me and for giving me the opportunity to pursue my scientific aspirations. I would not be graduating if it wasn't for you, I will forever be indebted.

To my labmates, Nawal Kassem, Brad Hancock, and Lateef Aregbe, I couldn't have done this without you guys. Thank you for enduring the many days of my frustration when things just didn't seem to work. Also, thank you for always being there, whether it was my talks or presentations or just needing someone to chat with, your presence was also so appreciated. Thank you for always making me laugh and always letting me know that I was loved by you all. The memories, I will never forget. Whether it was our trips to Washington, D.C., Denver, San Antonio, or Orlando, or just the many times we chatted in lab or at lunch, it is the camaraderie that made coming to lab everyday a joy.

To my collaborators Drs. Susan Clare, George Sledge, and Mircea Ivan, and also to Connie Rufenbarger, my project would have not been possible with you. Thank you for taking the risk and investing in a young graduate student with a crazy idea. I can't tell you how much it meant to me when you decided that I was worth your investment.

Thank you for the countless conversations and always lending yet another good scientific idea. I will always be deeply indebted, and I only hope that I can return many fold what you gave to me.

A special thanks goes to Dr. David A. Flockhart, I wouldn't have been in graduate school if it wasn't for you. I will forever be indebted to your kindness and generosity. I hope you will always know how appreciative I am of what you gave.

To my committee, Drs. Bryan Schneider, David Flockhart, Mircea Ivan, Brittney-Shea Herbert, Brenda Grimes, Harikrishna Nakshatri, thank you for your time and sacrifice in ensuring that I was becoming the best scientist that I can be. Thank you for your thoughtful suggestions and encouragement, and also thank you so much for making sure that I wasn't going too overboard in my scientific ambitions. I am indebted to the time that you gave while never expecting anything in return.

To my wife Betsy and my daughter Grace, thank you for being there for me through this entire time. I know it was so busy, and there wasn't always a ton of time, but we always knew it would be done soon. Now that I will finally have a real job, we can finally go on that family vacation. To my mother Desanka, brother Alex, and my In-laws, Jim & Micki Geise thank you for your love and unwavering support. It has always been one of my greatest joys to know that I have made you proud. I hope that I will continue to do so.

I know there are probably others that I am missing, but it goes without mentioning that so many were involved in making this project happen. To the volunteers and staff of the Susan G. Komen Tissue Bank who tirelessly collected and prepared the normal breast tissues, thank you for all you work and for the many many days of LCM! To the folks at Cofactor Genomics and Applied Biosystems whose technical expertise in next-generation sequencing made so much of this work possible, thank you for being such great collaborators and for being great friends. To the Cancer Biology Training Program

and the Indiana CTSI who provided the funding that allowed me to be a successful student. Finally, to the Indiana University Simon Cancer Center, the Department of Medical & Molecular Genetics, and the Department of Surgery, thank you for providing the resources to make this project happen… and ensuring that the work still continues.

ABSTRACT

Milan Radovich

DECODING THE TRANSCRIPTIONAL LANDSCAPE OF TRIPLE-NEGATIVE BREAST CANCER USING NEXT GENERATION WHOLE TRANSCRIPTOME SEQUENCING

Triple-negative breast cancers (TNBCs) are negative for the expression of estrogen (ER), progesterone (PR), and HER-2 receptors. TNBC accounts for 15% of all breast cancers and results in disproportionally higher mortality compared to ER & HER2-positive tumours. Moreover, there is a paucity of therapies for this subtype of breast cancer resulting primarily from an inadequate understanding of the transcriptional differences that differentiate TNBC from normal breast. To this end, we embarked on a comprehensive examination of the transcriptomes of TNBCs and normal breast tissues using next-generation whole transcriptome sequencing (RNA-Seq). By comparing RNA-seq data from these tissues, we report the presence of differentially expressed coding and non-coding genes, novel transcribed regions, and mutations not previously reported in breast cancer. From these data we have identified two major themes. First, BRCA1 mutations are well known to be associated with development of TNBC. From these data we have identified many genes that work in concert with BRCA1 that are dysregulated suggesting a role of BRCA1 associated genes with sporadic TNBC. In addition, we observe a mutational profile in genes also associated with BRCA1 and DNA repair that lend more evidence to its role. Second, we demonstrate that using microdissected normal epithelium maybe an optimal comparator when searching for novel therapeutic targets for TNBC. Previous studies have used other controls such as reduction mammoplasties, adjacent normal tissue, or other breast cancer subtypes, which may be sub-optimal and have lead to identifying ineffective therapeutic targets. Our data suggests that the comparison of microdissected ductal epithelium to TNBC can identify potential therapeutic targets that may lead to be better clinical efficacy. In summation, with these data, we provide a detailed transcriptional landscape of TNBC and normal breast that we believe will lead to a better understanding of this complex disease.

Bryan P. Schneider, MD, Chair

TABLE OF CONTENTS

List of Tables x

List of Figures xi

List of Abbreviations xiii

Chapter 1: Introduction

- 1.1 Triple-Negative Breast Cancer 1
  - 1.1.1 Clinical characteristics and treatment of triple-negative breast cancer 1
  - 1.1.2 Histological and molecular characterization of TNBC 8
  - 1.1.3 Hereditary and somatic mutation profile of TNBC 13
- 1.2 The normal breast in cancer research 17
  - 1.2.1 Overview of normal epithelium in the breast 17
  - 1.2.2 The use of normal breast tissue in cancer research 22
- 1.3 Next-generation whole transcriptome sequencing (RNA-Seq) 24
  - 1.3.1 Overview of next-generation sequencing and its role in cancer research 24
  - 1.3.2 RNA-sequencing overview and chemistry of the Applied Biosystems SOLiD sequencer 29
  - 1.3.3 Overview of SOLiD RNA-seq bioinformatics and data analysis 33
- 1.4 Statement of purpose 44

Chapter 2: Differential gene expression of the transcriptomes of TNBC and normal breast tissue

- 2.1 Introduction 45
- 2.2 Materials and methods 47
- 2.3 Results 58

2.3.1 RNA sequencing output of TNBCs and normals.............................58
2.3.2 Profiling and differential expression of known genes from the RefSeq database...59
2.3.3 Profiling and differential expression of noncoding RNAs.................80
2.3.4 Novel transcribed regions..82
2.4 Discussion 84

Chapter 3: Detection of mutations in TNBC from mapped RNA-seq data

3.1 Introduction..87
3.2 Materials and methods...90
3.3 Results...95
3.3.1 Point mutations 95
3.3.2 Small insertions and deletions...99
3.3.3 Gene fusions...101
3.4 Discussion 103
Chapter 4: Summary 105

Appendices

Appendix 1: Bioinformatics 110
Appendix 2: Lists of differentially expressed genes..144
References...145

Curriculum Vitae

LIST OF TABLES

Table 1. Ongoing clinical trials of target agents in TNBC 5

Table 2. Representative example of RPKM data 38

Table 3. Normal samples used for RNA-seq with corresponding information.................50

Table 4. TNBC samples used for RNA-seq with corresponding information 51

Table 5. Sequencing output from RNA-seq of 10 normal breast tissues and 10 TNBCs...60

Table 6. Differential gene expression of three drug targets clinically tested in enriched TNBC populations and their clinical trial outcomes...67

Table 7. 25 TNBC therapeutic targets identified by RNA-Seq.......................................78

Table 8. Gene fusions identified from mapped RNA-seq data 102

LIST OF FIGURES

Figure 1. Kaplan-Meier curve of breast-specific survival in triple-negative and other breast cancers 2
Figure 2. Overall survival as a function of response to neoadjuvant chemotherapy in TNBC 3
Figure 3. Mechanism of synthetic lethality by BRCA deficiency and PARP inhibition 7
Figure 4. Representative example of breast tumors classified by the intrinsic subtypes using unsupervised hierarchical clustering of gene expression array data 9
Figure 5. Shared characteristics of TNBC, basal-like breast cancer (BLBC), and BRCA1-associated breast cancer. 10
Figure 6. Mutated cancer genes in TNBC identified by the Wellcome Trust Sanger Institute 15
Figure 7. Low coverage next-generation whole genome sequencing of 24 breast cancers 16
Figure 8. Schematic of the normal cell types that comprise the human mammary duct 18
Figure 9. Illustration of the normal breast cancer and histology of the TDLU 19
Figure 10. Schematic of the normal breast stem cell hierarchy 20
Figure 11. Increase in sequencing output due to technological advances in the last decade 25
Figure 12. Current next-generation sequencing platforms 26
Figure 13. Schematic of RNA-seq library preparation for ABI SOLiD sequencing 31
Figure 14. Colorspace sequencing by ligation using the ABI SOLiD 34
Figure 15. Example of RNA-seq reads mapping to a gene 36
Figure 16. Example of non-uniformity of coverage when calling mutations from

mapped RNA-seq data 40
Figure 17. Example of a deletion detected from mapped RNA-seq data........................41
Figure 18. Diagram of detecting a gene fusion from reads crossing a
fusion junction 42
Figure 19. Picture of normal ductal epithelium from frozen normal breast tissue before
and after LCM...49
Figure 20. Principal components analysis (PCA) of normal samples 61
Figure 21. Principal components analysis (PCA) of TNBC and normal samples...........63
Figure 22. BRCA1 in DNA damage response pathway 65
Figure 23. Biological validation of next-generation sequencing results of EGFR, KIT, and
PARP1 using Taqman qPCR...68
Figure 24. Immunohistochemistry (IHC) of EGFR and KIT on normal and TNBC breast
tissues 69
Figure 25. Representative IHC stains for EGFR and KIT of normal and
TNBC tissues 72
Figure 26. Volcano plot of significant receptor tyrosine kinases (RTKs).........................74
Figure 27. The NEK2 interaction network 76
Figure 28. PCA of novel transcribed regions 83
Figure 29. Flow diagram of point mutation calling and filtering pipeline..........................92
Figure 30. Point mutations in the BRCA1 pathway 97
Figure 31. PCA of TNBC and normal samples with highlighted BRCA1
mutant sample..100
Figure 32. Example of output sequencing test files from the ABI SOLiD
sequencer 112
Figure 33. Schematic of the premise behind the SASR algorithm...............................143

LIST OF ABBREVIATIONS

AACR	American Association for Cancer Research
ABI	Applied Biosystems
ANOVA	Analysis of Variance
BAM	Binary Alignment/Mapping
BER	Base Excision Repair
BLBC	Basal-Like Breast Cancer
cDNA	Copy Deoxyribonucleic Acid
COSMIC	Catalogue of Somatic Mutations in Cancer
dbSNP	Database of Single Nucleotide Polymorphisms
DNA	Deoxyribonucleic Acid
DSB	Double Stranded Break
ePCR	Emulsion Polymerase Chain Reaction
ER	Estrogen Receptor
FDA	Food and Drug Administration
GWAS	Genome-Wide Association Study
IDC-NST	Infiltrating Ductal Carcinoma – No Special Type
IHC	Immunohistochemistry
LCM	Laser Capture Microdissection
lincRNA	Long Intergenic Non-Coding Ribonucleic Acid
miRNA	microRibonucleic Acid
mRNA	Messenger Ribonucleic Acid
mTOR	Mammalian Target of Rapamycin
NGS	Next-Generation Sequencing
NTR	Novel Transcribed Region

OS	Overall Survival
PAM	Prediction Analysis of Microarrays
PARP	Poly-ADP(ribose) polymerase
PCA	Principal Components Analysis
pCR	Pathological Complete Response
PCR	Polymerase Chain Reaction
PFS	Progression Free Survival
qPCR	Quantitative Polymerase Chain Reaction
RD	Residual Disease
RNA	Ribonucleic Acid
RNA-Seq	Ribonucleic Acid-Sequencing
RPKM	Reads per Kilobase of Exon per Million Mapped Reads
rRNA	Ribosomal Ribonucleic Acid
RT	Reverse Transcriptase
SASR	Suffix Array Single Read
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SSB	Single Stranded Break
TCGA	The Cancer Genome Atlas
TDLU	Terminal Duct Lobular Unit
TNBC	Triple-Negative Breast Cancer
TNM	Tumor Node Metastasis
UCR	Ultra Conserved Region
UCSC	University of California-Santa Cruz
WFA	Work Flow Analysis

Chapter 1: Introduction

1.1 Triple-Negative Breast Cancer

1.1.1 Clinical characteristics and treatment of triple-negative breast cancer

Triple-negative breast cancer (TNBC) preferentially affects pre-menopausal women and women of African-American descent and has been plagued by the absence of targeted therapies leading to poor survival (Figure 1) [1-3]. TNBC accounts for approximately 15% of cases of breast cancer in the United States [1]. Despite its lower frequency, it contributes to both breast cancer morbidity and mortality in a disproportionately high fashion compared to estrogen/progesterone receptor (ER/PR) or HER-2 positive breast cancers [1]. Because these tumors do not over-express the estrogen, progesterone, or HER-2 receptors (triple-negative), these patients do not respond to targeted therapies that are successfully used in patients who over express these proteins. Agents such as tamoxifen, the aromatase inhibitors, and Herceptin have made major advances in improving survival for hormone-positive and HER-2 positive breast cancers, but successful targeted therapies have been absent in TNBC. Current standard of care for TNBC utilizes standard chemotherapeutic agents including: anthracyclines; cyclophosphamide; and taxanes in the adjuvant and neoadjuvant settings, and platinums and nucleoside analogs for patients with metastatic disease [4]. When using neoadjuvant chemotherapy in the curative setting, patients with TNBC are more likely to experience a pathological complete response (pCR) compared to those with ER positive disease; however, those patients who do not experience a pCR are far more likely to have a poor prognosis compared with other subtypes (Figure 2) [5, 6]. This variability in response implies that TNBCs are a heterogeneous group of diseases as

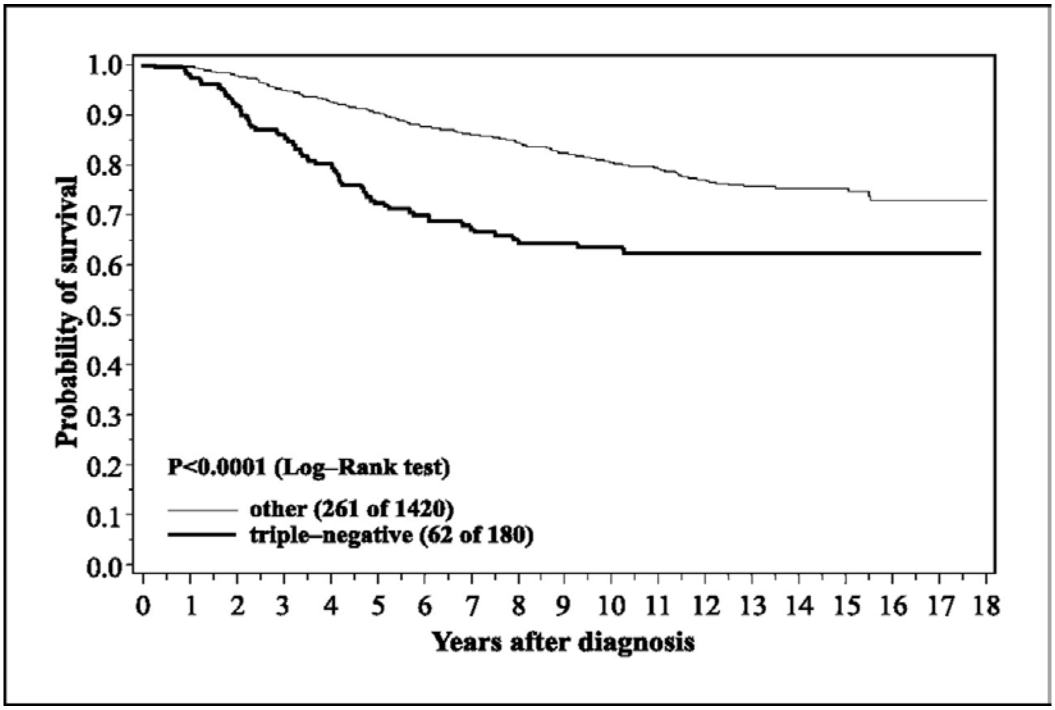


Figure 1. Kaplan-Meier curve of breast-specific survival in triple-negative and other breast cancers. This survival curve demonstrates the poorer prognosis of TNBC patients compared to other breast cancer subtypes. Adapted from Dent et al. [2].

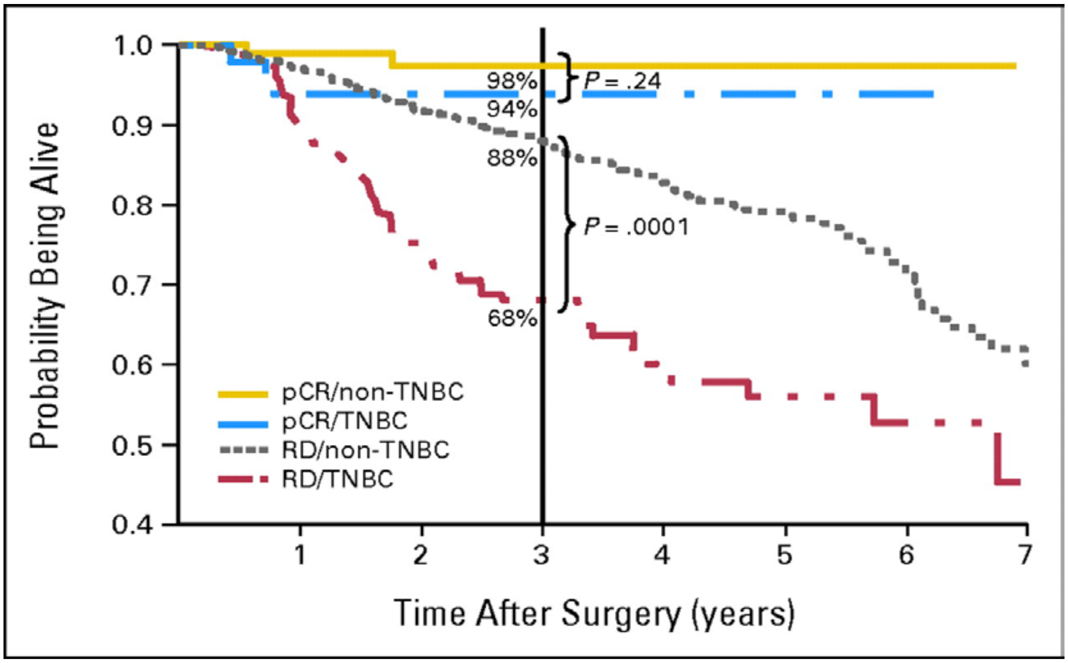


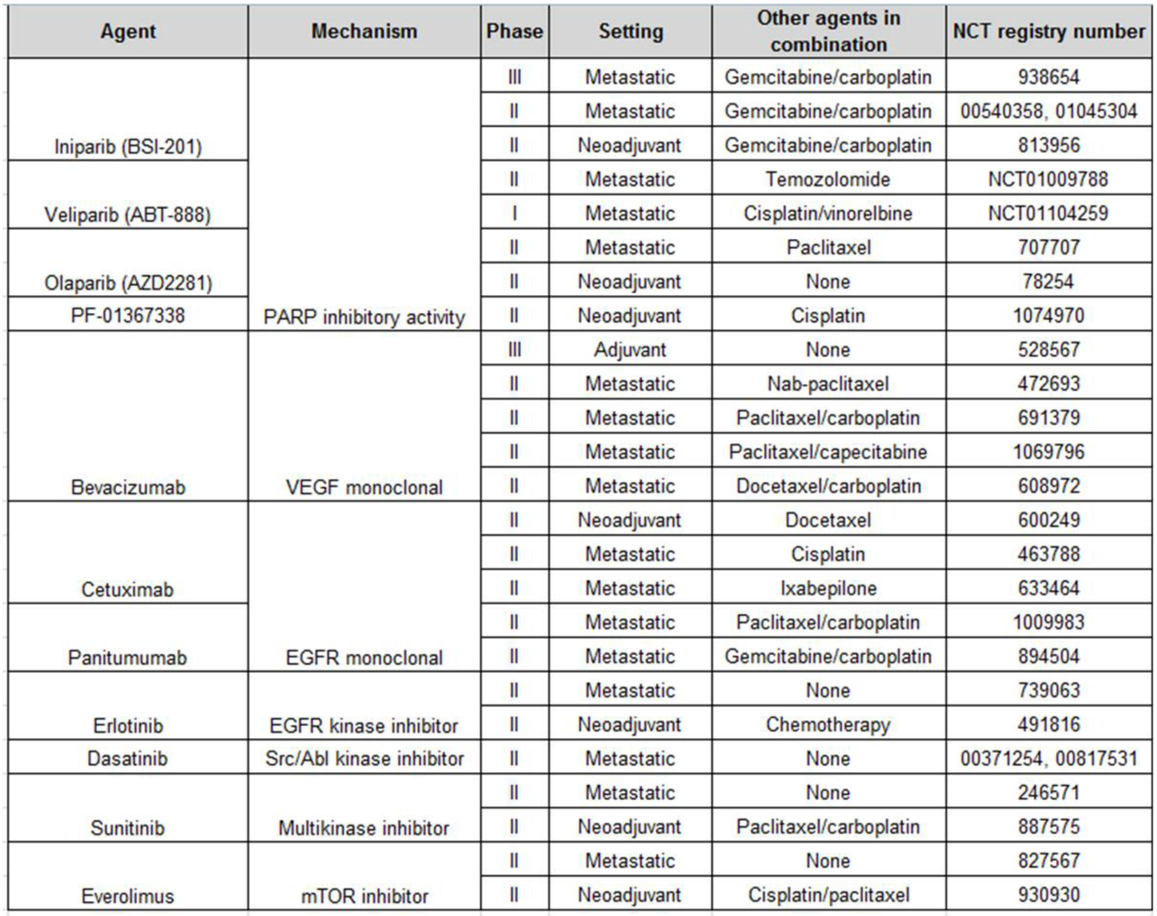
Figure 2. Overall survival as a function of response to neoadjuvant chemotherapy in TNBC. This figure demonstrates that TNBC patients who experience a pCR to neoadjuvant chemotherapy have a favorable prognosis that is nearly equivalent to other breast cancer patients who also experience a pCR. In contrast, TNBC patients who do not experience a pCR do considerably worse when compared to other breast cancer patients who do not experience a pCR. RD = Residual Disease. Adapted from Liedtke et al. [6].

opposed to a single subtype and that novel targeted agents are needed to improve outcomes.

The poor prognosis of TNBC patients coupled with the aforementioned paucity of effective therapies has led to substantial efforts to identify targeted agents for this disease (see Table 1 for current trials of targeted agents in TNBC) [7]. Some of the initial targeted agents tested in TNBC have focused on inhibiting the Epidermal Growth Factor Receptor (EGFR) and the c-KIT receptor which were both identified to be overexpressed by immunohistochemistry compared to other breast cancers [8]. In the case of the EGFR, both a monoclonal antibody (Cetuximab) and a small molecule (Gefitinib) along with standard chemotherapy have failed to produce statistically significant increases in progression free survival (PFS) or overall survival (OS) in several clinical trials [9-12]. Similarly, clinical trials testing the small molecules Imatinib and Dasatanib (both potent inhibitors of c-KIT) in patients with TNBC have also been negative [13, 14].

Anti-angiogenesis agents which target the growth and function of tumor vasculature, have also been implicated for use in treating TNBC. Results from the E2100 trial which tested paclitaxel vs. paclitaxel with the addition of the anti-Vascular Endothelial Growth Factor (VEGF) monoclonal antibody bevacizumab, reported a near doubling in PFS for all patients including triple-negatives [15]. Similarly, a Phase II clinical trial of the receptor tyrosine kinase inhibitor Sunitinib, which targets the VEGF receptor-2 (VEGFR-2/KDR), reported a slightly higher response rate in triple negative patients [16]. While modest clinical benefit has been reported with the use of anti-angiogenic agents in TNBC, an increase in overall survival has yet to be demonstrated.

More recently, the major focus of targeted therapy for TNBC has revolved around inhibiting DNA repair proteins, specifically the poly-ADP(ribose) polymerase (PARP). The premise of these trials revolve around the observations that sporadic and hereditary TNBC tumors are deficient in their DNA repair capacity, especially those mutated for the



Agent	Mechanism	Phase	Setting	Other agents in combination	NCT registry number
Iniparib (BSI-201)	PARP inhibitory activity	III	Metastatic	Gemcitabine/carboplatin	938654
		II	Metastatic	Gemcitabine/carboplatin	00540358, 01045304
		II	Neoadjuvant	Gemcitabine/carboplatin	813956
Veliparib (ABT-888)		II	Metastatic	Temozolomide	NCT01009788
		I	Metastatic	Cisplatin/vinorelbine	NCT01104259
Olaparib (AZD2281)		II	Metastatic	Paclitaxel	707707
		II	Neoadjuvant	None	78254
PF-01367338		II	Neoadjuvant	Cisplatin	1074970
Bevacizumab	VEGF monoclonal	III	Adjuvant	None	528567
		II	Metastatic	Nab-paclitaxel	472693
		II	Metastatic	Paclitaxel/carboplatin	691379
		II	Metastatic	Paclitaxel/capecitabine	1069796
		II	Metastatic	Docetaxel/carboplatin	608972
Cetuximab	EGFR monoclonal	II	Neoadjuvant	Docetaxel	600249
		II	Metastatic	Cisplatin	463788
		II	Metastatic	Ixabepilone	633464
Panitumumab		II	Metastatic	Paclitaxel/carboplatin	1009983
		II	Metastatic	Gemcitabine/carboplatin	894504
Erlotinib	EGFR kinase inhibitor	II	Metastatic	None	739063
		II	Neoadjuvant	Chemotherapy	491816
Dasatinib	Src/Abl kinase inhibitor	II	Metastatic	None	00371254, 00817531
Sunitinib	Multikinase inhibitor	II	Metastatic	None	246571
		II	Neoadjuvant	Paclitaxel/carboplatin	887575
Everolimus	mTOR inhibitor	II	Metastatic	None	827567
		II	Neoadjuvant	Cisplatin/paclitaxel	930930

Table 1. Ongoing clinical trials of targeted agents in TNBC. Current clinical trials of targeted agents are focusing particularly on PARP inhibitors, anti-angiogenics, kinase inhibitors, and mTOR inhibition. Adapted from Pal et al. [7].

DNA repair protein, BRCA1 (the role of BRCA1 mutations in TNBC is explained more in Section 1.1.3) [17, 18]. The majority of BRCA1 mutated tumors are triple-negative [19, 20], and because of these mutations these tumor cells cannot repair DNA via the homologous recombination pathway [21]. Previous preclinical data demonstrated that the inhibition of PARP, which is involved in DNA repair via the base excision repair pathway, can induce apoptosis in BRCA1 mutated cells [22]. This cytotoxicity is induced by “synthetic lethality” in which a single mutation in a gene or pathway does not induce cell death on its own, but upon introduction of a second mutation or inhibition of a complementary gene or pathway leads to cell death (Figure 3). To this end, trials of PARP inhibitors in TNBC are ongoing. Some of the first efficacy data was derived from a single-agent Phase II trial using the drug Olaparib in BRCA1/2 mutation carriers [23]. The trial showed a dramatic response rate of 41%, suggesting that the synthetic lethal approach is efficacious. A previous Phase I trial of Olaparib demonstrated that the benefit of the drug was restricted primarily to BRCA carriers [24]. A randomized Phase II trial of a second PARP inhibitor, Iniparib, along with Gemcitabine and Carboplatin in metastatic TNBC reported a significant increase in response rate (32% to 52%) and an increase in median overall survival (7.7 months to 12.3 months) [25]. This trial in particular has generated substantial excitement as the patient population was not selected for BRCA1/2 mutants and most likely contains a large proportion of women with sporadic TNBC. These data suggest that benefit from PARP inhibition may not be restricted to tumors defective in homologous recombination, but that PARP inhibition may act in concert with other defective DNA repair pathways in TNBC.

While significant advances in treating TNBC with chemotherapy and targeted agents has been made in the last decade, TNBC still remains very lethal with no FDA approved targeted agents to date. Further, the determinants of why TNBC preferentially affects young women and those of African-American descent remain unknown.

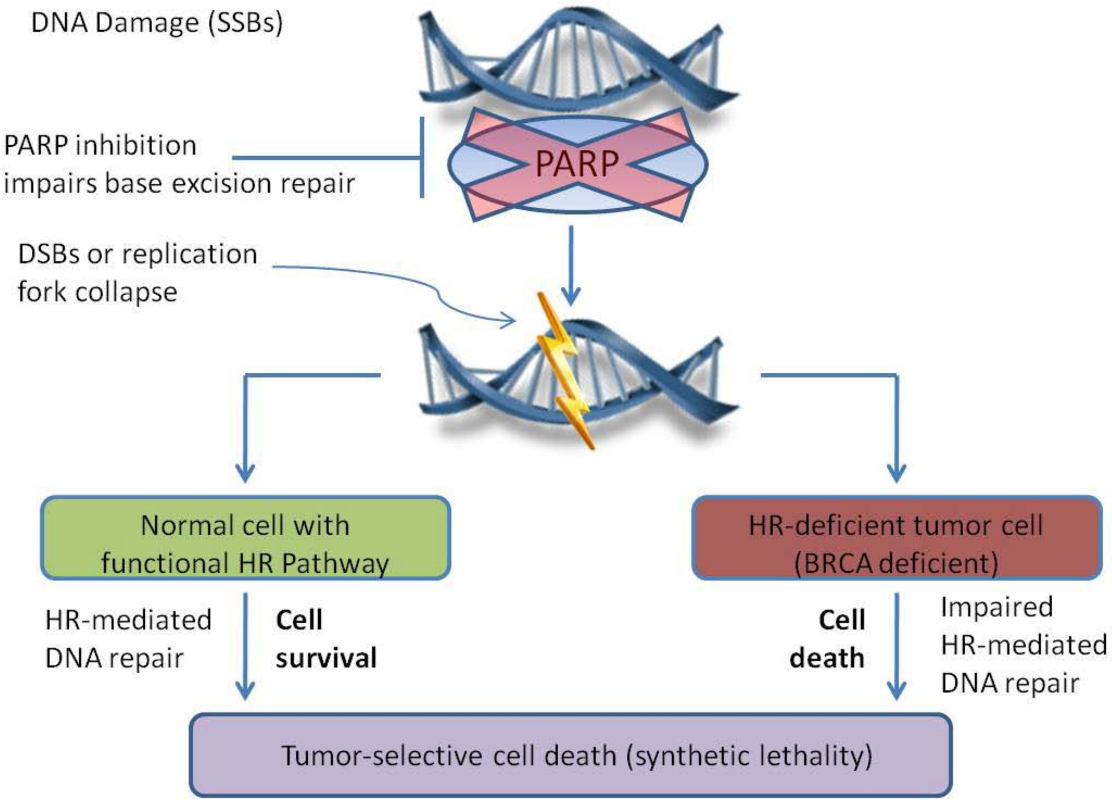


Figure 3. Mechanism of synthetic lethality between BRCA deficiency and PARP inhibition. Single stranded breaks (SSBs) that occur during regular DNA damage is repaired by base excision repair (BER). PARP is a key player in BER whose inhibition leads to unrepaired SSBs. These SSBs during DNA replication can cause collapse of the replication fork or turn in to double stranded breaks (DSBs). DSBs are normally repaired by homologous recombination (HR) which requires BRCA1 and BRCA2. Without functional BRCA1/2, the DNA is not repaired or repaired via alternative error-prone methods leading to gross genomic instability and eventually cell death. Redrawn and adapted from Banerjee et al. [26].

1.1.2 Histological and molecular characterization of TNBC

Histologically, TNBC is typically characterized as infiltrating ductal carcinoma of no special type (IDC-NST) with high grade, pushing borders, high mitotic index, lymphocytic infiltrate, central necrotic zones, and occasional medullary features [27, 28]. While other less common histological types of breast cancer including: medullary, metaplastic, secretory, and adenoid cystic carcinomas are also triple-negative, the vast majority of TNBC are IDC-NST [27]. This common histological type of TNBC will be the focus of this dissertation.

In 2000, Perou et al. had demonstrated that breast cancers can be characterized by their molecular profile (known as the intrinsic subtypes) using gene expression microarrays (Figure 4) [29]. This work stratified breast tumors into Luminal A and B (mostly ER-positive), HER-2, normal-like, and basal-like (mostly TNBC). Further work also demonstrated that the majority of tumors mutated for BRCA1 are also basal-like [19]. More recently, a subset of the basal-like tumors referred to as “claudin-low” (or Basal B) has been identified by microarray analysis, and these tumors are thought to be enriched with characteristics of stem cells and epithelial-to-mesenchymal transition [30, 31]. These tumors were coined “basal-like” to make reference of their origin from the basal/myoepithelial layer of the milk duct in the normal breast [29]. While the majority of sporadic TNBC and BRCA1 mutated tumors are basal-like, there is not a complete overlap and the terms should not be used synonymously (Figure 5). Indeed, 10-35% of TNBCs are not basal-like, where up to 45% of basal-like cancers are not triple-negative [32].

To further complicate the picture, recent controversy has called into question whether basal-like tumors really derive from the basal/myoepithelial layer and should the term “basal” even be used [33]. In the original Perou et al. paper, subsets of tumors were

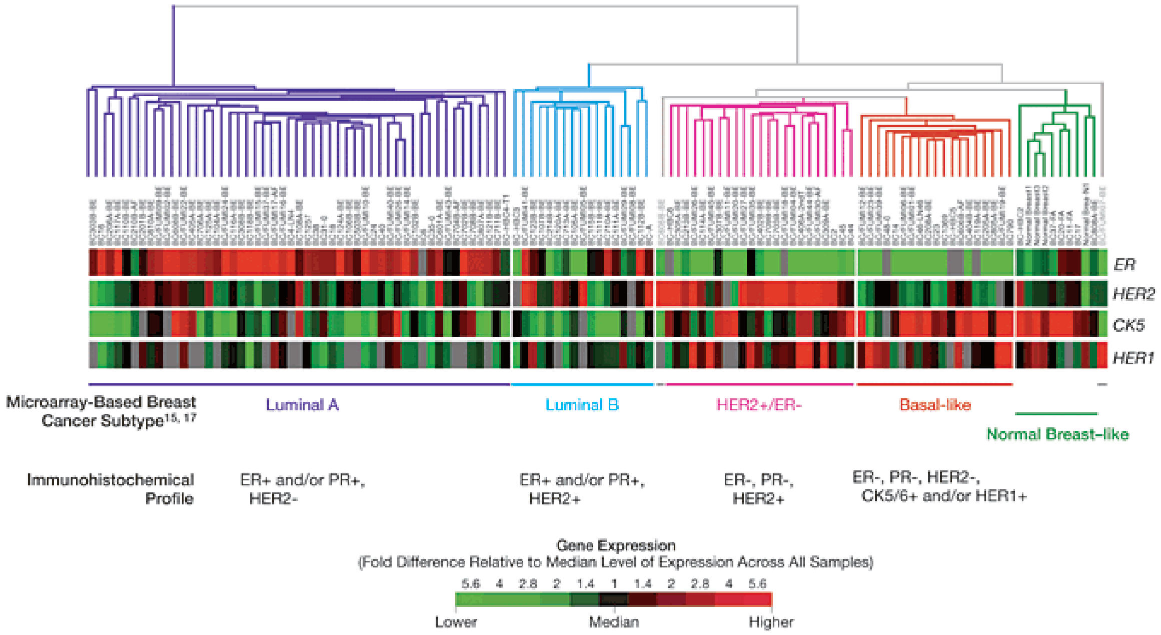


Figure 4. Representative example of breast tumors classified into intrinsic subtypes using unsupervised hierarchical clustering of gene expression array data. This clustering separates breast tumors into five subtypes which include Luminal A and B (predominately ER+), HER2, Basal-like (predominately TNBC), and Normal Breast-like. While there is some congruence of immunohistochemical profile with intrinsic subtype, a significant amount of non-overlap exists. For example, a significant proportion of basal-like tumors are not triple-negative and vice versa. Figure adapted from Carey et al. [3].

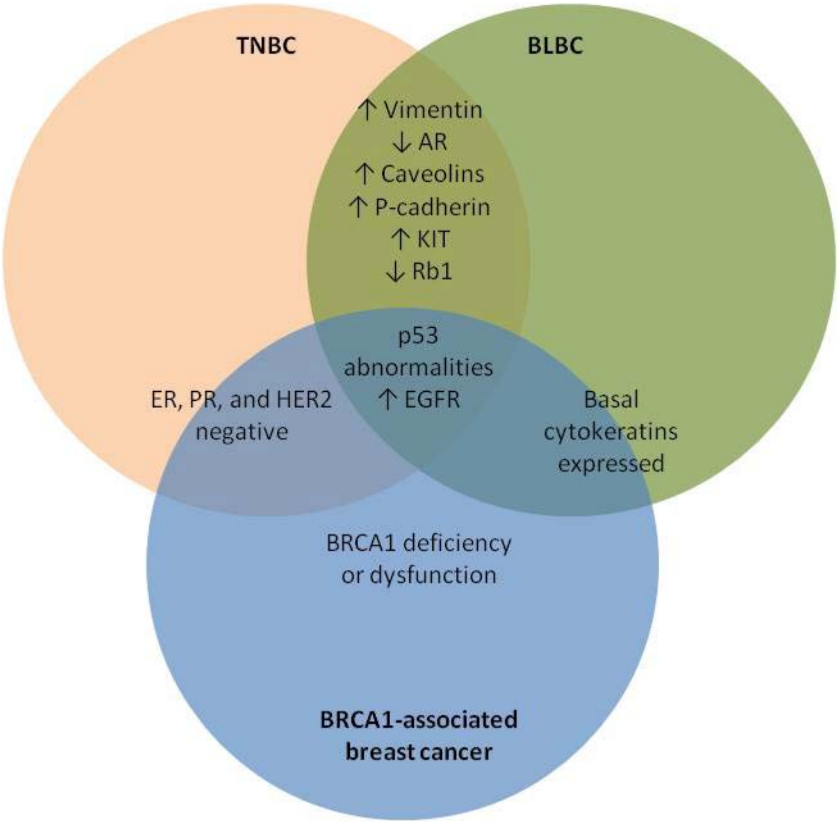

TNBC
BLBC
↑ Vimentin
↓ AR
↑ Caveolins
↑ P-cadherin
↑ KIT
↓ Rb1
p53 abnormalities
↑ EGFR
ER, PR, and HER2 negative
Basal cytokeratins expressed
BRCA1 deficiency or dysfunction
BRCA1-associated breast cancer

Figure 5. Shared characteristics of TNBC, Basal-like breast cancer (BLBC), and BRCA1-associated breast cancer. TNBC (by definition), BLBC, and BRCA1 cancers are predominately triple-negative. While they each have unique characteristics, they share substantial overlap in their gene expression, immunohistochemical profile, and DNA repair capacity. Redrawn and adapted from Carey et al. [32].

designated as either luminal or basal-like based on expression of Cytokeratins 8/18 for luminal cells and Cytokeratins 5/6 for basal cells [29]. Data from others have supported the fact that Cytokeratins 5/6 are also expressed in luminal cells [34]. Further, in a partial retraction, Perou and colleagues concluded in a paper in 2006, that the basal-like subtype most likely does not originate from the basal/myoepithelial after failing to detect consistent expression of the myoepithelial markers: CD10; alpha-smooth muscle actin; and p63; but in confirmation did detect positive expression for Cytokeratins 5/6 & 8/18 in tumors classified as basal-like [35]. The definitive work that basal-like/TNBC tumors most likely derive from a luminal progenitor versus a basal/myoepithelial progenitor were confirmed in two elegant studies which both used BRCA1 mutants as a means to determine the cell of origin. In the first study, Lim et al. isolated tissue from BRCA1 mutation carriers, and observed a factor-independent expansion of luminal progenitor cells followed by confirmation that breast tissue from BRCA mutants and basal-like tumors had gene expression profiles that were most similar to normal luminal progenitors [36]. In a second study, Molyneux et al. directly demonstrated that knocking out BRCA1 in luminal progenitor cells in mice produced tumors that histologically resemble BRCA1 and sporadic breast cancers whereas knockout of BRCA1 in basal/myoepithelial progenitor cells produced tumors that resemble rare adenomyoepitheliomas [37].

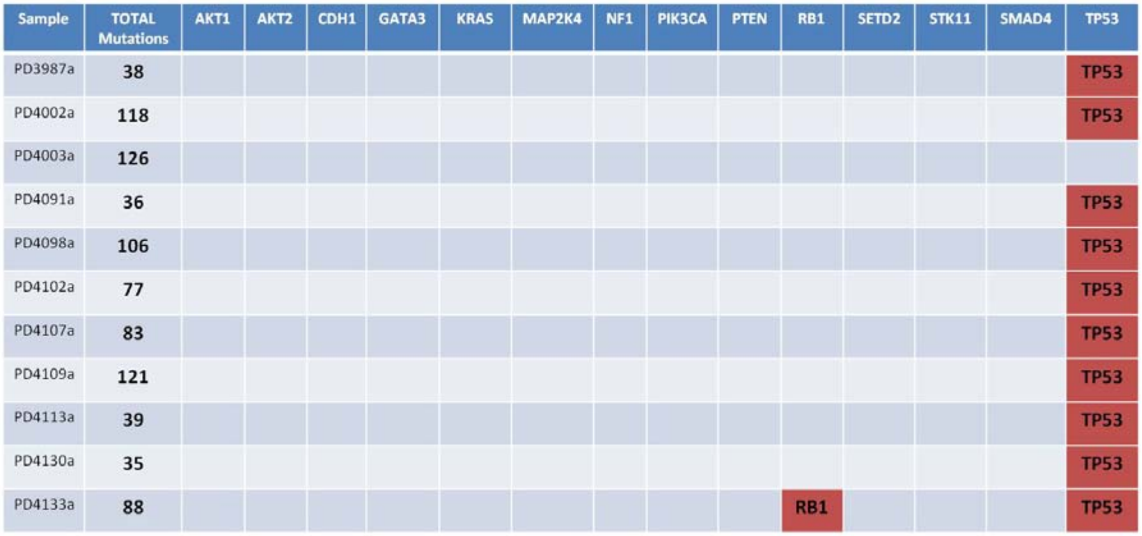
In summation of histological & molecular profiling of TNBC, standard histochemical methods of staining for ER, PR, and HER-2, and determining grade and mitotic index are still the mainstay in guiding treatment. Recent work using a 50-gene classifier (known as the PAM50) qPCR assay to determine intrinsic subtypes confirmed that basal-like tumors are at a higher risk of relapse [38], though the use of intrinsic subtypes even for prognosis has failed to translate into clinical utility. Whether better markers or therapeutic targets can be identified through gene expression profiling or by having *a priori* knowledge of the cell of origin of TNBC is still to be determined.

1.1.3 Hereditary and somatic mutation profile of TNBC

A significant amount of research has been published in relation to hereditary mutations and TNBC. In 1990, efforts let by Mary-Claire King using linkage analysis identified BRCA1 as a gene involved in hereditary breast cancer [39]. This gene was subsequently cloned by researchers at the University of Utah in 1994 and patented by Myriad Genetics [40]. BRCA1 has since become an archetypal example of inherited predisposition to cancer. BRCA1 is a tumor suppressor gene involved in DNA repair by homologous recombination, and loss-of-function mutations in BRCA1 leads to genomic instability [21]. The majority of BRCA1 mutated tumors are triple-negative and/or basal-like [19, 20], but BRCA1 mutants account for only a small percentage of TNBC cases [41]. Carriers of BRCA1 mutations are at a 60-85% lifetime risk for developing breast cancer and at a 15-40% risk of ovarian cancer. Because BRCA1 mutant TNBCs cluster by gene expression array with sporadic TNBCs, it has been suggested that sporadic TNBCs may have hallmarks of “BRCAness” or gene expression/mutations in BRCA related genes or its pathway that results in a similar tumor phenotype [17, 19]. Indeed, previous data has demonstrated that sporadic TNBCs that have wild-type BRCA1 do exhibit decreased BRCA1 expression which can occur through promoter hypermethylation or overexpression of negative regulators of BRCA1 such as the ID4 gene [42, 43]. More recent data in ovarian cancer has delineated a 60 gene set “BRCAness profile” that predicts responsiveness to platinum and PARP inhibitors for patients with BRCA-like sporadic ovarian cancer [44]. This data suggests that patients whose tumors are BRCA-like via a BRCAness gene expression profile may benefit from agents that damage DNA and inhibit DNA repair proteins, similar to the sensitivity seen in BRCA mutation carriers. While BRCA1 has dominated the landscape of germline susceptibility for TNBC, whether other germline variants exist (either in coding or non-coding regions) that predispose patients to TNBC is still to be determined.

In regards to somatic mutations in TNBC, the numbers of genes that are recurrently mutated are quite small. Data from the Vogelstein group demonstrated that the tumor suppressor p53 is by far the most recurrently mutated gene in TNBC [45]. Other data has also suggested the presence of Rb gene mutations in TNBC [46]. Very recent data presented by the Wellcome Trust Sanger Institute at the 2010 San Antonio Breast Cancer Symposium using next-generation exome sequencing in 11 TNBC cases, confirmed that p53 mutations dominate the mutational landscape of TNBC with Rb mutations coming in at a distant second (Figure 6) [47]. Also, very recent data presented at the 2011 American Association for Cancer Research Annual Meeting by researchers at the Translational Genomics Institute (TGen) using next-generation sequencing has identified recurrent somatic mutations in the ERBB4 gene in four African-American women with treatment resistant TNBC [48].

To further add to the somatic mutation landscape of TNBC, researchers at the Wellcome Trust Sanger Institute implemented low coverage next-generation DNA sequencing to examine genomic rearrangements in breast cancer [49]. This sequencing effort revealed that TNBCs have chaotic genomes compared to ER-positive or HER2-positive breast cancers as evidenced by significantly more intra- and inter-chromosomal fusions and tandem duplications (Figure 7). Of important note, the Sanger group did not identify any recurrent gene fusions in its dataset, suggesting that TNBC is not driven by a single gene fusion as its primary driver. These data combined reflect how little is known about the mutational landscape of TNBC, and suggest that there is a possibility of other germline or somatic mutations implicated in TNBC tumorigenesis that have yet to be discovered.



Sample	TOTAL Mutations	AKT1	AKT2	CDH1	GATA3	KRAS	MAP2K4	NF1	PIK3CA	PTEN	RB1	SETD2	STK11	SMAD4	TP53
PD3987a	38														TP53
PD4002a	118														TP53
PD4003a	126														
PD4091a	36														TP53
PD4098a	106														TP53
PD4102a	77														TP53
PD4107a	83														TP53
PD4109a	121														TP53
PD4113a	39														TP53
PD4130a	35														TP53
PD4133a	88										RB1				TP53

Figure 6. Mutated cancer genes in TNBC identified by the Wellcome Trust Sanger Institute. Eleven primary TNBCs were sequenced using next-generation DNA exome sequencing. Genes known to be mutated in cancer are listed across the top. TP53 was mutated in 10 of 11 cancers where RB1 was mutated in 1 of 11. This data confirms previous findings that TP53 is the most recurrently mutated gene in TNBC with RB1 coming in at a distant second. Figure redrawn and adapted from Futreal et al. presentation at the 2010 San Antonio Breast Cancer Symposium [47].

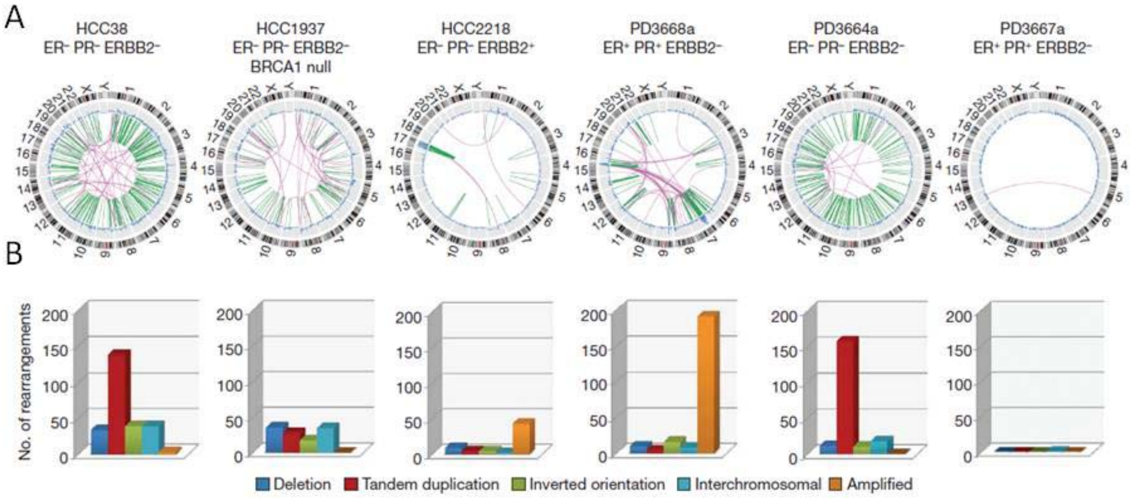


Figure 7. Low coverage next-generation whole genome sequencing of 24 breast cancers. **(A)** Circos plots of 6 representative samples. The outer ring is a circularized karyogram, the blue line represents copy number variation, green protrusions represent intrachromosomal fusions, and purple lines represent interchromosomal fusions. TNBC samples had considerably more inter- and intrachromosomal fusions compared to other subtypes. **(B)** Graphs of the number of genomic alterations present in each genome. TNBCs had more tandem duplications compared to other subtypes. Figure adapted from Stephens et al. [49].

1.2 The normal breast in cancer research

1.2.1 Overview of normal epithelium in the breast

The epithelial component of the human breast consists of a branching structure of ducts comprised of inner luminal epithelial cells which are surrounded by an outer layer of supporting myoepithelial cells (Figure 8) [50, 51]. The term “myoepithelial” is used as these cells have characteristics of both epithelium and of smooth muscle cells that can contract in order to move milk in response to oxytocin [33, 34, 51]. This layered structure of luminal and epithelial cells is maintained from the nipple through the lactiferous ducts to the lobules [34]. The ducts terminate at the terminal duct lobular unit (TDLU) which is the milk producing structure of the breast that consists of many acini (also known as alveoli) aggregated together to form lobules that are visually similar to a bunch of grapes (Figure 9) [52]. The TDLUs are dynamic structures that begin development during puberty [51, 53]; increases in number and size during the luteal phase of the menstrual cycle and pregnancy [53, 54]; and atrophies after menopause [53]. Of importance, it is in the TDLUs that the majority of breast cancers are formed [55]. Within the myoepithelial and luminal layers of the ducts and TDLUs is where breast stem cells reside (Figure 8) [50]. These stem cells are responsible for the growth of the ducts/TDLUs during the menstrual cycle and during pregnancy in preparation for lactation [53]. Many of these stem cells are highly pluripotent, and elegant experiments in mice have demonstrated that an entire functional mammary gland can be produced from a single mammary stem cell [56]. As mentioned in Section 1.1.2, in it has been experimentally determined that the source of BRCA1 induced TNBC is the luminal progenitor of the normal breast [36, 37]. Further work by others are elucidating the stem cell hierarchy of the normal breast and correlating this hierarchy to the origin of the various subtypes of breast cancer (Figure 10) [36]. While significant research is ongoing

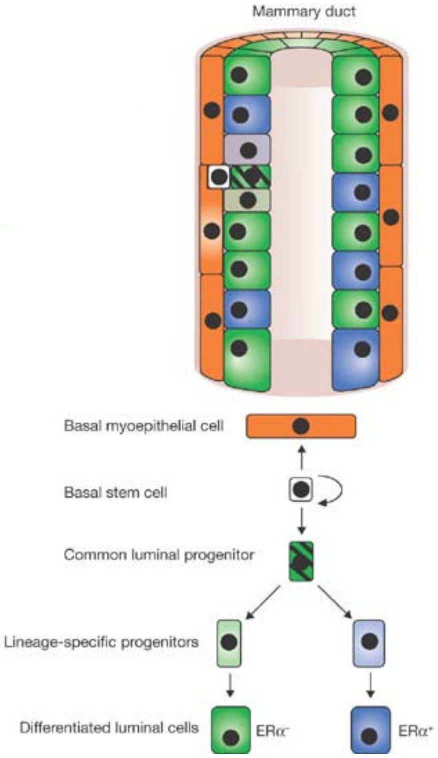

Mammary duct
Basal myoepithelial cell
Basal stem cell
Common luminal progenitor
Lineage-specific progenitors
Differentiated luminal cells
ERα⁻
ERα⁺

Figure 8. Schematic of the normal cell types that comprise the human mammary duct. The human mammary duct consists of two layers, an inner luminal layer primarily responsible for carrying and secreting milk, and an outer myoepithelial layer primarily responsible for contraction during lactation. Within the layers are embedded stem and progenitor cells responsible for cell renewal and the growth of the ducts/TDLUs during the menstrual cycle and pregnancy. Figure adapted from Smalley et al. [50].

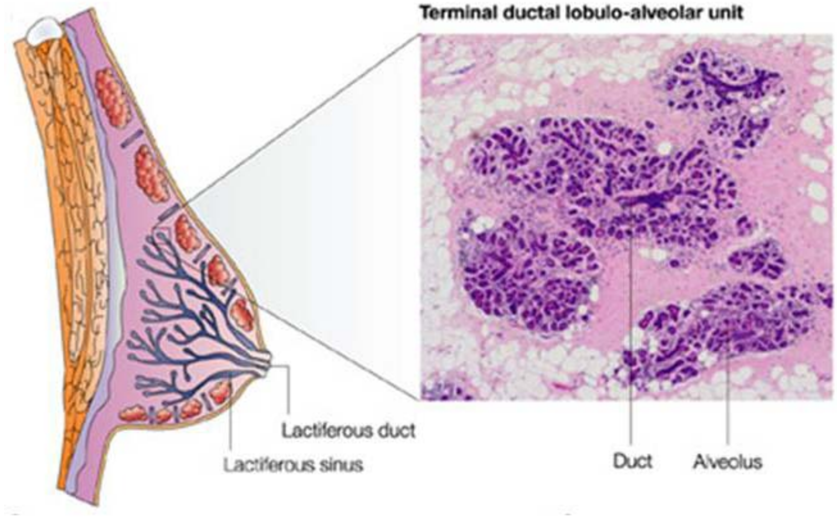


Figure 9. Illustration of the normal breast and histology of the TDLU. The normal human breast comprises of a network of lobules and ducts primarily tasked with lactation during pregnancy. Milk produced in the TDLUs travels through the lactiferous ducts and is aspirated through the nipple. The TDLUs, along with their milk secretion function, are also thought to be the site of origin of breast cancers. Figure adapted from Smalley et al. [52].

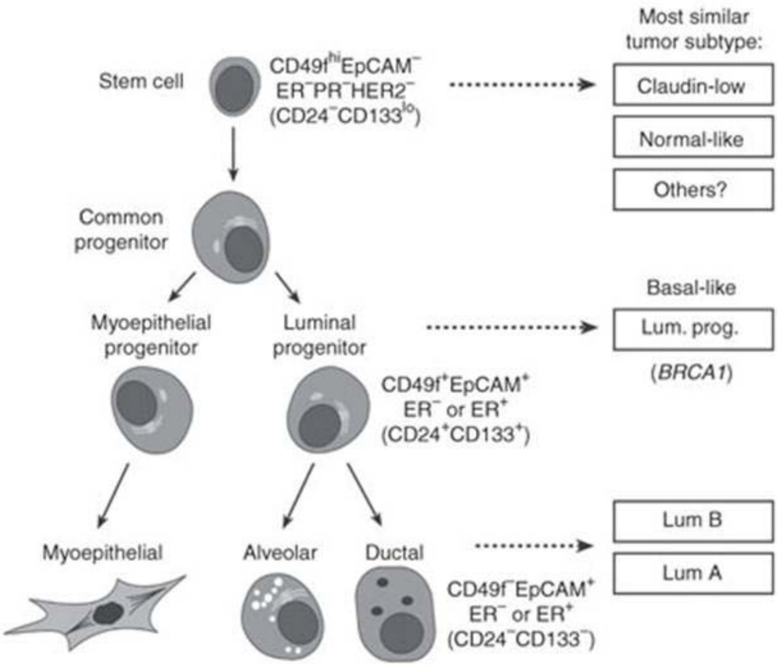

Most similar tumor subtype:
Stem cell
CD49f^hi EpCAM^− ER^−PR^−HER2^− (CD24^−CD133^lo)
Claudin-low
Normal-like
Others?
Common progenitor
Basal-like
Myoepithelial progenitor
Luminal progenitor
Lum. prog.
(BRCA1)
CD49f^+EpCAM^+ ER^− or ER^+ (CD24^+CD133^+)
Myoepithelial
Alveolar
Ductal
Lum B
Lum A
CD49f^−EpCAM^+ ER^− or ER^+ (CD24^−CD133^−)

Figure 10. Schematic of the normal breast stem cell hierarchy. This diagram demonstrates the adult stem cell hierarchy of the normal breast with associated cell surface markers, and its similarity to the intrinsic subtypes of breast cancer based on gene expression profiling. This schematic points out that basal-like (TNBC) is derived from the normal luminal progenitor, while the more stem-like “claudin-low” subtype derives from a more pluripotent cell and the luminal subtype (predominately ER-positive) arise from more differentiated cells. Figure adapted from Lim et al. [36].

into understanding the role of these stem cells in normal breast biology on breast cancer causation, it is accepted that the limited understanding of normal breast biology is a barrier to progress in breast cancer research.

1.2.2 The use of normal breast tissue in cancer research

A major impediment to therapeutic development in TNBC, and breast cancer in general, is an inadequate understanding of the transcriptional biology of the normal breast as a comparator. The use of microdissected ductal epithelium from healthy women as the optimal control is not commonly used secondary to sample availability from healthy volunteers and laborious sample preparation. Many prior gene expression studies have used undissected reduction mammoplasty or histologically “non-cancerous” tissue adjacent to the tumor. Both of these controls are fraught with problems. Specifically hyperplastic breasts that require surgical reduction are clearly not “normal” and may harbour neoplasms or pathological atypia [57-60]. Likewise, normal tissue adjacent to tumor may be significantly impacted by factors released near the tumor microenvironment. This includes pertubations in global gene expression [61, 62], changes in epigenetic markers [63], and loss of heterozygosity [64, 65]. Thus, the use of normal breast tissues from healthy volunteers is an optimal control for the biological study of breast cancer.

To address this, the research presented in this dissertation incorporated the use of fresh frozen biopsies of normal breast tissues from the Susan G. Komen for the Cure Tissue Bank at the Indiana University Simon Cancer Center. This unique collection of over a 1000 breast tissues with complementary germline DNA, serum, and clinical data provides a powerful resource for understanding normal breast biology but also to have an optimal comparator to better understand breast cancer. By using this type of resource, it empowers the researcher to understand the key differences that differentiate normal from cancer versus using sub-optimal controls which include reduction mammoplasties, adjacent normal, ductal carcinoma *in situ* cell lines, or even other breast cancer subtypes. In conclusion, as will be demonstrated in Chapter 2, the answers that are derived by comparing TNBC to normal versus other cancer subtypes can have significant biological and clinical ramifications.

1.3 Next-generation whole transcriptome sequencing (RNA-Seq)

1.3.1 Overview of next-generation sequencing and its role in cancer research

Next-generation sequencing (NGS) technology has revolutionized the study of genomics. This technology, also known as second-generation sequencing or massively parallel sequencing, allows for the interrogation of entire genomes, transcriptomes, and methylomes in an expedient and cost-effective manner. NGS is distinguished from its “first-generation” predecessor, Sanger/Capillary sequencing, by its shear sequencing output (Figure 11) [66]. In perspective, the first human genome reported in 2001 took ~10 years and cost nearly $1 billion. Sequencing of human genomes is now approaching $15,000 per genome and ~2 weeks of sequencing time. For the first time, the ability to interrogate and compare multiple genomes, transcriptomes, and methylomes across individuals and across cancers is now possible.

This explosion in sequencing capability has been fueled by advances in technology platforms. The current field of next-generation sequencing is dominated by three major systems: the Roche/454 Genome Sequencer FLX, the Illumina HiSeq 2000, and the Applied Biosystems (ABI) SOLiD 5500 XL (Figure 12). The Roche system is distinguished from the Illumina and ABI systems by its ability to sequence long reads (400-1000bp), but its low throughput (~1 GB/per run) suits this system for de novo sequencing applications where there is no reference genome or for lower order organisms with very small genomes. Unlike the Roche system, the Illumina and ABI systems are short-read technologies (50bp-150bp), but have considerable amount of throughput (~200-300GB/run). These platforms are best applied to projects where a reference genome is already available, in particular human sequencing. Because of the high throughput and high accuracy of the Illumina and ABI systems, these platforms

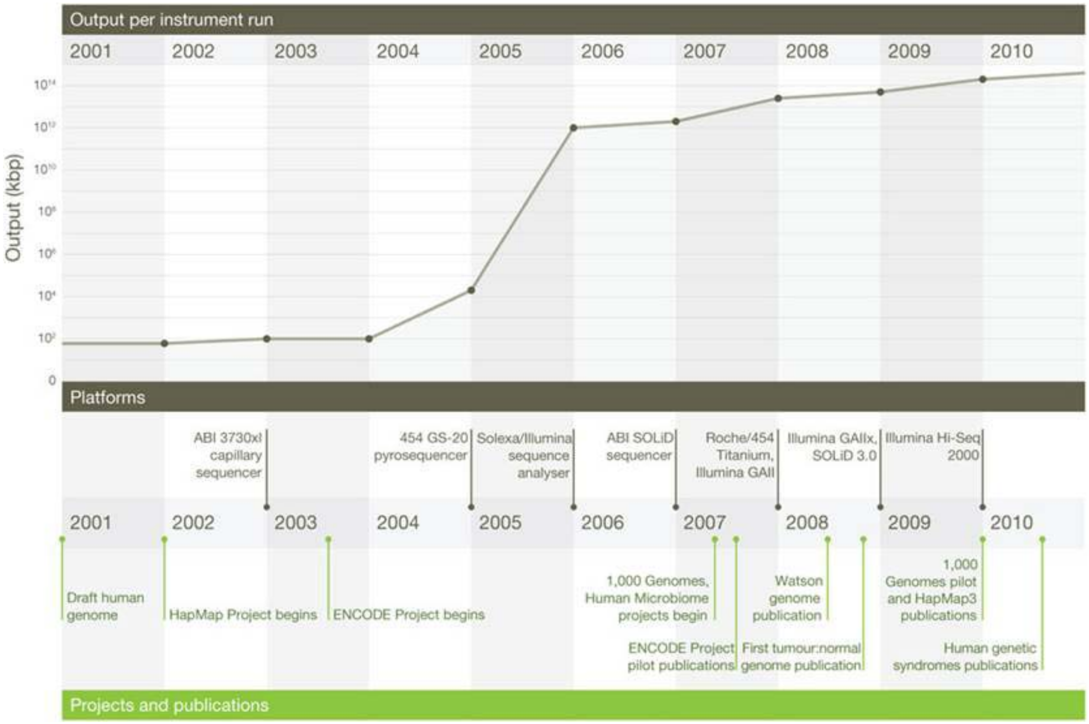


Figure 11. Increase in sequencing output due to technological advances in the last decade. Top: graph of sequencing output (logarithmic scale) due to technological advances over the last decade. The inflection of the curve in early 2005 coincides with the release of the first next-generation sequencer (Roche/454 GS-20). Since then, the increase in sequencing output has outpaced Moore's Law. Bottom: Milestones of technology releases and major publications in sequencing history. Figure adapted from Mardis et al. [66].

Figure 12. Current next-generation sequencing platforms. The Roche/454 GS FLX, Illumina HiSeq 2000, and ABI SOLiD 5500XL next-generation sequencing instruments. These platforms represent the most current versions of the top three next-generation sequencers.

have become the workhorses of the 1000 Genomes Project, the NIH Cancer Genome Atlas, and the International Cancer Genome Consortium.

One of the major applications of NGS is the detailed study of malignancy. NGS offers the capability to interrogate the DNA of entire cancer genomes in order to identify point mutations, insertion/deletions, fusion genes, amplifications, and duplications [67]. This technology has so far brought novel insights into understanding the mutational profiles of AML, melanoma, multiple myeloma, breast, and lung cancers [67]. Recently, in an innovative approach, DNA NGS was used as a means to monitor response to therapy for a cohort of patients with colorectal and breast cancer. This was achieved by detecting the levels of chromosomal translocations in the plasma of these patients, likely coming from cells of the primary tumor [68]. Likewise, NGS also offers the ability to understand the cancer transcriptome at an unprecedented scale and depth. The advent of next-generation whole transcriptome sequencing (RNA-seq) has revolutionized the way genes are studied in malignancy [69-72]. In contrast to microarrays, which use *a priori* gene selection, this technology can profile every mRNA regardless if it is known or unknown, coding or non-coding, and adenylated or unadenylated, in an expedient manner. It can obtain accurate expression of genes (including isoforms) while also obtaining sequence data for mutation and fusion detection. NGS also does not suffer from signal saturation, hybridization artifacts, and poor resolution of quantitation as seen in microarrays [73]. So far, RNA-seq has revealed the presence of recurrent gene fusions in prostate cancer [69, 70], and in breast cancer. Specifically for breast cancer, a fusion involving S6 kinase and VMP1 (Vacuole Membrane Protein 1) which is correlated with upregulation of the oncogenic microRNA miR-21, has been detected in 30% of breast cancers [74]. RNA-seq has also revealed recurrent mutations in ovarian cancer [71, 72], and allelic imbalances due to copy number variation in oral squamous cell carcinomas [75]. As evidenced by several presentations at the recent American

Association for Cancer Research 2011 Annual Meeting, the field is still very much in its infancy. The many insights that will be revealed through NGS technologies, in particular comparisons of cancer vs. normal, are still on the horizon.

1.3.2 RNA-sequencing overview and chemistry of the Applied Biosystems SOLiD sequencer.

As mentioned in the last section, RNA sequencing has empowered cancer research to reveal new insights into the perturbations that define the cancer transcriptome. This includes measuring differential gene expression, alternative splicing and isoform expression, non-coding RNA expression and discovery, point mutations, small insertions and deletions, gene fusions, RNA editing, and pathogen (viral or bacterial) detection/insertional mutagenesis. Having the ability to measure all these endpoints when comparing cancer to normal makes RNA-seq a powerful tool to discover novel biology.

In the following description, a brief overview of RNA-seq chemistry will be given. Because the primary research of this dissertation focuses on RNA-seq technology using the SOLiD system, the following description will be focused only on SOLiD and not the other platforms. RNA-seq beings with a somewhat complicated sample and library preparation. Total RNA is extracted from the tissues of interests (i.e. TNBCs and normals), using standard extraction techniques. Because the majority of RNA in any given cell is ribosomal (rRNA), a method to remove rRNA is required. Two major methods exist: Poly-A selection and ribosomal depletion. Poly-A selection employs the use of beads with attached Poly-T oligos. In a simple process similar to immunoprecipitation, poly-adenylated mRNA is hybridized to the Poly-T beads in a tube. Meanwhile, the non-poly-adenylated RNA (in particular the rRNA) is not hybridized and remains in solution. The beads are rinsed and the poly-A-mRNA is then precipitated and used downstream. The second (newer) method of ribosomal depletion, removes only the rRNA while retaining the rest of the mRNA including non-poly-A mRNAs. This method employes the use of biotinylated Locked Nucleic Acid (LNA) probes that hybridize to 5S, 5.8S, 18S, and 28S rRNA species. After the LNA probes are hybridized to the rRNA, streptavidin coated beads are added causing the biotinylated-LNA probes to be attached to the beads. The beads are removed leaving ribosomally depleted mRNA. This method is considered superior, as non-Poly-A species of mRNA are retained, in particular many non-poly-A noncoding RNAs.

After ribosomal depletion, mRNA is fragmented and universal adaptors are ligated to the mRNA prior to reverse transcription (RT). Adaptors are ligated to the RNA prior to RT in order to retain the strandedness of the RNA in later data analyses. The adaptors are referred to as the P1 & P2 adaptors, where the P1 is ligated to the 5’ end of the RNA, and the P2 to the 3’ end (Figure 13A). The mRNA is then reverse transcribed, size selected, and PCR amplified using primers specific for the universal adaptors. The amplified cDNA is then hybridized to microbeads via the P1 universal adaptor to a complementary P1 sequence on the bead. A single cDNA molecule is attached to the bead, and the bead then undergoes an emulsion PCR (Figure 13B). In the emulsion PCR, the fragment is amplified in an oil-in-water microreactor to ~30,000 copies. The beads that had successful amplification are enriched via the P2 adaptor and the beads are then covalently attached to a glass slide (Figure 13C).

The slide is then placed on the SOLiD system inside of a flowcell where the slide acts as the surface for the subsequent sequencing chemistry. Sequencing reagents are literally flowed over the slide cycle-by-cycle. The SOLiD system uses a unique sequencing-by-ligation chemistry in which fluorescently labeled octomers are ligated to the template strand (for details see Figure 14). Once ligated, the fluorophore is excited and measured by a sensitive camera, measuring fluorescence from over 1 billion beads on the slide simultaneously. A unique feature of the octomers is that each of the four flourophores represents two bases, known as dual-base encoding or colorspace (Figure 14). Also, as part of the sequencing process, each base is interrogated twice. The colorspace feature of the SOLiD allows for more accurate sequencing and more

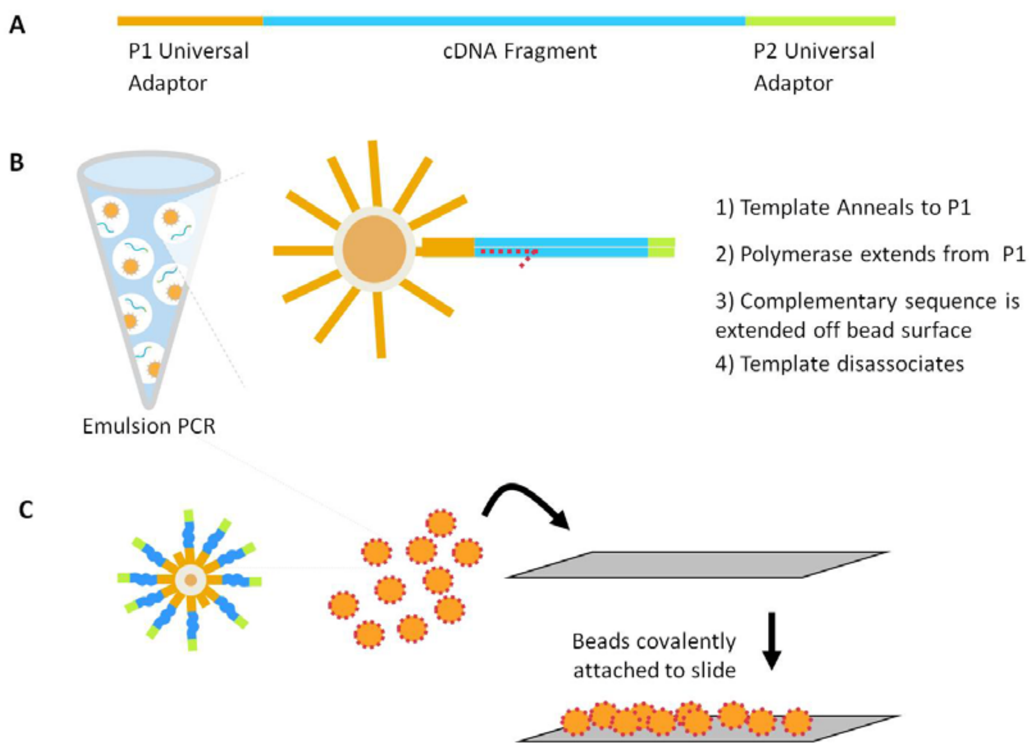

A
P1 Universal Adaptor
cDNA Fragment
P2 Universal Adaptor
B
1) Template Anneals to P1
2) Polymerase extends from P1
3) Complementary sequence is extended off bead surface
4) Template disassociates
Emulsion PCR
C
Beads covalently attached to slide

Figure 13. Schematic of RNA-Seq library preparation for ABI SOLiD sequencing. **(A)** RNA is ligated with two universal adaptors (P1 & P2) then reverse transcribed and amplified. **(B)** Illustration of Emulsion PCR with an inset of what occurs at the molecular level. A single fragment is hybridized to beads via the P1 adaptor. Emulsion PCR occurs in an oil-in-water reactor where up ~30,000 copies of the fragment is produced on the bead. **(C)** Beads with a multitude of copies of a fragment are then enriched and then covalently attached to a glass slide. Figure arranged by M. Radovich and graphics are courtesy of Applied Biosystems, Inc.

sensitive SNP/mutation detection. This process of ligation and excitation occurs for a total of 50 cycles producing sequence reads that are 50bp long. In the end, the user is left with large text files that are transferred from the instrument. These files contain the sequence data for any given sample and are subsequently bioinformatically analyzed.

1.3.3 Overview of SOLiD RNA-seq bioinformatics and data analysis

Analyzing data from short read technologies provides a unique challenge in analyzing vast amounts of data in a short amount of time. Because short read technologies rely on a reference genome, these short reads must be first mapped (also called alignment) to the reference genome. This presents a daunting task that requires over a billion reads that are 50 letters long to be matched to a genome that is 3 billion letters long. Also, the mapping algorithm has to be able to account for natural genetic variation, sequencing error, and in the case of RNA, reads that cross exon-exon junctions. Thankfully, intelligent software has been developed to help overcome this challenge. Because the field of mapping algorithm development is quite active, newer algorithms that map reads faster and more accurately are continually being developed. In addition, the amount of permutations involved in read mapping requires a significant amount of computing resources. Most mapping software handles the vast amount of data and permutations by running in “parallel” fashion. Meaning, the software splits the job of mapping the reads of a sample to a reference genome into many parts (usually between 24 to 64 parts) that are then distributed onto a high performance computing cluster to run in parallel. Subsequent to mapping, bioinformatics tools are then used to extract useful biological information. This includes gene expression, alternative splicing, mutation detection, gene fusions among others (details of this are explained in the Methods sections of Chapters 2, 3 and Appendix 1). One of the major drawbacks of next-generation sequencing is the paucity of effective downstream analysis tools. Because of this, a substantial amount of analysis is done through the development of homemade analysis pipelines to reach the endpoints required by the user. As mentioned, the first step of RNA-seq data analysis requires the mapping of reads to a reference genome (i.e. human genome, hg18) (Figure 15). For SOLiD data, the optimal mapping algorithm is the ABI software BioScope. BioScope is uniquely designed to

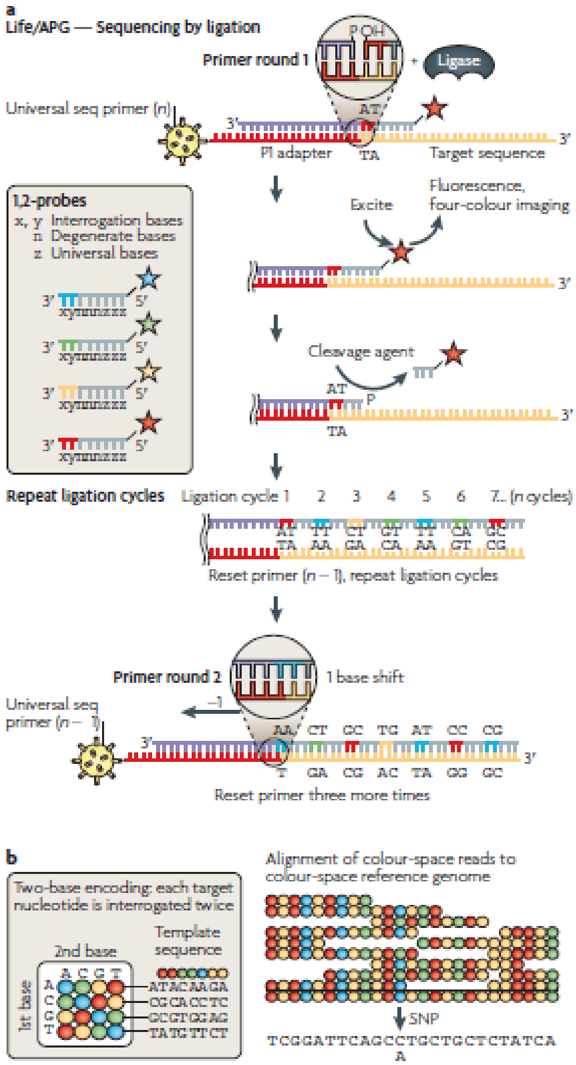


Figure 14. Colorspace sequencing by ligation using the ABI SOLiD. **(A)** A four-color sequencing by ligation method using the ABI SOLiD. Upon the annealing of a universal primer, a library of fluorescently labeled octomers is added. Appropriate conditions enable the selective hybridization and ligation of probes to complementary positions. Following four-color imaging, the ligated octomers are chemically cleaved with silver ions to generate a 5′-PO_4 group. The SOLiD cycle is repeated nine more times. The extended primer is then stripped and four more ligation rounds are performed, each with ten ligation cycles for a total of 50 cycles. The octomers are designed to interrogate the first (x) and second (y) positions adjacent to the hybridized primer, such that the 16 dinucleotides are encoded by four dyes (colored stars). The probes also contain inosine bases (z) to reduce the complexity of the octomer library and a phosphorothiolate linkage between the fifth and six nucleotides of the probe sequence, which is cleaved with silver ions. **(B)** A two-base encoding scheme in which four dinucleotide sequences are associated with one color (for example, AA, CC, GG and TT are coded with a blue dye). Each template base is interrogated twice and compiled into a string of colorspace data bits. The colorspace reads are aligned to a colorspace reference sequence to decode the sequence. Figure and caption adapted from Metzker et al. [76].

handle the mapping of colorspace data and takes advantages of its unique features. It begins by mapping reads to the reference genome. For those reads that do not map to the reference genome, it begins a process of determining whether those reads will map across an exon-exon junction (Figure 15). In order to do this, the software creates a compendium of all exon junctions that are possible for a given gene for all ~20,000 genes in the genome. This is done by taking the sequence of all exons of a gene and assembling the various combinations. Thus for each gene, both known exon junctions and “putative” exon junctions are included to account for the potential of novel alternative splicing. Reads that do not map to the genome are then mapped against this exon junction library.

After read mapping, it is the goal to being to extract useful biological information from the mapped data. One of the first analyses that can be performed is differential gene expression. Recently, bioinformatic tools to analyze differential gene expression using RNA-seq have become available. In particular, work by Mortvazi et. al. has established the RPKM model (reads per kilobase-exon per million mapped reads) as a means of standardized gene expression from RNA-seq [77]. The RPKM unit allows for a normalized absolute gene expression value that allows one to compare the expression of genes between samples. It also corrects for biases introduced by the gene length and variation in sequencing depth for each sample by accounting for the number of mapped reads (example of RPKM data in Table 2). The RPKM model has been widely adapted to commercially available software and serves as the basis of all differential gene expression calculations performed in this dissertation. Extraction of gene expression from mapped data requires an *a priori* gene annotation. Very simply, this is a database that consists mainly of the chromosome and position numbers of an area of interest. For example, if one is interested in the differential gene expression of hsa-mir-21, a gene annotation database of microRNAs would have in its entry for hsa-mir-21:

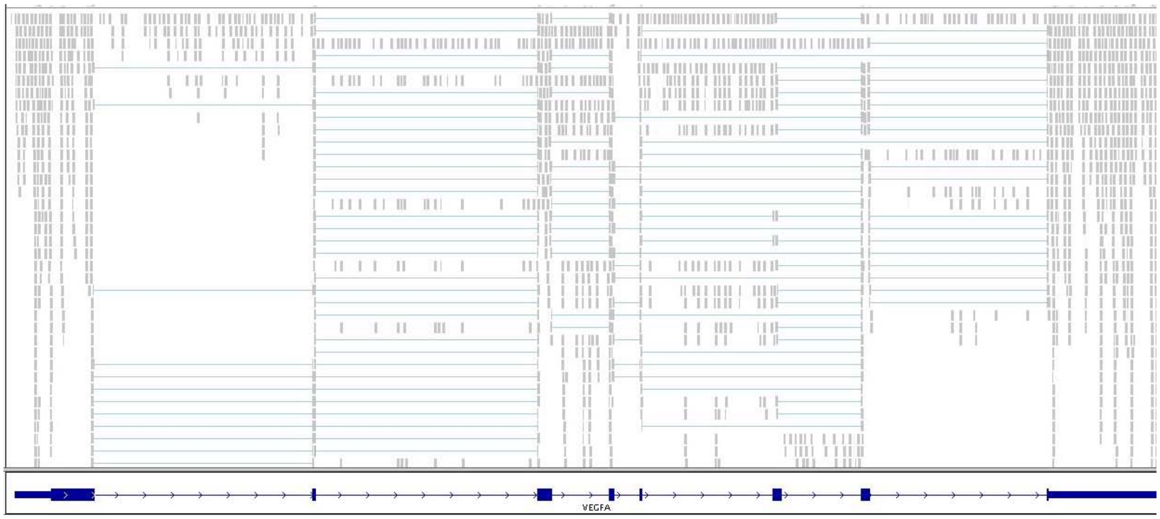
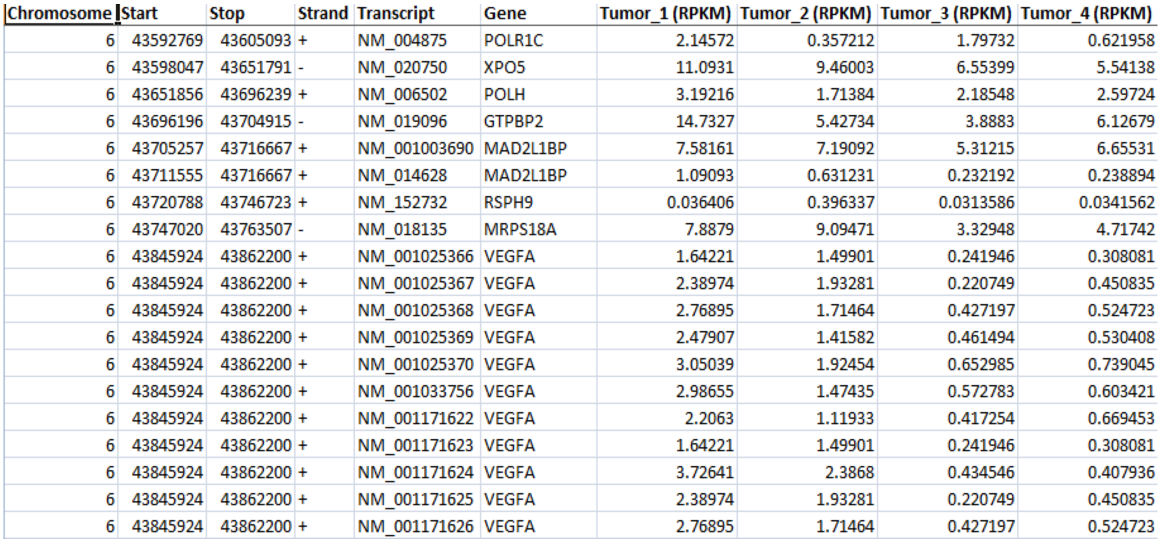


Figure 15. Example of RNA-seq reads mapping to a gene. This example shows read mapping to the Vascular Endothelial Growth Factor-A (VEGFA) gene. Each of the small boxes represents a read. Notice the majority of reads map to exons which are represented by thick boxes at the bottom. Reads in which two boxes are connected by a thin line represent reads that span an exon-exon junction. Reads that map to introns in many cases represent sequencing data derived from pre-spliced mRNA. In some cases these reads also represent non-coding RNA expression transcribed from the introns of protein-coding genes. This figure was generated using the Integrative Genomics Viewer [78].

chr17:55,273,409-55,273,480 (human genome, hg18). Most gene annotation database are available for download from the UCSC Genome Browser (http://genome.ucsc.edu). Gene annotations are also available from other private and public databases and from the literature. By combining the mapped data and the gene annotation database, RPKM values can then be calculated and then statistically compared between samples in order to derive differential gene expression. The sole exception to performing differential gene expression absent of a gene annotation database is when analyzing for novel transcribed regions (or areas of the genome where there is no annotated gene).
In this case, the mapped data is used to create a de novo gene annotation database, and then RPKMs are calculated from there.

In addition to differential gene expression, it could be desired to extract mutational information from mapped RNA-seq data. This includes point mutations, small insertion/deletions, and gene fusions. RNA-seq presents a slight challenge in calling mutations as the ability to derive base sequences is proportional to the expression of the gene. Thus, the coverage of mutations by sequencing reads is not uniform across the genome, and requires statistical models that can account for the variability (see Figure 16 for an example of coverage non-uniformity and point mutations). A recent algorithm named SNVMix2, has been developed to call point mutations specifically from RNA-seq data which account for the non-uniformity in coverage (more details presented in the Methods section of Chapter 3 and Appendix 1) [79].

Small insertion deletions (indels) are also called along the same line in terms of the need for coverage, but these mutations require special processing known as gapped alignment. Because indels by definition will have either missing bases (deletion) or extra bases (insertion) in the sequence read compared to the reference genome, the alignment software has to account for this sort of variation. In order to do this, the alignment software will break the read into smaller pieces and then begin to map the



Chromosome	Start	Stop	Strand	Transcript	Gene	Tumor_1 (RPKM)	Tumor_2 (RPKM)	Tumor_3 (RPKM)	Tumor_4 (RPKM)
6	43592769	43605093	+	NM_004875	POLR1C	2.14572	0.357212	1.79732	0.621958
6	43598047	43651791	-	NM_020750	XPO5	11.0931	9.46003	6.55399	5.54138
6	43651856	43696239	+	NM_006502	POLH	3.19216	1.71384	2.18548	2.59724
6	43696196	43704915	-	NM_019096	GTPBP2	14.7327	5.42734	3.8883	6.12679
6	43705257	43716667	+	NM_001003690	MAD2L1BP	7.58161	7.19092	5.31215	6.65531
6	43711555	43716667	+	NM_014628	MAD2L1BP	1.09093	0.631231	0.232192	0.238894
6	43720788	43746723	+	NM_152732	RSPH9	0.036406	0.396337	0.0313586	0.0341562
6	43747020	43763507	-	NM_018135	MRPS18A	7.8879	9.09471	3.32948	4.71742
6	43845924	43862200	+	NM_001025366	VEGFA	1.64221	1.49901	0.241946	0.308081
6	43845924	43862200	+	NM_001025367	VEGFA	2.38974	1.93281	0.220749	0.450835
6	43845924	43862200	+	NM_001025368	VEGFA	2.76895	1.71464	0.427197	0.524723
6	43845924	43862200	+	NM_001025369	VEGFA	2.47907	1.41582	0.461494	0.530408
6	43845924	43862200	+	NM_001025370	VEGFA	3.05039	1.92454	0.652985	0.739045
6	43845924	43862200	+	NM_001033756	VEGFA	2.98655	1.47435	0.572783	0.603421
6	43845924	43862200	+	NM_001171622	VEGFA	2.2063	1.11933	0.417254	0.669453
6	43845924	43862200	+	NM_001171623	VEGFA	1.64221	1.49901	0.241946	0.308081
6	43845924	43862200	+	NM_001171624	VEGFA	3.72641	2.3868	0.434546	0.407936
6	43845924	43862200	+	NM_001171625	VEGFA	2.38974	1.93281	0.220749	0.450835
6	43845924	43862200	+	NM_001171626	VEGFA	2.76895	1.71464	0.427197	0.524723

Table 2. Representative example of RPKM data. Gene expression of a portion of chromosome 6 derived from an RNA-seq experiment. The first 4 columns denote the read and position numbers of a gene isoform and the strand from which it is transcribed. The fifth column is the unique RefSeq ID for the isoform followed by the gene symbol in the sixth column. These first 6 columns were derived from a gene annotation database downloaded from the UCSC Genome Browser. Columns 7-10 illustrate the absolute gene expression values (RPKM) for each sample (in this case TNBCs) for each given gene isoform.

smaller pieces and extend the alignment from those pieces in order to merge them. The alignment software will allow for missing bases or extra bases in order to complete the merge, in essence mapping the read around the indel (Figure 17).

Finally, gene fusions present yet another unique bioinformatic challenge. In this case, the informatic search is for reads where one end of the read maps to one gene, and the other end of the read maps to a different gene (Figure 18). Because of sequence homology shared between like genes, a large degree of false positives are generated from fusion calling. That is why statistical confidence on whether a fusion truly exists depends on the number of reads that cross a fusion junction, and the uniqueness of the sequence reads that span the junction. Fusions can be between genes on the same chromosome (intrachromosomal) or genes on different chromosomes (interchromosomal). In RNA-seq, another level complexity is added in that intrachromosomal fusions can either be true fusions (usually caused by deletions in the underlying DNA between two genes) or can be the result of trans-splicing/read-through fusions in which two RNA transcripts are fused but there is no mutation in the DNA.

Data analysis of next-generation RNA sequencing is a complicated and arduous task. This section provides an overview of key points of what is involved in analysis of this data. In addition to these points, more downstream analyses such as network analysis, gene set enrichment, and gene ontologies which are also used in microarray data, can also be applied to RNA-seq data. As these tools are quite mature and commercially available, they are not detailed here. More granular details about RNA-seq bioinformatics are provided in the Methods sections of Chapters 2 and 3 and in Appendix 1 (Bioinformatics Appendix).

Recent comments by the co-director of the Washington University Genome Center and members of the NIH Cancer Genome Atlas at the AACR 2011 Annual Meeting has communicated that the ability to produce sequencing data is outpacing the

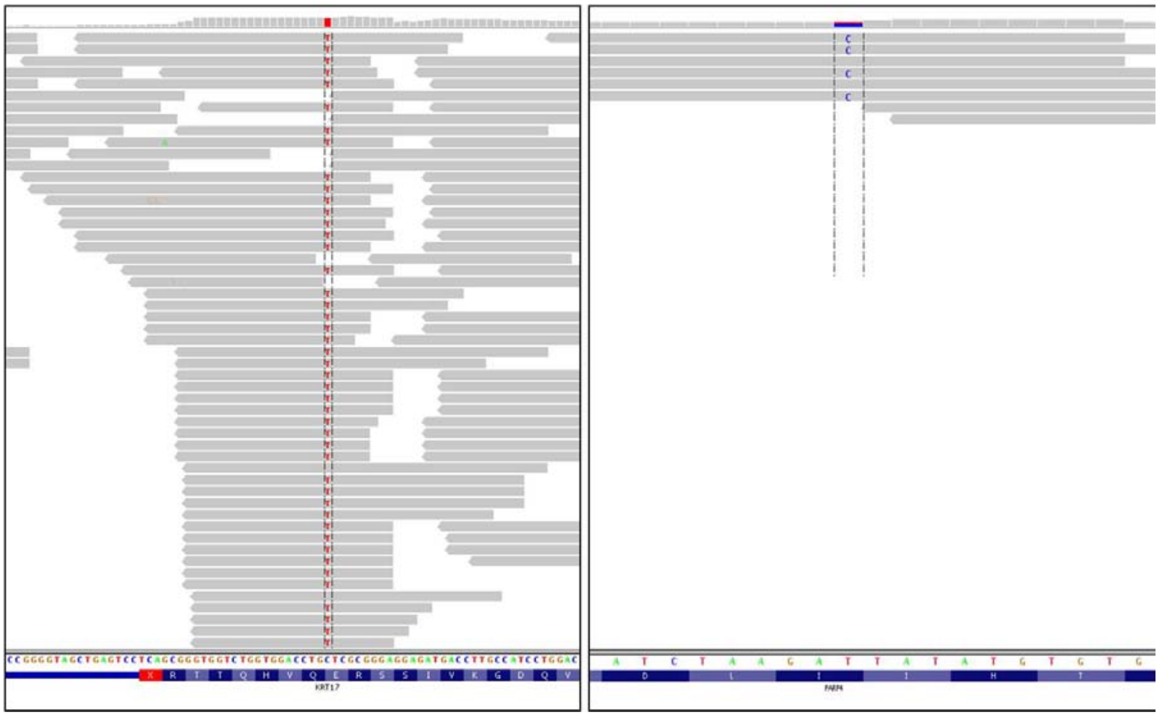


Figure 16. Example of non-uniformity in coverage when calling mutations from mapped RNA-seq data. **(LEFT)** A homozygous C-to-T mutation in a “highly expressed” gene, KRT17. Because of its high expression, a multitude of reads cover the mutation giving substantial statistical confidence that this is a true mutation. **(RIGHT)** A heterozygous T-to-C mutation in PARP4 which is expressed less than KRT17. This mutation is harder to call as there is less read coverage (in this case 6 reads) and it is also heterozygous (2 reads to the T allele and 4 reads to the C allele). In both cases these mutations are real (biologically verified), illustrating the need for bioinformatic algorithms that can account for differences in read coverage when calling mutations across the genome from mapped RNA-seq data while being vigilant against false positive mutations.

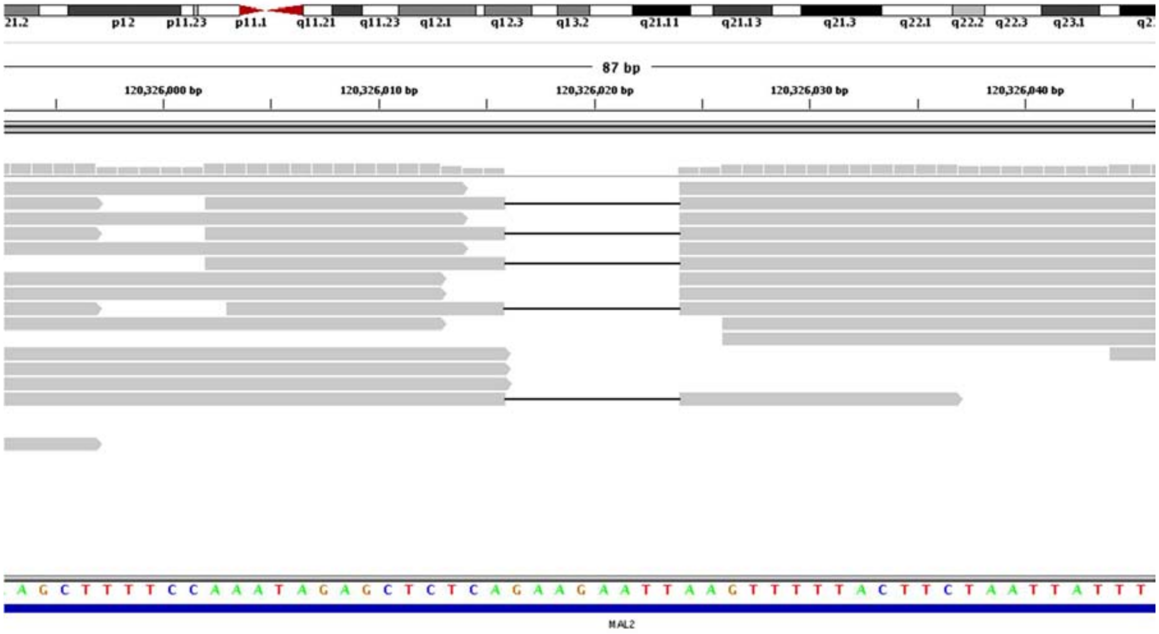


Figure 17. Example of a deletion (represented by the thick black line) detected from mapped RNA-seq data. In this case, an 8-bp deletion in the 3'-UTR of the MAL2 gene was detected from a TNBC tumor.

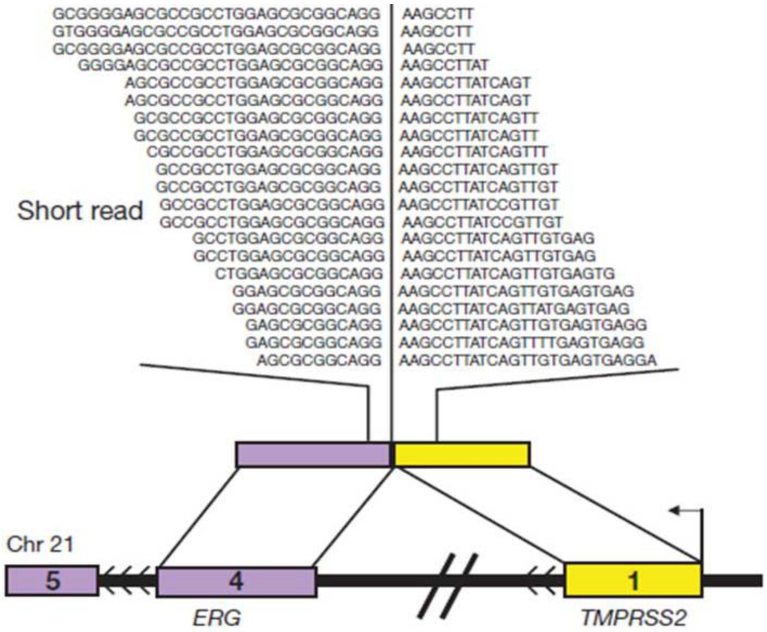

GCGGGGAGCGCCGCCTGGAGCGCGGCAGG AAGCCTT
GTGGGGAGCGCCGCCTGGAGCGCGGCAGG AAGCCTT
GCGGGGAGCGCCGCCTGGAGCGCGGCAGG AAGCCTT
GGGGAGCGCCGCCTGGAGCGCGGCAGG AAGCCTTAT
AGCGCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGT
AGCGCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGT
GCGCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGTT
GCGCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGTT
CGCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGTTT
GCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGTTGT
GCCGCCTGGAGCGCGGCAGG AAGCCTTATCAGTTGT
Short read
GCCGCCTGGAGCGCGGCAGG AAGCCTTATCCGTTGT
GCCGCCTGGAGCGCGGCAGG AAGCCTTATCCGTTGT
GCCTGGAGCGCGGCAGG AAGCCTTATCAGTTGTGAG
GCCTGGAGCGCGGCAGG AAGCCTTATCAGTTGTGAG
CTGGAGCGCGGCAGG AAGCCTTATCAGTTGTGAGTG
GGAGCGCGGCAGG AAGCCTTATCAGTTGTGAGTGAG
GGAGCGCGGCAGG AAGCCTTATCAGTTATGAGTGAG
GAGCGCGGCAGG AAGCCTTATCAGTTGTGAGTGAGG
GAGCGCGGCAGG AAGCCTTATCAGTTTTGAGTGAGG
AGCGCGGCAGG AAGCCTTATCAGTTGTGAGTGAGGA
Chr 21
5
4
1
ERG
TMPRSS2

Figure 18. Diagram of detecting a gene fusion from reads crossing a fusion junction. In this case, a representative example of using RNA-seq to detect the ERG-TMPRSS2 fusion in prostate cancer. Figure adapted from Maher et al. [69].

computational and human resources needed to analyze it. New methods to analyze RNA-seq are continuing to be developed. Specifically, advancements in assembling known and novel gene isoforms from RNA-seq data [80-82] as well as better statistical models to assess differential expression are at the forefront [83, 84]. As the field of cancer genomics using NGS is in its infancy, so is also the bioinformatic framework to analyze these large and powerful datasets.

1.4 Statement of purpose

The paucity of therapeutic targets in TNBC coupled with a lack of understanding of the global transcriptional differences between TNBC and normal breast has led us to use new technologies that can survey the transcriptome at an unprecedented depth. By using next-generation whole transcriptome sequencing (RNA-seq), we have sequenced the transcriptomes of 10 TNBCs and 10 normal breast tissues derived from healthy volunteers. We hypothesized that by sequencing the transcriptomes of TNBCs and microdissected normal tissues, we will identify novel biology, therapeutic targets, and recurrent mutations. This was accomplished through the following aims:

1. By sequencing and analyzing the transcriptomes of 10 TNBCs and 10 normal breast tissues derived from microdissected ductal epithelium, we observed key transcriptional differences between these tissues that have leant clues into the outcomes of current treatments for TNBC, and potentially identified novel therapeutic targets for future therapeutic development.

2. By using recently developed bioinformatic tools to identify mutations from mapped RNA-seq data, we have identified novel mutations in TNBC that may lead to a better understanding of disease causation.

The following chapters of this dissertation detail the methods and results of these two aims.

Chapter 2: Differential gene expression of the transcriptomes of TNBC and normal breast tissue

2.1 Introduction

Triple-negative breast cancer (TNBC) is a devastating disease that lacks effective therapeutic targets. A major impediment to discovering novel biology and developing new therapeutics for TNBC is a lack of understanding of the key transcriptional differences that differentiate TNBC from the normal breast. Microdissected ductal epithelium (the presumed origin of breast cancer) from normal breast tissues is not commonly used secondary to sample availability and difficulty in preparation. Previous gene expression studies have used either other subtypes of breast cancer (ER-positive or HER2-positive) or suboptimal normal controls which include reduction mammoplasties or adjacent normal tissue as comparators. As differential gene expression, by definition, is always relative to a control, the use of these suboptimal controls may lead to different conclusions than one would obtain when comparing TNBC to true normal tissue. In this study we sought to determine the key transcriptional differences between TNBC and normal for both coding and non-coding RNAs, and to determine whether these differences can explain the outcome of previous targeted therapies in TNBC and possibly identify new targets.

Our tool to perform this transcriptome wide comparison is next-generation whole transcriptome sequencing (RNA-seq). RNA-seq is a powerful technology that can survey the transcriptome at an unprecedented depth. In contrast to microarrays, which use *a priori* gene selection, this technology can profile every mRNA regardless if it is known or unknown, coding or non-coding, and adenylated or unadenylated, in a cost-effective and expedient manner. The ability to capture all transcripts, coupled with sequence and expression data, provides the input necessary to discover novel biology in an unbiased fashion. By analyzing RNA-seq data from TNBC tumors and normal breast tissues, we report a transcriptome-wide comparison of these tissues uncovering differentially expressed coding and non-coding genes that have not previously been implicated in triple-negative breast cancer.

2.2 Materials and methods

Selection of normal tissues for RNA-Seq and microdissection

Ten fresh-frozen normal breast tissues (core biopsies) from healthy pre-menopausal volunteers with no history of breast cancer were procured from the Susan G. Komen for the Cure® Tissue Bank at the Indiana University Simon Cancer Center. In choosing the tissues used for this study, Hematoxylin & Eosin (H&E) slides from core biopsies of various donors were reviewed in order to choose samples that contained high epithelial content, and thus produce more RNA. In an embedded pilot sub-study (outside of the main purview of this dissertation), five of the ten samples were from women in the follicular phase of the menstrual cycle, and five women were in the luteal phase. This was done in order to do determine the effects of the menstrual cycle on the gene expression of normal ductal epithelium. Because ductal epithelium (the presumed origin of breast cancer) comprises a minority of cells in the breast, normal tissues were laser capture microdissected (LCM) using the Arcturus Veritas and PixCell Microdissection Systems (Molecular Devices, Sunnyvale, CA). To perform the LCM, each frozen biopsy was sectioned (5µM thick) onto special membranous slides to create 35 slides each containing 2 sections for a total of 70 sections. These slides were then frozen at -80°C until it was time to LCM. When it was time to LCM, three slides were removed from the freezer at a time, and stained using the HistoGene LCM Frozen Section Staining Kit (Arcturus) which involves dehydration with ethanol and xylene followed by staining with a solution similar to hematoxylin. This allowed the epithelium to stain blue (Figure 19). Cells were then dissected using the LCM microscope within one hour in order to avoid RNA degradation. RNA from captured cells was then extracted using the PicoPure RNA Isolation Kit (Arcturus) and the purified RNA was quantified using the Qubit Fluorometer (Invitrogen, Carlsbad, CA) (See Table 3 for sample details and RNA yields). Because of the lower yields associated with LCM, the entire amount collected was used for the RNA-seq library preparation.

Selection of TNBC samples for RNA-seq

RNA from ten triple-negative breast cancers (TNBC) were procured from OriGene Technologies. All samples were pathologically verified for high tumor content and did not necessitate microdissection. TNBC samples were from pre-menopausal women in order to match our normal cohort (Table 4). All samples were quantified using the Qubit Fluorometer (Invitrogen, Carlsbad, CA). There was an abundance of available RNA, so 5µg was used for RNA-seq library preparation. The use of all samples including normals and TNBCs were approved for use by the Indiana University Institutional Review Board. After sample preparation, RNA was sent to Cofactor Genomics (St.Louis, MO), who performed the ribosomal depletion, library preparation and sequencing where subsequent analyses and validations were performed at Indiana University.

Ribosomal RNA depletion of samples

Because the majority of RNA in a cell is ribosomal, in order to profile the unique transcriptome, samples used for sequencing were depleted for ribosomal RNA (rRNA). Removal of rRNA was not achieved by traditional poly-A RNA selection, but by rRNA depletion via locked nucleic acid probes. This allowed for profiling of both poly-A and non-poly-A RNA species. Using the RiboMinus Eukaryote Kit (Invitrogen), samples were depleted for 5S, 5.8S, 18S, and 28S rRNA, per manufacturer's instructions. Briefly, total RNA was hybridized to rRNA-specific biotin labeled probes at 70°C degree for 5 minutes. The rRNA-probe complexes were then removed by incubating with streptavidin-coated magnetic beads. The rRNA free transcriptome RNA was precipitated with ethanol and concentrated.

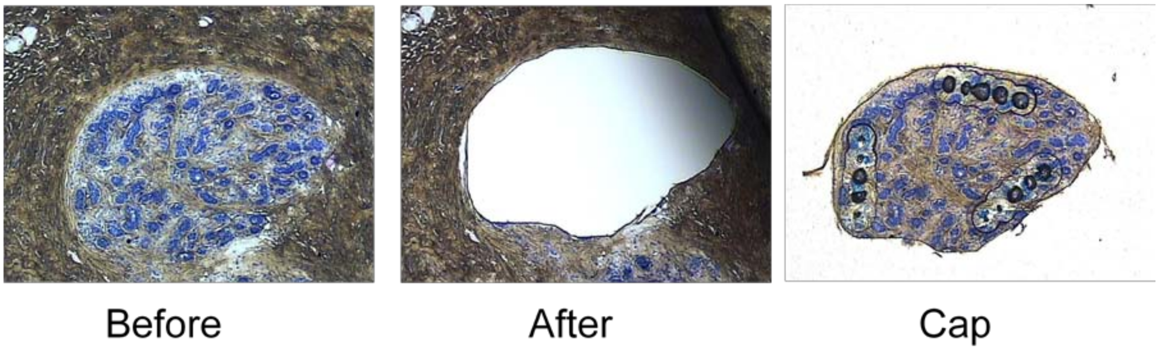


Figure 19. Pictures of normal ductal epithelium from frozen normal breast tissues before and after LCM. The normal ducts are stained blue. A laser was used to circumscribe the ducts, followed by a separate laser to melt a plastic film from a cap used in the LCM procedure in order to detach the tissue. LCM made it possible to focus sequencing efforts on RNA from ductal epithelium.

Sequencing ID	**Komen Sample #**	**Total RNA**	**Menstrual Phase**
Normal #1	104825	1.300ug	Follicular
Normal #2	104877	4.053ug	Follicular
Normal #3	102442	0.758ug	Follicular
Normal #4	102428	2.712ug	Luteal
Normal #5	104841	0.333ug	Follicular
Normal #6	102583	0.612ug	Luteal
Normal #7	102541	1.273ug	Luteal
Normal #8	104867	0.630ug	Luteal
Normal #9	102518	0.589ug	Follicular
Normal #10	102430	0.240ug	Luteal

Table 3. Normal samples used for RNA-seq with corresponding information. Information includes ID number from the Susan G. Komen Tissue Bank at the IUSCC, total RNA yield, and menstrual phase for each donor.

Sample ID	Age	Case Diagnosis from Donor Institution Pathology Report	Tumor Grade	TNM
Tumor #1	20	Adenocarcinoma of breast, ductal	Nottingham G3: 8-9 points High combined grade (unfavorable)	pT1bpNXpMX
Tumor #2	31	Adenocarcinoma of breast, ductal, recurrent	Nottingham G3: 8-9 points High combined grade (unfavorable)	pT2pN0pMX
Tumor #3	32	Adenocarcinoma of breast, ductal	Nottingham G3: 8-9 points High combined grade (unfavorable)	pT2pN0pMX
Tumor #4	36	Adenocarcinoma of breast, ductal	Nottingham G3: 8-9 points High combined grade (unfavorable)	pT1cpN1mipMX
Tumor #5	37	Adenocarcinoma of breast, ductal, medullary features	Nottingham G3: 8-9 points High combined grade (unfavorable)	pT2pN0 (i-)pMX
Tumor #6	38	Adenocarcinoma of breast, ductal	Nottingham G3: 8-9 points High combined grade (unfavorable)	pT2pN1apMX
Tumor #7	39	Adenocarcinoma of breast, ductal	Nottingham G3: 8-9 points High combined grade (unfavorable)	pT2pN1pMX
Tumor #8	40	Adenocarcinoma of breast, ductal, medullary features	Nottingham G3: 8-9 points High combined grade (unfavorable)	pT1cpN0pMX
Tumor #9	42	Adenocarcinoma of breast, ductal	Nottingham G3: 8-9 points High combined grade (unfavorable)	pT1cpN3bpMX
Tumor #10	49	Adenocarcinoma of breast, ductal	Nottingham G3: 8-9 points High combined grade (unfavorable)	pT3pN1apMX

Table 4. TNBC samples used for RNA-seq with corresponding information. Information includes age, diagnosis, grade and TNM (Tumor Node Metastasis) staging.

RNA fragmentation and adaptor ligation

The SOLiD Whole Transcriptome Analysis Kit (Applied Biosystems, Foster City, CA) was used to create the transcriptome libraries. Briefly, transcriptome RNA in 8µL was incubated with 1µL RNase III and 1µL 10X RNase III Reaction Buffer at 37°C for 10 minutes for fragmentation. 90µL of nuclease-free water was added into each reaction for cleanup using the RiboMinus Concentration Module (Invitrogen). Fragmented RNA was hybridized with 2µL Adaptor Mix A and 3µL Hybridization Solution at 65°C for 10 minutes followed by 16°C for 5 minutes. For RNA ligation, 10µL Ligation Buffer and 2µL Ligation Enzyme Mix were added to the hybridization reaction at 16°C and incubated for 16 hours.

cDNA synthesis

A 20µL reverse transcription master mix containing 13µL of nuclease-free water, 4µL of 10X RT buffer, 2µL of 2.5 mM dNTP Mix and1µL of ArrayScript Reverse Transcriptase was added to the previous 20µL ligation reaction and then incubated at 42°C for 30 minutes. Synthesized cDNA was purified with the Qiagen MinElute PCR Purification Kit (Qiagen, Valencia, CA) and eluted in 10µL EB buffer. The purified cDNA was run on a Novex 6% TBE-Urea Gel (Invitrogen) for size selection. The excised gel of 150-250nt was divided into 4 pieces and two in-gel PCR reactions were conducted to obtain enough material for subsequent emulsion PCR (ePCR). Meanwhile, each library is barcoded by using PCR primers containing different barcodes to allow multiple samples to be sequenced simultaneously on the SOLiD 3 system.

Emulsion PCR and sequencing

Emulsion PCR was conducted according to manufacturer’s instructions (Applied Biosystems SOLiD 3 System Templated Bead Preparation Guide). The amplified beads were first run on a Work Flow Analysis (WFA) slide to determine the quality and quantity of beads which was followed by two 50bp fragment sequencing runs using both slides. (Applied Biosystems SOLiD 3 System Instrument Operation Guide).

Read mapping

The SOLiD 3 system produces two primary output text files: the .csfasta file which contains the sequencing reads in colorspace format, and the .qual file which contains the quality values for each colorspace call. These primary output files from the SOLiD 3 were loaded onto a compute cluster and the reads were mapped in colorspace using the Applied Biosystems BioScope 1.2 software using default parameters (see Bioinformatics Appendix 1 for required input files). Reads were mapped to the human genome (hg18) downloaded from the UCSC Genome Bioinformatics Site (http://genome.ucsc.edu). The hg18 genome was slightly modified by deleting the Y chromosome in order to make a female genome. An hg18 exon reference file provided by Applied Biosystems was required by BioScope in order to create the exon junction libraries needed to map reads that cross exon boundries. This file was derived from the refGene database from UCSC. Also, a human filter reference file was required (provided by Applied Biosystems) that contains the sequences of repetitive regions of the genome. in order to filter reads that mapped to repetitive areas of the genome. Mapped reads were outputted from BioScope in the standard BAM (Binary Alignment/Map) format.

Gene expression analysis

BAM files were imported into Partek Genomics Suite 6.5 (Partek Incorporated, St. Louis, MO) for gene expression analysis. First, reads were cross-referenced against the RefSeq database (downloaded from UCSC) and RPKM values generated for each gene. Refseq is a database of ~20,000 highly annotated (mostly protein coding) genes.

Because the SOLiD 3 retains strandedness as part of its chemistry, we were capable of quantifying gene expression if two genes overlapped on opposing strands. The RPKM values of each gene for the 10 normals and the 10 TNBC tumors were then statistically compared using a 1-way ANOVA with a multiple test correction of FDR < 0.01. A one-way ANOVA was used as only one comparison (tumor vs. normal) was being considered. The same methods were then subsequently applied to pre-miRNAs, Ultra Conserved Regions (UCRs), and long intergenic non-coding RNAs (lincRNAs). Gene annotations for pre-miRNAs were downloaded from the wgRNA database at UCSC, UCRs from the uc16 database on the UCSC test-server, and lincRNAs from supplementary data in a publication from Khalil et al. [85]. In order to assess clustering of the samples, Principal Components Analysis of RPKM values from RefSeq genes was performed in Partek Genomics Suite. For network analysis, statistically significant genes were analyzed using the Ingenuity Pathway Analysis software (Ingenuity Systems, Redwood City, CA).

Novel transcribed regions (NTRs)

To identify differentially expressed novel transcribed regions (NTRs), we analyzed the entire genome for areas of significant expression for which there is no known annotated gene. We chose areas of novel transcription with the following criteria: 1) the NTR had to be a minimum of 10,000bp away from the nearest RefSeq gene to reduce the possibility of detection of a novel exon to a known gene, 2) the NTR must be at least 50bp long, 3) there must be at least 20 supporting reads mapping to the NTR. We then further filtered out loci that overlapped known miRNAs, lincRNAs, UCRs, snoRNAs, snRNAs, scRNAs, rRNAs, tRNAs, and mtRNAs. Regions that satisfied these criteria were then compiled using GALAXY (http://galaxy.psu.edu), a online bioinformatics software suite, to determine start and end coordinates of each region and to form a new “NTR reference.” This method of determining region boundaries is not the optimal method (compared to traditional cloning methods), but provides an initial screen for differential expression. Reads were then cross-referenced against the NTR reference to calculate RPKM values with subsequent statistical comparison between TNBC and normal samples using 1-way ANOVA (FDR < 0.01 as the significance cutoff).

Validation cohort of samples used for qPCR validation of differential gene expression

A separate cohort of ten fresh-frozen normal breast tissues from healthy pre-menopausal volunteers with no history of breast cancer was procured from the Susan G. Komen for the Cure® Tissue Bank at the Indiana University Simon Cancer Center. In this validation cohort, five samples were from women in the follicular phase of the menstrual cycle, and five women in the luteal phase, the same as the sequencing cohort. These samples were also laser capture microdissected for ductal epithelium as described previously, but with the following exceptions. The LCM was performed using a Leica System (Leica Microsystems, Buffalo Grove, IL) and the RNA was extracted with the miRNeasy Mini Kit (Qiagen) which extracts both full-length and microRNA.

In addition, a separate cohort of 26 fresh-frozen TNBC tumors were procured from the Indiana University Simon Cancer Center Tissue Bank and from Asterand plc. The 26 TNBC tumors represented a mixture of pre- and post-menopausal patients. RNA was extracted from the 26 TNBC samples using the miRNeasy Mini Kit (Qiagen). All samples were quantified using the Qubit Fluorometer (Invitrogen).

qPCR validation of EGFR, KIT, & PARP1 differential gene expression

TaqMan qPCR was performed using the following inventoried TaqMan assays from Applied Biosystems:

Gene Symbol	Gene Name	Assay Type	ABI Assay ID
EGFR	epidermal growth factor receptor	Target	Hs01076091_m1
KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	Target	Hs00174029_m1
PARP1	poly (ADP-ribose) polymerase 1	Target	Hs00242302_m1
HNRNPH1	heterogeneous nuclear ribonucleoprotein H1 (H)	Housekeeper	Hs00800662_sH
IPO8	importin 8	Housekeeper	Hs00183533_m1

The housekeeping genes, HNRNPH1 & IPO8, were chosen for their low coefficient of variation among the samples in the next-generation sequencing data.

Briefly, 20ng of RNA from each sample of the validation cohort was reversed transcribed using the High Capacity RNA-to-cDNA kit (Applied Biosystems). cDNA was then pre-amplified using TaqMan PreAmp Master Mix (Applied Biosystems) and the TaqMan assays per manufacturer’s instructions. qPCR of target genes and housekeepers was performed in triplicate using TaqMan Gene Expression Master Mix (Applied Biosystems) and the above listed assays. qPCR reactions were run on an ABI 7900HT Real-Time PCR System and data analyzed using the SDS2.3 and DataAssist v2.0 software from Applied Biosystems. Fold change was calculated using the standard ΔΔCt method incorporating the geometric mean of the housekeepers. Error bars represent the Standard Error of the Mean (S.E.M). P-values were calculated using t-test.

Immunohistochemistry (IHC) of EGFR and KIT

IHC was performed for EGFR using the Dako EGFR PharmDX Kit per manufacturer’s instructions (Dako, Denmark). For KIT, slides were deparaffinized using successive ethanol and xylene washes followed by antigen retrieval in a pressure cooker using TRS low pH buffer (Dako). Slides were incubated in 3% hydrogen peroxide for 10 minutes and then rinsed. The slides were then incubated with an anti-human c-Kit antibody (Dako) at a 1:200 dilution for 10 minutes and then rinsed. This slides were then incubated with Flex reagent + Rabbit linker (Dako) for 10 minutes then rinsed, followed by and incubation with Flex reagent HRP for 10 minutes and rinsed, followed by incubation with Flex reagent DAB for 5 minutes and then rinsed. Slides were then counterstained, dehydrated, and then coversliped. Slides were then read and scored by an experienced breast pathologist (Dr. Sunil Badve). Data analysis was performed by calculating a H-score for each sample (Intensity x % cellular staining) followed by statistical comparison using t-test.

2.3 Results

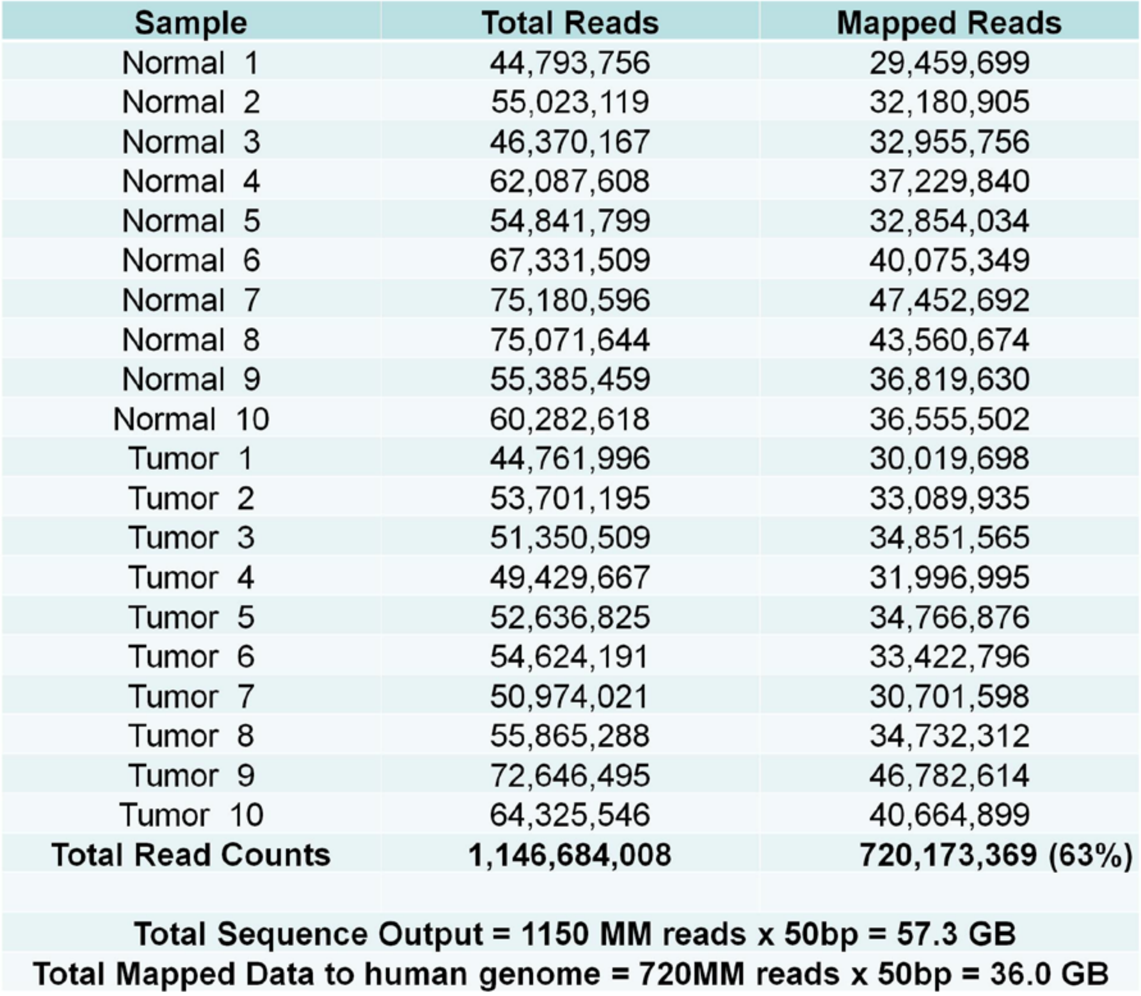
2.3.1 RNA sequencing output of TNBCs and normals

RNA sequencing of the 10 TNBCs and 10 normal breast tissues was performed on the Applied Biosystems SOLiD 3 next-generation sequencer using 50bp fragment runs. A total of 2 sequencer runs across 4 flowcell slides was performed. The sequencing produced a total of 1.1 billion reads equaling 57.3 GB of data of which 36.0GB (63%) mapped to the human genome (Table 5). A 63% mapping rate is reflective of the standard 60-70% average mapping rate seen in RNA-seq experiments [77]. Unmapped reads can occur for a variety of reasons including sequencing error, inability to uniquely map a read, and transcription from gapped or unassembled areas of the genome. Total output of RNA-sequencing was equivalent to 18 human genomes. On average, each sample had 36 million mapped reads for which gene expression values were interrogated.

2.3.2 Profiling and differential expression of known genes from the RefSeq database

Data from mapped reads were first cross-referenced against 20,839 known genes in the RefSeq database and RPKM values calculated for each gene. In our first analysis, we performed unsupervised Principal Components Analysis (PCA) of the data using RPKM values for the normal samples only (Figure 20). A PCA analysis reduces the dimensionality of the data (in this case 20,839 axes) to a three dimensional space which allows one to determine how different or alike a group of samples are based on their gene expression. The PCA clustering shows a significant separation of the normal samples between follicular and luteal samples. The analysis of the genes that differentiate the normal breast samples is outside the purview of this dissertation, but do include genes involved in chromosomal separation, mitosis, and gland development (data not shown).

In our next analysis, we compared TNBC to normal breast epithelium using PCA. The PCA demonstrates a significant separation of TNBC and normal samples illustrating the vast differences in their respective transcriptome profiles (Figure 21). To better understand the individual genes that differentiate TNBC and normal breast, we compared the RPKM expression values between the 10 TNBCs and 10 normal breast tissues for each gene using 1-way ANOVA. In Appendix 2, we report the differentially expressed genes between TNBC and normal breast sorted by p-value. When considering a false discovery rate (FDR) < 0.01, we report 7,140 differentially expressed RefSeq genes. In a first- pass analysis of the individual genes that differentiate TNBC from normal, we sought to confirm the findings of other groups that TNBC is not of basal/myoepithelial origin, but instead derives from a luminal origin (reviewed in Section 1.1.2). To do this, we looked at the expression of some key basal and luminal markers. Congruent with past literature, we observed a downregulation of basal markers CD10



Sample	Total Reads	Mapped Reads
Normal 1	44,793,756	29,459,699
Normal 2	55,023,119	32,180,905
Normal 3	46,370,167	32,955,756
Normal 4	62,087,608	37,229,840
Normal 5	54,841,799	32,854,034
Normal 6	67,331,509	40,075,349
Normal 7	75,180,596	47,452,692
Normal 8	75,071,644	43,560,674
Normal 9	55,385,459	36,819,630
Normal 10	60,282,618	36,555,502
Tumor 1	44,761,996	30,019,698
Tumor 2	53,701,195	33,089,935
Tumor 3	51,350,509	34,851,565
Tumor 4	49,429,667	31,996,995
Tumor 5	52,636,825	34,766,876
Tumor 6	54,624,191	33,422,796
Tumor 7	50,974,021	30,701,598
Tumor 8	55,865,288	34,732,312
Tumor 9	72,646,495	46,782,614
Tumor 10	64,325,546	40,664,899
Total Read Counts	**1,146,684,008**	**720,173,369 (63%)**

Total Sequence Output = 1150 MM reads x 50bp = 57.3 GB

Total Mapped Data to human genome = 720MM reads x 50bp = 36.0 GB

Table 5. Sequencing output from RNA-seq of 10 normal breast tissues and 10 TNBCs. The table show the number of reads produced for each sample, and the number of the reads that mapped to human genome. Summary statistics are at the bottom.

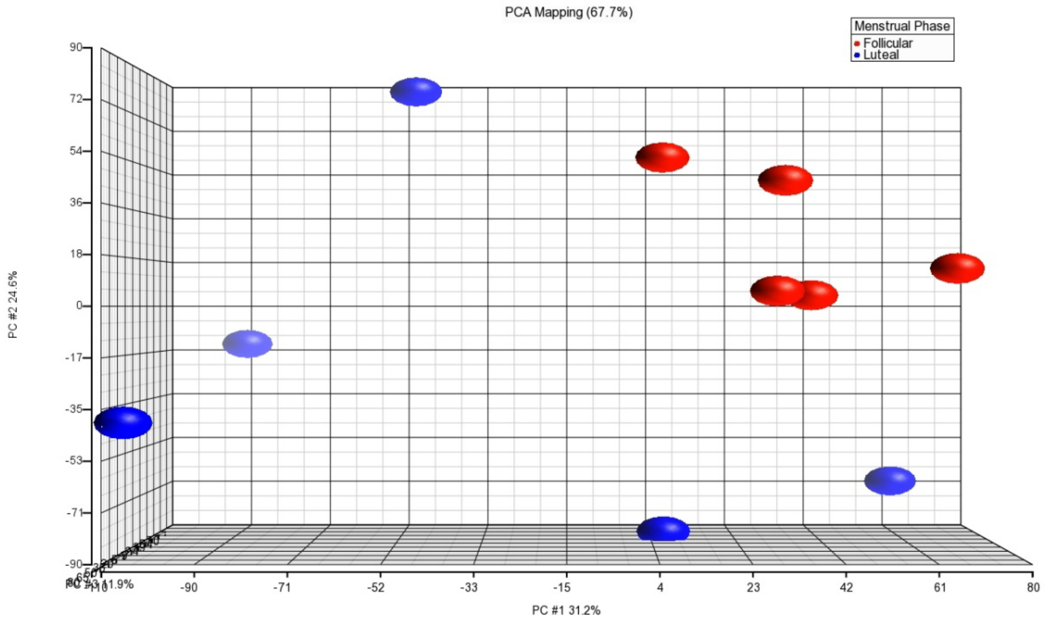


Figure 20. Principal components analysis (PCA) of normal samples. The PCA illustrates in three dimensional space the global gene expression clustering of samples. The diagrams show that normal ductal epithelium derived from women in the follicular phase of their menstrual cycle nicely separate from samples derived from women in the luteal phase of their menstrual cycle suggesting an effect of the menstrual cycle on the transcriptomes of ductal epithetlium. This unsupervised PCA was created using $\log_2$ transformed RPKM data from RefSeq genes with very low expressing genes (max RPKM < 1) omitted.

(RefSeq gene symbol: MMF, -6.5 fold downregulated), p63 (RefSeq gene symbol: TP63, -8.8 fold downregulated), and alpha-smooth muscle actin (RefSeq gene symbol: ACTA2, -2.32 fold downregulated), in TNBCs when compared to normal (all genes statistically significant at $p < 0.005$). We also observed a significant upregulation of luminal Cytokeratin 8 (RefSeq gene symbol KRT8, 5.3 fold upregulation, p=0.002), and significant expression of Cytokeratins 5,6, & 18 that were not significantly different from normal. These findings together support the work of others of a luminal phenotype for TNBC [33-37], and further confirm as a positive control the validity of the bioinformatic and statistical framework for the differential gene expression.

The #1 most statistically significant differentially expressed gene was COBRA1 (cofactor of BRCA1, 3.67 fold upregulated, p=7.42e-12). This gene is quite interesting in the context of TNBC in that is a negative regulator ER-alpha activity [86, 87], and physically associates with BRCA1. COBRA1 is part of the negative elongation factor complex and is known to suppress ER-alpha mediated transcription by RNA Polymerase II [86]. In work by Aiyar et al., shRNA knockdown of either COBRA1, BRCA1 or the combination of both in T47D human breast cancer cells (ER-positive) followed by microarray analysis showed a significant overlap of the genes that are regulated by both proteins [88]. A total of 287 genes overlapped, and gene ontology analysis showed an enrichment of genes involved in cell cycle control, proliferation, development, cell death, and cancer. A review of the literature shows that the overwhelming majority of research on COBRA1 has been performed in ER+ cell lines, and thus the implications of COBRA1 in TNBC is unknown. Further, the role of COBRA1 in breast cancers already defective for BRCA1 is still to be determined. Nonetheless, a protein involved in inhibition of ER-alpha working in concert with a protein well known to be implicated in TNBC (BRCA1) warrants further study.

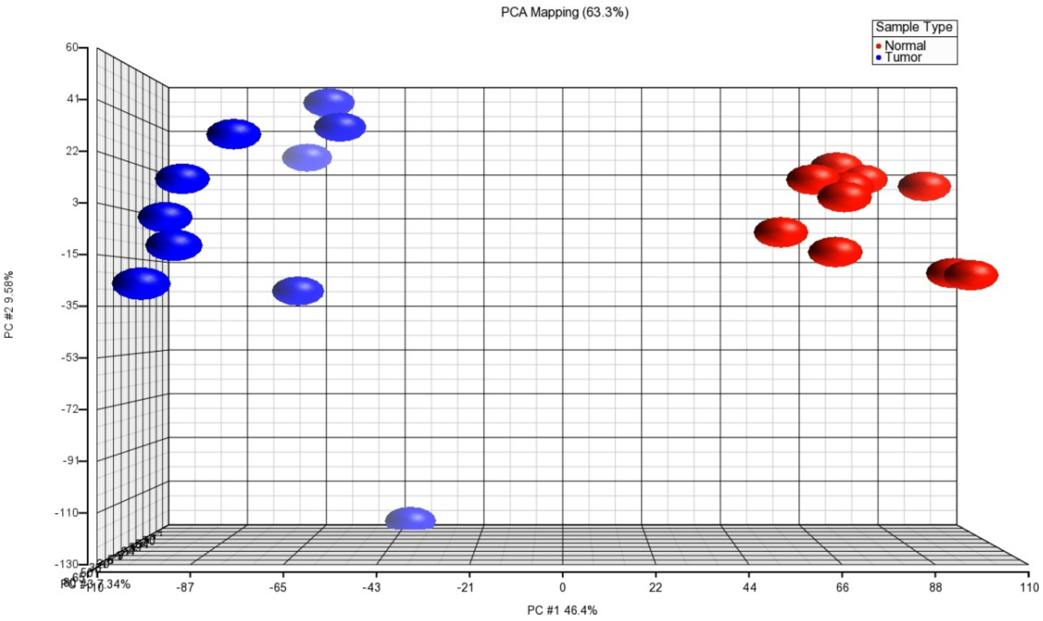


Figure 21. Principal components analysis (PCA) of TNBC and normal samples. The PCA analysis demonstrates a significant separation of TNBC and normal samples illustrating drastically different transcriptional profiles. This unsupervised PCA was created using $\log_2$ transformed RPKM data from RefSeq genes with very low expressing genes (max RPKM < 1) omitted.

To better understand the differential expression of genes at a more genome-wide level we performed a pathway analysis of the 7,140 genes which revealed many networks known to be involved in tumorigenesis. One of the statistically significant pathways identified was “Role of BRCA1 in DNA Damage Response” (Figure 22). This pathway contains many genes involved in the BRCA pathway and its downstream effectors. As mentioned in Section 1.1.1, TNBCs are defective in their DNA repair capacity [17, 18], and it is not surprising to see many genes in the BRCA/DNA repair pathway to be significantly upregulated in TNBC. Interestingly, BRCA1 itself is not differentially expressed, and its raw expression is very low both in normals and TNBCs. As seen in Figure 22, genes such as the Fanconi Anemia family (including BRCA2), MSH2, MSH6, ATR, CHK1, CHK2, and PLK1 are all significantly upregulated in TNBC.

Of interest, inhibitors of CHK1 & CHK2, are in several early clinical trials led by Eli Lilly, Pfizer, AstraZeneca, and Exelixis [89]. Similarly, several inhibitors of PLK1 (polo-like kinase 1) are in early clinical trials led by GlaxoSmithKline, Boehringer Ingelheim, and Tekmira (clinicaltrials.gov). The effectiveness of these drugs in TNBC is still yet to be determined.

We then examined the differential expression of genes that have been targeted in late stage clinical trials. Recently, inhibition of the DNA repair protein PARP has demonstrated clinical activity for patients with sporadic TNBC [25]. In our study, PARP expression was indeed significantly upregulated when compared to normal breast (Table 6). In addition to PARP, there were some equally important and interesting genes/targets that were not overexpressed as expected. EGFR and KIT, which have previously been shown to be overexpressed in TNBC [8] (and unsuccessfully tested as drug targets [9-14]) were not differentially expressed or even down-regulated, respectively, when compared to normal breast in our study (Table 6). To further validate these findings, we

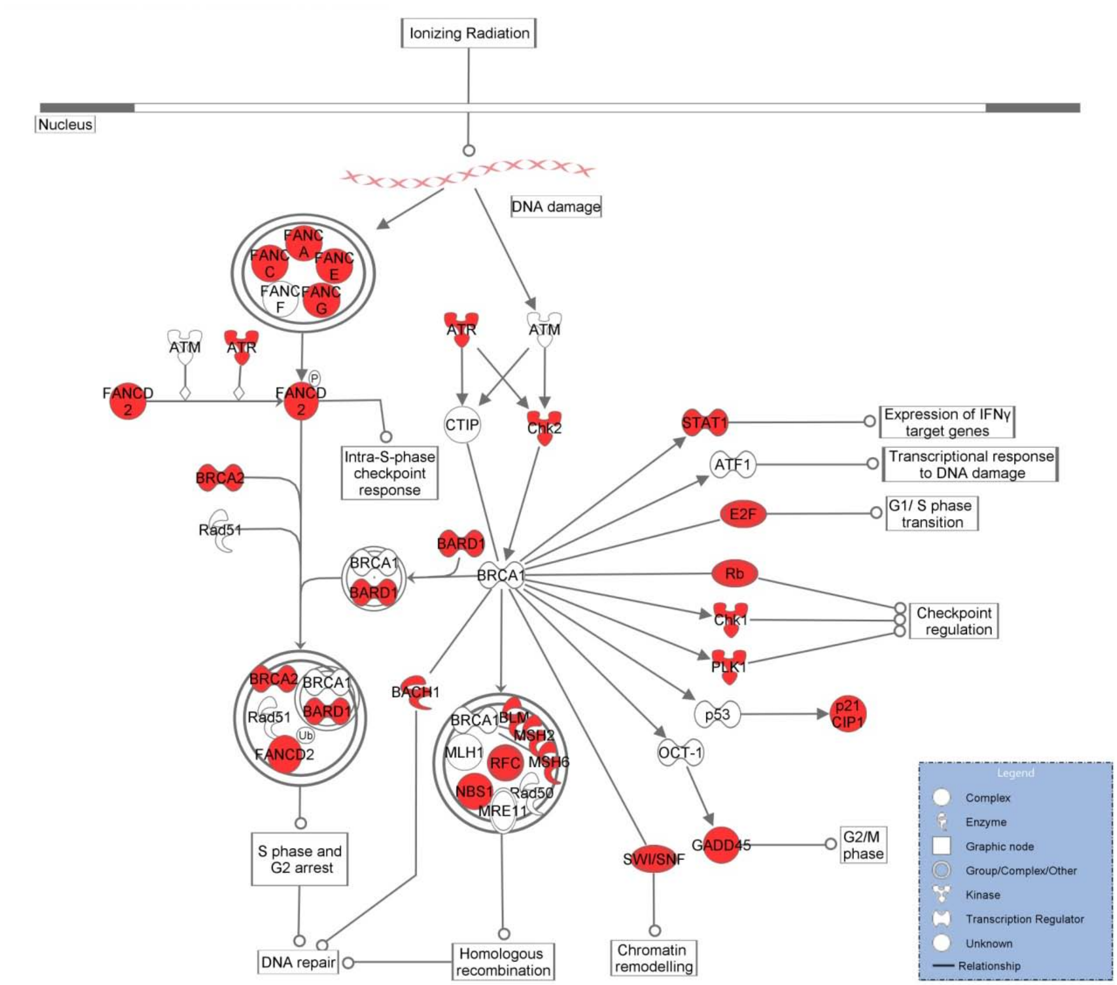
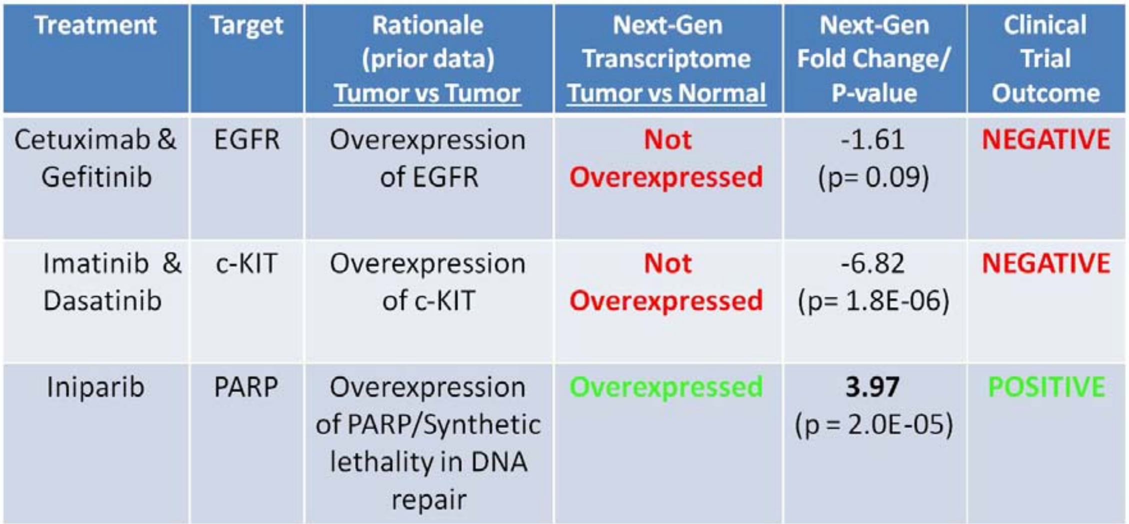


Figure 22. BRCA1 in DNA damage response pathway. This pathway shows many genes that are upregulated in the BRCA1 pathway and illustrates the importance of this pathway in TNBC. This figure was generated using Ingenuity Pathway Analysis of statistically significant differentially expressed genes.

assessed the gene expression of EGFR, KIT, and PARP1 in a separate cohort of 26 frozen TNBCs and 10 normal samples by qPCR (Figure 23). The qPCR data from the validation cohort confirmed the findings from the next-generation sequencing of a lack of differential expression of EGFR, downregulation of KIT, and upregulation of PARP1 (Figure 23). To further confirm at the protein level, we performed immunohistochemistry (IHC) for EGFR and KIT on 20 normal breast tissues and 11 TNBCs (Figures 24 and 25). EGFR and KIT IHC are standard clinical stains in our hospital laboratory. The IHC also demonstrates no difference in EGFR expression and downregulation of KIT in TNBC compared to normal (Figures 24 and 25). Interestingly, the lack of upregulation of EGFR and KIT when comparing TNBC to normal is in contradiction to previous reports of overexpression of these genes in TNBC [8]. Strikingly, our data is in line with clinical trial outcomes of agents that target these proteins, and suggests that comparing TNBC to microdissected ductal epithelium versus other suboptimal comparators may yield better therapeutic targets.

Building upon that observation, we analyzed the dataset to see if we could identify novel therapeutic targets. In addition to the previous thought that EGFR and c-KIT were overexpressed in TNBC, these proteins were attractive clinical targets because of their intrinsic nature as receptor tyrosine kinases (RTKs). RTKs are potent activators of key signaling cascades important for tumor cell proliferation and survival [90]. Indeed, RTK inhibition has had a history of success in cancer, for example EGFR in lung cancer [91], HER2 in breast cancer [92], and VEGFR2 in renal cell carcinomas [93]. With our dataset suggesting that EGFR and c-KIT are not the RTKs overexpressed in TNBC, we obtained a list of 58 known RTKs in the human genome [94]. 32 of these 58 were significantly expressed in our samples. We then statistically compared these 32 RTKs between the 10 TNBCs and 10 normals to form a volcano plot (Figure 26). This volcano plot identifies significant RTKs by considering both fold change and statistical



Treatment	Target	Rationale (prior data) Tumor vs Tumor	Next-Gen Transcriptome Tumor vs Normal	Next-Gen Fold Change/ P-value	Clinical Trial Outcome
Cetuximab & Gefitinib	EGFR	Overexpression of EGFR	Not Overexpressed	-1.61 (p= 0.09)	NEGATIVE
Imatinib & Dasatinib	c-KIT	Overexpression of c-KIT	Not Overexpressed	-6.82 (p= 1.8E-06)	NEGATIVE
Iniparib	PARP	Overexpression of PARP/Synthetic lethality in DNA repair	Overexpressed	**3.97** (p = 2.0E-05)	POSITIVE

Table 6. Differential gene expression of three drug targets clinically tested in enriched TNBC patient populations and their clinical trial outcomes. The table reports the prior rationale for inhibiting these targets in TNBC and then compares it to our data using RNA-seq of TNBC vs Normal. When compared to normals, we observe no difference in EGFR expression, downregulation of KIT, and upregulation of PARP1 in TNBC. These results are congruent with clinical trial outcomes testing targeted agents against these proteins. This data suggests that comparing TNBC to normal may explain the outcome of previous clinical trials, and may potentially identify novel therapeutic targets.

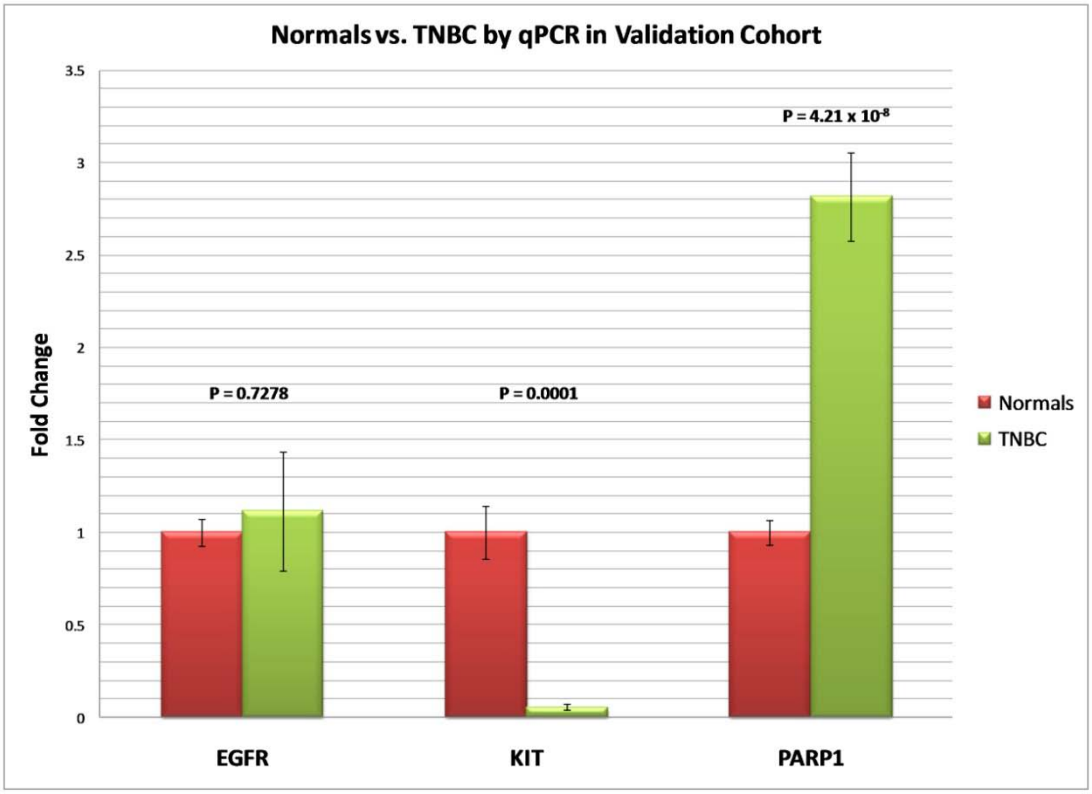


Figure 23. Biological validation of next-generation sequencing results of EGFR, KIT, and PARP1 using Taqman qPCR. qPCR was performed in a separate validation cohort of 26 TNBCs and 10 microdissected normals. Error bars reflect standard error of the mean and p-value is from t-test. HNRNPH1 (Heterogeneous nuclear ribonucleoprotein H) and IPO8 (Importin 8) were used as housekeepers. The qPCR data from the validation cohort confirmed the findings from the next-generation sequencing of a lack of differential expression of EGFR, downregulation of KIT, and upregulation of PARP1.

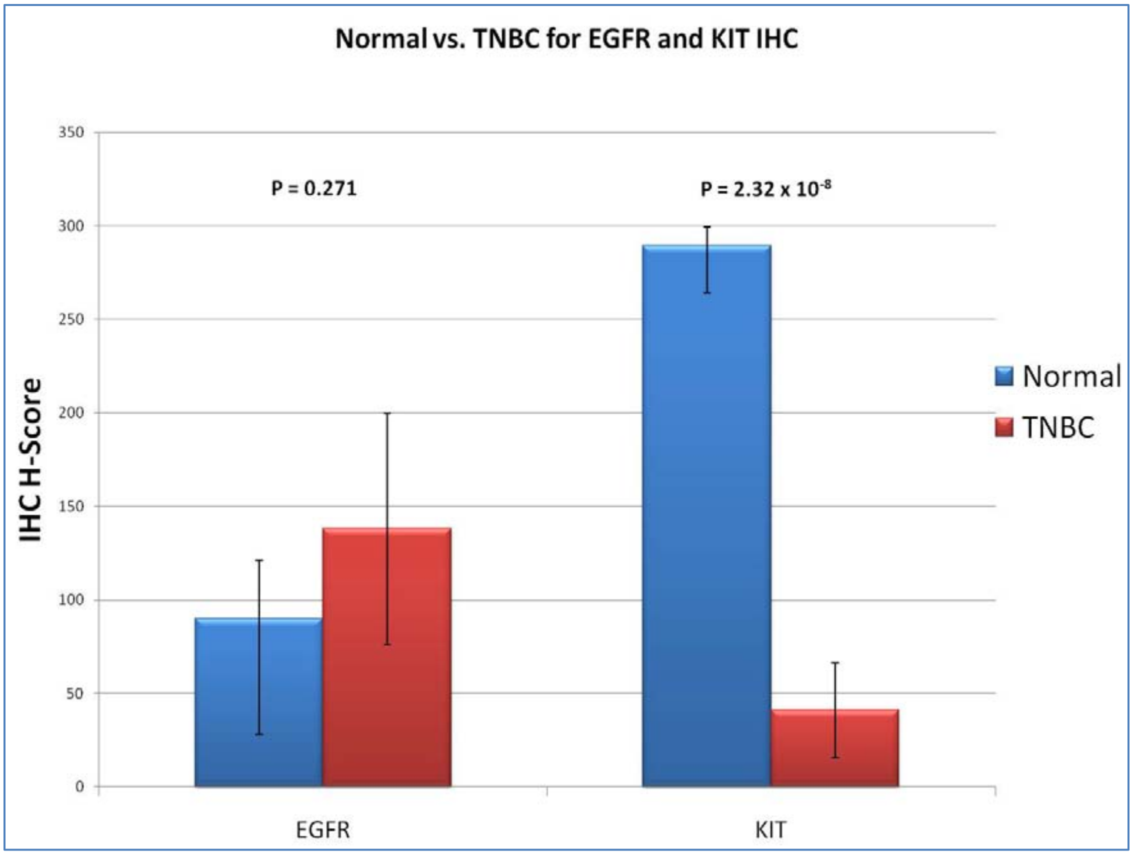
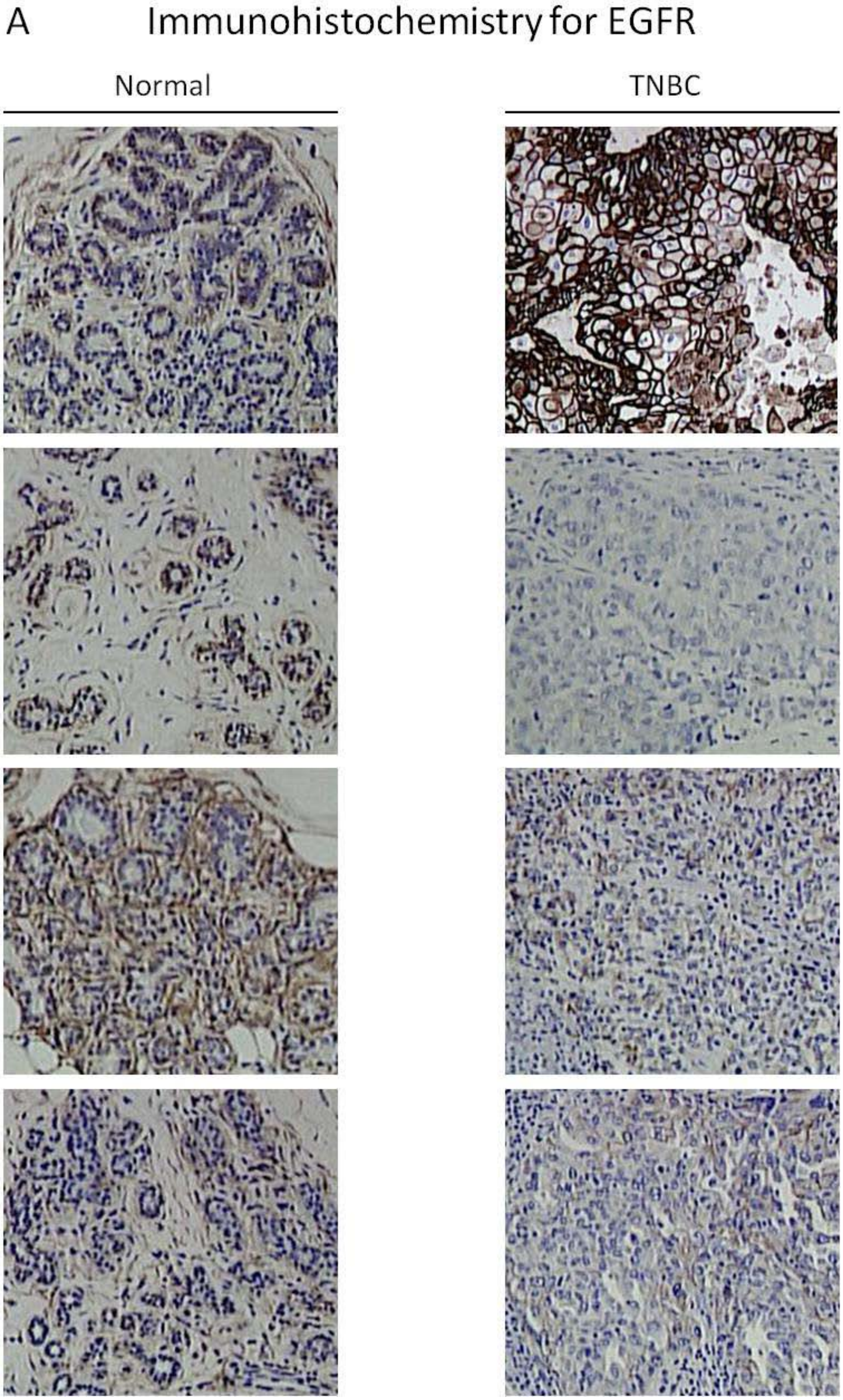


Figure 24. Immunohistochemistry (IHC) of EGFR and KIT on normal and TNBC breast tissues. Clinically used IHC stains were used on 20 normal and 11 TNBC breast tissues and then scored by an experienced breast pathologist. Each tissue was given a staining intensity (0+,1+,2+,3+) and a percent cell staining (0-100%). The H-score is calculated by multiplying the intensity with the percent cell staining, giving a possible range of 0-300. This data demonstrates that at the protein level, there is no statistical difference in EGFR between normal and TNBC, and a significant downregulation of KIT in TNBC compared to normal.


A
Immunohistochemistry for EGFR
Normal
TNBC

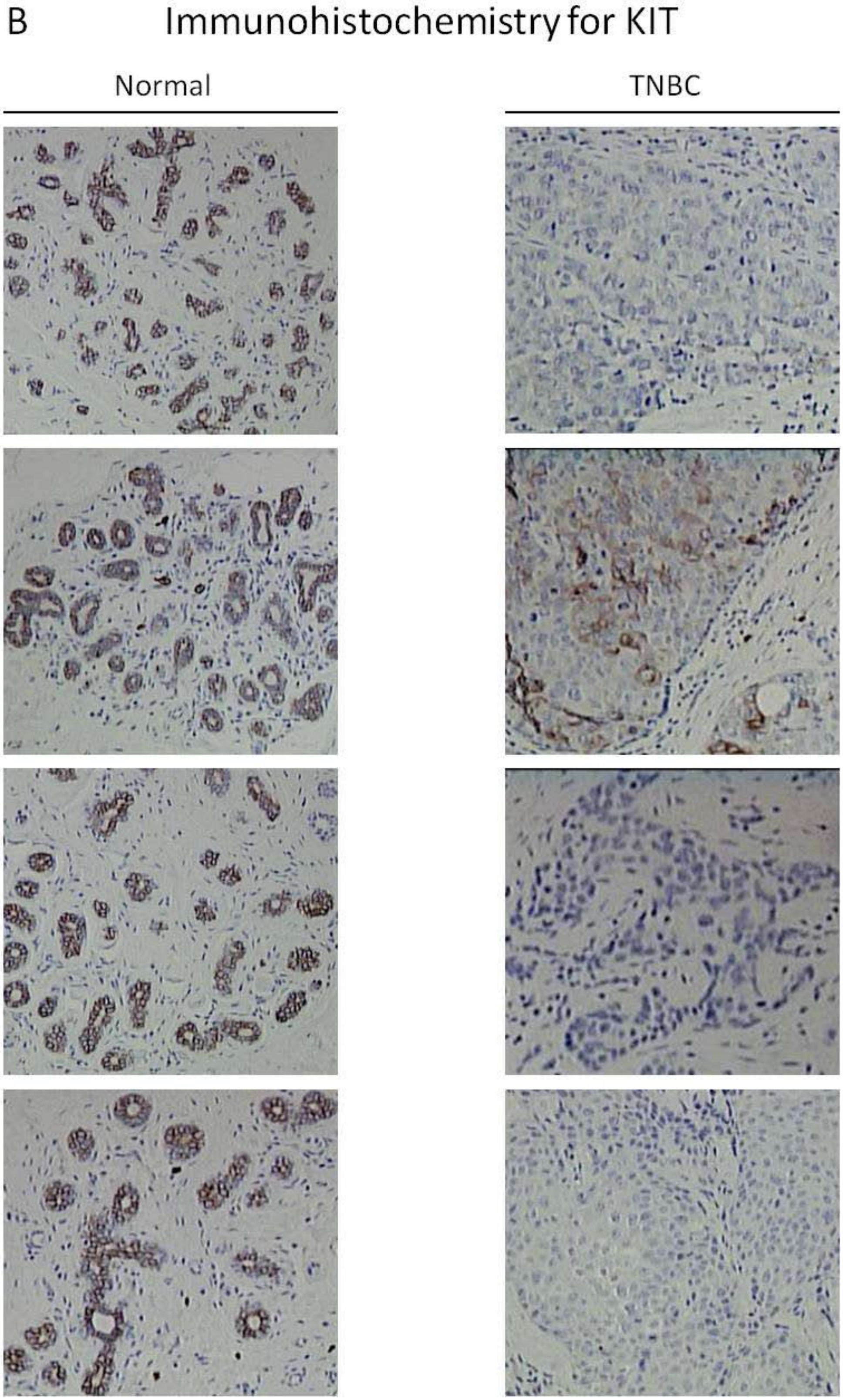

B
Immunohistochemistry for KIT
Normal
TNBC

Figure 25. Representative IHC stains for EGFR and KIT of normal and TNBC tissues. **(A)** IHC of EGFR shows variable expression across normal tissues. Appreciable staining can be seen in the myoepithelial cells. IHC of TNBCs also shows variable expression. The top panel demonstrates a TNBC with very strong staining, where the second panel shows no staining, while there is weak staining in the third and fourth panels. This pattern is consistent with entire the entire IHC sample set, as well as the RNA gene expression datasets. **(B)** IHC of KIT shows strong staining in normals that very nicely stain the ducts. In stark contrast, staining is either very weak or non-existent in the TNBC samples. This IHC data is also consistent with the RNA gene expression datasets.

significance of differentially expressed genes. The volcano plot identified a set of highly upregulated RTKs including PTK7, TIE1, CSF1R, EPHB4, and EPHB6.

Our top hit, PTK7 (protein tyrosine kinase 7), was first discovered as CCK-4 (colon carcinoma kinase 4) as a protein that was expressed in colon carcinoma but not in normal colon tissue [95]. Interestingly, it has been previously observed using gene expression microarrays to be upregulated in ER-negative cancers, and siRNA knockdown in TNBC cell lines causes inhibition of proliferation which does not occur in ER-positive cell lines [96]. In addition to this single paper in breast cancer, siRNA inhibition of PTK7 in HCT116 colon carcinoma cells resulted in inhibition of cell proliferation and induction of apoptosis [97]. Work in acute myleogenous leukemia (AML) has shown that PTK7 promotes anthracyline resistance and PTK7-positive AML patients have a decreased disease free survival [98]. PTK7 also plays a major role in development as a key regulator of planar cell polarity in epithelial tissues [99]. Mouse embryos that contained a mutation of PTK7 resulted in the failure of neural tube closure and defects in heart, lung, and ear development [100]. Because of PTK7's roles in embryonic development, we were concerned that its overexpression maybe an artifact of TNBCs deriving from normal breast luminal progenitor cells. As mentioned in Section 1.1.2, recent data has demonstrated that BRCA1 basal-like breast cancers are most likely derived from breast luminal progenitor cells [36, 37]. To assess this, we used publicly available data from Lim, et al.[36] and determined whether PTK7 had any differential expression between luminal progenitors and mature luminal cells. We found no difference in expression of PTK7 between the two cell types ($p=0.34$). To further validate the importance of PTK7, we wanted to determine if PTK7 was overexpressed in TNBC when compared to other subtypes of breast cancer. Using recent publically available data from the Perou lab [101], PTK7 was indeed significantly overexpressed in basal breast cancer compared to the other intrinsic subtypes (2.5-fold, $FDR<0.001$).

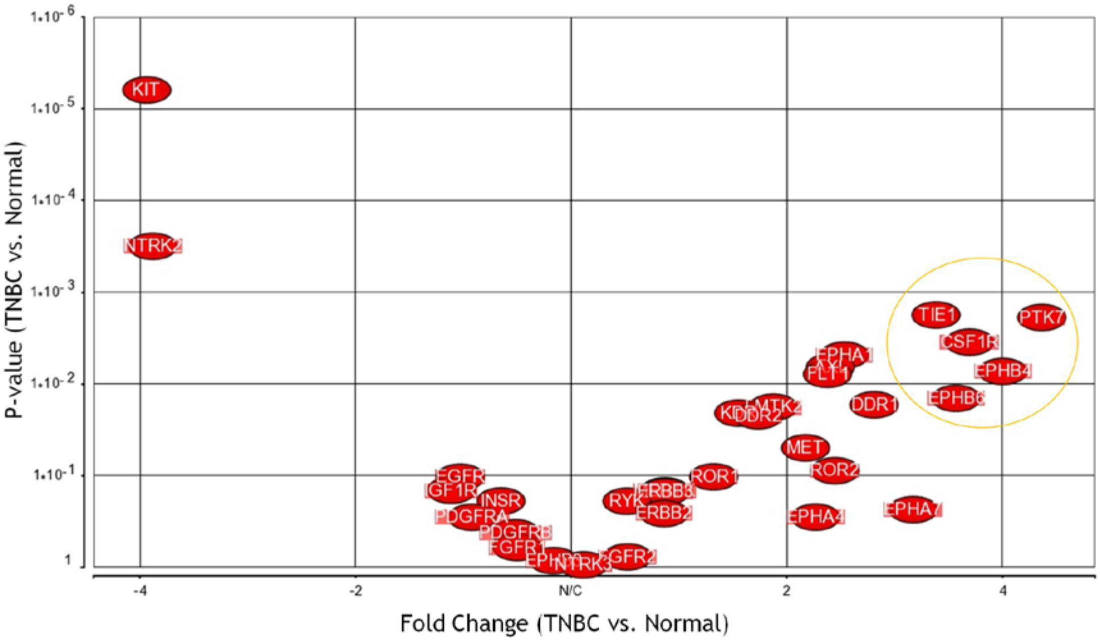


Figure 26. Volcano plot of significant receptor tyrosine kinases (RTKs). RTKs are critical proteins for the initiation of signaling cascades leading to cell proliferation and survival. RTKs hold the potential to be potent therapeutic targets. This volcano plot identifies significant RTKs by considering both fold change and statistical significance of differentially expressed genes. In our dataset, 32 RTKs are expressed in these tissues. This analysis reveals PTK7, TIE1, CSF1R, EPHB4, & EPHB6 as significant RTKs overexpressed in TNBC. PTK7 is extensively explained in the text. In addition, TIE1 is an RTK well known for its role in angiogenesis and expression in endothelial cells, but it has also been demonstrated to be overexpressed in breast cancers and not in corresponding normal breast tissues [102]. CSF1R has also been previously demonstrated to be expressed in breast cancer [103], and is inhibited by Sunitinib [104], which has potentially increased activity in TNBC patients [16]. The Ephrin B receptors, EPHB4 & EPHB6, are both known to be expressed in breast cancer, with EPHB4 important for breast cancer cell survival [105], and EPHB6 has interestingly been shown to decrease invasiveness in TNBC cell lines [106].

PTK7 was also overexpressed in HER2+ positive breast cancers (2.4-fold, FDR<0.001). An interesting aspect of PTK7 is that there is no known ligand of PTK7 and its kinase domain is catalytically inactive [97], yet there is no doubt that PTK7 is actively involved in many important cellular processes. This is very analogous to HER3 which lacks a catalytically active domain but partners with HER2 to transduce signaling. All these data combined reflects a strong rationale for targeting PTK7 in TNBC.

We also then performed a similar analysis looking at all kinases. The most statistically significant (and most overexpressed) kinase was NEK2 (NIMA-related kinase 2). NEK2 is a serine/threonine protein kinase involved in centrosome separation during mitotic entry. The expression of NEK2 is 34-fold overexpressed in TNBC vs. normal and is highly statistically significant ($p = 1.25x10^{-6}$). Previous studies have shown that siRNA knockdown of NEK2 in breast cancer cell lines inhibits cell growth, colony formation, and *in*-vitro invasiveness [107]. In a breast cancer cell line, overexpression of NEK2 resulted in chromosomal instability with aneuploidy, multinucleated cells and multiple centrosomes [108]. Like PTK7, NEK2 was not differentially expressed between luminal progenitors and mature luminal cells (p=0.13). NEK2 was also significantly overexpressed in basal-like breast cancer compared to other subtypes (5-fold, p<0.001). Its protein network as elucidated by Ingenuity Pathway Analysis is depicted in Figure 27. Review of the network analysis reveals involvement in several pathways including: centrosome duplication, cell division, MAPK pathway, among others. Because of its involvement and pervasive role in tumor biology, it has been proposed as a potential drug target in cancer [109, 110]. Based on current evidence, we also agree that NEK2 is a potential therapeutic target for TNBC.

In a further in-depth search for novel therapeutic targets, we sought to take advantage of our group's expertise in computational *in-silico* drug discovery of enzymatic proteins. To perform this search we undertook a comprehensive approach to target

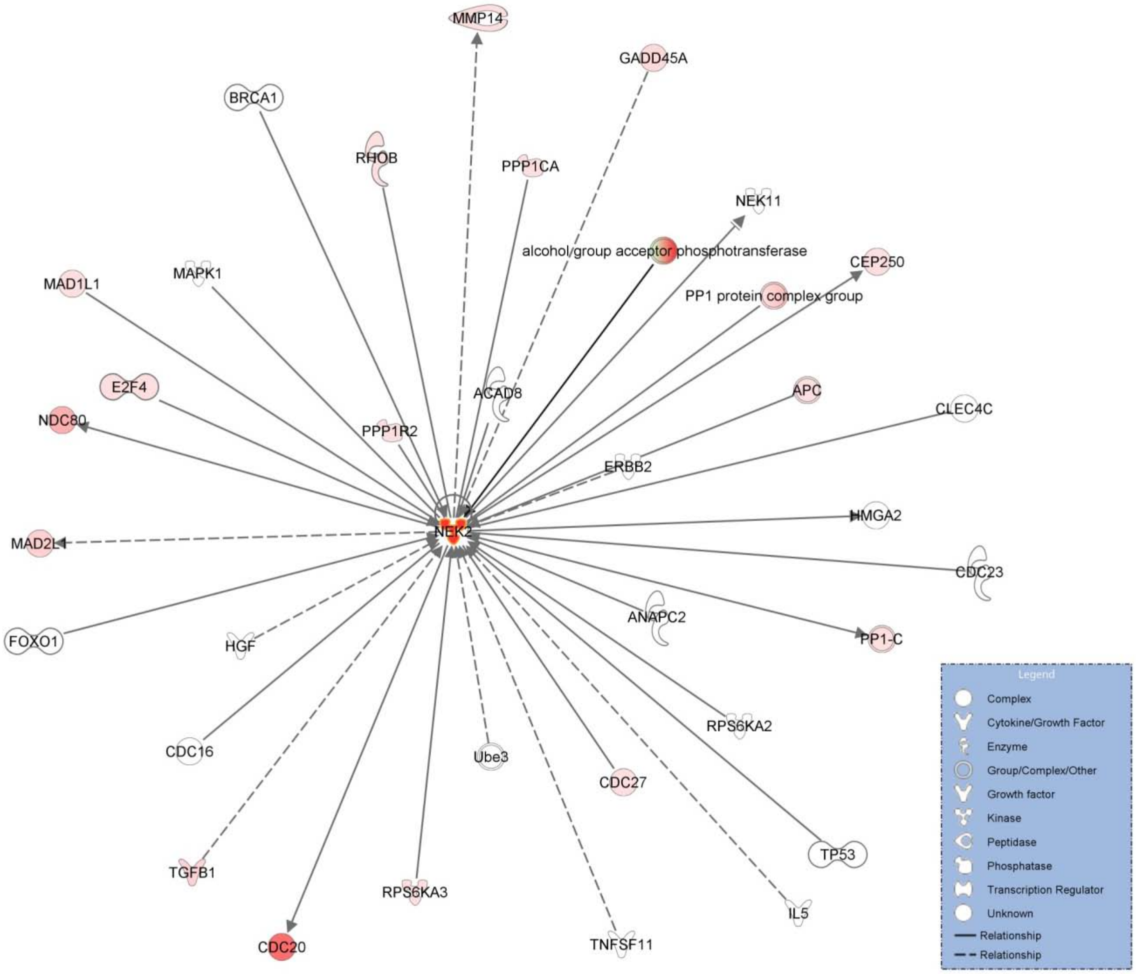


Figure 27. The NEK2 interaction network. This network was elucidated by Ingenuity Pathway analysis. The red color indicated up-regulation in our next-generation sequencing dataset and those proteins in white indicate that they were not differentially expressed in our RNA-seq dataset.

identification which includes: differential RNA-seq expression data, pathway analysis, information from the Protein Data Bank, and existing literature. With this approach, we have identified 25 high-priority targets to further characterize for therapeutic development. Of importance, none of these targets have been pursued in previous clinical trials in cancer. We first started with 7,140 genes that the RNA-seq identified as being significantly differentially expressed between TNBC and normal (FDR<0.01). From these, 1,694 genes were upregulated and contained a 3D protein structure in the Protein Data Bank that is suitable for *in-silico* molecular docking. Of the 1,694, we narrowed our focus on enzymes, as these proteins have the best likelihood of success to find a small molecule inhibitor with *in-silico* molecular docking. This is because an enzyme with a 3D structure has a defined active site, and many cases additional allosteric binding sites that can assist with making our compounds more selective. By focusing on enzymes, this narrowed our list to 687 proteins. Of these, 60 already had prior inhibitors that have been developed and tested in clinical trial, and were thus excluded to avoid a replicative effort. This left a list of 626 potential targets. We then analyzed the gene expression data for the 626 targets, and set an exclusion criteria, where the expression of the target has to be low in the normal samples (defined as average RPKM<2) and whose expression is considerably higher in the TNBC samples (defined by statistical comparison). We believe this filter enables the target to be highly specific for TNBC. With this filter, the list was narrowed to 268 potential targets. We then rank ordered the list by statistical p-value and fold change, and performed pathway analysis and literature searches to determine the function and significance of the target in tumor biology. In this manual search, we excluded genes that are well known to be pro-apoptotic, and we also excluded targets that were already the focus of extensive preclinical efforts, though not yet tested in clinical trial. From this effort we identified the 25 targets listed in Table 7. This list represent a diverse set of enzymatic targets with nuclear, cytoplasmic, and

Gene Symbol	Entrez Gene Name	Fold Change	p-value	Location	Type(s)
ADPRHL2	ADP-ribosylhydrolase like 2	3.48	3.36E-09	Cytoplasm	enzyme
BCAT1	branched chain amino-acid transaminase 1, cytosolic	8.76	3.86E-05	Cytoplasm	enzyme
CKS2	CDC28 protein kinase regulatory subunit 2	10.94	9.59E-05	unknown	kinase
CLPP	ClpP caseinolytic peptidase	3.52	1.59E-05	Cytoplasm	peptidase
CSNK1G2	casein kinase 1, gamma 2	4.30	3.62E-10	Cytoplasm	kinase
CTPS	CTP synthase	5.98	5.37E-06	Nucleus	enzyme
DPP3	dipeptidyl-peptidase 3	7.48	1.18E-07	Cytoplasm	peptidase
DTYMK	deoxythymidylate kinase (thymidylate kinase)	5.86	6.39E-06	unknown	kinase
ECE2	endothelin converting enzyme 2	6.77	8.27E-05	Plasma Membrane	peptidase
GSG2	germ cell associated 2 (haspin)	9.27	6.72E-06	Nucleus	kinase
HAGH	hydroxyacylglutathione hydrolase	4.73	1.01E-05	Cytoplasm	enzyme
HCK	hemopoietic cell kinase	7.10	3.17E-04	Cytoplasm	kinase
MICAL1	microtubule associated monoxygenase, calponin and LIM domain	5.72	4.29E-07	Cytoplasm	enzyme
NEK2	NIMA (never in mitosis gene a)-related kinase 2	34.54	1.25E-06	Nucleus	kinase
PACSIN1	protein kinase C and casein kinase substrate in neurons 1	13.34	4.16E-04	Cytoplasm	kinase
PASK	PAS domain containing serine/threonine kinase	3.68	9.03E-06	Cytoplasm	kinase
PLK4	polo-like kinase 4 (Drosophila)	4.35	1.31E-05	Cytoplasm	kinase
PRPS1	phosphoribosyl pyrophosphate synthetase 1	4.32	2.89E-08	unknown	kinase
RECQL4	RecQ protein-like 4	13.29	3.01E-05	Nucleus	enzyme
RND1	Rho family GTPase 1	7.90	8.52E-04	Cytoplasm	enzyme
SEPX1	selenoprotein X, 1	5.51	3.84E-05	unknown	enzyme
TK1	thymidine kinase 1, soluble	13.59	2.10E-05	Cytoplasm	kinase
TYRO3	TYRO3 protein tyrosine kinase	3.09	4.66E-06	Plasma Membrane	kinase
UBE2C	ubiquitin-conjugating enzyme E2C	16.11	5.58E-08	Cytoplasm	enzyme
UPP1	uridine phosphorylase 1	5.85	5.61E-07	Cytoplasm	enzyme

Table 7. 25 TNBC therapeutic targets identified by RNA-Seq.These targets were chosen from analysis of next-generation RNA seq data, pathway analysis, information from the Protein Data Bank, and literature searching. In the case of all targets, the expression of the gene is low in normal samples, and is considerably higher in TNBC samples helping to enforce TNBC specific targets.

plasma membrane co-localization. The targets include genes involved in DNA replication, cell cycle, cytoskeleton regulation, and cell survival. These targets are now being actively pursued by our group using siRNA knockdown to see their phenotypic effect on TNBC and normal cell lines. Knockdown of targets that show an anti-tumor response are then pursued using *in-silico* molecular docking to identify small molecular inhibitors that will be used for further *in-vitro* testing.

2.3.3 Profiling and differential expression of noncoding RNAs

A significant advantage of RNA-seq is the ability to profile a diverse species of RNA including non-coding RNAs. We interrogated reads mapping to known precursor miRNAs (pre-miRNAs), lincRNAs, and ultra conserved regions (UCRs). While miRNAs are well known as master RNA regulators, lincRNAs and UCRs have recently joined the cadre of non-coding RNAs species that elicit regulatory function [111, 112]. Statistical comparison of 705 pre-miRNAs revealed 22 differentially expressed pre-miRNAs at a FDR < 0.01 (Appendix 2). Recently, mir-146a and mir-146b have been shown to downregulate the expression of BRCA1 [113]. In our dataset, miR-146a was upregulated in TNBCs 2.6-fold compared to normal (p=0.046). In addition, we observed several differentially expressed microRNAs involved in angiogenesis. Specifically, miR-93 and -210 (both upregulated) are known to be inducible by hypoxia [114]; miR-31 (upregulated) is a HIF-1a inducer [115]; and miR-205 (downregulated), is an inhibitor of VEGF [116]. We also observed significant upregulation of miR-221&-222 which are known to negatively regulate ER expression in breast cancer cell lines [117]. Finally, known oncogenic miRNAs including miR-16, miR-21, and miR-31 were also upregulated (Appendix 2). Several of our most significant pre-miRNAs have no prior literature in cancer. This underscores the enigmatic role of pre-miRNAs in TNBC biological processes.

LincRNAs, such as Xist and HOTAIR, have been recently reported to play a major role in chromatin regulation and gene expression [85, 111]. Khalil et. al. has recently annotated the exons of lincRNAs in 6 human cell lines [85]. Using their annotation of 4860 lincRNA exons, we sought to determine if lincRNA expression is also dysregulated in TNBC. In our analysis, we report 109 lincRNA exons that are differentially expressed at an FDR < 0.01 (Appendix 2). An interesting first-pass observation of the data is the generalized down-regulation of statistically significant lincRNAs. 75/109 of the top lincRNAs are downregulated. The most significantly downregualted lincRNA is XIST located at chrX:72,957,220-72,989,313 (Appendix 2). We also analyzed non-coding RNAs to identify expression differences in UCRs. UCRs are areas of the genome at a minimum of 200bp where there is 100% conservation of the sequence between human, rat, and mouse. These areas have been found to have functional significance in cancer [112]. In our analysis, we searched 481 known UCRs and found 15 to be differentially expressed at FDR < 0.01 (Appendix 2). One interesting UCR, uc.63, has been recently shown to be induced by hypoxia (personal communication, M. Ivan). The functions elucidated by the differentially expressed lincRNAs and UCRs is largely unknown, but these data suggest a possible role in TNBC biology. Future investigation will determine their contributions to TNBC.

2.3.4 Novel transcribed regions

Previous literature has reported the presence of pervasive transcription across the genome, particularly in intergenic regions [118]. In an attempt to understand the presence and possible function of novel transcribed regions (NTR) in TNBC, we analyzed the entire genome for areas of significant expression for which there is no known annotated gene. We chose areas of novel transcription with the following criteria: 1) the NTR had to be a minimum of 10,000bp away from the nearest RefSeq gene to reduce the possibility of detection of a novel exon to a known gene, 2) the NTR must be at least 50bp long, 3) there must be at least 20 supporting reads mapping to the NTR. We then further filtered out loci that overlapped known miRNAs, lincRNAs, UCRs, snoRNAs, snRNAs, scRNAs, rRNAs, tRNAs, and mtRNAs. Regions that satisfied these criteria were then analyzed to determine start and end coordinates of each novel transcript in order to determine RPKM values for each NTR. This method of determining region boundaries is not the optimal method (compared to traditional cloning methods), but provides an initial screen for differential expression. In total we report 43,351 NTRs (Appendix 2). Using 1-way ANOVA, we then compared the gene expression values of the 10 TNBC tumors and 10 normals and identified 6,408 differentially expressed NTRs between TNBC and normal at an FDR < 0.01 (Appendix 2). To further support the presence of the NTRs, we cross-referenced the 43,351 NTRs against the AceView database [119], which is a comprehensive, non-redundant database of mRNA sequences including EST libraries. 17,082 NTRs overlapped AceView sequences lending biological evidence to these NTRs (Appendix 2). Of further note, PCA of NTRs revealed a separation of TNBC and normal samples indicating that these regions alone can define the disease phenotype (Figure 28). The function of these NTRs is still to be determined, but the data suggests a large undiscovered territory of transcribed RNAs with no known function that may be implicated in TNBC.

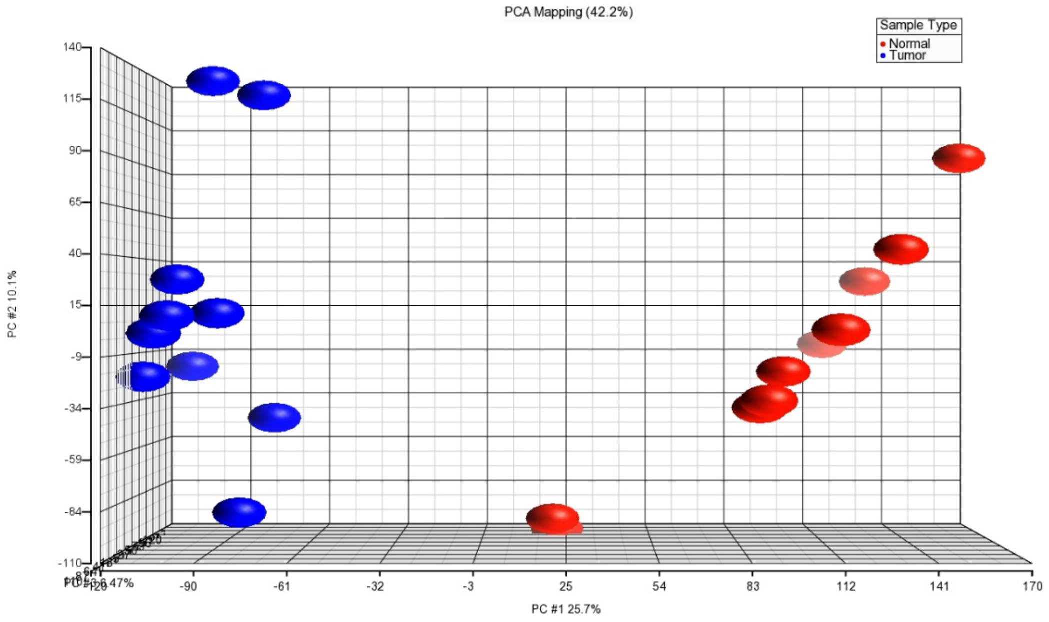


Figure 28. PCA of Novel Transcribed Regions. PCA of log2 transformed RPKM data shows separation of TNBC and normal sample phenotypes using novel transcribed regions. The separation is quite stark, in particular, when considering that this separation is independent of all known coding and non-coding genes.

2.**4 Discussion**

In this chapter, we have described the application of next-generation whole transcriptome sequencing in TNBC and normal breast to study differential gene expression. A critical strength of this study includes the use of microdissected ductal epithelium from normal breast tissue as a comparator. RNA-seq is a powerful technology whose strength in accuracy over traditional microarrays, and the ability to survey the entire transcriptome with no *a* priori gene selection suits it as the optimal platform for this study.

In our analysis, over 7200 known genes were found to be differentially expressed between TNBC and normal breast tissue. Two major themes were illustrated by this data. The first major theme revolves around differentially expressed genes associated with BRCA or its pathway. Many of these genes were highly significant and pervasive through several of our analyses. The most striking illustration was the observation that our most statistically significant gene was COBRA1 (cofactor of BRCA1), a gene that plays a dual role in both inhibiting estrogen receptor activity, and modulating many genes that are also regulated by BRCA1 [86, 88]. This gene is quite fitting to have the top spot with TNBCs being negative for estrogen receptor coupled with the well known link between BRCA1 mutations and development of TNBC. We also observed several genes in the BRCA DNA repair pathway that were upregulated. Inhibitors of several of these genes including CHK1, CHK2, and PLK1 are all in clinical trial. This suggests that the targeted inhibition of DNA repair proteins and the “synthetic lethal” approach led by Ashworth and colleagues may not be limited to PARP only. Upcoming clinical trials will soon tell whether targeted agents against DNA repair proteins, particularly when administered with DNA damaging agents, are an effective means of treating TNBC. Further we observed upregulation of a microRNA that has been very recently implicated in downregulating BRCA1 [113]. While previous work has shown that mutations, loss of heterozygosity, and methylation can all lead to downregulation of BRCA1 in TNBC, noncoding RNAs are now new players. The role of noncoding RNAs in BRCA1 regulation is still largely unknown, but it could be speculated that other microRNAs (either direct or indirect) or other forms of non-coding regulation including lincRNAs could be implicated.

The second major theme demonstrated that using normal ductal epithelium from healthy volunteers is an optimal control for discovering therapeutic targets in TNBC. This was most strikingly illustrated by the fact that some genes previously reported to be over-expressed in TNBC by microarray (e.g. EGFR and c-kit) were not upregulated in this study [29, 120]. The lack of transcriptional upregulation (compared with normal breast) might explain the disappointing outcomes to several clinical trials implementing agents designed to target these pathways [11-14]. In contradistinction, the only positive randomized clinical trial to date testing a targeted agent in an enriched triple negative population used Iniparib (BSI-201), a PARP inhibitor [25]. The target for BSI-201 (PARP1/2) is 3-fold over expressed in TNBC compared to normal in our study (Figure 23). Further, we identified additional targets that will serve as the basis for future drug discovery work. This included PTK7, our most significant receptor tyrosine kinase, a developmental gene whose knockdown by siRNAs has previously demonstrated inhibition of proliferation only in ER-negative cells and not ER-positive [96]. Another target, NEK2, our most differentially expressed kinase, has also had extensive preclinical work to demonstrate that inhibition has a significant anti-tumor effect. Some very recent work by our group has developed small molecule inhibitors of NEK2 that when tested *in-vitro* inhibited the proliferation of cells at the micromolar level. These data are quite preliminary, but it does begin to demonstrate the ability of comparing TNBC to normal breast to identify actionable targets. All together, these data imply that developing drugs against targets that are actually differentially expressed when compared to “true normal” breast tissue is of high value.

In addition to the two major themes, the power of next-generation sequencing is highly leveraged in this study through the profiling of non-coding RNAs and novel transcribed regions. These areas are of special interest as they have the potential to elucidate previously undiscovered loci important in tumorigenesis, and may ultimately provide insight into novel therapeutic targets. Indeed several microRNAs were detected that upregulate angiogenesis, and anti-angiogenic therapy seems to be moderately effective in TNBC patients. A large limitation to noncoding RNA analyses is the limited knowledge of the function of many of these noncoding RNAs, especially lincRNAs and UCRs. While we are able to detect them, this limited knowledge makes it difficult to associate lincRNAs and UCRs with the functions of individual coding genes. In the original paper that described lincRNAs, the group used a “guilt by association” method which involved statistical correlation of the expression of lincRNAs with protein coding genes followed by gene set enrichment analysis and hierarchical clustering in order to begin deciphering a putative function [111]. This method identified that lincRNAs have a role in modulating many cancer associated gene sets including: cell cycle regulation, proliferation, and chromatin remodelling complexes. Unfortunately this method does not provide evidence for how these lincRNAs work directly with cancer genes, and thus remains solely as a means to understand this RNA species globally.

In conclusion this chapter presents a comprehensive and novel characterization of the differential expression of a lethal disease with no FDA-approved targeted therapies using cutting edge next-generation sequencing technology. Through the use of tissue controls from healthy pre-menopausal women we describe the landscape of transcriptional perturbations that comprise TNBC. The differential biology outlined here may now be used as a framework for the future development of targeted therapies.

Chapter 3: Detection of mutations in TNBC from mapped RNA-seq data

3.1 Introduction

One of the most powerful aspects of next-generation RNA sequencing outside of gene expression is the ability to interrogate the sequence itself. This ability is one of the significant advantages of RNA-seq over traditional gene expression microarrays. Some recent examples of the successful application of calling mutations from RNA-seq have come from ovarian cancer. Two articles published in the *New England Journal of Medicine*, both used RNA-seq to call highly recurrent mutations in the FOXL2 gene and the ARID1A gene in Granulosa-Cell and Endometriosis-Associated ovarian carcinomas, respectively [71, 72]. Another publication from the same group also used RNA-seq to identify mutational differences from a metastatsis of a lobular breast cancer and compared these mutations to the primary cancer that occurred 9 years earlier [121].

As mentioned in Section 1.3.1, there is still a paucity of knowledge in regards to the mutational profile of TNBC. In regards to inherited mutations, it is well known that BRCA1 mutations predispose women to developing TNBC, but these mutations represent only a minority of all TNBC cases [41]. Also, because not all BRCA1 mutation carriers develop breast cancer, it has led to speculation of potential modifier genes that increase the risk of breast cancer of breast cancer for BRCA1 carriers. A recent study tested this hypothesis by performing a Genome Wide Association Study (GWAS) of BRCA1 carriers with breast cancer compared to carriers with no breast cancer diagnosis [122]. This study identified five SNPs on chr19p13 with two of them increasing risk and three of them decreasing risk for development of breast cancer. When the same five SNPs were applied to a cohort of triple-negative breast cancer and controls, the same SNPs correlated with increased or decreased risk. While these findings were highly statistically significant, the highest odds ratio for increased risk was 1.28, and the lowest odds ratio for decreased risk was 0.79, demonstrating the very modest effect of these SNPs as genetic modifiers [122]. Epidemiological data also supports the premise that inherited variation can play a role in TNBC. The observation that TNBC has an earlier age of onset coupled with its greater frequency in women of African descent are classic clues for the potential presence of other germline mutations. Technologies that can survey the entire genome, like next-generation sequencing, will be able to tease out novel germline mutations in sporadic TNBC if they are present.

In regards to somatic mutations, p53 has dominated the mutational landscape [45], with mutations in Rb and ERBB4 also reported to a lesser extent [46-48]. Also, a recent study looking at somatic mutations in a variety of cancers, also identified an abundance of TP53 mutations in TNBC (19/53), but also somatic mutations in PIK3CA (9/53) [123]. Outside of these genes, somatic mutations in TNBC tend to be patient specific [45, 49, 123]. This is best illustrated in the context of gene fusions, where low-coverage next-generation sequencing of DNA illustrated the presence of gene fusions in TNBC, but none were recurrent [49]. Whether gene fusions in TNBC are involved in causation, or are merely by-products of deficient DNA repair coupled with highly mitotic cells is still to be determined.

With the need for a complete understanding of the genomic perturbations that cause cancer, several large consortia are now actively working to identify the mutational landscape of all common malignancies. Without a doubt, breast cancer (and specifically TNBC) is a major target for all of them. These groups include the ICGC (International Cancer Genome Consortium), TCGA (The Cancer Genome Atlas), and METABRIC (Molecular Taxonomy of Breast Cancer International Consortium). These groups combined will sequence the genomes and transcriptomes of thousands of breast cancers and matched normal tissues, including many TNBCs. But as with all large consortia, some pitfalls exist. As of April 2011, neither the ICGC nor the TCGA have released any breast cancer sequencing data nor do their databases indicate when it will occur. METABRIC does not maintain an online accessible database. In addition to logistics, another major pitfall of the consortia is that the designs of these studies are primarily driven to identify somatic mutations and do not include germline. The general workflow is to sequence both the tumor and matching normal, identify all genetic variation in both, and then subtract the variation found in the normal from the tumor. This in essence subtracts any potential germline causative variants. Whether these groups will eventually compare cancer genomes to a large group of normal genomes, such as the 1000 genomes project, is unknown. Also, as already illustrated in the previous chapter, the use of matched normal tissue (usually normal adjacent), for RNA-seq by the large consortia most likely will be fraught with gene expression changes that do not reflect the use of a truly normal control. While problems do exist with the major consortia, it is well accepted that a comprehensive compendia of the mutations that cause cancer will lead to a better approach to treating disease.

With the lack of a comprehensive database of mutation data in TNBC, along with the previously mentioned success of determining mutations from RNA-seq data in ovarian and lobular breast cancers, we sought to apply the same approach to our data of TNBC and normal. Using newer algorithms and parsing pipelines developed in-house, we sought to identify point mutations, small insertions/deletions, and gene fusions that could give us clues to the causation of this disease.

3.2 Materials and methods

Bioinformatic detection and validation of point mutations

In order to identify point mutations from mapped RNA-seq data, we took a multistep approach that allowed for mutations to be called with high statistical confidence. We then developed a pipeline that also allowed us to remove known common genetic variation. (Extensive details are provided in Appendix 1). Starting with the mapped RNA-seq data, mutations were called using the SNVMix2 algorithm which is designed to call mutations specifically from RNA-seq data [79]. As illustrated in Section 1.3.3, there is particular difficulty of calling mutations from RNA-seq data as the ability to call mutations at any given region of the genome is proportional to the RNA expression of that region. This means that algorithms that assume a uniform coverage, which is commonly used for DNA sequencing studies, cannot be used. SNVMix2 accounts for this non-uniform coverage by using a probabilistic model which accounts for both the number of reads that cover a variant and the base and mapping qualities of the reads that cover the variant. While SNVMix2 is a powerful algorithm that has been used successfully in several cancer RNA-seq studies, our first use of the algorithm with default parameters yielded a slew of false positives. This was evidenced by the fact that most of the variants were not in the NCBI dbSNP database as would be expected. After many iterations, we finally tuned the algorithm such that the majority of the single nucleotide variants called were true positives as evidence by the majority of the hits being in the dbSNP database. This included setting quality thresholds for the base and mapping qualities to a PHRED score of 20, and also requiring the algorithm to consider heterozygous and homozygous evidence for a variant separately versus in a combined fashion. Also setting the statistical probability threshold to 90% for any given heterozygous or homozygous call helped reduce the false positive rate tremendously.

The mutations for each sample (10 TNBCs and 10 normals) were then called and independently run through a custom made pipeline to discover novel mutations in coding regions (Figure 29). To quickly annotate each variant, we used the Annovar software which allowed for a quick identification of whether a variant was in a coding region, and if the variant was synonymous, nonsynonymous, stop gain, or stop loss [124]. The synonymous variation was removed, and the rest of the variants were then subjected to several rounds of filtering to remove known genetic variation. This included filtering out variation identified through the 1000 genomes projects, specifically data from the Caucasian, Yoruban, Japanese, and Chinese populations. Variants that remained were then filtered for variants present in dbSNP (version 130). After all the variants were called and filtered for the 10 TNBCs and 10 Normals, the data was collated into a single master file of mutations. This master file was then parsed to look for mutations that were recurrent at the exact same base position, and to look for genes that were recurrently mutated across samples but not the same base position. Mutated genes were also put into Ingenuity Pathway Analysis in order to discover pathways that are mutated as a third level of processing above recurrent base mutations and cross-gene mutations.

Mutations of interest were then individually assessed by manually inspecting each read that covered a mutation using the Integrative Genomics Viewer [78]. As the PCR during library preparation can introduce mutations, a visual inspection can catch these errors by observing reads where the mutation occurs on reads that contain the same exact start site (known in the field as PCR duplicates). Those mutations that were called solely on evidence from PCR duplicates were not considered. In essence, the read support for any given mutation had to have two unique start sites in order to be considered. Once the visual inspection was complete, validation of mutations were carried out by designing PCR primers that flanked the mutation. PCR reactions were setup consisting of Amplitaq Gold 360 Master Mix (Applied Biosystems), 200nM of each

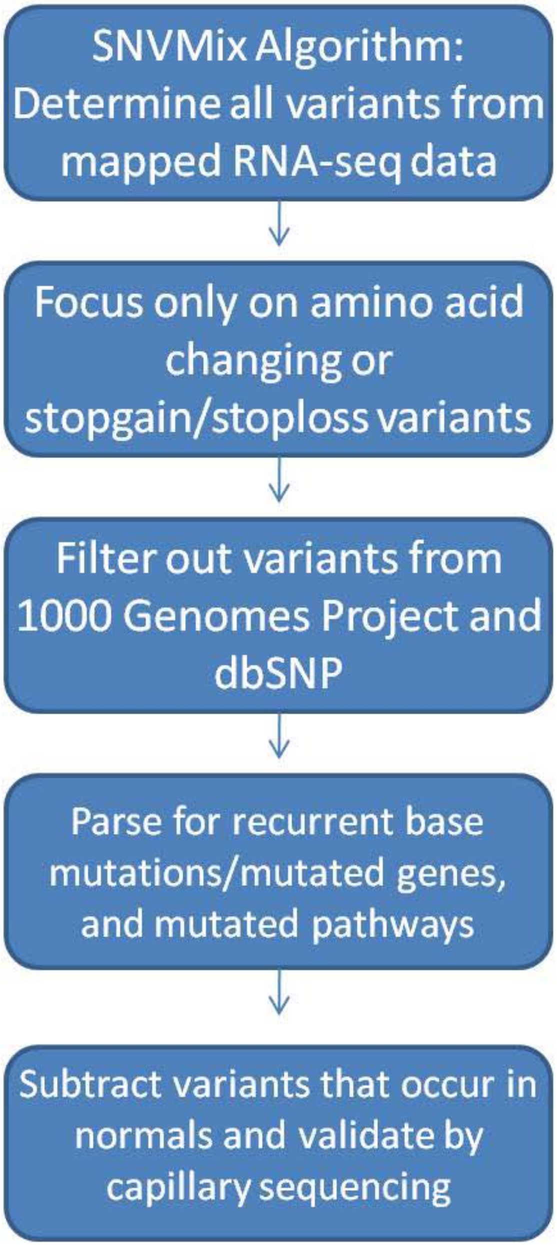

SNVMix Algorithm: Determine all variants from mapped RNA-seq data
Focus only on amino acid changing or stopgain/stoploss variants
Filter out variants from 1000 Genomes Project and dbSNP
Parse for recurrent base mutations/mutated genes, and mutated pathways
Subtract variants that occur in normals and validate by capillary sequencing

Figure 29. Flow diagram of point mutation calling and filtering pipeline. The pipeline begins with mapped RNA-seq data followed by calling of variants with SNVMix2, followed by annotating with Annovar, and then filtering for variants that occur in the 1000 genomes project, dbSNP, and in the normal samples. Variants are then searched for recurrence at the same position, in the same gene, or in the same pathway. Mutations are then subjected to manual inspection and then validated by PCR and sequencing.

primer (IDT, Coralville, IA), and 5ng of DNA (or cDNA). A gradient PCR was performed followed by agarose gel electrophoresis in order to identify the optimal annealing temperature that creates a single band. The optimized PCR product was then sequenced using capillary sequencing at the Indiana University DNA Sequencing Core Facility. Analysis of the capillary sequencing either validated or invalidated the presence of the mutation.

Bioinformatic detection of small insertions and deletions

To detect small insertions and deletions (indels) requires special processing of the mapped RNA-seq data. As described in Section 1.3.3, mapped reads have to be re-aligned in order to do a gapped alignment which will tolerate the presence of the small indels. In order to do this, we used a custom modified pipeline of the ABI BioScope software. This is not a standardized pipeline for the software and requires different input files and commands to run. Very briefly and simply, intermediate files that corresponded to the raw mapping of the reads to the genome along with the quality files and other inputs were locally re-aligned to the genome using a gapped alignment feature in BioScope. The local realignment allowed for insertions up to 3bp and deletions up to 11bp to be detected. After alignment, the output mapping files were converted to the standard BAM format. Small indels were then called from the BAM file and outputted into a highly annotated text file containing all the information available for the support of an indel (Appendix 1). Small indels were then annotated using the same pipeline used for point mutations except the software annotates the indel as either frameshift insertion, frameshift deletion, nonframeshift insertion, nonframeshift deletion, stopgain, or stoploss. The indels were then parsed for recurrent mutations at the same base positions, recurrent mutations occurring in the same gene but not the same positions, and across a pathway.

Bioinformatic detection and validation of gene fusions

In order to determine the presence of gene fusions from 50bp fragment reads, we searched for reads spanning exons from two different genes. Only reads that partially mapped to the human genome in the original read mapping were used for gene fusion discovery. To reduce the search space for gene fusion junctions, only the ends of known exons derived from the RefSeq database were considered. Using the SASR Junction Finder (a pipeline of ABI BioScope v1.2.1), we identified putative fusion junctions that had at least 2 supporting reads with 2 unique starting points (2 unique reads). To filter for false positives, any candidate fusion must only appear in the TNBC samples and not in the normals. We further filtered the output by removing any gene & exon involved in more than 2 fusions as these are most likely false positives. In our final list, we considered only fusions that had at least 3 supporting reads (with at least 2 unique starting points).

Primers were designed against the exons that were bioinformatically deduced to be involved in a fusion junction. The RNA from the 10 sequenced TNBC samples was reverse transcribed using the High Capacity RNA-to-cDNA kit (Applied Biosystems). PCR reactions were then setup consisting of Amplitaq Gold 360 Master Mix (Applied Biosystems), 200nM of each primer (IDT, Coralville, IA), and 10ng of cDNA. PCR products were then visualized by agarose gel electrophoresis to determine the formation of a PCR product in the expected sample.

3.3 Results

3.3.1 Point mutations

In interrogating point mutations from the data, we took a bottom-up approach that started with analyzing first for recurrent mutations that occurred at the same base pair in the same gene. We then looked for recurrent mutations in the same gene but not the same base pair. This was the followed by looking for mutations that occurred in separate genes but in the same pathway.

In our analysis of genes with recurrent mutations at the same base pair, our best hit was in PARP4 (also known as vPARP). PARP4 is similar to PARP1 in that it can catalyze a poly(ADP-ribosyl)ation reaction, but it does not bind DNA directly [125]. It has also been shown to be associated with telomerase activity by interacting with telomerase-associated protein 1 (TEP1) [126]. PARP4 knockout mice are susceptible to chemical induced carcinogenesis [127]. It is known to be a component of the Major Vault Protein (MVP) complex which includes MVP, TEP1, and PARP4. The MVP complex is an extremely large (12.9 mDa) protein important in nuclear-cytoplasmic transport, signal transduction, and immune responses [128]. Interestingly, MVP has also been implicated in chemoresistance in cervical cancer [129], and is a significant prognostic factor in ovarian cancer, melanoma, and osteosarcoma [130]. Another study also showed that overexpression of MVP may suppress DNA repair by non-homologous end joining by downregulating Ku70/80 [131].

In our analysis, we identified an amino acid changing mutation in PARP4 at chr13:23973960 that converted the 49th amino acid from Isoleucine to Valine in exon 3. At first, this conversion seems indolent as the mutation converts a branched chain amino acid to another branched amino acid. But two aspects of this mutation are interesting. First, using PolyPhen, a popular bioinformatic tool to assess the effect of single basepair changes on protein structure, predicted this change to be possibly damaging. Also the location of this mutation occurs in the BRCA1 carboxy-terminal domain (BRCT) a domain similar to the c-terminus of BRCA1 which mediates phospho-protein binding [132]. If validated to be recurrent in a larger sample set, studies to determine its effect on mediating DNA repair, or sensitivity to DNA damaging agents would be warranted.

In a second analysis of mutations, we looked for mutations in the same gene but not necessarily the same base pair. We identified FAT1 (FAT tumor suppressor homolog 1) has having strong bioinformatic evidence for recurrent mutations. FAT1 is a tumor suppressor whose knockout causes excessive cell proliferation in drosophila [133].Homozygous deletions of FAT1 have been detected in 80% of primary oral cancers by CGH [134]. A survey of 326 breast cancers showed expression of FAT1 across tumor grades, but with grade 3 tumors have significantly less FAT1 compared to Grade 1. TNBCs are almost all Grade 3 tumors.

In our analysis we validated two mutations (one mutation in each of two samples) in FAT1. One at chr4:187821524 which created a Serine 1168 to Leucine change in exon 3. Also another at chr4:187775265 which created a Lysine 2988 to Isoleucine change. While point mutations in FAT1 in other breast cancers have not been reported, a search of the COSMIC database, (a comprehensive database of all known somatic mutations) revealed 12 reported somatic mutations for FAT1 (www.sanger.ac.uk/genetics/CGP/cosmic/) [135-137]. 10 of the 12 were in ovarian carcinomas. This data along with the previous report of deletions in oral cancers shows that FAT1 seems to be a repetitively mutated gene in multiple cancers.

In a third analysis, we analyzed the mutations to determine if a particular pathway was overrepresented using Ingenuity Pathway Analysis. In our analysis, our top hit was the BRCA1 pathway. This pathway is dedicated to genes involved with BRCA1 and its upstream and downstream effectors. We validated mutations in 5 genes:

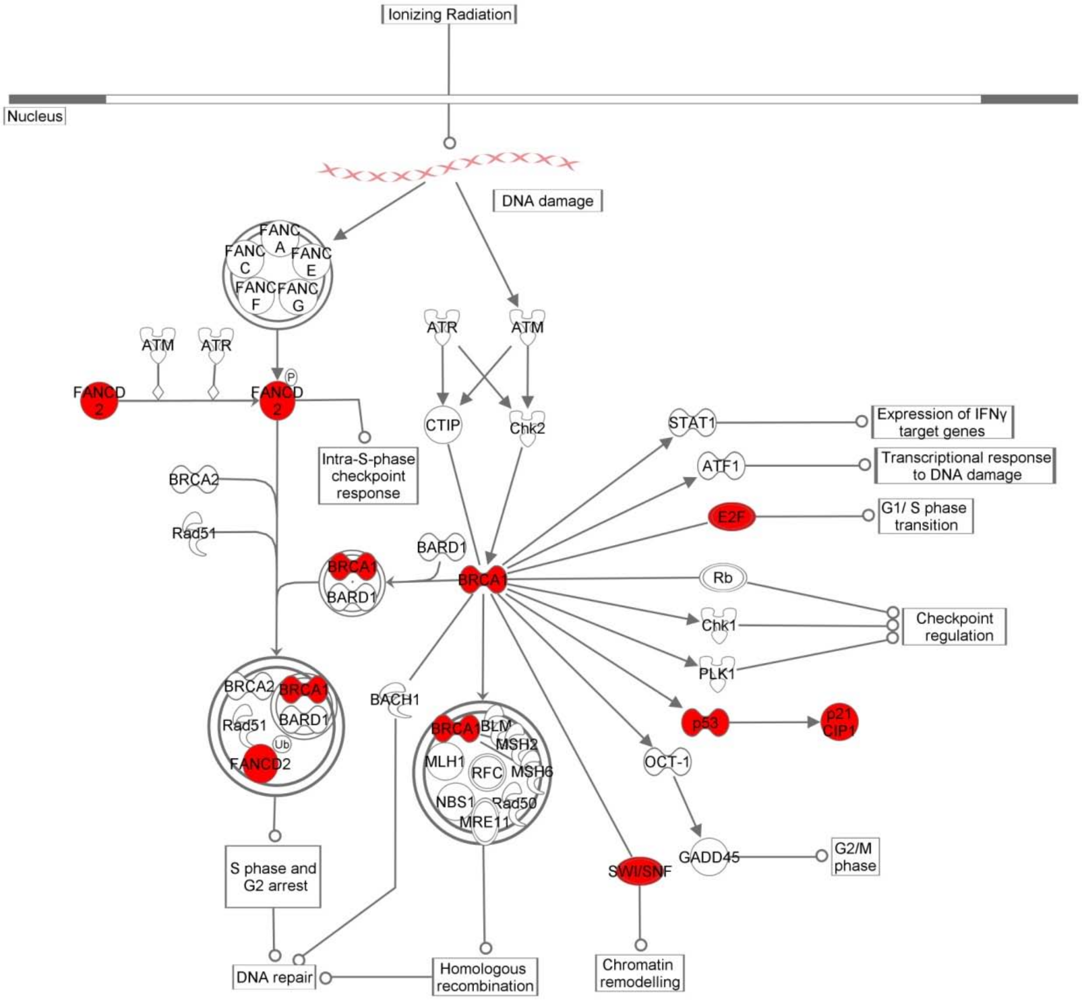


Figure 30. Point mutations in the BRCA1 pathway. Genes in red indicate those that have validated mutations in their coding regions. For this figure, due to genes having different gene symbols for the same gene, SWI/SNF = SMARCA4, and p21CIP21 = CDKN1A. BRCA1 is also shown mutated in this figure because of a detected indel in a sample described in the next section.

FANCD2, E2F3, CDKN1A, SMARCA4, and TP53 (Figure 30). In the figure BRCA1 is also shown to be mutated as this was detected in an indel analysis as will be explained in the next section. The data suggests a “BRCAness” mutation pattern, where genes that are involved with BRCA1 are mutated, but not necessarily the gene itself.

3.3.2 Small insertions and deletions

Small indels are potent mutations secondary to their ability to induce a frameshift in the reading frame of coding genes. Their deleterious nature is well known in cancer. For example, the overwhelming majority of known mutations in BRCA1 and BRCA2 are small indels. Also, the recent work identifying recurrent mutations of the ARID1A gene in Endometriosis-associated ovarian carcinomas using RNA-seq revealed that the majority of mutations are also indels. Even though deleterious point mutations do occur in BRCA1/2 and ARID1A, their prevalence is considerably less.

Because the majority of BRCA1 mutations are small indels, we scoured the indel data to determine if any of our samples contained known mutations for BRCA1. Interestingly, one of our samples, has a single base insertion at amino acid 1450, a known mutated site for BRCA1. What was more intriguing, that upon inspection of the gene expression PCA from Section 2.3.2, this sample is an outlier based on gene expression compared to the other samples (Figure 31). While conclusions from one sample cannot be made, it is nonetheless provocative for the lone outlier sample in our dataset to be mutated for BRCA1, and suggests a possible effect of BRCA1 mutations on the global gene expression profile.

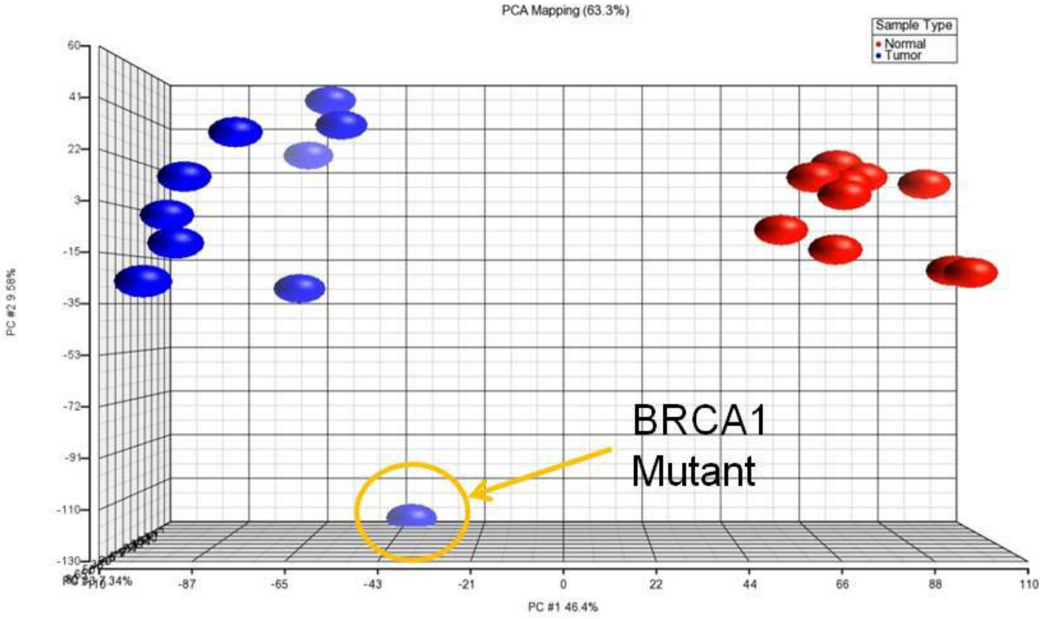


Figure 31. PCA of TNBC and normal samples with highlighted BRCA1 mutant sample. The PCA demonstrates that the BRCA1 mutated sample is an outlier when compared to other sporadic TNBCs. While conclusions cannot be made from one sample, the difference in gene expression profile between this BRCA1 mutated sample and the others suggests a possible effect on global gene expression.

3.3.3 Gene fusions

It is well known that translocations in DNA can result in the production of oncogenic fusions proteins [138]. Work by Maher et al. at has demonstrated the utility of using RNA-seq to discover *de novo* expressed fusion transcripts [69, 70]. We used a custom developed bioinformatic method known as the SASR (Suffix Array Single Read) which is designed to identify gene fusions from 50bp colorspace reads. By using reads that partially map to the human genome as input, the SASR determined reads that span exon junctions derived from two different genes (for SASR details, see Appendix 1). Candidate fusions from the SASR were then sorted by confidence and high confidence calls were subsequently validated by RT-PCR. In Table 8, we report 6 validated gene fusions identified in our dataset. 2 of 6 fusions are interchromsomal, whereas the other 4 are intrachromosomal with the corresponding genes in close proximity. These latter fusions most likely represent read-through events. Of important note, each fusion was specific for an individual sample, and no recurring fusion was detected in multiple samples.

Gene 1	Exon	Gene 2	Exon	Fusion Type	*Total Reads	*Unique Reads
AFAP1L2	1	MYCBP	3	Inter-	30	4
PDLIM5	2	GRID2	15	Intra-	8	2
PLCB4	3	PLCB1	3	Intra-	7	2
INHA	1	STK11IP	2	Intra-	4	2
CHADL	5	RANGAP1	16	Intra-	3	2
RUNX1	10	TDRG1	2	Inter-	3	2

Table 8. Gene fusions identified from mapped RNA-seq data. Using bioinformatic analysis followed by PCR we identified and validated six gene fusions in our TNBC samples. None of our gene fusions were recurrent in multiple samples. Total reads refers to the number of reads that spanned the fusion junction. Unique read evidence refers to the number of different unique sequences (or start points) that span the fusion junction. Notice the range in the number of reads of each fusion indicating the differences in the expression level of each one.

3.4 Discussion

The data presented in this chapter demonstrates the ability to call mutations from RNA-seq data that are specific for TNBC. There are some particular advantages of using mutations derived from RNA-Seq. 1) Using mutations only in expressed areas of the genome helps to focus the number of mutational candidates. Genome sequencing does suffer from a deluge of detecting a multitude of variants in non-expressed regions, in particular intergenic regions. While the data may potentially be useful, current bioinformatic tools to understand those variants do not currently exist. 2) With enough depth of sequencing, variants can be integrated with gene expression to do allele-specific expression analysis. While this study was not designed with the depth to do this analysis, future studies can determine the effect of variants detected in TNBC on the expression of the entire gene. 3) Calling mutations from RNA-seq is cost effective [139]. Genome sequencing requires a considerable amount of sequencing coverage. While these costs have come down dramatically, running one sample for RNA-seq is still considerably cheaper than running one sample for whole genome sequencing. Thus, in a cost-restrictive environment, one gains a dual benefit in RNA-seq by obtaining expression and mutational data.

In regards to the mutations detected, the point mutations and indels fit an evolving paradigm of a “BRCAness” pattern that goes along with the gene expression dysregulation from our data and others [17, 44]. Of interest, the PARP4 mutations, and the BRCA pathway mutations all fit with the realm of BRCA1 and DNA repair. This would suggest that TNBC causation could in part be a result of a combination of mutations and gene expression dysregulation that all revolve around BRCA1. Further, the observation of recurrent mutations in FAT1, a lesser known tumor suppressor, is quite interesting. Our data supports other reports in oral and ovarian cancers of recurrent mutations in this gene [134-137]. A complete knowledge of all mutated tumor suppressors in TNBC is not known, but it is quite possible that FAT1 mutations could join the cadre of TP53 and Rb tumor suppressor mutations. Further studies identifying the genes that are dysregulated by introducing FAT1 mutations could provide interesting clues to understanding tumorigenesis of TNBC and potentially identify therapeutic targets.

In regards to gene fusions, our data supports the work of others that no recurrent gene fusion in TNBC exists. While TNBC genomes are in general chaotic, thus raising the possibility of a recurrent fusion, it would seem that these translocations are more random. It could be postulated that these random gene fusions are merely by-products of deficient DNA repair in highly mitotic cells, and whether they confer a growth or survival advantage could be equally random. This would also suggest that identification of a critical, single fusion for drug targeting (such as that seen with imatinib for the BCR-ABL in chronic myelogenous leukemia) is highly unlikely. Nonetheless, it also possible that sequencing of much larger cohorts, such as those currently being sequenced by the major cancer genome consortia will reveal some kind of pattern of translocations in TNBC.

In summation of this chapter, RNA-seq has revealed several mutations present in TNBC that were not detected in normal samples or in genetic variation databases. The genes where these mutations occur are involved in BRCA and DNA repair, and in a lesser-known tumor suppressor. How these mutations play a role in TNBC causation will be the focus of future work utilizing this data. By understanding causation, one can hope it will point to proper therapeutic targeting as been recently demonstrated with PARP inhibition and BRCA mutations using the “synthetic lethal” approach. By integrating mutations with gene expression data, a process that is most capably done by next-generation sequencing, could paint a global picture of the interplay of gene expression perturbations caused by germline and somatic mutations.

Chapter 4: Summary

Triple-negative breast cancer (TNBC) is a disease that is unfortunately defined by what it lacks (ER-,PR-,HER2-) versus being defined by actionable therapeutic targets. Data from our group and others have demonstrated that there are significant shared features that define TNBC as a whole, including histological features [27, 28], gene expression changes, and mutational patterns [45-47]. But there is also extensive heterogeneity. Clinically, about half of TNBC patients will experience a pathological complete response to neoadjuvant chemotherapy, while the other half will not [5]. In addition, molecular subtyping suggests that while the majority of TNBCs belong to a single intrinsic subtype (the “basal-like” subtype), some TNBCs will also cluster with the other subtypes (Luminal A and B, HER2, and Normal-like) [32]. Also, a further subset of the basal-like subtype, known as the claudin-low (or Basal B) subtype, has been identified as consisting primarily of TNBCs [30, 101]. Interestingly, a recent article by the group of Reis-Filho et al. have questioned the validity and accuracy of the intrinsic subtypes, and convincingly demonstrated a lack of reproducibility between algorithmic predictors [140]. Even though the intrinsic subtypes have been known for 10 years, this work would suggest that the ability to use microarrays to define these cancers in a way that is clinically applicable is still a distant reality.

To further complicate the picture of TNBC, many of the genes that have been deemed to be under- or over-expressed in TNBC have been based on comparisons with suboptimal controls. This is well illustrated in the original paper by Perou et al. of the identification of the intrinsic subtypes, where the normal samples were primarily defined by the expression of adipose genes [29]. This suggests that the normal controls used were un-dissected breast tissues consisting mostly of fat. Further, data by others have demonstrated that suboptimal comparators such as reduction mammoplasties or adjacent normal tissue is fraught with problems that include benign neoplasia and pathological atypia [57-60], altered gene expression [61, 62], altered epigenetic markers [63], and loss of heterozygosity [64, 65]. Some have recently used cell separation techniques to separate ductal epithelium from stroma using live fresh biopsied tissues. But this also can be problematic in that the very process of digesting the stroma and using antibodies to separate the cells (especially if the purified cells enter a culture environment) would most likely induce large changes in expression in possibly the same genes that are dysregulated in cancer (proliferation, p53, apoptosis, etc.)

As a whole, TNBC is an undefined cancer who changes in gene expression and mutational profile from normal breast is not understood. This problem was the primer for the utilization of a powerful technology, next-generation whole transcriptome sequencing, to perform the most comprehensive comparison to date of the gene expression changes that occur between TNBC and normal breast. This comparison to normal breast was made possible by the Susan G. Komen Tissue Bank at the Indiana University Simon Cancer Center, which allowed us to obtain microdissected ductal epithelium from healthy volunteers as our normal control. While it was a laborious effort to tediously microdissect the frozen normal tissues for the epithelium, its results provided a genuine comparator to better understand the expression perturbations that comprises TNBC.

Using RNA-seq, we identified large sets of coding and non-coding RNAs that differentiate TNBC from normal breast. Genes involved with BRCA1 and DNA repair were highly present. As the link between BRCA1 mutations and development of TNBC is well known [19, 20], it is fitting to see genes associated with BRCA1 as some of our top hits. This included our #1 most differentially expressed gene COBRA1, which is a gene known to both downregulate estrogen receptor activity, and to regulate many genes that are also known to be regulated by BRCA1. As the majority of the work of the COBRA1 gene has been done in ER+ cancers, its role it TNBC is still undefined. We also observed a marked upregulation of DNA repair genes in the BRCA1 pathway and a microRNA known to suppress BRCA1 expression. Together, these data suggest that BRCA associated genes are highly implicated in sporadic TNBC, and not just in hereditary TNBC.

In a further unexpected observation, we noticed a discordance between genes targeted in previous clinical trials and their gene expression in TNBC compared to normal. In particular, the EGFR and KIT receptors have been previously implicated to be overexpressed in TNBC [8]. But we did not see this overexpression, in particular we noticed a lack of differential expression of the EGFR and a downregulation of KIT. Our data was congruent with several clinical trials showing a lack of clinical benefit for inhibiting these proteins [9-14]. The bigger picture that is portrayed, is the possibility that needless clinical trials have been performed based on microarray data that compared TNBC to other breast cancer subtypes and suboptimal normal controls [19, 29, 120]. Even the establishment of the Translational Breast Cancer Research Consortium (TBCRC) and its first trial TBCRC001: EGFR inhibition with cetuximab added to carboplatin in triple-negative breast cancer, was based on this data (personal communication and [11]). To further illustrate our point, a targeted agent that has shown clinical activity in TNBC, Iniparib, a PARP inhibitor, was significantly overexpressed in TNBC compared to normal. This would suggest that genes that are significantly unregulated compared to a “true normal” control can be potential therapeutic targets.

Indeed, novel therapeutic targets are desperately needed for a disease that preferentially affects young women and carriers a poor prognosis [1-3]. To this end, we performed an exhaustive search of our data and accompanying databases to identify prime targets. From our analysis we identified PTK7, the most significantly upregulated receptor tyrosine kinase in our dataset. Interestingly, previous work has already demonstrated that inhibition of this protein results in diminished proliferation specifically for TNBC cell lines and not in ER+ cell lines [96]. This protein plays an important role in epithelial development [99], which is fitting for a cancer known to be derived from a luminal epithelial progenitor [36, 37]. We also identified NEK2, our most overexpressed kinase in our dataset (34-fold), a gene important in centrosome separation during cell division. Previous studies have shown that siRNA knockdown of NEK2 in breast cancer cell lines inhibits cell growth, colony formation, and *in-vitro* invasiveness [107]. It has already been suggested as a potentially new therapeutic target in cancer [109, 110], and to this end, our group has already begun small molecule development of NEK2 with very early data indicating inhibition of proliferation using two small molecules identified from a compound library. We also scoured our data to identify other targets based on differential gene expression, pathway analysis, existing literature, and ability to target using *in-silico* molecular docking. This list of 25 targets is now being evaluated using an RNAi approach to determine their significance as therapeutic targets in a panel of TNBC and normal cell lines.

From this extensive dataset, we were also able to interrogate the noncoding regions of the genome. We were able to profile pre-miRNAs, lincRNAs, UCRs, and novel transcribed regions. From this data we observed upregulation of pre-miRNAs that downregulate estrogen and progesterone receptor, induce angiogenesis, and downregulate BRCA1. We also identified a large group of lincRNAs, UCRs, and novel transcribed regions that were significantly differentially expressed between TNBC and normal but have no known function. The NTRs were especially intriguing as they were able to separate the TNBC and normals by PCA, but were completely devoid of any known coding or noncoding genes.

In our final analyses, we leveraged the power of next-generation RNA-sequencing to uncover the mutational profile of TNBC. Of interest, we observed genes involved in the BRCA pathway to be mutated in samples that did not have BRCA1 mutations. Along with the gene expression data, this suggested a strong presence of a “BRCAness” profile underpinning TNBC. We also identified a recurrently mutated gene, FAT1, a novel tumor suppressor that has been previously demonstrated to be recurrently mutated in oral and ovarian cancers. Tumor suppressor mutations of p53 and Rb have dominated the landscape of TNBC, and this data suggest new players in the realm of mutated tumor suppressor genes. We also scanned our mutation data to identify gene fusions. Unfortunately, congruent with work by others [49], we did not identify any recurrent gene fusions in TNBC. Though, TNBC genomes tend to be quite chaotic, the development of gene fusions maybe a random event versus bonafide tumor initiators like BCR-Abl.

In summation, this dissertation provides an in-depth analysis of the transcriptomes of TNBC and normal breast using next-generation sequencing technology. This technology was highly leveraged in this study by taking advantage of measuring gene expression across the genome in an accurate and digital manner. We were also able to interrogate the base sequence of TNBCs in order to derive mutations and gene fusions. The largest hurdle encountered in this study was the constant development and update of bioinformatic tools to analyze the data. This entailed a substantial learning curve to understand and use the tools effectively to arrive at the desired endpoints. Nonetheless, new bioinformatic tools are continually being developed by our group and others, and soon RNA-seq will become a mainstay for the study of tumor transcriptomes. With this work, it is our hope that we have laid the framework to not only better understand and define this devastating disease, but to eventually uncover the key therapeutic targets needed to discover the cure.

Appendix 1: Bioinformatics

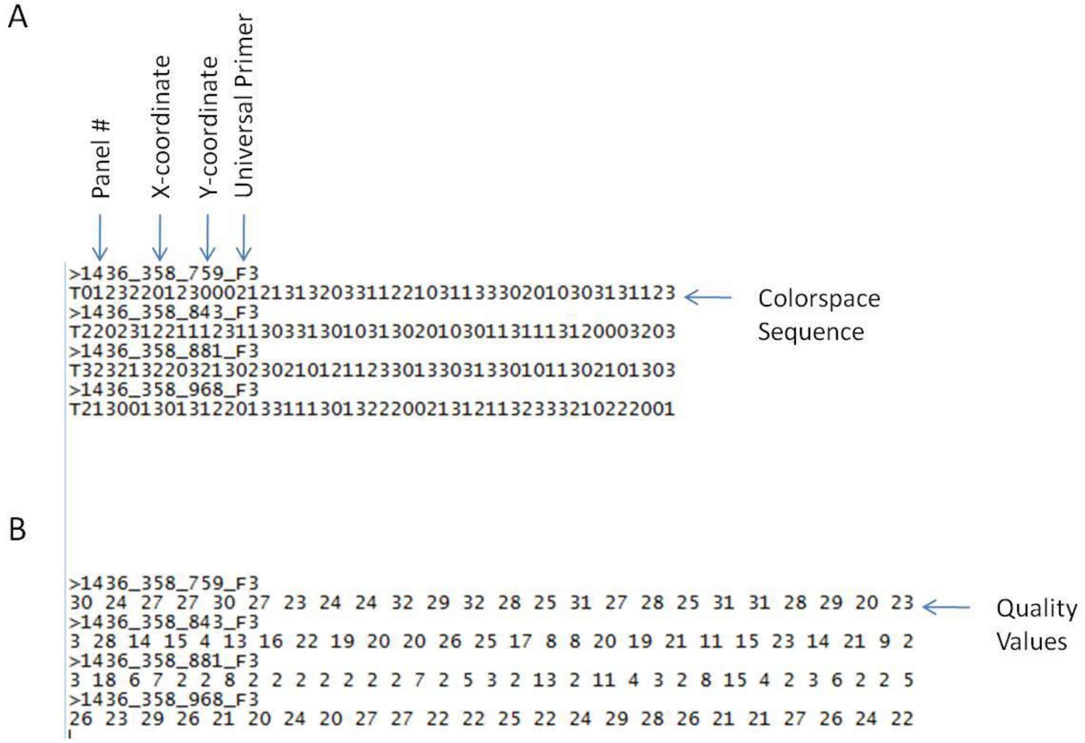
A1.1 Introduction

The following appendix provides in detail the bioinformatic steps taken to analyze the RNA-seq data presented in this dissertation. The majority of the analyses are performed in a Linux computing environment with some analyses in Windows. Unless stated, the computer scripting that is presented is Linux script (Red Hat Enterprise Linux 6, Bash Shell). This appendix will begin with the basic output files from the Applied Biosystems SOLiD sequencer and will walk through the software and scripts necessary to map sequencing data; derive gene expression (RPKM) values; perform downstream gene expression analyses; and detect point mutations, small indels, and gene fusions. Substantial effort is placed so that a well-trained bioinformatician would be able to exactly reproduce results given the same data, software, and computing environment. In regards to input files, scripting, and code, each file is explained and provided in detail.

A1.2 Sequencing data and read mapping

A1.2.1 Output files of the Applied Biosystems SOLiD sequencer

As mentioned in the introduction of this dissertation, next-generation sequencers produce vast amounts of sequencing data. These data are normally transferred from the sequencer either to large external hard drives, or to servers with large storage capacities. The ABI SOLiD produces two important files necessary for analysis. The first is the .csfasta file (short for colorspace FASTA) which is a text base file containing the read identifier and colorspace sequence for each read (Figure 32). The second file is the .qual file (short for quality) which is a text based file that contains the read identifier and the quality score for each color called for each read (Figure 32). The size of these files will always be proportional to sequencing output of the machine. Also the number of files will also be proportional to the number of samples run per slide (flowcells). In the case of this project, 20 samples were run in total, where the first 10 samples were run across two flowcells of the first instrument run, and the second 10 samples were run across two flowcells of the second instrument run. For this project, it results in a total of 20 samples x 2 flowcells = 40 .csfasta and 40 .qual files. Each of these files ranging in the 2-5 GB in size. The files were labeled in accordance to the sample, for example, “Tumor_1_FC1.csfasta” would mean a .csfasta file corresponding to TNBC sample #1 run on Flowcell 1. This nomenclature is used throughout the scripting examples in this appendix.



```
>1436_358_759_F3
T012322012300021213132033112210311333020103031311 23
>1436_358_843_F3
T22023122111231130331301031302010301131111312000 3203
>1436_358_881_F3
T32321322032130230210121123301330313301011302101303
>1436_358_968_F3
T21300130131220133111301322200213121132333210222001
```

```
>1436_358_759_F3
30 24 27 27 30 27 23 24 24 32 29 32 28 25 31 27 28 25 31 31 28 29 20 23
>1436_358_843_F3
3 28 14 15 4 13 16 22 19 20 20 26 25 17 8 8 20 19 21 11 15 23 14 21 9 2
>1436_358_881_F3
3 18 6 7 2 2 8 2 2 2 2 2 2 2 7 2 5 3 2 13 2 11 4 3 2 8 15 4 2 3 6 2 2 5
>1436_358_968_F3
26 23 29 26 21 20 24 20 27 27 22 22 25 22 24 29 28 26 21 21 27 26 24 22
```

Figure 32. Example of output sequencing text files from the ABI SOLiD sequencer. **(A)** An example of 4 sequencing reads from .csfasta file. The read identifier beings with a carat symbol ">" followed by 4 pieces of information. The first is the panel number. Each slide on a SOLiD sequencer is divided into 2357 panels from which the sequencer will image the beads on each panel to derive the sequence. The second and third pieces of information are the X- and Y- coordinate of the bead on the panel. The fourth piece of information is the designation of the universal primer used. In the case of SOLiD RNA-seq, "F3" refers to forward primer. Below the read identifier is the read sequence in colorspace. The read begins always begins with a "T" which is the last base of the P1 adaptor. The colorspace sequence then follows (see Figure 14 for details). **(B)** An example of quality information from a .qual file of the same 4 reads shown in Panel A. A .qual file is identical to a .csfasta file, that instead of colorspace sequence the PHRED quality score is shown. The PHRED score is the negative log-odds of the base being incorrectly called.

A1.2.2 Alignment of RNA-seq reads using ABI BioScope software

The first step in analyzing RNA-seq data is to map (also known as alignment) the reads to the human genome. For this, the ABI BioScope software is used as it contains a specially designed module to map RNA-seq colorspace data. Some favorable features of this software includes: relative efficient speed of mapping, the ability to map reads across splice junctions, filtering of RNA transcribed from ribosomal and repetitive elements, detection of gene fusions, and takes advantage of increased sequence accuracy afforded by the use of colorspace. The RNA-seq portion (referred to as Whole Transcriptome) in the software uses several input files.

Required reference files:

1) Human genome (version NCBI Build 36/hg18). This was downloaded from the UCSC Genome Bioinformatics Site (www.genome.ucsc.edu). To prepare the genome, the chromFa.zip file containing hg18 is unpacked, and only the full chromosome contigs are merged (cat) into a single file (the other .fa files including *.random.fa and alternate haplotypes are excluded). In order to create a female genome, the Y chromosome was simply omitted.
2) Human filter reference file. This file contains the fasta sequences of ribosomal RNA and repetitive elements. This file is provided by ABI.
3) Exon file. This is a gene transfer file (.gtf) that contains the positions and identifying information of all known exons in the human genome. This file is used to create the exon-junction library for mapping splice spanning reads. This file is provided by ABI, but is originally derived from the Refseq (specifically refGene) database of the UCSC Genome Bioinformatics Site. All exons from the Y chromosome were omitted to create a female specific file.

Required read files:

1) .csfasta. This file corresponds to the colorspace reads.
2) .qual. This file corresponds to the quality values of the colorspace reads.

To further run BioScope, three input files (known as .ini files) are needed. These files give BioScope the required parameters to map the RNA-seq data.

Required .ini files:

1) global.ini . This file provides the locations of all key files needed to run the whole transciptome module.
2) wt.single.read.workflow.ini . This file sets the various parameters for the whole transcriptome module and its various plug-ins. This includes whether specific plug-ins are run or not, and the number of reads needed to call specific events (alternative splice, fusion, etc.) (see example).
3) analysis.plan . This file is used to initate BioScope and simply calls the wt.single.read.workflow.ini file.

To run bioscope, the following usage is used:

```
>nohup bioscope.sh analysis.plan &
```

This will initiate BioScope using the parameters in the .ini files. BioScope requires a minimum of 1 head node, 3 compute nodes with 16GB of RAM, 8 processors, and 1TB of scratch space, along with sufficient I/O between the compute nodes and the central storage. BioScope uses a Java Messaging Service to run jobs on a PBS/Torque scheduler to parallel process each .csfasta/.qual file. Briefly, BioScope will perform three simultaneous mapping. This includes: mapping reads to the human genome, mapping reads to an exon-junction library, and mapping reads to the human filter reference. These three mappings are then merged into the standardized single output file (.bam, binary alignment mapping). The .bam files are then used in downstream applications including differential gene expression, discovery of novel alternative splicing, discovery of novel genes, gene fusions, mutation calling, etc. For more detailed information, see the BioScope user manual.

A1.2.3 Examples of required input files need for BioScope

Filename: human_filter_reference.fa

Description: Provides bioscope with a multi-fasta sequence file of adaptor sequences, ribosomal RNA, and repetitive elements in order for BioScope to remove reads deriving from the repetitive elements of the genome and sequencing artifact.

Contents (significantly truncated due to length):

```
>gi|124517659|ref|NR_003286.1| Homo sapiens 18S ribosomal RNA (LOC100008588)
TACCTGGTTGATCCTGCCAGTAGCATATGCTTGTCTCAAAGATTAAGCCATGCATGTCTAAGTACGCACG
GCCGGTACAGTGAAACTGCGAATGGCTCATTAAATCAGTTATGGTTCCTTTGGTCGCTCGCTCCTCTCCT
ACTTGGATAACTGTGGTAATTCTAGAGCTAATACATGCCGACGGGCGCTGACCCCCTTCGCGGGGGGGAT
GCGTGCATTTATCAGATCAAAACCAACCCGGTCAGCCCCTCTCCGGCCCCGGCCGGGGGGCGGGCGCCGG
CGGCTTTGGTGACTCTAGATAACCTCGGGCCGATCGCACGCCCCCCGTGGCGGCGACGACCCATTCGAAC
GTCTGCCCTATCAACTTTCGATGGTAGTCGCCGTGCCTACCATGGTGACCACGGGTGACGGGGAATCAGG
GTTCGATTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTAC
CCACTCCCGACCCGGGGAGGTAGTGACGAAAAATAACAATACAGGACTCTTTCGAGGCCCTGTAATTGGA
ATGAGTCCACTTTAAATCCTTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAAT
TCCAGCTCCAATAGCGTATATTAAAGTTGCTGCAGTTAAAAAGCTCGTAGTTGGATCTTGGGAGCGGGCG
GGCGGTCCGCCGCGAGGCGAGCCACCGCCCGTCCCCGCCCCTTGCCTCTCGGCGCCCCCTCGATGCTCTT
AGCTGAGTGTCCCGCGGGGCCCGAAGCGTTTACTTTGAAAAAATTAGAGTGTTCAAAGCAGGCCCGAGCC
GCCTGGATACCGCAGCTAGGAATAATGGAATAGGACCGCGGTTCTATTTTGTTGGTTTTCGGAACTGAGG
CCATGATTAAGAGGGACGGCCGGGGGCATTCGTATTGCGCCGCTAGAGGTGAAATTCTTGGACCGGCGCA
AGACGGACCAGAGCGAAAGCATTTGCCAAGAATGTTTTCATTAATCAAGAACGAAAGTCGGAGGTTCGAA
GACGATCAGATACCGTCGTAGTTCCGACCATAAACGATGCCGACCGGCGATGCGGCGGCGTTATTCCCAT
GACCCGCCGGGCAGCTTCCGGGAAACCAAAGTCTTTGGGTTCCGGGGGGAGTATGGTTGCAAAGCTGAAA
CTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAA
CCTCACCCGGCCCGGACACGGACAGGATTGACAGATTGATAGCTCTTTCTCGATTCCGTGGGTGGTGGTG
CATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTCTGGCATGCT
AACTAGTTACGCGACCCCCGAGCGGTCGGCGTCCCCCAACTTCTTAGAGGGACAAGTGGCGTTCAGCCAC
CCGAGATTGAGCAATAACAGGTCTGTGATGCCCTTAGATGTCCGGGGCTGCACGCGCGCTACACTGACTG
GCTCAGCGTGTGCCTACCCTACGCCGGCAGGCGCGGGTAACCCGTTGAACCCCATTCGTGATGGGGATCG
GGGATTGCAATTATTCCCCATGAACGAGGGAATTCCCGAGTAAGTGCGGGTCATAAGCTTGCGTTGATTA
AGTCCCTGCCCTTTGTACACACCGCCCGTCGCTACTACCGATTGGATGGTTTAGTGAGGCCCTCGGATCG
GCCCCGCCGGGGTCGGCCCACGGCCCTGGCGGAGCGCTGAGAAGACGGTCGAACTTGACTATCTAGAGGA
AGTAAAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTA
>3000072055893=AluYb8#SINE/Alu
GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCATGAGGTCAGGAGATCG
AGACCATCCTGGCTAACAAGGTGAAACCCCGTCTCTACTAAAAATACAAAAAATTAGCCGGGCGCAGTGGCGGGCGCCT
GTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAAGCGGAGCTTGCAGTGAGCCGAGATTG
CGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAA
>3000072055878=FAM#SINE/Alu
GCCGGGCGCGGTGGCGCGTGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGTGGGAGGATCGCTTGAGCCCAGGAGTTC
GAGGCTGTAGTGCGCTATGATCGCGCCTGTGAATAGCCACTGCACTCCAGCCTGAGCAACATAGCGAGACCCCGTCTCT
TAAAAAAAAAAAAAAAA
>3000072055879=FLAM_A#SINE/Alu
GCCGGGCGCGGTGGCGCGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCGGGAGGATCGCTTGAGCCCAGGAGTTC
GAGACCAGCCTGGGCAACATAGCGAGACCCCGTCTCTAAAAAAAAAAAAAAAAA
>3000072055880=FLAM_C#SINE/Alu
GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGAGGATCGCTTGAGCCCAGGAGTT
CGAGACCAGCCTGGGCAACATAGCGAGACCCCGTCTCTAAAAAAAAAAAAAAAA
>3000072055881=FRAM#SINE/Alu
```

Filename: human_refGene.090513.female.gtf

Description: This file contains the positions and identifying information of all known exons in order to create an exon-junction library for mapping splice spanning reads. The important files (from left to right) include: Field 1: chromosome; Field 4: exon start position; Field 5: exon end position; and Field 9: gene symbol and RefSeq accession ID.

Contents (significantly truncated due to length):

```
## gff-version 2
## gtf
## source-version refgene2gff.sh 1.2dev
## source-file /home/mullermw/test/1.2_testing/refGene.txt
## date 2009-05-14
## This file is a transformation of the refGene.txt file from the
## UCSC genome browser FTP site.
## Example:
## http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/refGene.txt.gz
##
## This file is similar to the gtf files available through the UCSC genome
## browser, with a few differnces.
## The gene_id attribute contains the HUGO name of the associated gene.
## A gene_id_repeat attribute appears when a gene_id appears at multiple loci.
## The value of gene_id_repeat is a unique integer for each locus.
## The transcript_id_repeat attribute appears when a transcript_id appears
## multiple times in the refGene.txt file.  The value is a unique integer
## for each occurence of the transcript_id.
##
##
chr1   refGene       exon   4225   4692   0.000000      -      .      gene_id
"WASH5P"; transcript_id "NR_024540";
chr1   refGene       exon   4833   4901   0.000000      -      .      gene_id
"WASH5P"; transcript_id "NR_024540";
chr1   refGene       exon   5659   5810   0.000000      -      .      gene_id
"WASH5P"; transcript_id "NR_024540";
chr1   refGene       exon   6470   6628   0.000000      -      .      gene_id
"WASH5P"; transcript_id "NR_024540";
chr1   refGene       exon   6721   6918   0.000000      -      .      gene_id
"WASH5P"; transcript_id "NR_024540";
chr1   refGene       exon   7096   7231   0.000000      -      .      gene_id
"WASH5P"; transcript_id "NR_024540";
chr1   refGene       exon   7469   7605   0.000000      -      .      gene_id
"WASH5P"; transcript_id "NR_024540";
chr1   refGene       exon   7778   7924   0.000000      -      .      gene_id
"WASH5P"; transcript_id "NR_024540";
chr1   refGene       exon   8131   8229   0.000000      -      .      gene_id
"WASH5P"; transcript_id "NR_024540";
chr1   refGene       exon   14601  14754  0.000000      -      .      gene_id
"WASH5P"; transcript_id "NR_024540";
chr1   refGene       exon   19184  19233  0.000000      -      .      gene_id
"WASH5P"; transcript_id "NR_024540";
chr1   refGene       exon   24475  25037  0.000000      -      .      gene_id
"FAM138A"; transcript_id "NR_026818";
chr1   refGene       exon   25140  25344  0.000000      -      .      gene_id
"FAM138A"; transcript_id "NR_026818";
chr1   refGene       exon   25584  25944  0.000000      -      .      gene_id
"FAM138A"; transcript_id "NR_026818";
```

Filename: global.ini

Description: This file provides the locations of all the needed files for bioscope to run the whole transcriptome pipeline.

Contents:

```
# © 2010 Life Technologies Corporation. All rights reserved.
################################################################################
#######################
#      Global settings for the pipeline run

examples.dir = /N/home/mradovic/bioscope/examples
reference.file=/N/home/mradovic/bioscope/BioScope-1.3.rBS131-
55029_20101119113500/etc/files/hg18/reference/human_hg18_female.fa
filter.reference.file=${examples.dir}/demos/wholeTranscriptome/references/human
_filter_reference.fasta
exons.gtf.file=/N/home/mradovic/bioscope/BioScope-1.3.rBS131-
55029_20101119113500/etc/files/hg18/GTF/human_refGene.090513.female.gtf
mapping.tagfiles = /N/home/mradovic/TNBC_BioScope-
1.3_Results/FC1/Tumor_1_FC1/Tumor_1_FC1.csfasta
qual.file = /N/home/mradovic/TNBC_BioScope-
1.3_Results/FC1/Tumor_1_FC1/Tumor_1_FC1.qual
junction.reference.file =
${intermediate.dir}/spljunctionextraction/junction.fasta
read.length = 50
base.dir = /N/home/mradovic/TNBC_BioScope-1.3_Results/FC1/Tumor_1_FC1

#####################################################
####################
#      Analysis settings

output.dir = ${base.dir}/output/single_read
intermediate.dir = ${base.dir}/intermediate
tmp.dir = ${base.dir}/temp
merge.output.directory = ${output.dir}/mapping
merge.output.bam.file = Tumor_1_FC1.bam
```

Filename: wt.single.read.workflow.ini

Description: This file sets the various parameters for the whole transcriptome module and its various plug-ins. This includes whether specific plug-ins are run or not (1=yes, 0=no), and the number of reads needed to call specific events (alternative splice, fusion, etc.)

Contents:

```
# © 2010 Life Technologies Corporation. All rights reserved.
###############################################################################
#######################

import ./global.ini

####################################################
####################
#      plugin run statements

wt.spljunctionextractor.run = 1
wt.junction.mapping.run = 1
wt.filter.mapping.run = 1
wt.genomic.mapping.run = 1
wt.merge.run = 1
wt.sam2wig.run = 1
wt.counttag.run = 1
wt.exon.sequence.extractor.run = 1
wt.junction.finder.run = 1

###########################################################
#####################
#       Splice Junction Extractor

wt.splext.genegtf.file = ${exons.gtf.file}
wt.splext.reference.file = ${reference.file}

############################################################
######################
#     Filter Mapping Plugin

wt.filter.mapping.reference = ${filter.reference.file}



###########################################################
######################
#      Genomic Mapping Plugin

wt.genomic.mapping.reference = ${reference.file}


###########################################################
######################
#      Merge Plugin
wt.merge.reference.file = ${reference.file}
wt.merge.filter.reference.file = ${filter.reference.file}
wt.merge.junction.reference.file = ${junction.reference.file}
wt.merge.qual.file = ${qual.file}
wt.merge.tmpdir = ${tmp.dir}
wt.merge.output.dir = ${merge.output.directory}
wt.merge.output.bam.file = ${merge.output.bam.file}

#wt.merge.known.juntion.penalty = 0
#wt.merge.putative.junction.penalty = 1
#wt.merge.score.clear.zone = 5
#wt.merge.min.junction.overhang = 8
#wt.merge.num.alignments.to.store = 1

#############################################################
########################
#      Sam2wig Plugin

wt.sam2wig.input.bam.file = ${merge.output.directory}/${merge.output.bam.file}
wt.sam2wig.output.dir = ${output.dir}/sam2wig
wt.sam2wig.basefilename = coverage

#wt.sam2wig.alignment.score = 0
#wt.sam2wig.min.coverage = 10
#wt.sam2wig.wigperchromosome = true
#wt.sam2wig.alignment.filter.mode = primary
#wt.sam2wig.score.clear.zone = 5
#wt.sam2wig.min.mapq = 10

#############################################################
########################
#      Count Tag Plugin


wt.counttag.exon.reference = ${exons.gtf.file}
wt.counttag.input.bam.file = ${merge.output.directory}/${merge.output.bam.file}
wt.counttag.output.dir = ${output.dir}/counttag
wt.counttag.output.file.name = countagresult.txt

#wt.counttag.score.clear.zone = 5
#wt.counttag.alignment.filter.mode = primary
#wt.counttag.min.alignment.score = 0
#wt.counttag.min.mapq = 10

##############################################################
########################
#      Junction Finder Plugins
# WARNING: Fusion caller is designed to work primarily with paired end
datasets.
# It is not suggested for use only the single read split evidence for calling
# gene fusions. However, calling and quantifying already known junctions is
# fine but fewer junctions will be found.

wt.genome.reference = ${reference.file}
wt.gtf.file = ${exons.gtf.file}
wt.f5.exseqext.output.reference =
${intermediate.dir}/exonsequenceextraction/exons_reference.fasta
wt.junction.finder.gtf.file = ${exons.gtf.file}
wt.junction.finder.input.exon.reference = ${wt.f5.exseqext.output.reference}
wt.junction.finder.input.bam =
${merge.output.directory}/${merge.output.bam.file}
wt.junction.finder.output.dir = ${output.dir}/junction_finder

#wt.junction.finder.min.exon.length = 25
#wt.junction.finder.first.read.max.read.length = 50
#wt.junction.finder.second.read.max.read.length = 25
```

```
#wt.junction.finder.single.read = 1
#wt.junction.finder.single.read.min.mapq = 0
#wt.junction.finder.single.read.min.overlap = 10
#wt.junction.finder.single.read.max.mismatches = 2
#wt.junction.finder.single.read.clip.size = 2
#wt.junction.finder.single.read.clip.total = 10
#wt.junction.finder.single.read.ReportMultihit = 0
#wt.junction.finder.single.read.remap = 0
#wt.junction.finder.single.read.clip.5.prime = 1
#wt.junction.finder.single.read.min.read.length = 37
#wt.junction.finder.paired.read = 0
#wt.junction.finder.paired.read.min.mapq = 10
#wt.junction.finder.paired.read.avg.insert.size = 120
#wt.junction.finder.paired.read.std.insert.size = 60
#wt.junction.finder.single.read.min.evidence.for.junction = 2
#wt.junction.finder.paired.read.min.evidence.for.junction = 0
#wt.junction.finder.combined.min.evidence.for.junction = 2
#wt.junction.finder.single.read.min.evidence.for.alt.splice = 2
#wt.junction.finder.paired.read.min.evidence.for.alt.splice = 0
#wt.junction.finder.combined.min.evidence.for.alt.splice = 2
#wt.junction.finder.single.read.min.evidence.for.fusion = 2
#wt.junction.finder.paired.read.min.evidence.for.fusion = 0
#wt.junction.finder.combined.evidence.for.fusion = 2
#wt.junction.finder.show.same.exon.pairs = 0
#wt.junction.finder.output.format = 3
```

Filename: analysis.plan

Description: This file calls all the required .ini files need to run a specific BioScope pipeline. In the case of the whole transcriptome pipeline, only the wt.single.read.workflow.ini file is called. The wt.single.read.workflow.ini file will in turn call the global.ini file upon initiation.

Contents:

```
./wt.single.read.workflow.ini
```

Filename: ***.**bam

Description: This file contains the information of mapped reads. The filename is usually preceded by the name of the sample, for example, Tumor_1_FC1.bam. The file is encoded in binary in order to increase the speed of processing and reduce file size. In order to visualize it, the software package samtools is required (http://samtools.sourceforge.net). For extensive details about BAM format, see the samtools site and the BioScope user manual. For brevity, the important fields for RNA-seq (from left to right) are described. Field 1:Read ID; Field 3: chromosomal location of mapped read; Field 4: position of mapped read; Field 6: CIGAR string where numbers refer to bases and letters refer to specific codes, where M=match, N=gap due to intron, H=hard-clip, I=insertion, D=deletion; Field 10: sequence in basespace; Field 11: mapping quality in Ascii-33 format; Field 18: colorspace quality values in Ascii-33 format; Field 19: colorspace sequence of the read.

Contents: (Example of four mapped reads to chr1 from a BAM file)

```
2327_122_1766   16      chr1    4269    1       12H38M  *       0       0
TCTGCTCAGTTCTTTATTGATTGGTGTGCCGTTTTCTC  IIIIIIIIIIIIIIIIDDIIIIIIIIIIIIIIIIIIII
MD:Z:38 RG:Z:2011033111424185   IH:i:1  NH:i:6  HI:i:1  XN:i:37
CQ:Z:???=<9:@:9;A=?;9?7==;*;@:<:59=><;<7@3=9537,594,)43 AS:i:37
CS:Z:T12220001303111101032103300220121223122220133020103
1047_224_626    16      chr1    4281    2       50M     *       0       0
TTTATTGATTGGTGTGCCGTTTTCTCTGGAAGCCTCTTAAGAACACAGTG
!IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIICA      MD:Z:50
RG:Z:2011033111424185   IH:i:1  NH:i:6  HI:i:1  XN:i:49
CQ:Z:6,8988:>>=:?>?=:?;>@8;;?==?:>6;>::4>=>;<:>5;</;4>= AS:i:49
CS:Z:T21121111022030222032020122220001303111101032103300
1615_1687_46    16      chr1    4292    2       50M     *       0       0
GTGTGCCGTTTTCTCTGGAAGCCTCTTAACAACACAGTGGCGCAGGCTGG
!IFIII>.FIII=HIIIA=IEFIIIIE:IIIIIIIIIIIIIIIIIIIIII      MD:Z:29G20
RG:Z:2011033111424185   IH:i:1  NH:i:7  HI:i:1  XN:i:43
CQ:Z:=<9858;;4:9=3956;88=9<0+;;;2:-97';:8://=>=*%:5;164 AS:i:43
CS:Z:T20123021333011211110110302220320201222200013031111
502_556_894     16      chr1    4297    2       12H38M  *       0       0
CCGTTTTCTCTGGAAGCCTCTTAAGAACACAGTGGCGC  IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
MD:Z:38 RG:Z:2011033111424185   IH:i:1  NH:i:3  HI:i:1  XN:i:37
CQ:Z::>>=?698A>96<8>@=>>@7?<<;;8<<?;8@:<=8;<999=:;+6:-< AS:i:37
CS:Z:T13330112111102203022203202012222000130131110033020
```

A1.3 Derivation of differential gene expression

A1.3.1 Performing differential gene expression analysis

One of the powerful applications of RNA-seq is to derive differential gene expression. Unlike microarrays, which use arbitrary signal intensity, RNA-seq is a digital counting of mapped reads allowing for an accurate assessment of gene expression. In order to perform this digital counting, mapped reads from the BAM files are cross-referenced against known genes present in the NCBI RefSeq database. To obtain the database of all known curated genes in hg18, the RefSeq database in the form of the refFlat (refFlat.txt.gz) file is downloaded from the UCSC Genome Bioinformatics Site. The refFlat file contains the exon position numbers of all known genes in the human genome with corresponding identifying information (see file example). While several different software programs exist in order to count reads mapped to known genes and derive RPKM expression values, Partek Genomics Suite (Windows) performs this function in the most efficient manner. In Partek, BAM files of mapped RNA-seq data are imported into the software. The refFlat file is then downloaded and converted into a Partek compatible annotation file. The software then cross-references the mapped RNA-seq from the BAM file using the position numbers of the mapped reads against the position numbers of all known exons in the refFlat database in order to assign each read to a gene. The counts of the mapped reads to each gene are then determined and the RPKM value is calculated by taking total number of reads mapping to the gene divided by the length of the gene in kilobasepairs divided by the number of mapped reads (in millions).

Once the RPKM values for each gene and sample are determined, differential expression can then be calculated and statistically analyzed. For basic differential expression, fold change = average RPKM value of sample set 1/average RPKM value of sample set 2. In the case of this dissertation, this was the average RPKM values of the 10 TNBC samples divided by the average RPKM values of the 10 normal samples. Because we are comparing 10 values vs. 10 values (in essence 10 biolological replicates vs. 10 biological replicates), a t-based statistic can then be used to determine significance. In the case a 1-way ANOVA (equivalent to a 2-sample, 2-tailed t-test, 1-way refers to the sample type, TNBC or normal) is performed in order to determine a p-value of the differential expression for each gene. Because thousands of genes are being considered, a correction for multiple comparisons must be performed. The accepted standard for gene expression data is FDR (false discovery rate) which when given a range of p-values, assesses the percentage of genes of a dataset that are likely to be false positives and do not reflect true differential expression.

Along with differential gene expression, downstream analyses that describe global gene expression can be performed in Partek. This includes principal components analysis and hierarchical clustering of the RPKM data, and gene ontology/gene set enrichment analysis of the differential expression data. Also, differential gene expression data can be exported to Ingenuity Pathway Analysis (IPA) to perform network and pathway analyses.

A1.3.2 Example refFlat file used for differential gene expression

Filename: refFlat.txt

Description: The refFlat file is a file detailing the position numbers of all genes (and their exons) in the human genome along with gene identifier information. This file is important in the use of cross-referencing mapped RNA-seq reads to genes in order to derive RPKM values and subsequent differential gene expression. The fields (from left to right) include: Field 1: gene symbol; Field 2: RefSeq accession number; Field 3: chromosome; Field 4: strand; Field 5: transcription start position; Field 6: transcription end position; Field 7: coding start position; Field 8: coding end position; Field 9: exon count; Field 10: start positions of all exons; Field 11: end positions of all exons.

Contents (significantly truncated due to length):

```
KCNMB3 NM_171829     chr3   -      180443247     180460373     180443385
       180459431     4      180443247,180444976,180451224,180459429,
       180443766,180445175,180451416,180460373,
HSCB   NM_172002     chr22  +      27468042      27483496      27468083
       27483157      6
       27468042,27469869,27470602,27471851,27477228,27483065,
       27468319,27469966,27470692,27471996,27477276,27483496,
TEDDM1 NM_172000     chr1   -      180633874     180636374     180635421
       180636243     1      180633874,      180636374,
LACTB  NM_171846     chr15  +      61201051      61209079      61201123
       61208906      5      61201051,61201887,61206110,61206604,61208736,
       61201480,61201954,61206301,61206941,61209079,
SERF2  NR_037672     chr15  +      41871465      41875579      41875579
       41875579      4      41871465,41872464,41873199,41873863,
       41871873,41872573,41873426,41875579,
FAM138A       NR_026818     chr1   -      24473  25944  25944  25944  3
       24473,25139,25583,   25037,25344,25944,
FAM138F       NR_026820     chr1   -      24473  25944  25944  25944  3
       24473,25139,25583,   25037,25344,25944,
MIR183 NR_029615     chr7   -      129201980     129202090     129202090
       129202090     1      129201980,      129202090,
MIR221 NR_029635     chrX   -      45490528      45490638      45490638
       45490638      1      45490528,     45490638,
MIR939 NR_030635     chr8   -      145590171     145590253     145590253
       145590253     1      145590171,      145590253,
SPRY2  NM_005842     chr13  -      79808112      79813087      79808893
       79809841      2      79808112,79812757,   79809892,79813087,
SPRY1  NM_005841     chr4   +      124540132     124544359     124542196
       124543156     2      124540132,124542141, 124540399,124544359,
```

A1.4 Detection of point mutations from mapped RNA-seq data

A1.4.1 Bioinformatic methods to call point mutations from RNA-seq data

Another powerful application of RNA-seq is the ability to interrogate the raw sequence of mapped reads in order to search for mutations. As explained in detail in Section 1.3.3, RNA-seq has unique challenges in accurately determining base pair changes. In order to derive accurate and biologically meaningful mutational data, a four-step process is employed:

1) Preparation of BAM files for use in point mutation calling
2) Detection of point mutations from BAM data
3) Annotation of point mutations
4) Parsing of the point mutation data

First, the BAM files are prepared for use in point mutation calling. This first involves merging all .bam files for each sample into a single file. As mentioned previously, each sample may have multiple .bam files because of the sample being run across multiple flowcells. In order to merge the .bam files, the software package Picard is used (http://picard.sourceforge.net). A java script MergeSamFiles.jar is a component of Picard and is used in the following example to merge two .bam files from sample Tumor_1 Where I = input files and O = output file:

```
[Usage]: > java –jar MergeSamFiles.jar I=<input files> O=<output file>
[Example]: > java -jar MergeSamFiles.jar
I=/panasas/milan/milan_TNBC_Data/FC1/Tumor_1_FC1/ _1_FC1.bam
I=/panasas/milan/milan_TNBC_Data/FC2/Tumor_1_FC2/Tumor_1_FC2.bam
O=/panasas/milan/IUPUI_data/milan_TNBC_Data/merged_bam/Tumor_1/Tumor_1.bam
```

Once the BAM files are merged for each sample, the BAM files are then "piledup" using samtools which is a format that reports the base sequence, base quality value, and mapping quality value for every position in the genome (line-by-line) that is covered by a read(s) (see file example in Section A1.4.2). Samtools extracts this information from the .bam file. To do this the following command is run:

```
[Usage]: >samtools pileup –s –f <reference.fa> <.bam file> > <output file>
[Example]: >samtools pileup –s –f human_hg18_female.fa Tumor_1_FC1.bam >
Tumor_1_FC1.bam.pileup
```

In the second step, the .bam.pileup files are then used to call point mutations. As described in Section 1.3.3, unique statistical challenges are presented when calling point mutations from mapped RNA-seq data. To overcome this, the SNVMix2 (Single nucleotide variants Mix 2, http://compbio.bccrc.ca/?page_id=204) software is used which is specifically designed to account for the non-uniformity in coverage when calling point mutations. SNVMix2 is run using the following:

```
[Usage]: >./SNVMix2 –i <input pileup file> -o <output SNVMix2 file>
[Example]: > ./SNVMix2 -i
/N/gpfs/mradovic/merged_TNBC_bam_files/Normal_1.bam.pileup -o
/N/gpfs/mradovic/merged_TNBC_bam_files/Normal_1.bam.pileup.SNVMix2
```

By default, SNVMix will only consider bases that have a minimum base and mapping PHRED score of Q20, thus eliminating a substantial number of false positives due to poor sequencing quality. The output file from SNVMix2 contains information for each position of the genome where a potential point mutation has occurred. This includes the reference allele and its read count, the variant allele and its read count, the probabilities of each genotype (homozygous wildtype, heterozygous, homozygous variant), and the genotype call based on the genotype with the highest probability. The SNVMix2 output file is then further filtered in order reduce false positives by considering only those mutations that had enough read support to result in a 90% probability of the call being real. This is done using the following command in SNVMix2:

```
[Usage]: > perl snvmix2summary.pl –i <input .bam.pileup.SNVMix2 file> -c <int> -t
<probability>
[Example]:> perl snvmix2summary.pl -i
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/Tumor_1.bam.pileup.SNVMix2 -c 3 -t
0.9 >
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/Tumor_1.bam.pileup.SNVMix2.SNVfilt
ered0.9
```

In the third step, the called mutations need to be annotated in order to determine their gene location and potential function. To do this, the ANNOVAR software (http://www.openbioinformatics.org/annovar/) is used as an efficient means of annotating genes. In order to use ANNOVAR, output from SNVMix2 has to be reformatted in order to be used in annovar. This can be done using the following command:

```
[Usage]: >awk '{print $1,$2,$3}' <SNVfiltered file> | sed 's/:/ /g' | awk '{print
$1,$2,$2,$3,$4}' > <output file>
[Example]: >awk '{print $1,$2,$3}'
Tumor_1.bam.pileup.SNVMix2.SNVfiltered0.9.delrandom | sed 's/:/ /g' | awk '{print
$1,$2,$2,$3,$4}' > Tumor_1.annovar
```

ANNOVAR can then be used to filter mutations through a custom pipeline (described in Section 3.2, and Figure 27) that focuses only on nonsynonymous and stopgain/stoploss mutations while also filtering for known variants from the 1000 genomes project and dBSNP. To do this a custom bash script pipeline was created (see Section A1.4.2) for example file. At the end of this, a file is created called .filtered that is then used for subsequent parsing (see Section A1.4.2 for example and details). This

.filtered file contains the chromosome and position number and gene for each called mutation that is nonsynonymous and not present in 1000 genomes/dbSNP.

In the fourth step, the .filtered files are combined and parsed in order to find recurrent patterns of mutations. To combine the files, a custom perl script, consolidatereport.pl is used as follows:

```
[Usage]: > perl consolidateReport.pl <path to directory of filtered files> > <output file>
[Example]: > perl consolidateReport.pl
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/annovar/filtered/ >
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/annovar/filtered/consolidated.TNBC_N
ew_Mutation.output
```

The output of this perl script is a single file that combines all the data from the .filtered files in order to report which mutations occur in which sample (se Section A1.4.2 for example and details.). This file can then be imported into Excel for further sorting.

A1.4.2 Examples of files used in point mutation calling

Filename: *.bam.pileup

Description: A file that is derived from a .bam file which reports the base sequence, base quality value, and mapping quality value for every position in the genome (line-by-line) that is covered by a read(s). The contents are the following (left to right): Field 1: chromosome; Field 2: position number; Field 3: reference allele; Field 4: number of reads covering the allele; Field 5: a dot stands for a match to the reference base on the forward strand, a comma for a match on the reverse strand, `ACGTN' for a mismatch on the forward strand and ‘acgtn' for a mismatch on the reverse strand, carat symbol ‘^' marks the start of a read segment which is a contiguous subsequence on the read separated by ‘N/S/H' CIGAR operation, and a symbol ‘$' marks the end of a read segment; Field 6: Base qualities in Ascii-33; Field 7: Mapping qualities in Ascii-33.

Contents: (Example)

```
chr1   4860   A      7      ,,,,,,,       IEIIIII       """""""
chr1   4861   G      7      ,,,,,,,       FIIIIII       """""""
chr1   4862   A      7      ,,,,,,,       IIIIHII       """""""
chr1   4863   G      7      ,,,,,,,       IIIIIII       """""""
chr1   4864   A      7      ,,,,,,,       IIIIIII       """""""
chr1   4865   T      7      ,,,,,,,       IHIIDII       """""""
chr1   4866   C      7      ,,,,,,,       IIIIEII       """""""
chr1   4867   C      7      ,,,,,,,       IIII@II       """""""
chr1   4868   G      7      ,,,,,,,       IIIIAII       """""""
chr1   4869   A      7      ,$,,,,,,      IIII9II       """""""
chr1   4870   C      6      ,,,,,, III'II """"""
chr1   4871   A      6      ,,,,,, III'II """"""
chr1   4872   T      7      ,,,,,,^",     IIIIII!       """""""
chr1   4873   C      7      ,,,,,,,       IIIIIC@       """""""
chr1   4874   A      7      ,,,,,,,       IEIIIIB       """""""
chr1   4875   A      7      ,,,,,,,       HGIIIII       """""""
chr1   4876   G      7      ,,,,,,,       IDIAIII       """""""
chr1   4877   T      7      ,,,,,,,       IHI.IHI       """""""
chr1   4878   G      7      ,,c,$,,,      III8II)       """""""
```

Filename: *.SNVMix2 (SNVMix2 output)

Description: Output file after .bam.pileup is run through SNVMix2 for mutation calling.

SNVMix outputs 4 columns:
1: coordinate in "chromosome:position" format
2: reference base
3: non-reference base
4: comma separated field:

REF:#, NREF:#, p(AA), p(AB), p(BB), maxP

REF:#	reference base and number of occurrences that passed quality settings
NREF:#	non-reference base and number of occurrences that passed quality settings
p(AA)	probability assigned to homozygous to reference
p(AB)	probability assigned to heterozygous genotype
p(BB)	probability assigned to homozygous to the non-reference
maxP	class with max probability (1=AA, 2=AB, 3=BB)

Contents: (Example)

```
chr10:176795  A      G      A:0,G:4,0.0000163823,0.0712955092,0.9286881085,3
chr10:284953  A      G      A:0,G:10,0.0000000000,0.0021432927,0.9978567073,3
chr10:353785  C      G      C:0,G:4,0.0000163827,0.0712955701,0.9286880472,3
chr10:458310  T      G      T:0,G:4,0.0000164198,0.0713014847,0.9286820955,3
chr10:489191  G      A      G:0,A:6,0.0000000380,0.0228081436,0.9771918184,3
chr10:671950  G      T      G:2,T:3,0.0119679627,0.9851925401,0.0028394972,2
chr10:846385  C      T      C:0,T:6,0.0000000375,0.0227975797,0.9772023828,3
chr10:846918  A      T      A:0,T:63,0.0000000000,0.0000000000,1.0000000000,3
chr10:847151  A      G      A:0,G:4,0.0000164615,0.0713081315,0.9286754070,3
chr10:863978  G      A      G:0,A:5,0.0000007840,0.0405909275,0.9594082885,3
chr10:868359  a      G      a:0,G:4,0.0000163827,0.0712955694,0.9286880479,3
chr10:868895  g      A      g:2,A:5,0.0000840398,0.9905118400,0.0094041202,2
chr10:877382  C      G      C:0,G:4,0.0000165223,0.0713177683,0.9286657095,3
chr10:889804  c      T      c:0,T:6,0.0000000366,0.0227758256,0.9772241378,3
chr10:1053441 T      G      T:0,G:7,0.0000000017,0.0126831403,0.9873168580,3
chr10:1053447 A      G      A:0,G:7,0.0000000017,0.0126829171,0.9873170812,3
chr10:3144461 G      A      G:0,A:4,0.0000163847,0.0712958896,0.9286877257,3
chr10:3170227 T      C      T:0,C:12,0.0000000000,0.0006517710,0.9993482290,3
chr10:3170298 C      T      C:0,T:10,0.0000000000,0.0021433504,0.9978566496,3
chr10:3170316 T      C      T:0,C:15,0.0000000000,0.0001094150,0.9998905850,3
chr10:3170689 C      A      C:0,A:5,0.0000008596,0.0407348580,0.9592642824,3
chr10:3175237 A      G      A:0,G:5,0.0000008012,0.0406247274,0.9593744713,3
chr10:3179380 C      T      C:0,T:12,0.0000000000,0.0006515631,0.9993484369,3
chr10:3190292 G      A      G:0,A:14,0.0000000000,0.0001979034,0.9998020966,3
chr10:3192065 T      C      T:0,C:6,0.0000000370,0.0227862114,0.9772137516,3
chr10:3196027 A      G      A:0,G:32,0.0000000000,0.0000000043,0.9999999957,3
chr10:3197705 G      C      G:0,C:6,0.0000000378,0.0228035383,0.9771964239,3
chr10:3198512 G      A      G:0,A:19,0.0000000000,0.0000100735,0.9999899265,3
chr10:3198557 A      G      A:0,G:5,0.0000007846,0.0405922066,0.9594070088,3
```

Filename: Custom ANNOVAR pipeline bash script

Description: The bash script run a set of sequential commands that will input the .annovar file modified from the SNVMix2 output and outputs a .filtered file for further downstream use in data parsing. This set of command will isolate those called mutations that are nonsynonymous and are not present in the 1000 genomes project or dbSNP and the output the mutations in chromosome:position format with the associated gene symbol.

Contents: (example of script written for sample Tumor_1)

```
/N/gpfs/mradovic/annovar/annotate_variation.pl -geneanno
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/annovar/Tumor_1.annovar
/N/gpfs/mradovic/annovar/humandb/
sed '/\<synonymous\>/d'
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/annovar/Tumor_1.annovar.exonic_vari
ant_function >
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/annovar/Tumor_1.annovar.exonic_vari
ant_function.delsyn
awk '{print $5,$6,$7,$8,$9,$4}'
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/annovar/Tumor_1.annovar.exonic_vari
ant_function.delsyn | sed 's/:/ /g' | awk '{print $1,$2,$3,$4,$5,$6}' >
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/annovar/Tumor_1.annovar.exonic_vari
ant_function.delsyn.varlist
/N/gpfs/mradovic/annovar/annotate_variation.pl -filter -dbtype 1000g2010jul_ceu
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/annovar/Tumor_1.annovar.exonic_vari
ant_function.delsyn.varlist /N/gpfs/mradovic/annovar/humandb/
/N/gpfs/mradovic/annovar/annotate_variation.pl -filter -dbtype 1000g2010jul_yri
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/annovar/Tumor_1.annovar.exonic_vari
ant_function.delsyn.varlist.hg18_CEU.sites.2010_07_filtered
/N/gpfs/mradovic/annovar/humandb/
/N/gpfs/mradovic/annovar/annotate_variation.pl -filter -dbtype
1000g2010jul_jptchb
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/annovar/Tumor_1.annovar.exonic_vari
ant_function.delsyn.varlist.hg18_CEU.sites.2010_07_filtered.hg18_YRI.sites.2010
_07_filtered /N/gpfs/mradovic/annovar/humandb/
/N/gpfs/mradovic/annovar/annotate_variation.pl -filter -dbtype snp130
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/annovar/Tumor_1.annovar.exonic_vari
ant_function.delsyn.varlist.hg18_CEU.sites.2010_07_filtered.hg18_YRI.sites.2010
_07_filtered.hg18_JPTCHB.sites.2010_07_filtered
/N/gpfs/mradovic/annovar/humandb/
awk '{print $1":"$2"&"$6,1}'
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/annovar/Tumor_1.annovar.exonic_vari
ant_function.delsyn.varlist.hg18_CEU.sites.2010_07_filtered.hg18_YRI.sites.2010
_07_filtered.hg18_JPTCHB.sites.2010_07_filtered.hg18_snp130_filtered >
/N/gpfs/mradovic/TNBC_New_Mutation_Analysis/annovar/Tumor_1.filtered
```

Filename: *.filtered

Description: This is the output file from the ANNOVAR pipeline bash script. Each sample has its own .filtered file. The .filtered file consists of two columns: Column 1: chromosome:position”&”gene symbol; Column 2: is always the number “1” in reference that a mutation has occurred. Column 2 is used in the subsequent perl script, consolidateReport.pl.

Contents: (Example, truncated due to length)

```
chr3:123902102&PARP14 1
chr3:126662353&SNX4 1
chr17:19515743&ALDH3A2 1
chr1:240108883&EXO1 1
chr4:48217871&FRYL 1
chr2:202212016&ALS2CR4 1
chr1:107401632&PRMT6 1
chrX:155944&PLCXD1 1
chr1:17470024&PADI3 1
chr13:47705589&ITM2B 1
chr12:6794398&CD4 1
chr10:126663441&ZRANB1 1
chr17:20850825&USP22 1
chr8:67650912&MYBL1 1
chr12:9116703&A2M 1
chr1:54190557&LRRC42 1
chr17:74311632&USP36 1
chr6:90410593&MDN1 1
chr16:30999807&ZNF646 1
chr13:95037910&DZIP1 1
chr14:87974248&SPATA7 1
chr17:7438743&FXR2 1
chr2:227370087&IRS1 1
chr3:137499597&PCCB 1
chr1:225909455&ZNF678 1
chr7:69866145&AUTS2 1
chr17:26577818&NF1 1
chr1:202033790&ZBED6 1
chrX:128520589&OCRL 1
chr7:30797453&FAM188B 1
chr2:237948178&COL6A3 1
chr9:32624577&TAF1L 1
chr3:194815258&OPA1 1
chr3:13500037&HDAC11 1
chr17:53928094&MTMR4 1
chr9:2828479&KIAA0020 1
chr5:96165111&ERAP1 1
chr19:8894662&MUC16 1
```

Filename: consolidateReport.pl

Description: This perl script will combine multiple .filtered files into a single output file. The script also produces a log file in order to inform the user of the order of the columns. The output file starts with first column which is chromosome:position”&”gene symbol followed by columns for each sample which denotes whether the mutation is present “1” or absent “0”.

Contents:

```
#!/usr/bin/env perl

use strict;
use warnings;
use File::Basename;

my $directory = shift;
my %Clusters;
my @files;

##process each file in the directory
### Process each file in the From Directory
opendir(IMPORTANTE, $directory) or die "can't opendir $directory $!\n";
while (defined (my $file = readdir(IMPORTANTE))) {
        # do something with "$directory/$file"
              next unless ($file =~ /filtered/);
       my $filebase = basename($file, ".filtered");
       print STDERR "Processing: ", $filebase, "\n";

       push @files, $file;
       my $path = $directory . "/" . $file;
       open (FILE, $path) || die "could not open file ", $directory, "/",
$file;
       while(<FILE>) {
              if ($_ =~ /^(\S+)\s+(\S+)/){
                     $Clusters{$1}{$file} = $2;
              }
       }
       close FILE;
}

#end going through each file
open (ORDER, ">Log.order.txt") || die "could not open log file $!";
for (my $i = 0; $i < @files; $i++) {
       print ORDER $files[$i], "\n";
}
close ORDER;

###print out the report
foreach my $key (keys %Clusters) {
       print $key, "\t\t";
       for (my $i = 0; $i < @files; $i++) {
              if (exists $Clusters{$key}{$files[$i]}) {
                     print $Clusters{$key}{$files[$i]}, "\t";
              }else {
                     print "0", "\t";
```

```
                }
        }
        print "\n";
}
```

Output file (Example, truncated due to length):

```
chr11:102699863&DYNC2H1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
chr3:197078169&TNK2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
chr7:5376830&TNRC18 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
chr16:179203&LUC7L 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
chr18:26840888&DSC3 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
chr1:93430183&CCDC18 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
chr12:111829274&OAS1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
chr16:55092429&BBS2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
chr14:105029074&C14orf80 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
chr19:3013825&AES 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
chr22:22903565&CABIN1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
chr14:67321677&ZFYVE26 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
chr15:43448936&GATM 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
chr8:22034552&HR 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
chr17:17640432&RAI1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
chr1:54025439&TMEM48 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
chr14:35406937&BRMS1L 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
chr5:141338098&RNF14 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
chr1:22086638&HSPG2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
chr7:91795051&ANKIB1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
chr3:171681547&SLC7A14 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
chr2:27458920&PPM1G 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
chr14:67311570&ZFYVE26 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
chr4:119881302&SEC24D 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
chr10:75190063&SEC24C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
```

Log file output (Example):

```
Normal_1.filtered
Normal_10.filtered
Normal_2.filtered
Normal_3.filtered
Normal_4.filtered
Normal_5.filtered
Normal_6.filtered
Normal_7.filtered
Normal_8.filtered
Normal_9.filtered
Tumor_1.filtered
Tumor_10.filtered
Tumor_2.filtered
Tumor_3.filtered
Tumor_4.filtered
Tumor_5.filtered
Tumor_6.filtered
Tumor_7.filtered
Tumor_8.filtered
Tumor_9.filtered
```

A1.5 Detection of small indels from mapped RNA-seq data

A1.5.1 Bioinformatic methods to call point mutations from RNA-seq data

Another powerful application of RNA-seq is the ability to interrogate the raw sequence of mapped reads in order to search for small insertion/deletions (indels). As explained in detail in Section 1.3.3, small indels require the use of gapped alignment in order to be interrogated. There currently no existing standardized tools to call small indels from colorspace RNA-seq. But a non-standard custom pipeline exists that takes advantage of BioScope's gapped alignment capability in its DNA resequencing module. To perform this analysis, the whole transcriptome module is run as normal (described in Section A1.2.2). After its run, a three step process is run involving the use of three .ini files. The .ini files are run in the following order and are provided as part of the BioScope package: 1) example.smallIndelFrag.ini; 2) example.matobam.ini; 3) example.smallIndel.ini.

The first .ini file, example.smallIndelFrag.ini, uses the intermediate match file of the RNA-seq genomic mapping from the whole transcriptome pipeline. This file is found in the intermediate/s_mapping/genomic_map directory with the filename extension .csfasta.ma . This file also requires the original .qual file and the reference. This .ini file will perform the gapped alignments of the reads in order to find the indels. The output is written to the output directory with the file indel-evidence-list.pas. The second .ini file, example.matobam.ini will simply convert the gapped alignment output (.pas file) into BAM format. The third .ini file, example.smallIndel.ini, will take the BAM file (and actually multiple BAM files if a sample has multiple BAM files) and output the indels in spreadsheet format. The indels are then parsed exactly the same was as performed for point mutations as described in Section A1.4.1.

A1.5.2 Examples of files used in small indel calling

Filename: example.smallIndelFrag.ini

Description: This is a BioScope .ini file that will perform the gapped alignment needed for detecting small indels. The file uses an intermediate .csfasta.ma file from the genomic mapping portion of the whole transcriptome pipeline of BioScope. It also requires a .qual file, reference genome, and a cmap file that is simply the locations of individual chromosomal fasta files (see BioScope software for example).

Contents: (Example)

```
####################################
####################################
##
##  global parameters
##
import ./global.ini
reference = /N/home/mradovic/bioscope/BioScope-1.3.rBS131-
55029_20101119113500/etc/files/hg18/reference/human_hg18_female.fa
output.dir = ${base.dir}/output

run.name = Tumor_1_FC1
sample.name = Tumor_1_FC1
primer.set = F3
read.length = 50

mapping.output.dir = ${output.dir}/s_mapping

#
#      small indel fragment pipeline
#
small.indel.frag.run = 1

cmap = /N/home/mradovic/bioscope/BioScope-1.3.rBS131-
55029_20101119113500/etc/files/hg18/reference/human_hg18_female.cmap
small.indel.frag.match = /N/home/mradovic/TNBC_BioScope-
1.3_Results/FC1/Tumor_1_FC1/intermediate/s_mapping/genomic_map/Tumor_1_FC1.csfa
sta.ma
small.indel.frag.qual = /N/home/mradovic/TNBC_BioScope-
1.3_Results/FC1/Tumor_1_FC1/Tumor_1_FC1.qual

small.indel.frag.output.dir = ${output.dir}/smallIndelFrag
small.indel.frag.job.script.dir = ${output.dir}/smallIndelFrag/job
small.indel.frag.intermediate.dir = ${output.dir}/smallIndelFrag/intermediate
small.indel.frag.log.dir = ${base.dir}/smallIndelFrag-log-dir

#small.indel.frag.indel.preset =
```

Filename: example.matobam.ini

Description: This is a BioScope .ini file that will convert the gapped alignment output into a BAM file. This .ini file requires the .pas file, the original intermediate .csfasta.ma file, the reference genome and the .qual file.

Contents: (Example)

```
####################################
####################################
##
##  global parameters
##
import ./global.ini
output.dir = ${base.dir}/output
reference = /N/home/mradovic/bioscope/BioScope-1.3.rBS131-
55029_20101119113500/etc/files/hg18/reference/human_hg18_female.fa

run.name = Tumor_1_FC1
sample.name = Tumor_1_FC1
primer.set = F3
read.length = 50

#pipeline.cleanup.middle.files = 0
#job.cleanup.temp.files = 0


#
#     mapping pipeline
#
#mapping.output.dir = ${output.dir}/s_mapping/


#
#       ma to bam pipeline
#
ma.to.bam.run = 1
### depends on output of mapping
ma.to.bam.match.file = /N/home/mradovic/TNBC_BioScope-
1.3_Results/FC1/Tumor_1_FC1/intermediate/s_mapping/genomic_map/Tumor_1_FC1.csfa
sta.ma
ma.to.bam.pas.file = /N/home/mradovic/TNBC_BioScope-
1.3_Results/FC1/Tumor_1_FC1/output/smallIndelFrag/indel-evidence-list.pas
ma.to.bam.output.dir = ${output.dir}/maToBam

ma.to.bam.reference =  ${reference}
ma.to.bam.qual.file =  /N/home/mradovic/TNBC_BioScope-
1.3_Results/FC1/Tumor_1_FC1/Tumor_1_FC1.qual
```

Filename: example.smallIndel.ini

Description: This is a BioScope .ini file that will take the BAM files created using the example.matobam.ini and output the small indels in spreadsheet format. This .ini file requires the reference genome, the BAM file, and cmap file.

Contents: (Example)

```
####################################
####################################
##
##  global parameters
##
import ./global.ini
reference = /N/home/mradovic/bioscope/BioScope-1.3.rBS131-
55029_20101119113500/etc/files/hg18/reference/human_hg18_female.fa
output.dir = ${base.dir}/output

run.name = Tumor_1
sample.name = Tumor_1
primer.set = F3
read.length = 50


#
#      small indel pipeline
#
small.indel.run = 1

small.indel.bam.file = /N/home/mradovic/TNBC_BioScope-
1.3_Results/FC1/Tumor_1_FC1/output/maToBam/Tumor_1_FC1.csfasta.ma.bam,/N/home/m
radovic/TNBC_BioScope-
1.3_Results/FC2/Tumor_1_FC2/output/maToBam/Tumor_1_FC2.csfasta.ma.bam

small.indel.candidate.dir = ${output.dir}/smallIndel
cmap = /N/home/mradovic/bioscope/BioScope-1.3.rBS131-
55029_20101119113500/etc/files/hg18/reference/human_hg18_female.cmap


small.indel.log.dir = ${base.dir}/smallIndel-log-dir

#set to 1 for small test data
small.indel.min.num.evid = 1

small.indel.output.prefix=Tumor_1
```

A1.6 Detection of gene fusions from mapped RNA-seq data

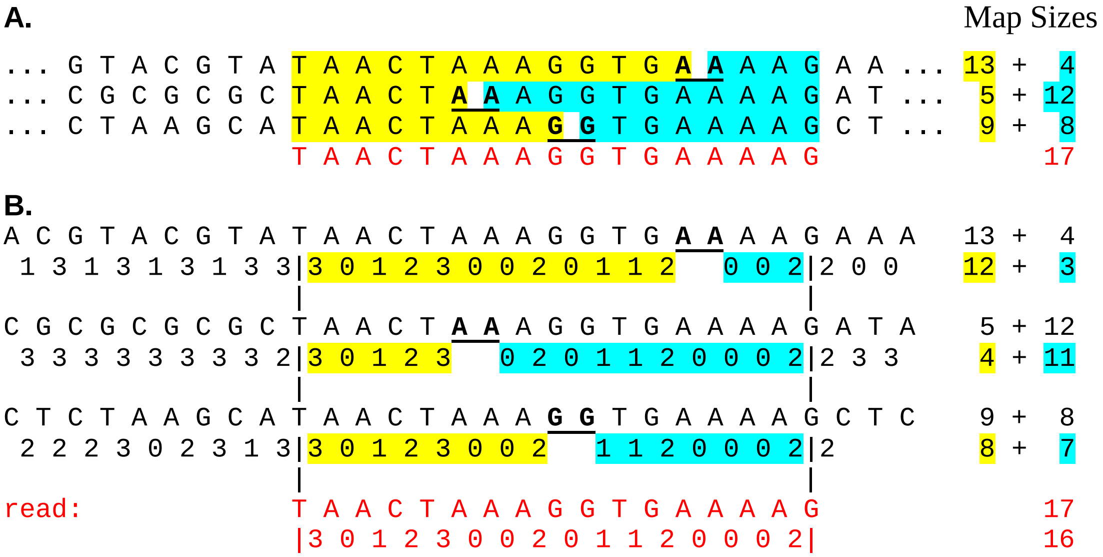
As described in Section 1.3.3, gene fusions are yet another output that can be interrogated from mapped RNA-seq data. To perform this detection, reads that map to exons of two different genes have to be identified. Only reads that partially mapped to the human genome in the original read mapping are used for gene fusion discovery. To reduce the search space for gene fusion junctions, only the ends of known exons derived from the RefSeq database are considered. Using the SASR Junction Finder (an plug-in of BioScope), putative fusion junctions are identified that have at least 2 supporting reads with 2 unique starting points (2 unique reads). To filter for false positives, any candidate fusion must only appear in the TNBC samples and not in the normals. The fusion output is further filtered by removing any gene & exon involved in more than 2 fusions as these are most likely false positives. In our final list, we consider only fusions that have at least 3 supporting reads (with at least 2 unique starting points).

Suffix Array Single Read (SASR) Junction Finder

A read provides evidence of a junction between an exon *e* and exon *f* if and only if (1) exon *e* maps to the prefix of the read, (2) exon f maps to the suffix of the read and (3) the sum of the two map lengths is equal to the length of the read (Figure 33A). SASR junction finder is adapted to work with di-base (color) reads. In this space, two fused exons introduce an additional color between them which does not map to reference genome. The color is not internal to either exon in the pair so there needs to be a plus 1 in condition (3) for color space (Figure 33B). For an exon to map to the prefix of the read, the exon must contain a suffix that starts with the prefix of the read. To avoid spurious maps, i.e. false positives, we require that, for an exon to map to a read, it must do so in at least 10 positions. The process for mapping exons to read suffixes is similar to that for mapping exons to prefixes of the reads. Once the list of exons that mapped to the prefix and suffix of the read are identified, it can be determined whether the read provides evidence for a unique junction. Multiple evidences with the same start position are not stored as separate evidence; only one of them is kept.

Running and parsing the gene fusion pipeline

As calling gene fusions is a plug-in of the BioScope whole transcritome pipeline, no special additional files are needed. The number of read evidences required to call a fusion are stipulated in the wt.single.read.workflow.ini file of the original transcriptome mapping. Fusion output is written to the standard BioScope output directory. The text files outputs can then be manipulated in Excel.



```
A.                                                                   Map Sizes
... G T A C G T A T A A C T A A A G G T G A A A A G A A ...   13 +  4
... C G C G C G C T A A C T A A A G G T G A A A A G A T ...    5 + 12
... C T A A G C A T A A C T A A A G G T G A A A A G C T ...    9 +  8
                  T A A C T A A A G G T G A A A A G               17

B.
A C G T A C G T A T A A C T A A A G G T G A A A A G A A A     13 +  4
 1 3 1 3 1 3 1 3 3|3 0 1 2 3 0 0 2 0 1 1 2   0 0 2|2 0 0      12 +  3
                  |                                |
C G C G C G C G C T A A C T A A A G G T G A A A A G A T A      5 + 12
 3 3 3 3 3 3 3 3 2|3 0 1 2 3   0 2 0 1 1 2 0 0 0 2|2 3 3       4 + 11
                  |                                |
C T C T A A G C A T A A C T A A A G G T G A A A A G C T C      9 +  8
 2 2 2 3 0 2 3 1 3|3 0 1 2 3 0 0 2   1 1 2 0 0 0 2|2           8 +  7
                  |                                |
read:             T A A C T A A A G G T G A A A A G               17
                  |3 0 1 2 3 0 0 2 0 1 1 2 0 0 0 2|               16
```

Figure 33. Schematic of the premise behind the SASR algorithm. **(A)** Three (artificial) exon pairs aligned with a read. The junction is marked in **<u>bold underline</u>**. The read serves as evidence for all three. The lengths of the aligning suffix and prefix from each pair must add up to the length of the read, 17 bases. **(B)** The color alignment corresponding to A. When two exons are spliced together, they introduce an additional color between them. This color is present in the read, but not in the aligned exons.

Appendix 2: Lists of differentially expressed genes

Please see supplemental file for lists of differentially expressed genes identified in this dissertation.

REFERENCES

1. Schneider, B.P., et al., *Triple-negative breast cancer: risk factors to potential targets.* Clin Cancer Res, 2008. **14**(24): p. 8010-8.
2. Dent, R., et al., *Triple-negative breast cancer: clinical features and patterns of recurrence.* Clin Cancer Res, 2007. **13**(15 Pt 1): p. 4429-34.
3. Carey, L.A., et al., *Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study.* JAMA, 2006. **295**(21): p. 2492-502.
4. Hudis, C.A. and L. Gianni, *Triple-negative breast cancer: an unmet medical need.* Oncologist, 2011. **16 Suppl 1**: p. 1-11.
5. Rouzier, R., et al., *Breast cancer molecular subtypes respond differently to preoperative chemotherapy.* Clin Cancer Res, 2005. **11**(16): p. 5678-85.
6. Liedtke, C., et al., *Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer.* J Clin Oncol, 2008. **26**(8): p. 1275-81.
7. Pal, S.K., B.H. Childs, and M. Pegram, *Triple negative breast cancer: unmet medical needs.* Breast Cancer Res Treat, 2011. **125**(3): p. 627-36.
8. Nielsen, T.O., et al., *Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma.* Clin Cancer Res, 2004. **10**(16): p. 5367-74.
9. Modi, S., et al., *A phase I study of cetuximab/paclitaxel in patients with advanced-stage breast cancer.* Clin Breast Cancer, 2006. **7**(3): p. 270-7.
10. O'Shaughnessy, J., et al., *Preliminary results of a randomized phase II study of weekly irinotecan/carboplatin with or without cetuximab in patients with metastatic breast cancer.* San Antonio Breast Cancer Symposium, 2007(Abstract 308).
11. Carey, L.A., et al., *TBCRC 001: EGFR inhibition with cetuximab added to carboplatin in metastatic triple-negative (basal-like) breast cancer.* J Clin Oncol (Meeting Abstracts), 2008. **26**(15_suppl): p. 1009-.
12. Baselga, J., et al., *Phase II and tumor pharmacodynamic study of gefitinib in patients with advanced breast cancer.* J Clin Oncol, 2005. **23**(23): p. 5323-33.
13. Modi, S., et al., *A phase II trial of imatinib mesylate monotherapy in patients with metastatic breast cancer.* Breast Cancer Res Treat, 2005. **90**(2): p. 157-63.
14. Finn, R.S., et al., *Phase II trial of dasatinib in triple-negative breast cancer: results of study CA180059.* Cancer Res (Meeting Abstracts), 2009. **69** (2 Suppl): p. Abstract nr 3118.
15. Miller, K., et al., *Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer.* N Engl J Med, 2007. **357**(26): p. 2666-76.
16. Burstein, H.J., et al., *Phase II study of sunitinib malate, an oral multitargeted tyrosine kinase inhibitor, in patients with metastatic breast cancer previously treated with an anthracycline and a taxane.* J Clin Oncol, 2008. **26**(11): p. 1810-6.
17. Turner, N., A. Tutt, and A. Ashworth, *Hallmarks of 'BRCAness' in sporadic cancers.* Nat Rev Cancer, 2004. **4**(10): p. 814-9.
18. Alli, E., et al., *Defective repair of oxidative dna damage in triple-negative breast cancer confers sensitivity to inhibition of poly(ADP-ribose) polymerase.* Cancer Res, 2009. **69**(8): p. 3589-96.
19. 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.

20. Foulkes, W.D., et al., *Germline BRCA1 mutations and a basal epithelial phenotype in breast cancer.* J Natl Cancer Inst, 2003. **95**(19): p. 1482-5.
21. Venkitaraman, A.R., *Cancer susceptibility and the functions of BRCA1 and BRCA2.* Cell, 2002. **108**(2): p. 171-82.
22. Farmer, H., et al., *Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy.* Nature, 2005. **434**(7035): p. 917-21.
23. Tutt, A., et al., *Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and advanced breast cancer: a proof-of-concept trial.* Lancet, 2010. **376**(9737): p. 235-44.
24. Fong, P.C., et al., *Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers.* N Engl J Med, 2009. **361**(2): p. 123-34.
25. O'Shaughnessy, J., et al., *Iniparib plus chemotherapy in metastatic triple-negative breast cancer.* N Engl J Med, 2011. **364**(3): p. 205-14.
26. Banerjee, S., S.B. Kaye, and A. Ashworth, *Making the best of PARP inhibitors in ovarian cancer.* Nat Rev Clin Oncol, 2010. **7**(9): p. 508-19.
27. Badve, S., et al., *Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists.* Mod Pathol, 2011. **24**(2): p. 157-67.
28. Foulkes, W.D., I.E. Smith, and J.S. Reis-Filho, *Triple-negative breast cancer.* N Engl J Med, 2010. **363**(20): p. 1938-48.
29. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.
30. Herschkowitz, J.I., et al., *Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors.* Genome Biol, 2007. **8**(5): p. R76.
31. Creighton, C.J., et al., *Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features.* Proc Natl Acad Sci U S A, 2009. **106**(33): p. 13820-5.
32. Carey, L., et al., *Triple-negative breast cancer: disease entity or title of convenience?* Nat Rev Clin Oncol, 2010. **7**(12): p. 683-92.
33. Gusterson, B., *Do 'basal-like' breast cancers really exist?* Nat Rev Cancer, 2009. **9**(2): p. 128-34.
34. Gusterson, B.A., et al., *Basal cytokeratins and their relationship to the cellular origin and functional classification of breast cancer.* Breast Cancer Res, 2005. **7**(4): p. 143-8.
35. Livasy, C.A., et al., *Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma.* Mod Pathol, 2006. **19**(2): p. 264-71.
36. Lim, E., et al., *Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers.* Nat Med, 2009. **15**(8): p. 907-13.
37. Molyneux, G., et al., *BRCA1 basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells.* Cell Stem Cell, 2010. **7**(3): p. 403-17.
38. Parker, J.S., et al., *Supervised risk predictor of breast cancer based on intrinsic subtypes.* J Clin Oncol, 2009. **27**(8): p. 1160-7.
39. Hall, J.M., et al., *Linkage of early-onset familial breast cancer to chromosome 17q21.* Science, 1990. **250**(4988): p. 1684-9.
40. Miki, Y., et al., *A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.* Science, 1994. **266**(5182): p. 66-71.
41. Wooster, R. and B.L. Weber, *Breast and ovarian cancer.* N Engl J Med, 2003. **348**(23): p. 2339-47.

42. Turner, N.C., et al., *BRCA1 dysfunction in sporadic basal-like breast cancer.* Oncogene, 2007. **26**(14): p. 2126-32.
43. Beger, C., et al., *Identification of Id4 as a regulator of BRCA1 expression by using a ribozyme-library-based inverse genomics approach.* Proc Natl Acad Sci U S A, 2001. **98**(1): p. 130-5.
44. Konstantinopoulos, P.A., et al., *Gene expression profile of BRCAness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer.* J Clin Oncol, 2010. **28**(22): p. 3555-61.
45. Wood, L.D., et al., *The genomic landscapes of human breast and colorectal cancers.* Science, 2007. **318**(5853): p. 1108-13.
46. Hu, X., et al., *Genetic alterations and oncogenic pathways associated with breast cancer subtypes.* Mol Cancer Res, 2009. **7**(4): p. 511-22.
47. Futreal, P.A., *Leveraging next generation sequencing in the study of breast cancer somatic genetics.* Oral presentation at the 2010 San Antonio Breast Cancer Symposium, 2010.
48. Craig, D.W., et al., *Whole-genome and transcriptome interrogation of metastatic chemo-resistant triple negative breast cancer from African American patients.* Abstract from the 2011 American Association for Cancer Research Annual Meeting, Abstract Number LB-267, 2011.
49. Stephens, P.J., et al., *Complex landscapes of somatic rearrangement in human breast cancer genomes.* Nature, 2009. **462**(7276): p. 1005-10.
50. Smalley, M.J., J.S. Reis-Filho, and A. Ashworth, *BRCA1 and stem cells: tumour typecasting.* Nat Cell Biol, 2008. **10**(4): p. 377-9.
51. Britt, K., A. Ashworth, and M. Smalley, *Pregnancy and the risk of breast cancer.* Endocr Relat Cancer, 2007. **14**(4): p. 907-33.
52. Smalley, M. and A. Ashworth, *Stem cells and breast cancer: A field in transit.* Nat Rev Cancer, 2003. **3**(11): p. 832-44.
53. Donegan, W.L. and J.S. Spratt, *Cancer of the breast*. 5th ed. 2002, Philadelphia: Saunders. xvi, 1050 p.
54. Pike, M.C., et al., *Estrogens, progestogens, normal breast cell proliferation, and breast cancer risk.* Epidemiol Rev, 1993. **15**(1): p. 17-35.
55. Wellings, S.R. and H.M. Jensen, *On the origin and progression of ductal carcinoma in the human breast.* J Natl Cancer Inst, 1973. **50**(5): p. 1111-8.
56. Shackleton, M., et al., *Generation of a functional mammary gland from a single stem cell.* Nature, 2006. **439**(7072): p. 84-8.
57. Ambaye, A.B., et al., *Carcinoma and atypical hyperplasia in reduction mammaplasty: increased sampling leads to increased detection. A prospective study.* Plast Reconstr Surg, 2009. **124**(5): p. 1386-92.
58. Clark, C.J., S. Whang, and K.T. Paige, *Incidence of precancerous lesions in breast reduction tissue: a pathologic review of 562 consecutive patients.* Plast Reconstr Surg, 2009. **124**(4): p. 1033-9.
59. Ishag, M.T., et al., *Pathologic findings in reduction mammaplasty specimens.* Am J Clin Pathol, 2003. **120**(3): p. 377-80.
60. Ramakrishnan, R., et al., *Pathologic findings in contralateral reduction mammaplasty specimens in patients with breast cancer.* Breast J, 2005. **11**(5): p. 372-3.
61. Tripathi, A., et al., *Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients.* Int J Cancer, 2008. **122**(7): p. 1557-66.
62. Graham, K., et al., *Gene expression profiles of estrogen receptor-positive and estrogen receptor-negative breast cancers are detectable in histologically normal breast epithelium.* Clin Cancer Res, 2011. **17**(2): p. 236-46.

63. Yan, P.S., et al., *Mapping geographic zones of cancer risk with epigenetic biomarkers in normal breast tissue.* Clin Cancer Res, 2006. **12**(22): p. 6626-36.
64. Deng, G., et al., *Loss of heterozygosity in normal tissue adjacent to breast carcinomas.* Science, 1996. **274**(5295): p. 2057-9.
65. Lakhani, S.R., et al., *Genetic alterations in 'normal' luminal and myoepithelial cells of the breast.* J Pathol, 1999. **189**(4): p. 496-503.
66. Mardis, E.R., *A decade's perspective on DNA sequencing technology.* Nature, 2011. **470**(7333): p. 198-203.
67. Meyerson, M., S. Gabriel, and G. Getz, *Advances in understanding cancer genomes through second-generation sequencing.* Nat Rev Genet, 2010. **11**(10): p. 685-96.
68. Leary, R.J., et al., *Development of personalized tumor biomarkers using massively parallel sequencing.* Sci Transl Med, 2010. **2**(20): p. 20ra14.
69. Maher, C.A., et al., *Transcriptome sequencing to detect gene fusions in cancer.* Nature, 2009. **458**(7234): p. 97-101.
70. Maher, C.A., et al., *Chimeric transcript discovery by paired-end transcriptome sequencing.* Proc Natl Acad Sci U S A, 2009. **106**(30): p. 12353-8.
71. Shah, S.P., et al., *Mutation of FOXL2 in granulosa-cell tumors of the ovary.* N Engl J Med, 2009. **360**(26): p. 2719-29.
72. Wiegand, K.C., et al., *ARID1A mutations in endometriosis-associated ovarian carcinomas.* N Engl J Med, 2010. **363**(16): p. 1532-43.
73. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
74. Inaki, K., et al., *Transcriptional consequences of genomic structural aberrations in breast cancer.* Genome Res, 2011. **21**(5): p. 676-87.
75. Tuch, B.B., et al., *Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations.* PLoS One, 2010. **5**(2): p. e9317.
76. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
77. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.
78. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.
79. Goya, R., et al., *SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors.* Bioinformatics, 2010. **26**(6): p. 730-6.
80. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
81. 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.
82. Katz, Y., et al., *Analysis and design of RNA sequencing experiments for identifying isoform regulation.* Nat Methods, 2010. **7**(12): p. 1009-15.
83. Bullard, J.H., et al., *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.* BMC Bioinformatics, 2010. **11**: p. 94.
84. Auer, P.L. and R.W. Doerge, *Statistical design and analysis of RNA sequencing data.* Genetics, 2010. **185**(2): p. 405-16.
85. 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.

86. Aiyar, S.E., et al., *Attenuation of estrogen receptor alpha-mediated transcription through estrogen-stimulated recruitment of a negative elongation factor.* Genes Dev, 2004. **18**(17): p. 2134-46.
87. Ye, Q., et al., *BRCA1-induced large-scale chromatin unfolding and allele-specific effects of cancer-predisposing mutations.* J Cell Biol, 2001. **155**(6): p. 911-21.
88. Aiyar, S.E., et al., *Concerted transcriptional regulation by BRCA1 and COBRA1 in breast cancer cells.* Int J Biol Sci, 2007. **3**(7): p. 486-92.
89. Wagner, J.M. and S.H. Kaufmann, *Prospects for the Use of ATR Inhibitors to Treat Cancer.* Pharmaceuticals, 2010(3): p. 1311-1334.
90. Gschwind, A., O.M. Fischer, and A. Ullrich, *The discovery of receptor tyrosine kinases: targets for cancer therapy.* Nat Rev Cancer, 2004. **4**(5): p. 361-70.
91. Shepherd, F.A., et al., *Erlotinib in previously treated non-small-cell lung cancer.* N Engl J Med, 2005. **353**(2): p. 123-32.
92. Piccart-Gebhart, M.J., et al., *Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer.* N Engl J Med, 2005. **353**(16): p. 1659-72.
93. Motzer, R.J., et al., *Sunitinib versus interferon alfa in metastatic renal-cell carcinoma.* N Engl J Med, 2007. **356**(2): p. 115-24.
94. Robinson, D.R., Y.M. Wu, and S.F. Lin, *The protein tyrosine kinase family of the human genome.* Oncogene, 2000. **19**(49): p. 5548-57.
95. Mossie, K., et al., *Colon carcinoma kinase-4 defines a new subclass of the receptor tyrosine kinase family.* Oncogene, 1995. **11**(10): p. 2179-84.
96. Speers, C., et al., *Identification of novel kinase targets for the treatment of estrogen receptor-negative breast cancer.* Clin Cancer Res, 2009. **15**(20): p. 6327-40.
97. Meng, L., et al., *Silencing of PTK7 in colon cancer cells: caspase-10-dependent apoptosis via mitochondrial pathway.* PLoS One, 2010. **5**(11): p. e14018.
98. Prebet, T., et al., *The cell polarity PTK7 receptor acts as a modulator of the chemotherapeutic response in acute myeloid leukemia and impairs clinical outcome.* Blood, 2010. **116**(13): p. 2315-23.
99. Lu, X., et al., *PTK7/CCK-4 is a novel regulator of planar cell polarity in vertebrates.* Nature, 2004. **430**(6995): p. 93-8.
100. Paudyal, A., et al., *The novel mouse mutant, chuzhoi, has disruption of Ptk7 protein and exhibits defects in neural tube, heart and lung development and abnormal planar cell polarity in the ear.* BMC Dev Biol, 2010. **10**: p. 87.
101. Prat, A., et al., *Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer.* Breast Cancer Res, 2010. **12**(5): p. R68.
102. Yang, X.H., et al., *Overexpression of the receptor tyrosine kinase Tie-1 intracellular domain in breast cancer.* Tumour Biol, 2003. **24**(2): p. 61-9.
103. Maher, M.G., et al., *Prognostic significance of colony-stimulating factor receptor expression in ipsilateral breast cancer recurrence.* Clin Cancer Res, 1998. **4**(8): p. 1851-6.
104. Roskoski, R., Jr., *Sunitinib: a VEGF and PDGF receptor protein kinase and angiogenesis inhibitor.* Biochem Biophys Res Commun, 2007. **356**(2): p. 323-8.
105. Kumar, S.R., et al., *Receptor tyrosine kinase EphB4 is a survival factor in breast cancer.* Am J Pathol, 2006. **169**(1): p. 279-93.
106. Fox, B.P. and R.P. Kandpal, *EphB6 receptor significantly alters invasiveness and other phenotypic characteristics of human breast carcinoma cells.* Oncogene, 2009. **28**(14): p. 1706-13.
107. Tsunoda, N., et al., *Nek2 as a novel molecular target for the treatment of breast carcinoma.* Cancer Sci, 2009. **100**(1): p. 111-6.

108. Hayward, D.G., et al., *The centrosomal kinase Nek2 displays elevated levels of protein expression in human breast cancer.* Cancer Res, 2004. **64**(20): p. 7370-6.
109. Hayward, D.G. and A.M. Fry, *Nek2 kinase in chromosome instability and cancer.* Cancer Lett, 2006. **237**(2): p. 155-66.
110. Hayward, D.G., et al., *Identification by high-throughput screening of viridin analogs as biochemical and cell-based inhibitors of the cell cycle-regulated nek2 kinase.* J Biomol Screen, 2010. **15**(8): p. 918-27.
111. 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.
112. Calin, G.A., et al., *Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas.* Cancer Cell, 2007. **12**(3): p. 215-29.
113. Garcia, A.I., et al., *Down-regulation of BRCA1 expression by miR-146a and miR-146b-5p in triple negative sporadic breast cancers.* EMBO Mol Med, 2011.
114. Kulshreshtha, R., et al., *A microRNA signature of hypoxia.* Mol Cell Biol, 2007. **27**(5): p. 1859-67.
115. Liu, C.J., et al., *miR-31 ablates expression of the HIF regulatory factor FIH to activate the HIF pathway in head and neck carcinoma.* Cancer Res, 2010. **70**(4): p. 1635-44.
116. Wu, H., S. Zhu, and Y.Y. Mo, *Suppression of cell growth and invasion by miR-205 in breast cancer.* Cell Res, 2009. **19**(4): p. 439-48.
117. Zhao, J.J., et al., *MicroRNA-221/222 negatively regulates estrogen receptor alpha and is associated with tamoxifen resistance in breast cancer.* J Biol Chem, 2008. **283**(45): p. 31079-86.
118. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
119. Thierry-Mieg, D. and J. Thierry-Mieg, *AceView: a comprehensive cDNA-supported gene and transcripts annotation.* Genome Biol, 2006. **7 Suppl 1**: p. S12 1-14.
120. 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.
121. Shah, S.P., et al., *Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution.* Nature, 2009. **461**(7265): p. 809-13.
122. Antoniou, A.C., et al., *A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population.* Nat Genet, 2010. **42**(10): p. 885-92.
123. Kan, Z., et al., *Diverse somatic mutation patterns and pathway alterations in human cancers.* Nature, 2010. **466**(7308): p. 869-73.
124. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.
125. Kickhoefer, V.A., et al., *The 193-kD vault protein, VPARP, is a novel poly(ADP-ribose) polymerase.* J Cell Biol, 1999. **146**(5): p. 917-28.
126. Liu, Y., et al., *Vault poly(ADP-ribose) polymerase is associated with mammalian telomerase and is dispensable for telomerase function and vault structure in vivo.* Mol Cell Biol, 2004. **24**(12): p. 5314-23.

127. Raval-Fernandes, S., et al., *Increased susceptibility of vault poly(ADP-ribose) polymerase-deficient mice to carcinogen-induced tumorigenesis.* Cancer Res, 2005. **65**(19): p. 8846-52.
128. Berger, W., et al., *Vaults and the major vault protein: novel roles in signal pathway regulation and immunity.* Cell Mol Life Sci, 2009. **66**(1): p. 43-61.
129. Lara, P.C., et al., *Severe hypoxia induces chemo-resistance in clinical cervical tumors through MVP over-expression.* Radiat Oncol, 2009. **4**: p. 29.
130. Mossink, M.H., et al., *Vaults: a ribonucleoprotein particle involved in drug resistance?* Oncogene, 2003. **22**(47): p. 7458-67.
131. Lloret, M., et al., *Major vault protein may affect nonhomologous end-joining repair and apoptosis through Ku70/80 and bax downregulation in cervical carcinoma tumors.* Int J Radiat Oncol Biol Phys, 2009. **73**(4): p. 976-9.
132. Yu, X., et al., *The BRCT domain is a phospho-protein binding domain.* Science, 2003. **302**(5645): p. 639-42.
133. Bryant, P.J., et al., *Mutations at the fat locus interfere with cell proliferation control and epithelial morphogenesis in Drosophila.* Dev Biol, 1988. **129**(2): p. 541-54.
134. Nakaya, K., et al., *Identification of homozygous deletions of tumor suppressor gene FAT in oral cancer using CGH-array.* Oncogene, 2007. **26**(36): p. 5300-8.
135. Forbes, S.A., et al., *The Catalogue of Somatic Mutations in Cancer (COSMIC).* Curr Protoc Hum Genet, 2008. **Chapter 10**: p. Unit 10 11.
136. Forbes, S.A., et al., *COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.* Nucleic Acids Res, 2011. **39**(Database issue): p. D945-50.
137. Forbes, S.A., et al., *COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer.* Nucleic Acids Res, 2010. **38**(Database issue): p. D652-7.
138. Mitelman, F., B. Johansson, and F. Mertens, *The impact of translocations and gene fusions on cancer causation.* Nat Rev Cancer, 2007. **7**(4): p. 233-45.
139. Shendure, J. and C.J. Stewart, *Cancer genomes on a shoestring budget.* N Engl J Med, 2009. **360**(26): p. 2781-3.
140. Weigelt, B., et al., *Breast cancer molecular profiling with single sample predictors: a retrospective analysis.* Lancet Oncol, 2010. **11**(4): p. 339-49.

CURRICULUM VITAE
Milan Radovich

EDUCATION

Graduate:	Indiana University, IUPUI, Indianapolis, IN (August 2011) Major: Medical & Molecular Genetics, Doctor of Philosophy Minor: Cancer Biology
Undergraduate:	Purdue University, West Lafayette, IN, (May 2004) Major: Biochemistry, Bachelors of Science Minors: Religious Studies, History

PROFESSIONAL ORGANIZATIONS

1) American Society of Clinical Oncology (ASCO), Affiliate Member. 2007 – Present

2) American Association for Cancer Research (AACR), Associate Member. 2008 – Present

HONORS

1) Indiana University Simon Cancer Center 2010 AACR Travel Award. Awarded 04/14/2010.

2) AACR-Aflac Scholar-in-Training Award for 2010 AACR Annual Meeting. Awarded 04/16/2010.

3) 1st Place – Translational/Clinical Research by Graduate Student. Indiana University Simon Cancer Center Annual Cancer Research Day. Awarded 05/05/2010.

4) AACR Translational Research Scholarship from Susan G. Komen for the Cure for the 2010 San Antonio Breast Cancer Symposium. Awarded 12/08/2010.

5) Indiana University Simon Cancer Center 2011 AACR Travel Award. Awarded 03/23/2011.

PATENTS

1) PCT/US2008/084933 (Filed Nov 26, 2008). “VEGF Polymorphisms and Anti-Angiogenesis Therapy.” Inventors: Bryan P. Schneider, **Milan Radovich**, George W. Sledge.

PREVIOUS SCIENTIFIC WORK EXPERIENCE

1) Undergraduate Research Assistant. Purdue University Cancer Center, Dept. of Biochemistry, West Lafayette, IN. October 2000 – May 2004

2) Research Technician. Indiana University School of Medicine, Dept. of Microbiology & Immunology, Indianapolis, IN. October 2004 – June 2005

3) Lab Director. Indiana University School of Medicine, Div. of Hematology/Oncology, Indianapolis, IN. June 2005 – August 2008

CERTIFICATIONS

1) Bioinformatics on the Applied Biosystems SOLiD Next Generation Sequencer (November 2008)
   Applied Biosystems, Inc. Foster City, CA.

FUNDING

1) 07/01/2009 – 06/30/2010. Cancer Biology Training Program Fellowship Trainee. National Cancer Institute. NRSA 1 T32 CA111198.

2) 07/01/2010 – 06/30/2011. Indiana Clinical & Translational Sciences Institute Pre-doctoral Fellowship. National Institutes of Health. NIH TL1 RR025759

SEMINARS & INVITED TALKS

1) “Cancer Research 101.” Invited Speaker, Decatur Township American Cancer Society Relay for Life Volunteer Reception. Valley Mills Christian Church, Indianapolis, IN. 10/05/2007.

2) “Genetic Variability in Angiogenesis Genes on Disease Risk, Function, & Pharmacogenetics.” Hematology/Oncology Research Meeting. Indiana University School of Medicine, Indianapolis, IN. 12/13/2007.

3) “Finding New Ways to Attack Triple-Negative Breast Cancer.” Seminar to the Vera Bradley Foundation for Breast Cancer Research. Indiana University School of Medicine, Indianapolis, IN. 02/11/2010.

4) “Next-Generation Transcriptome Sequencing of Triple-Negative Breast Tumors and Normal Tissues.” The Amelia Project – Giving Wings to Research. Indianapolis, IN. 02/13/2010.

5) “Decoding the Transcriptional Landscape of Cancer using Next-Generation Sequencing.” Indiana University Simon Cancer Center Grand Rounds. Indianapolis, IN. 04/23/2010.

6) “Next-Generation Whole Transcriptome Sequencing in Triple-Negative Breast Cancer.” Dept. of Medical & Molecular Genetics Seminar. Indiana University School of Medicine. Indianapolis, IN. 05/04/2010.

7) “Decoding the Transcriptional Landscape of Cancer using Next Generation Sequencing.” Life Technologies, Inc. Foster City, CA. 07/19/2010.

8) “Cancer Genomics at AACR 2010.” Indiana University Simon Cancer Center Grand Rounds. Indianapolis, IN. 08/20/2010.

9) “Personalized Medicine for Triple Negative Breast Cancer – New Dimensions in Therapeutic Individualization.” Metastatic Breast Cancer Network 2010 Conference. Indianapolis, IN. 10/16/2010.

10) “Decoding the Transcriptional Landscape of Triple-Negative Breast Cancer Using Next-Generation Sequencing.” Advances and Applications in Cancer & Medical Research Using Next-Generation Sequencing. Northwestern University Center for Genetic Medicine. Chicago, IL. 10/20/2010.

11) “Decoding the Transcriptional Landscape of Thymic Malignancies Using Next-Generation Sequencing.” International Thymic Malignancy Interest Group Workshop. Yale University. New Haven, Connecticut. 11/15/2010.

12) “Decoding the Transcriptional Landscape of Triple-Negative Breast Cancer Using Next-Generation Sequencing.” Sequencing at Tipping Point Meeting. San Diego, CA. 12/01/2010.

13) “Decoding the Transcriptional Landscape of Triple-Negative Breast Cancer Using Next-Generation Sequencing.” Eli Lilly Corporation. Indianapolis, IN. 12/06/2010.

MEETING PRESENTATIONS

1) Poster: “Analysis of angiogenesis genes from paraffin-embedded breast tumor and lymph nodes.” Bryan Schneider, **Milan Radovich**, Todd Skaar, Sunil Badve, George Sledge, Lang Li, and David Flockhart. *Pharmacogenetics Research Network Meeting.* St. Louis, MO. December 2005.

2) Poster: “Exploratory study evaluating the association of polymorphisms of angiogenesis genes with hot flashes.” Bryan P Schneider, MD, **Milan Radovich**, David A Flockhart, MD,PhD, Janet Carpenter, PhD,RN, Lang Li, PhD, Jason Robarge, Anna M Storniolo, MD, Suzanne Lemler, RN, Anne Nguyen, Todd Skaar, PhD and George W Sledge, MD. *San Antonio Breast Cancer Symposium. San Antonio, TX. December 2006.* Published in: *Breast Cancer Research and Treatment* (2006 Dec) 100(S1): S241.

3) Poster: “Association of genetic polymorphisms of angiogenesis genes and breast cancer risk.” Bryan P Schneider, MD, **Milan Radovich**, George W Sledge, MD, Lang Li, PhD, Jason D Robarge, Todd Skaar, PhD, Anna M

Storniolo, MD and David A Flockhart, MD, PhD. *San Antonio Breast Cancer Symposium. San Antonio, TX. December 2006.* Published in: *Breast Cancer Research and Treatment* (2006 Dec) 100(S1): S241.

4) Poster: "Association of genetic polymorphisms of VEGF and VEGFR-2 with outcome in E2100." Bryan Schneider, Molin Wang, **Milan Radovich**, George Sledge, Sunil Badve, Ann Thor, David Flockhart, Bradley Hancock, Nancy Davidson, Kathy Miller. *San Antonio Breast Cancer Symposium. San Antonio, TX. December 2007.* Published in: *Breast Cancer Research and Treatment* (2007 Dec) 106(S1): S78.

5) Poster: "Association of single nucleotide polymorphisms in angiogenesis genes with expression of VEGF and VEGFR-2 in breast carcinoma." Bradley A. Hancock, **Milan Radovich**, Sunil Badve, Mangesh A. Thorat, Faouzi Azzouz, Bryan P. Schneider. *American Association for Cancer Research 2008 Annual Meeting.* San Diego, CA. April 2008.

6) Poster: "Association of single nucleotide polymorphisms in angiogenesis genes with expression of VEGF and VEGFR-2 in breast carcinoma." Bradley A. Hancock, **Milan Radovich**, Sunil Badve, Mangesh A. Thorat, Faouzi Azzouz, Bryan P. Schneider. *Pharmacogenetics Research Network Meeting.* Nashville, TN. April 2008.

7) Poster: "Association of genetic polymorphisms of VEGF and VEGFR-2 with outcome in E2100." Bryan Schneider, Molin Wang, **Milan Radovich**, George Sledge, Sunil Badve, Ann Thor, David Flockhart, Bradley Hancock, Nancy Davidson, Kathy Miller. *Pharmacogenetics Research Network Meeting.* Nashville, TN. April 2008.

8) Poster: "Resequencing of the Vascular Endothelial Growth Factor Promoter Reveals Haplotype Structure and Functional Diversity." Brad A. Hancock, **Milan Radovich**, Nawal Kassem, Deming Mi, Todd C. Skaar, Bryan P. Schneider. *San Antonio Breast Cancer Symposium.* San Antonio, TX. December 2009.

9) Poster: "Next-Generation Whole Transcriptome Sequencing of Triple-Negative Breast Tumors and Normal Tissues." **Milan Radovich**, Susan E. Clare, Ivanesa Pardo, Bradley A. Hancock, George W. Sledge, Connie Rufenbarger, Anna Maria V. Storniolo, Theresa Mathieson, Jie Sun, Jill E. Henry, Eric E. Hilligoss, James S. Elliott, Ryan Richt, Matthew Hickenbotham, Jarret Glasscock, Yunlong Liu, Bryan P. Schneider. *San Antonio Breast Cancer Symposium.* San Antonio, TX. December 2009.

10) Poster: "Next-Generation whole transcriptome sequencing of triple-negative breast tumors and normal tissues." **Milan Radovich**, Susan E. Clare, Ivanesa Pardo, Bradley A. Hancock, Nawal Kassem, George W. Sledge, Connie Rufenbarger, Anna Maria V. Storniolo, Theresa Mathieson, Jie Sun, Jill E. Henry, Heather A. Lillemoe, Eric E. Hilligoss, James S. Elliott, Ryan Richt, Matthew Hickenbotham, Jarret Glasscock, Yunlong Liu, Bryan P. Schneider. *American Association for Cancer Research 2010 Annual Meeting.* Washington, D.C. April 2010.

11) Poster Discussion: “Decoding the Transcriptional Landscape of Triple-Negative Breast Cancer Using Next-Generation Whole Transcriptome Sequencing.” **Milan Radovich**, Susan E. Clare, George W. Sledge, Ivanesa Pardo, Theresa Mathieson, Nawal Kassem, Bradley A. Hancock, Anna Maria V. Storniolo, Connie Rufenbarger, Heather A. Lillemoe, Jie Sun, Jill E. Henry, Robert Goulet, Eric E. Hilligoss, Asim S. Siddiqui, Heinz Breu, Onur Sakarya, Fiona C. Hyland, Matthew W. Muller, Liviu Popescu, Jin Zhu, Matthew Hickenbotham, Jarret Glasscock, Mircea Ivan, Yunlong Liu, Bryan P. Schneider. *San Antonio Breast Cancer Symposium.* San Antonio, TX. December 2010.

12) Poster: “Next-generation whole transcriptome sequencing of thymic malignancies.” **Milan Radovich**, Bradley A. Hancock, Nawal Kassem, Jin Zhu, Jarret Glasscock, Sunil Badve, Yunlong Liu, Kenneth A. Kesler, Patrick J. Loehrer, Bryan P. Schneider. *American Association for Cancer Research 2011 Annual Meeting*. Orlando, FL. April 2011.

PUBLICATIONS

1) Pei Zhang, Olivier Schwartz, Milica Pantelic, Geling Li, Quita Knazze, Cinzia Nobile, **Milan Radovich**, Johnny He, Soon-Cheol Hong, John Klena, and Tie Chen. “DC-SIGN (CD209) recognition of *Neisseria gonorrhoeae* is circumvented by lipooligosaccharide variation.” *Journal of Leukocyte Biology.* 2006 Apr;79(4): 731-8. Epub 2006 Feb 3.

2) Bryan P. Schneider, **Milan Radovich**, George W. Sledge, Jason D. Robarge, Lang Li, Anna M. Storniolo, Suzanne Lemler, Anne T. Nguyen, Bradley A. Hancock, Michael Stout, Todd Skaar, David A. Flockhart. “Association of polymorphisms of angiogenesis genes with breast cancer.” *Breast Cancer Research and Treatment.* 2008 Sep;111(1):157-63. Epub 2007 Sep 20.

3) Bryan P. Schneider, **Milan Radovich**, David A. Flockhart, Janet S. Carpenter, Lang Li, Jason D. Robarge, Anna M. Storniolo, Bradley A. Hancock, Todd C. Skaar, George W. Sledge. “Exploratory study evaluating the association of polymorphisms of angiogenesis genes with hot flashes.” *Breast Cancer Research and Treatment.* 2009 Aug;116(3):543-9. Epub 2008 Sep 11.

4) Bryan P. Schneider, Molin Wang, **Milan Radovich**, George W. Sledge, Sunil Badve, Ann Thor, David A. Flockhart, Bradley Hancock, Nancy Davidson, Julie Gralow, Maura Dickler, Edith A. Perez, Melody Cobleigh, Tamara Shenkier, Susan Edgerton, Kathy D. Miller. “Association of Vascular Endothelial Growth Factor and Vascular Endothelial Growth Factor Receptor-2 Genetic Polymorphisms With Outcome in a Trial of Paclitaxel Compared with Paclitaxel Plus Bevacizumab in Advanced Breast Cancer: ECOG 2100.” *Journal of Clinical Oncology.* 2008 Oct 1;26(28):4672-8.

5) Jia Miao, Yan Jin, Rita L. Marunde, Seongho Kim, Sara K. Quinney, **Milan Radovich,** Lang Li, Stephen D. Hall. “Association of genotypes of the CYP3A

cluster with midazolam disposition in vivo." *The Pharmacogenomics Journal.* 2009 Oct;9(5):319-26. Epub 2009 Jun 9.

6) Xin Wang, Kejun Wang, **Milan Radovich**, Guohua Wang, Weixing Feng, Jeremy R. Sanford, and Yunlong Liu. "Genome-wide prediction of *cis*-acting RNA elements regulating tissue-specific pre-mRNA alternative splicing." *BMC Genomics.* 2009 Jul 7;10 Suppl 1:S4.

7) Bryan P. Schneider, **Milan Radovich**, Kathy D. Miller. "The Role of VEGF Genetic Variability in Cancer." *Clinical Cancer Research.* 2009 Sep 1;15(17):5297-302. Epub 2009 Aug 25.

8) **Milan Radovich,** Bradley A. Hancock, Nawal Kassem, Deming Mi, Todd C. Skaar, Bryan P. Schneider. "Resequencing of the Vascular Endothelial Growth Factor promoter reveals haplotype structure and functional diversity." *Angiogenesis*. 2010 Sep;13(3):211-8. Epub 2010 Jun 16.