Sparsity and scarcity:
Multiomic studies in a low resource setting
(A study in archival FFPE cancer tissue)

by
Soonweng Cho

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
October, 2016

Abstract

Archival formalin-fixed, paraffin-embedded (FFPE) tissues are an invaluable resource for biomarker discovery due to their vast number in pathology laboratories and availability of high-quality, long-term clinical information. These materials are even more precious for studies on diseases with rare events and long time-to-event intervals. However, these tissues are usually available in limited amounts and formalin treatment of tissue rendersrecovery of nucleic acids difficult, leading to a setting with low resource. At the time of writing, no robust methodologies have been developed for genomic analysis of FFPE material. To that end, this thesis addresses fundamental questions in genomic profiling in a low resource setting, identified best practice workflows for doing so, and demonstrated their application in three different tumor types.

The first section of this dissertation, spanning three chapters, considers the technical challenges and extends the repertoire of methods in performing genomics analysis in low resource settings. We begin by reviewing the nucleic acid modifications and technological advancements for molecular profiling of FFPE tissues. We developed workflows for genomic analysis of FFPE materials and showed application of these findings in a series of microarray experiments. Extending the tools for genomics analysis in a low resource setting, we developed Epicopy, a method to obtain copy number variation (CNV) data from Illumina 450K methylation microarrays and demonstrated its ability to make concordant CNV calls. We also showed comparable, if not better, performance by Epicopy compared to two previously published methods, CHAMP-CNV and CopyNumber450K. We next validated the use of these methods in answering biological questions across three tumor types.

First, we analyzed a series of ductal carcinoma in situ (DCIS) samples in the context of disease progression into IDC. We observed the presence of global methylation changes in DCIS-adjacent normal tissue and identified the presence of four epitypes of DCIS samples, associated with grade and a CIMP-like phenotype. The CNV profiles of these DCIS samples reflect those of previous studies and differential CNV changes were identified between DCIS that progressed to invasive cancer and DCIS that did not recur.

Second, multiomic analysis performed on a series of ER-negative breast cancers identified three functionally relevant clusters of androgen receptor-driven, immune infiltration high, and CNV rich disease. In the clinical context of recurrence in the absence of adjuvant chemotherapy, we discovered a 130-gene panel of markers that was able to predict recurrence in both the institutional ER-negative cohort and validated it in an independent external dataset.

Finally, in a total RNA-seq profiling of follicular thyroid cancer (FTC) to study molecular landscapes associated with distant metastasis, we identified a set of biomarkers of distant metastasis, enriched in epithelial-mesenchymal transition genes, and predicted a distant metastasis event in an FTC tumor, 10-years before the fact. These genes were validated in the TCGA thyroid cancer dataset for their ability to predict distant metastasis in follicular variant papillary thyroid cancer (FVPTC), a tumor type which molecularly resembles FTC.

Taken together, this work establishes our ability to molecularly profile FFPE tissues using microarray and NGS platforms, unlocking the potential of using archival materials with high quality clinical follow-up information to address important clinical questions.

Thesis Advisors

Saraswati Sukumar, Ph.D.

Christopher B. Umbricht, M.D., Ph.D.

Leslie M. Cope, Ph.D.

Thesis Committee Members

Edward Gabrielson, M.D.

Luigi Marchionni, M.D., Ph.D.

Acknowledgements

The body of work presented in this dissertation would not be possible without an amazing team of scientists, clinicians, and, most importantly, patients. My growth as an independent researcher would not be possible without my mentors and peers. My time in graduate school would not be as rich without my friends and family. Words alone are not enough to express my heartfelt gratitude to everyone.

First and foremost, I would like to extend my gratitude to my advisors; Saraswati Sukumar, Christopher Umbricht, and Leslie Cope for their unwavering support, patience, and guidance through my Ph.D. studies. Sara has the uncanny ability to recognize one's potential, and will work tirelessly to bring your best self forward. Chris, between his humor and his adeptness at asking fundamental questions, does everything he can, even through lean times, to provide the resources, expertise, and collaborations necessary for us to achieve our goals in lab and beyond. Leslie, amid the witty philosophical digressions, is an amazing statistician and scientist who have taught me everything I know about statistics and programming.

My advisors collectively taught me how to perform comprehensive interdisciplinary scientific inquiries, often by example and drove me to strive for the scientific rigor and standards they hold. They ask difficult questions, critically assess ideas, while maintaining a hospitable atmosphere. Aristotle once said, "the whole is greater than the sum of all parts", and, in my case, they are giants in their field. Thank you all for being such a great team of tremendous mentors.

I would also like to extend my thanks to Mary Jo Fackler, "Mama Jo", who has been instrumental in many parts of my thesis and scientific career. I am grateful for your

willingness to go out of your way to help, wisdom, and infectious humor. I also thank Martha Zeiger, whose advice on my work is always appreciated. My gratitude extends to Liliana Florea, whose advice and assistance are critical for many studies. This work would also not be possible without my thesis committee members, Edward Gabrielson and Luigi Marchionni. Ed's clinicopathological ideas and patience in teaching me breast histopathology enriched my training as a scientist. Luigi devises the most innovative analysis schemes for critical genomics questions and is always generous with his code.

Beyond that, I would also like to thank Antonio Wolff, Kala Visvanathan, Ashley Cimino-Matthews, and Justin Bishop whose help and clinical discussions were important in understanding the biology of disease. I also wish to extend my appreciation to Kristen Wagner-Smith, Crystal Graham, and Cindy Morin, whose swift work and exceptional support made graduate school a smoother journey.

Equally important are our collaborators. I thank Charles Lynch and Freda Selk of University of Iowa; William Grizzle of University of Alabama Birmingham; Ann Hamilton of University of Southern California; Jeff Marks of Duke University; Brenda Hernandez of University of Hawaii; Rosita Camilla, Meridith Reagan, and Giuseppe Viale of the International Breast Cancer Surgical Group, for their exceptional work and discussion in the identification and collection of patient samples for the various studies. I thank Wayne Yu, Jinshui Fan, Connie Talbot, and Haiping Hao from the core facilities at Johns Hopkins. I extend my gratitude to Charles Perou, Katie Hoadley, Neil Hayes, and Matthew Revilla of University of North Carolina, Chapel Hill for RNA-seq expertise.

My appreciation also goes out to my peers in the Sukumar lab. I thank Helen Sadik, whose brilliant talk prodded me to interview with Sara. Wei Wen Teo, thank you

for all the advice and stimulating scientific discussions. Thank you Vanessa Merino, whose scientific brilliance is matched by her diligence and resolve. Leigh-Ann Cruz, thank you for always being there, all the advice, and pop culture references. Thank you Liangfen Han, Kideok Jin, Preethi Korangath, Sunju Park, Bradley Downs, and Danielle Meir-Levi for discussions and help.

I extend this deep gratitude to members of the Umbricht lab. I thank Nivedita Chowdhury for her significant help with all the projects; Brandon Kim, who developed Epicopy with me; Yongchun Wang, for all the scientific advice; Kathleen Wilsbach, thank you for your work with the TNBC project. I also thank Cheria Jelita, Lauren Sangenario, Aurelien Marti, Alireza Najafian, and Patricia Aragon Han for the invaluable lab support and discussions.

I am extremely lucky to have the opportunity to interact with the brilliant scientists in the SKCCC Oncology Bioinformatics group, whose advice and discussions are valued. I thank Elana Fertig who, despite the many bugs that existed with Epicopy, helped me troubleshoot, worked through them, and used it in her studies. Ludmilla Danilova, thank you for your help in implementing various methylation approaches.

I am continuously grateful towards the Cellular and Molecular Medicine program for taking a chance in me. I thank Colleen Graham. Leslie Lichter, and Rajini Rao for their unwavering support and kindness.

I also wish to thank my mentors Wei Jen Lin of Cal Poly Pomona, Christine Brown, Wen Chung Chang, and Megan Prosser of City of Hope, and Michael Jensen of Seattle Children's Institute, whose kindness and mentorship prior to graduate school built my understanding of scientific research and drove me to pursue my graduate studies.

My experience through my PhD would have been less rich, and less wonderful without my classmates and friends. I appreciate my classmates, with whom I share my knowledge, struggles, joys, and friendship. I am especially grateful for Allison Galanis-Moloney for her infectious positivity, study marathons, and her tutelage of American sports. Sorry for being such a poor student. I am also thankful for Alexis Norris for the study sessions, amazing parties and strong support. Steven Wang, thank you for the camaraderie and consulting experiences. Brittany Avin, thank you for sharing in the joys and labors of graduate school, for the yoga kittens and the laughs. Your passion is contagious, determination inspiring, and friendship treasured. Thank you Kimberly Yang for answering all my immunology questions. Thank you Iris Chen, David Chu, Nina Hosmane, and Donna Dang for all the good food and fun times.

I am extremely thankful for Jessica Yang, my loving and supportive girlfriend who has made my experience in my PhD studies more memorable. Thank you for always being there, for teaching me to be a better person, making my life full, and for standing by my decisions, no matter where they may take me.

My deepest gratitude goes to my family. I cannot be more grateful to my father Chee Seng, and mother, Siew Lim, who believed in me and allowed me to pursue my passion. I thank my sisters, Rochelle and Faye, who are always willing listeners and my supportive extended family. Most importantly, my grandmother, Ah Ma, who had always believed in me, thank you.

Last but not least, I thank the patients who agreed to provide samples for these studies. This work and the discoveries will not be possible without you.

Table of Contents

List of Figures

List of Tables

Chapter 1: Genomic analysis of archival tissues

1.1: Overview

1.1.1: Rationale: Clinical utility of genomic data

Recent advances in molecular biology have allowed the scientific community to perform high-throughput profiling of tissues to study molecular alterations of disease states on a global level. One of the major diseases to have its genome, epigenome, transcriptome, and proteome studied at such a level is cancer.

Efforts from independent groups and international consortiums, such as The Cancer Genome Atlas (TCGA) [1] and the International Cancer Genome Consortium (ICGC), have profiled the molecular phenotypes of multiple types of cancer, including invasive breast cancer (IBC).

These studies have generated an unprecedented abundance of information and understanding of cancer, but remain limited in the clinical utility of this information. Clinical utility is a function of therapeutic and prognostic utility [2]. Therapeutic utility is exemplified by the discovery of molecular alterations that identify possible available interventions and can inform clinical decisions, such as a genetic alteration in an actionable target gene, *e.g.*, the FLT3 tyrosine kinase in acute myelogenous leukemia (AML) [3] or an ESR1 mutation in refractory, hormone-resistant ER-positive breast cancer (BCa) [4]. An example of prognostic utility is the use of molecular data, with or without clinical data, to predict patient survival. While TCGA and ICGC have identified

many potential therapeutic targets, they fare less well in prognosticating disease. Developing prognostic markers will be the focus of this thesis.

Prognostic utility using these public datasets is limited by inadequate follow-up time and lack of concerted effort to collect high-quality, longitudinal clinical and treatment information. Furthermore, since samples acquired by these efforts were samples of convenience, these datasets lack the follow-up and case-control pairs controlled for other estimates of poor prognosis such as grade, subtype, and treatment.

Using BCa as an example, at the time of writing, in the TCGA BRCA cohort of 1085 samples [5], median time to event is 2.08 years (Figure 1.1), and only 20% of patients have follow-up of more than 5 years. This is inadequate for the development of prognostic indicators in breast cancer; in which median recurrences occur at 5- or 10-years depending on the subtype [6-8]. Prospective studies for diseases with long time to event, while optimal, have a long wait time to completion, and information often continues to be gathered years after the study end date.

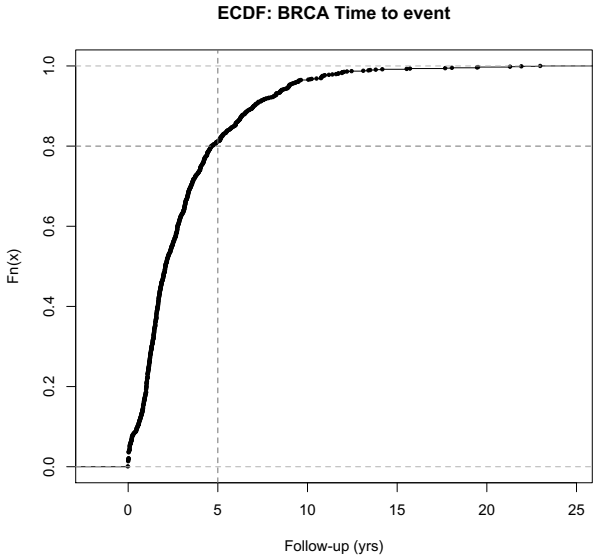


Figure 1–1: Empirical cumulative distribution function of time to event for TCGA BRCA dataset

Horizontal line at 80% ECDF and vertical line at 5 years of follow-up.

These problems can be addressed by using archival materials collected by medical institutions over many years, with the advantages of almost uniform methods of

preserving the tissues and access to high-quality clinical information with long-term follow-up [9]. This enables retrospective studies to identify prognostic markers of disease, especially in diseases that have a long time to event such as breast cancer and its premalignant precursor lesions [10]. Additionally, the ability to perform macro- or microdissection of tissues to enrich for disease lesions allows for the study of such tissues with minimal signal contamination of surrounding normal tissues, or to separately analyze the effects in tumor and stromal tissue.

1.1.2: Tissue fixation

The main objective of fixation had been, and remains, to preserve the microarchitecture of the tissue for visualization during initial assessment as well as after long-term storage [11]. The fixative used should minimize the loss of cellular components; peptides, sugars, nucleic acids, and lipids – with the ultimate purpose of preserving the architecture of intra- and extracellular structures, such as the nucleus, mitochondria, cell-cell interactions, and basement membrane. This is achieved through preventing autolysis by catabolic enzymes, minimizing the diffusion of soluble materials, and neutralization of microbial agents.

While the goal of fixation is to minimize the loss of as many components as possible, different fixatives preserve different components with various efficiencies, depending on the tissue processing protocol. Furthermore, any method will lead to artifacts, both in the appearance of the tissue and alterations on molecular components.

Therefore, the selection of the fixative to use has to take into account the features of interest, whether they are relevant cellular architecture or molecular components.

There are two major types of fixation; physical and chemical, each with their own advantages and disadvantages. The most common physical method is cryopreservation, snap-freezing for example allows immediate morphological assessment of surgical tissue samples as cryosections; another is dry heat fixation of organisms for Gram staining [11]. Chemical fixation can be broken down into denaturation, cross-linking, or a combination of the two.

Denaturing fixatives cause the coagulation of proteins, rendering them insoluble. Since the architecture of the tissue is maintained primarily by lipoproteins and fibrous proteins, this method will maintain the tissue morphology. Unfortunately, coagulant fixatives cause cytoplasmic flocculation and preserve mitochondria and secretory granules poorly [12].

There are two types of coagulant fixatives; dehydrant and acidic. Dehydrant coagulant fixatives, such as ethanol/methanol and acetone, remove water molecules and thus destabilize hydrophobic interactions and hydrogen bonding, allowing for denaturation of hydrophobic regions. The rate of reversal to a soluble state is slow and most proteins remain insoluble after such treatment even if reintroduced to an aqueous environment [12]. Acidic coagulant fixatives, like picric acid, change the charge on ionizable side chains of proteins, which disrupts electrostatic and hydrogen bonding to allow for coagulation. Since many acidic fixatives may lead to loss of nucleic acids, they are often used in conjunction with acetic acid, which coagulates nucleic acid but not proteins. Of note, picric acid fixation, such as in the use of Bouin's solution, causes

hydrolysis and destruction of membranes at low pH, leading to loss of intact nuclei and nucleic acids [11].

Cross-linking fixatives function by forming cross-links in proteins and nucleic acids. Examples of cross-linking reagents include 1) aldehydes (formaldehyde, glutaraldehyde, glyoxal), 2) metal salts (mercuric oxide, zinc chloride), and 3) other metallic compounds (osmium tetroxide). Historically, the most important characteristic of the fixation process is to support high quality and consistent staining with hematoxylin and eosin (H&E) [11]. To that end, the most popular and widely adopted fixative in diagnostic pathology is formalin, in the form of a 10% solution of neutral buffered formaldehyde. Formalin fixation is typically followed by embedding in paraffin, which provides external, structural support during sectioning on a microtome for light microscopy. I will focus the discussion on formalin-fixed, paraffin-embedded (FFPE) materials.

1.1.3: Effect of formaldehyde fixation on macromolecules

Formaldehyde reacts with many macromolecules in the tissue, and the resulting interactions are numerous and complex [11-14].

In a series of simple and well-planned studies, Frankel-Conrat and his colleagues investigated the effects of formaldehyde on proteins, and showed that formaldehyde created numerous intra- and intermolecular cross-links between side chains, amino groups, and primary amide chains [15-19]. The main action of formaldehyde on protein

involves the reaction of methylene hydrate with several protein side chains, forming methylol adducts in form of reactive hydroxymethyl side groups (—CH_2—OH) [20].

In an equally meticulous series of studies, McGhee and von Hippel examined the reactions between formaldehyde and free DNA [21-24]. Their studies revealed that formation of hydroxymethyl groups and dihydroxymethyl adducts on adenine and cytosine, while endocyclic imino groups form at guanine residues. Notably, the formation of these methylol adducts occurs rapidly and is reversible, but methylol adducts can condense in a slower second reaction, creating very stable crosslinks across molecules that are difficult to reverse. Furthermore, cross-linking occurs at AT-rich regions and has a direct correlation with increasing temperature.

With its ability to cross-link both protein and DNA, formaldehyde also reacts with nuclear proteins and nucleic acids. It penetrates spaces between these molecules and stabilizes the nucleic acid-protein shell, leading to extensive cross-linking between them [11].

This extensive cross-linking and modification of cellular macromolecules suggests a need for the reversal of some of these alterations to allow for optimal molecular profiling and characterization of FFPE tissues.

1.1.4: Recovery of macromolecules from FFPE material

In the same series of studies mentioned above, Frankel-Conrat et al. showed that the addition and condensation reactions in proteins were unstable and reversible by dilution or dialysis [15-19]. Cross-links, however, are more stable and have to be

hydrolyzed either in acidic or basic conditions at elevated temperatures, with unavoidable damage to molecules of interest. In some cases, they are completely irreversible.

With alteration of temperature, pH, and constituents of extraction buffers, partial recovery of nucleic acids is possible, and its success and extent depends of the intensity and duration of preceding fixation procedures. Nucleic acids in the cell associate closely with proteins and often form DNA-protein and RNA-protein cross-links following treatment with formaldehyde. Under extremely favorable formalin-treatment conditions, protein-DNA cross-linking can be reversed by incubating for 2 days in 0.1% SDS/50 mM Tris/pH 8.8 at 37°C, or 2 hours in the same solution at 60°C [25, 26]. Enzymatic digestion with proteinase K doubles the amount of RNA/DNA yield from FFPE materials, presumably due to a release of nucleic acids from cross-links with proteins [27]. Nucleic acid-nucleic acid cross-links can be hydrolyzed by heating at 60°C for 5 hours or 90°C for an hour, or at 60°C under alkaline conditions for an hour [28]. RNA recovery from FFPE materials has been difficult, since RNA lacks the inherent stability of double-stranded DNA, and is attributed to fragmentation and degradation, cross-linking with proteins, and modifications preventing reverse transcription and PCR reactions [14].

1.1.5: DNA/RNA extraction from FFPE material

Recovery of RNA and DNA from FFPE material include variations of deparaffinization using a solvent like xylene, liberation of nucleic acids through digestion of the tissue using proteinase K (PK) or Trizol [29], removal of methylol adducts by

heating or extended suspension in buffer, depletion of cross-linking metals via heating in the presence of chelating agents, and nucleic acid purification by precipitation or utilizing columns [30-46].

A non-exhaustive list of the work done in this area is summarized in Table 1. Many of these studies compared different extraction methods; commercial kits and in-house methods, while altering certain conditions such as temperature, digestion time, and deparaffinization. These studies collectively showed that increased deparaffinization, extended PK digestion times, and elevated temperature incubations post-extraction increased the yield and/or quality of the nucleic acids. Interestingly, there was no method that consistently performed better in terms of yield and quality. This may be attributed to differences in fixation of tissues, age of the FFPE block, PK formulation used, and tissue type [30-46].

Table 1: Summary of DNA/RNA extraction studies from literature

Study	Year	Type	Extraction method	Metric	Notes	Ref
Masuda et al.	1999	Total RNA	Trizol, PK	PCR	Temperature elevation removed methylol adducts	[41]
Chung et al.	2006	Total RNA	3 day PK, RNAlater	Bioanalyzer, PCR	RNAlater improved yield, but not quality	[33]
Abramovitz et al.	2008	Total RNA	ARA, RHR, QR, extended digestion	Transcriptome (DASL, Illumina)	Roche FFPE Highpure kit performed the best, extended PK digestion increased yield, quality	[30]
Doleshal et al.	2008	miRNA	QR, SAR, RHR, RHM, ARA, IPR	QPCR, comparison with FF tissues	ARA had best yield, with good FF and FFPE correlation of miRNA C_T	[34]
Roberts et al.	2009	Total RNA	Various commercial kits	QPCR, transcriptome (Affy U133Av2)	ARA had best yield and QPCR results, gene expression correlated between FF and FFPE	[44]
Bonin et al.	2010	DNA/RNA	In-house+/- purification	PCR	Extended digestion time critical for RNA recovery. Adsorption silica extraction method best for DNA.	[31]
Munoz-Cadavid et al.	2010	DNA	Various commercial kits	PCR	TDP followed by QD performed the best	[42]
Funabashi et al.	2012	DNA	In-house+/- salting out	PCR	Salting out improved DNA quality, at the cost of lower yield	[35]
Kotorashvili et al.	2012	DNA/RNA/ miRNA	Various commercial kits, in-house method (Trizol-based, +RNAlater)	Transcription (DASL, Illumina); Mirnome (Illumina); methylation (EpiTyper)	Co-extraction of all nucleic acids, home-grown method performed the best, followed by Qiagen Allprep FFPE kit	[37]
Ludyga et al.	2012	DNA/RNA	QD, QR, NDR, PCI	Bioanalyzer, PCR	PCI and Qiagen kits performed the best. PCI was inconsistent in quality and yield.	[39]
Ton et al.	2012	RNA	Modified QR, RHR	Transcriptome (DASL, Illumina)	Modified QR performed better, attributed to additional deparaffinization steps, longer PK digestion, and high temperature	[45]
Turashvili et al.	2012	DNA/RNA	QR, QD, ARA, TWR, TWD, in-house	PCR, QPCR	In-house method, with overnight digestion performed the best. Increase fixation time in NBF led to poorer quality nucleic acid.	[46]
Potluri et al.	2015	DNA	QD, AFA, PCI	PCR	QD had best yields, AFA had better quality DNA	[43]

AFA: adaptive focused acoustics ; ARA: Ambion RecoverAll; IPR: Invitrogen PureLink RNA FFPE; QD: Qiagen QiaAmp FFPE DNA; QDR: Qiagen Allprep DNA/RNA; QR: Qiagen RNeasy; RHR: Roche High Pure RNA; RHM: Roche High Pure miRNA; SAR: Strategene Absolutely RNA FFPE; TDP: TaKaRa DexPat; TWD: Trimgen WaxFree DNA; TWR: Trimgen WaxFree RNA

1.1.6: Genomic profiling of FFPE material

1.1.6.1: Challenges

Various studies have attempted to obtain genomic information from FFPE material, starting with PCR-based methods to second-generation sequencing. This dissertation will focus on studying the ability to use RNA and DNA from FFPE material.

There are many challenges in profiling RNA and DNA from FFPE tissue. Depending on the fixation process and the degree of over-fixation, nucleic acids could have been modified to varying extents, with over-fixed FFPE material being of the poorest quality. Regardless of the fixation time, any degree of alteration will lead to improper application of methods optimized for high quality nucleic acid preparations.

Nucleic acid degradation by formalin treatment reduces the amount of nucleic acids available for profiling, leading to the need for increasingly efficient extraction and purification methods from FFPE material. This also speaks to the need for genomics methods that require less input material, often in the realm of pico- to nanograms. Some groups have applied whole transcriptome amplification (WTA) and whole genome amplification (WGA) methods to globally amplify whole RNA and DNA in an attempt to rescue signal from limited material [47].

Hydrolysis of nucleic acid chains by high temperature, especially around AT-rich regions, lead to extensive fragmentation, restricting the extent of amplification especially in the case of RNA. Indeed, Bioanalyzer RNA integrity (RIN) scores from many studies consistently showed poor RNA integrity, even in mock, freshly-prepared FFPE samples.

Furthermore, with regards to mRNA, fragmentation could lead to the loss of the 3'-UTR and polyA-tails, rendering canonical reverse transcription reactions or enrichment protocols using oligo-dT useless. Non-sequencing studies have circumvented this, to varying degrees of success, using random hexamer primers and limiting the amplicon size. To select for coding mRNA, sequencing studies have used a negative-enrichment method, through rRNA depletion, instead of a positive-enrichment for mRNA, but have yet to overcome the problem of over-fragmentation of RNA.

Cross-linking reactions and methylol adduct formation introduce covalent modifications on nucleic acids that prevent the activity of enzymes and annealing reactions, critical steps in recovery and methods of analyzing nucleic acids. This is often rescued by the prolonged incubation in specialized buffers and heating protocols, methods that will often introduce additional hydrolysis, compounding the problem. Moreover, there are specific modifications that cannot be overcome by such extraction methods. One such modification is a high frequency of non-reproducible sequence alteration, often C-T or G-A transitions, speculated to be a result of DNA polymerase inability to recognize cytosine residues [48, 49]. In this instance, artifactual mutations were inversely correlated with the amount of input DNA, and required either increased input material or independent validation by sequencing.

Despite the challenges, many groups have successfully profiled FFPE-derived RNA and DNA using high-throughput methods such as microarrays and second-generation sequencing. Additional challenges are associated with DNA methylation profiling of FFPE material, and will be detailed in the DNA methylation profiling section.

1.1.6.2: Transcriptome profiling

Bibikova and colleagues from Illumina Inc. developed the cDNA-mediated annealing, selection and ligation (DASL) assay, suitable for analysis of the transcriptome of FFPE materials [50, 51]. This method overcomes the limitations of loss of polyA tails by using random priming in cDNA synthesis and has probes with target sequences of 50bp in length. Using FFPE tissues stored between 1- to 10-years of age, they showed technical reproducibility in FFPE with an average R^2 of 0.95. Interestingly, even when the correlation between matched frozen and FFPE pairs were lower at an average R^2 of 0.69, there was considerable overlap ($p = 1 \times 10^{-9}$, Fisher's Exact test) in differentially expressed genes between cancer and normal. This suggests that although FFPE treatment globally alters gene expression profile, either by degradation or base modification, these effects are not selective, allowing the detection of tumor specific changes using FFPE material.

Following that, multiple studies have validated the reproducibility of the DASL assay [30, 52] and have used it in analyzing the transcriptome of archival FFPE tissues [53]. Other studies have used a combination of WTA techniques with other microarray technologies such as the Affymetrix HGU133v2 array [47]. Affymetrix released a combination of WTA sample preparation followed by microarray analysis for FFPE material in the form of the WT Pico kit [54] and the Human Gene ST 2.0 array [55], the

latter of which contain 25bp exon specific, junction, microRNA, and non-coding RNA (ncRNA) probes.

More recently, FFPE-derived RNA has been profiled using second-generation sequencing. Sinicropi et al. [56] successfully generated RNA-seq data on a cohort of 136 BCa patients using a modified RNA-seq protocol with an average of 43 million reads per patient. They increased the input RNA into library generation, used ribosomal RNA depletion instead of a positive-selection for polyA tail, and extended the time of the cDNA synthesis step to increase library yield. They compared hazard ratios of RNA-seq generated reads with hazard ratios of the OncotypeDX RT-PCR panel [57], a molecular test developed by Genomics Health Inc. which funded the study. Using the same samples, they found a Pearson correlation of 0.81. Furthermore, when considering transcripts with high read counts, this study identified a series of recurrent markers, which significantly overlapped, at 27.0% of 11659 RefSeq genes (a set of markers identified from an independent, microarray-based study).

Norton et al. [58] were able to identify expressed single nucleotide variants (eSNVs) and fusion transcripts from nine paired fresh frozen (FF)/FFPE BCa pairs using the RiboZeroGold rRNA depletion and the ScriptSeq V2 library generation kits (Illumina, San Diego, CA). Consistent with Sinicropi et al's findings, they noted that while the correlation between FF and FFPE paired samples were moderately strong, there was considerably better agreement between differentially expressed genes between cancer and normal sample pairs. On average, only 20% of all reads of FFPE samples mapped to genes, compared to 50% in FF samples. Longer insert sizes resulted in high

eSNV detection sensitivity. Unfortunately, increased read depth did not increase fusion transcript detection.

Another approach used in FFPE samples is a target capture protocol. Cieslik et al. [59] developed an exome capture RNA-seq approach and applied it to FFPE samples. Their experience showed comparable alignment, strandedness, gene detection, and variant calling between exome capture and polyA selection using RNA from cell lines as a proof of concept. Exome capture had a better performance in identifying variants and junction spanning reads compared to a polyA selection method. Exome capture was able to better deplete for rRNA and align to protein coding regions compared to an rRNA depletion protocol. Extending it to FFPE samples, they found moderate correlation between FFPE and FF samples at an average Pearson correlation of 0.8, irrespective of library type. Virtually all splice junctions and fusions were detected in FFPE samples.

Overall, the findings in existing literature agreed that FFPE derived RNA-seq data when compared to FF data were 1) moderately correlated in terms of transcript expression values, 2) concordant in genes and relative difference between disease states, 3) less precise with increased RNA degradation due to higher technical variability. These shorcomings could be rescued by increasing read depth, 4) less sensitive in detecting eSNVs and gene fusions regardless of read depth, and 5) poorer in terms of fragment diversity, with smaller insert sizes and less unique fragments.

1.1.6.3: Genetic profiling

Using an Illumina SNP BeadArray, Lips et al. [60] profiled a series of matched FF and FFPE colorectal tumors and discovered identical genotype and loss of heterozygosity (LOH) profiles between the paired samples. In a follow-up study, Oosting et al. [61] showed high concordance in copy number profiles between FFPE samples profiled using high density SNP microarrays and arrayCGH platforms. Their observations were supported by studies from other groups using FF-FFPE pairs in various platforms, even in microsatellite regions [62-65]. Interestingly, while the copy number profiles remain consistent between FF and FFPE pairs, the amount of variation in log R ratios were much higher, while the overall log R ratio signals were lower in FFPE samples compared to their FF counterparts, which may be caused by degradation and covalent modifications on the DNA molecules [60-65].

Thompson et al. [66] evaluated the use of SNP arrays in FFPE tissues for making genotype calls, LOH identification, and CNV profiling, and found good concordance between FF and FFPE tissues in all assessments, despite a higher level of noise in the FFPE samples.

Whole genome sequencing (WGS) technologies had also been applied on FFPE samples to obtain genetic information. Schweiger et al. [67] performed a small, three-sample experiment varying ischemia and fixation times of breast tissues prior to sequencing. While FFPE tissue had lower mappable reads and higher variability, the fragments were distributed equally across the genome and they were able to obtain comparable copy number information between FFPE and FF samples. Yost et al. [68]

performed a more thorough analysis of a two-sample experiment; comparing WGS results from paired FFPE TNBC tumors and FF normal genomic DNA. Consistent with literature of DNA modifications following formalin fixation, they identified a large number of C/G to T/A substitutions in the FFPE samples; an artifact observed with as few as a million reads. However, this artifact was rescued by implementing stringent filters; incorporating information on read diversity, local mismatch rates, and global mismatch rates.

The application of whole exome sequencing (WES), where exonic regions of the genome are selected for using a variety of target enrichment strategies similar to that of exome capturing for RNA-seq, has had more success in FFPE material compared to WGS. Whereas WGS applications had mostly been technical and comparative, WES experiments in FFPE tissue have been used for discovery purposes.

Kerick et al. [69] performed WES on 3 FF-FFPE sample pairs and were able to detect concordant SNVs and InDels in pairs. They highlight, however, a larger coefficient of variation of about 2-fold in FFPE vs FF samples. The same study detected false positive and false negative results in FFPE vs FF comparisons that were rescued with increased coverage. Wagle et al. [70] adapted an exon capture approach to enrich for cancer-relevant genomic alterations and were able to identify copy number gains and losses, validated using quantitative PCR (QPCR), in archival breast cancer samples. Furthermore, they were able to detect point mutations using their approach, validated by hME genotyping; some of which were missed by another mass spectrometry-based approach, Oncomap. Oh et al. [71] performed WES on 4 FF-FFPE pairs, and showed that FFPE samples had shorter insert sizes, contained artificial base alterations (C/G>T/A),

and resulted in increased duplicate reads. They noted that about 30% of the bases in FFPE samples were soft clipped, which is a read filtering step where the alignment is performed only on clipped regions but the whole read is retained. The study attributed to non-specific annealing of degraded DNA fragments during library construction. However, with appropriate processing of the data, high-confidence mutation calls in FFPE samples were validated in FF samples.

Using a solution hybrid selection exome capture approach for target enrichment, Van Allen et al. [72] demonstrated no difference in WES coverage metrics between 99 FFPE samples and 768 non-FFPE samples. This study noted that while an amount of input DNA as low as 1 ng was acceptable, there was an increase in the number of duplicate reads. In 11 paired lung adenocarcinoma samples, WES data showed extremely good concordance between FFPE and FF tissue in identifying mutations and copy number profiles (average $r^2 = 0.79$). Using only the FFPE data, they identified clinically actionable targets, and in one demonstrative case enrolled a patient with a $KRAS^{A146V}$ mutation in a CDK4 inhibitor trial to which the patient achieved their only clinical response to cancer; stable disease for 16 weeks.

In summary, analysis of DNA in FFPE using microarray and sequencing is prone to 1) the introduction of C/G > T/A artifacts [68, 71], 2) shorter insert lengths [68, 69, 71], 3) increased duplicates and lower diversity [67-72], 4) non-specific annealing of DNA fragments [71], 5) lower mappability [69], and 6) higher variability in read count compared to FF samples [66-69]. However, this can be rescued with appropriate workflows; 1) increasing amount of input DNA, 2) increasing read depth, 3) use of soft

clipping during alignment, and 4) defining appropriate filters for making high-confidence calls.

1.1.6.4: DNA methylation profiling

The most commonly used methylation microarray technology is the Illumina Infinium Methylation microarray, which profiles DNA methylation changes in bisulfite treated genomic DNA. Briefly, bisulfite treatment converts unmethylated C into U, while 5'-methyl-C (5meC) remains unmodified. Upon WGA, the U is then converted into a T. The Illumina Infinium Methylation microarrays were designed based on Illumina's BeadChip technology, which measures either a C or T at a CpG locus of interest in bisulfite treated DNA, and reports the presence of converted Ts as either beta-values, or more intuitively percent methylation of a site, or M-values, a logit transformation of beta-values which approximates a normal distribution that better fits assumptions of certain models for differential methylation analysis [73]. The first of these arrays was the 27K microarray which probes over 27,000 CpG loci in the human genome. In 2010, Illumina released the 450K microarray, which measures methylation across more than 470,000 CpG loci, and remains the most abundantly used methylation array with the most publicly available datasets at the time of writing. More recently, Illumina has released the MethylationEPIC microarray, which profiles over 850,000 CpG loci.

The first documented use of methylation microarrays on FFPE material was described by Thirlwell et al. (2010) [74] with Illumina's Infinium Human Methylation 27K microarray. The Infinium technology contains a WGA step, which can be disrupted by fragmented and low molecular weight DNA. The authors introduced a ligation step

prior to bisulfite treatment to increase the molecular weights of DNA present, which they hypothesize would allow for better WGA. Indeed, when comparing FFPE samples that went through a ligation step with those which did not, there was improvement in WGA yield and better correlation between FFPE and FF beta-values in the ligated samples. Perhaps unsurprisingly, the beta-value histograms of the ligated sample resemble that of FF samples, but the beta-values of the unligated sample differed, with the mean regressing closer to 0.5, which is expected of samples with non-specific binding, and therefore signal intensity measurement, of probes.

At the time of writing, the Illumina Human 450K Methylation Microarray, in conjunction with the Illumina FFPE restoration kit, was the most used DNA methylation profiling technology for FFPE material. The Illumina FFPE DNA Restoration kit is marketed as a two-enzyme protocol that restores partially degraded DNA for use in Illumina's HD assays, including their HD SNP and methylation microarrays. While the technology is proprietary, I hypothesize that a ligation and global amplification step are involved, similar to the innovation performed by Thirlwell et al.

Dumenil et al. (2014) [75] studied a series of 21 paired FF-FFPE colorectal cancer tissues using the 450K microarray with DNA restoration and showed that they were able to identify the CpG island methylator phenotype (CIMP) status of samples across tissue type with concordant changes in methylation status of CIMP-related genes. Interestingly, they showed that higher age of the FFPE block correlated with increased Euclidean distance between FF and FFPE material. Unfortunately, sample-wise correlations of high quality probes were not reported. De Ruijter et al. (2015) [76] performed pairwise analysis of FF and FFPE tissue, where after sample collection, tumors were separated

into two parts, with one undergoing formalin fixation. Their study showed improvement in pairwise Spearman correlation of FF and FFPE samples with restoration, from a mean of 0.896 in un-restored FFPE samples to 0.989 in restored FFPE samples.

Methylome profiling of FFPE tissue by bisulfite sequencing has been limited, as challenges extend beyond problems affecting FFPE-derived DNA. The conditions required for bisulfite conversion result in a high degree of fragmentation and depurination of DNA, which leads to poor quality templates for downstream analyses, including sequencing technologies. Further complicating the analysis is the potential for false positive 5meC calls in regions of incomplete bisulfite conversion, which are common in denaturation-resistant regions. This is due to the fact that formation of ssDNA is vital for bisulfite conversion of C, and in dsDNA unmethylated C will remain unconverted, leading to a false positive call. Furthermore, presence of residual protein in the gDNA preparation will lead to incomplete conversion of C. The presence of cross-linked dsDNA fragments combined with inefficient removal of crosslinked DNA binding proteins in FFPE material can lead to poor denaturation and only partial conversion of Cs.

Perhaps due to these reasons, there has only been one poorly designed study published on bisulfite sequencing of FFPE-derived material. Li et al (2014) performed bisulfite sequencing following WGA of bisulfite converted DNA from two FFPE ovarian tissues, which were fixed in 1999 (O1999) and 2011 (O2011). The authors observed bisulfite conversion efficiencies of 96.7% & 88.8% and unique mapping rates of 19.9% & 7.0% in O2011 and O1999 respectively. Unfortunately, no paired FF samples were run, and we are unable to evaluate the reproducibility of methylation of these samples. The authors did not ligate fragments pre-WGA, which has been shown in other

technologies to improve amplification efficiency, bisulfite conversion rates, and reproducibility in FFPE materials. While the authors conclude differently, we conclude that bisulfite sequencing is currently not optimized for DNA methylation profiling in FFPE material.

1.2: Optimization of protocols for genomic analysis of FFPE tissue

1.2.1: Effect of light H&E staining and O-phenylenediamine on DASL performance

Based on available literature, Roche Highpure FFPET RNA Kit (Roche, Cat# 06650775001) was used as starting point for my optimization of RNA extraction from FFPE tissue [40], with the comparison metrics being *yield* by Nanodrop, *quality* by a GAPDH-based quantitative PCR (QPCR), and *performance* on the Illumina DASL microarray. I assessed the effect of light hematoxylin & eosin (H&E) staining, as well as the use of O-phenylenediamine (OPD) in the extraction buffer on RNA extraction.

The presence of non-cancerous tissue in the RNA/DNA preparation leads to dampening of any cancer-specific signal. This results in difficulty in detecting transcripts present in low copies, or have small changes in gene expression levels compared to normal tissue. One of the advantages of starting with tissue sections, common with FFPE material, is the ability to stain the tissue and identify malignant cells. While serial sections can be used to identify areas of interest, cells may shift through the sections and the ability to stain the tissue section to be analyzed will allow better enrichment or direct microdissection of the tissue.

Given the instability of RNA compared to dsDNA, the usual steps to improve yield and quality that are applicable to dsDNA preparation, such as prolonged proteinase K digestion at elevated temperature are problematic for RNA extraction. O-phenylenediamine is an aromatic compound, we hypothesized, would be an efficient acceptor of methylol groups, and would serve as a methylol sink in the RNA extraction procedure allowing for more efficient reversal of methylol adducts under milder conditions.

A pilot experiment was performed using 2 FFPE samples, AFB15 and AFB21. AFB21 is a smaller sample and was used to exemplify a sample with low tumor content, such as ductal carcinoma in situ (DCIS). A total of twenty slides were serially sectioned from both samples and every other serial section is randomized into different protocols (Table 2). I further performed limiting dilution experiments on the amount of input RNA into DASL to estimate the lower threshold of RNA inputs into DASL that may be important for samples with poor yields.

Table 2: RNA yield from Highpure FFPET kit comparing OPD and staining

Sample	Treatment	Area (mm2)	Slides	Total RNA Yield (ng)	Yield/Area (ng/mm2)
AFB15	+OPD, Stained	21	10	3138	14.94
AFB15	-OPD, Stained	21	10	2664	12.69
AFB21	+OPD, Stained	6	5	314	10.47
AFB21	-OPD, Stained	6	5	302	10.07
AFB21	+OPD, Unstained	6	5	284	9.47

1.2.1.1: No statistically significant difference was observed in RNA yield and QPCR CT between different extraction protocols

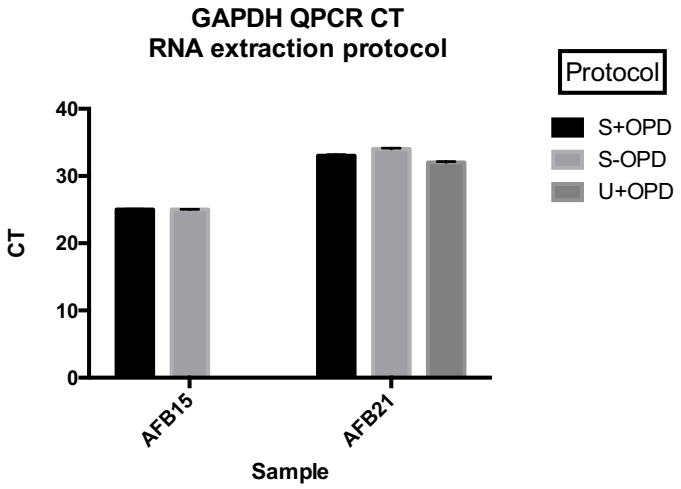


Figure 1–2: GAPDH QPCR CT values

There is no statistically significant difference between GAPDH QPCR C_T values across different extraction protocols for both samples.

There was no statistically significant difference between RNA yield and quality as assessed by GAPDH QPCR across the different protocols. Of note however, was the lower yield and quality of AFB21 compared to AFB15, which may translate to differences in DASL performance as neither assesses the ability to globally profile transcript expression.

1.2: Overall DASL performance

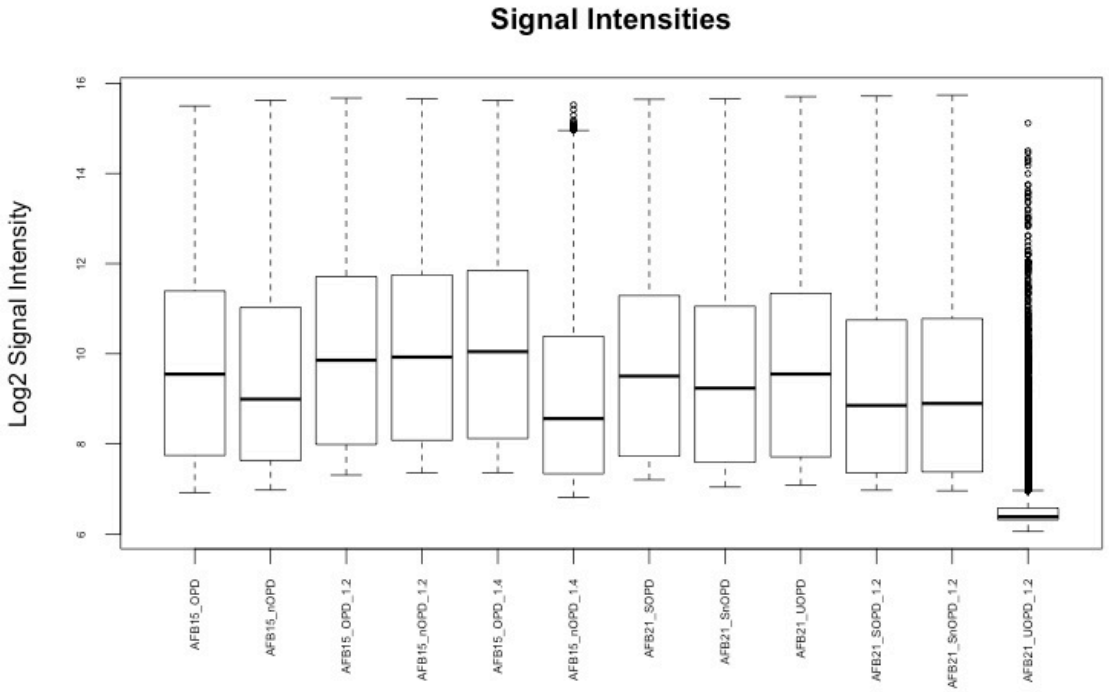


Figure 1–3: Signal intensities for all probes measured in pilot DASL experiment

Signal intensities for all the probes in each array are used as a measure of overall success of the DASL assay given the sample. OPD: Highpure +OPD. nOPD: Highpure –OPD. S: stained. U: unstained. Note that all assessments in AFB15 were on stained slides.

One sample, AFB21 U+OPD had no signal intensities for majority of the probes, suggesting that it failed to perform on the DASL chip. There were no differences in signal intensities across protocols of the same dilution. However, decreased signal intensities were observed with increasing dilutions, especially in AFB21, which is the sample with poorer amplification of GAPDH.

1.2.1.3: Light H&E staining of section does not alter overall DASL performance or gene expression profile

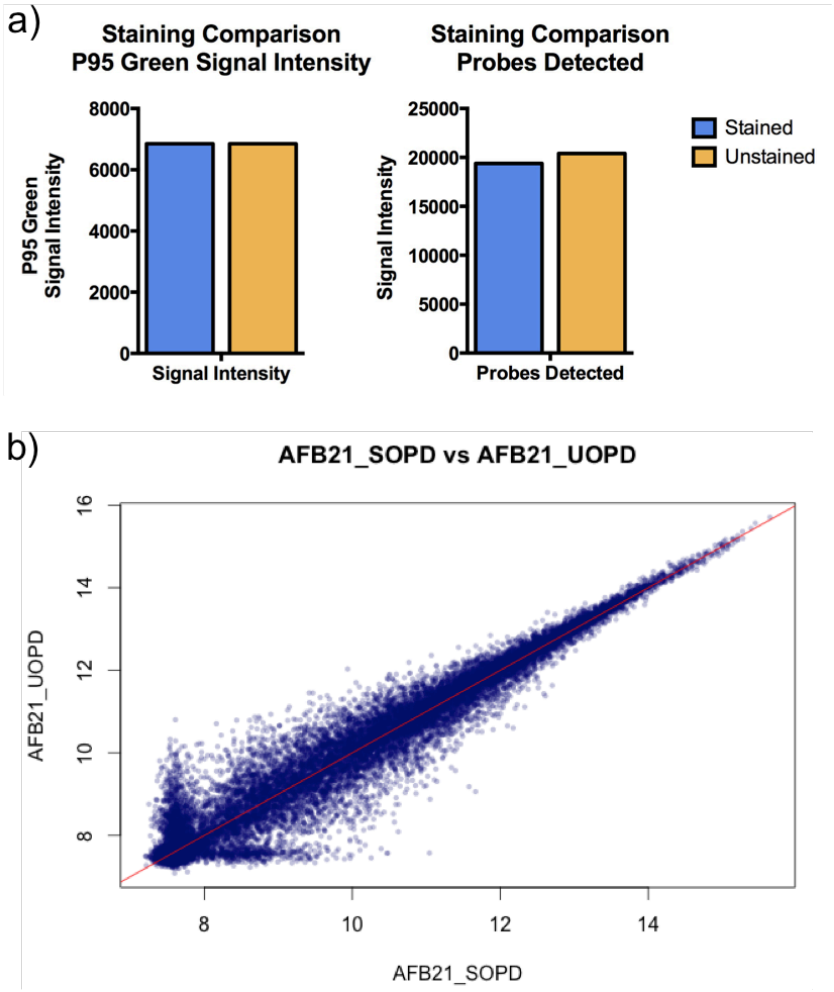


Figure 1–4: Comparison of DASL performance between stained and unstained AFB21

a) P95 green signal intensity and number of probes detected above a threshold compared to negative control probes. b) Correlation between probes in stained versus unstained AFB21 (Pearson R = 0.97). Note the vertical and horizontal trend of a subset of probes from the lower left hand corner, suggesting probes detected in one experiment but not the other.

A direct comparison between P95 green signal intensity and number of probes detected between stained and unstained AFB21 sample revealed no difference between these two metrics (Figure 1-4a). There was also good correlation between all the probes

of the stained and unstained samples, with a Pearson correlation R of 0.97. Interestingly however, both samples contained probes that were not detected in the other, as observed as a series of probes with a vertical or horizontal trend from the lowest effective signal (Figure 1-4b).

1.2.1.4: O-phenylenediamine treatment improves DASL performance in poor quality and low yield samples

Introduction of OPD during RNA extraction of FFPE tissue revealed no changes to the rank-based gene expression profile of these tissues. In samples where RNA is abundant (Figure 1-5a), overall signal intensities were unchanged. Interestingly, in samples where input RNA into DASL is lower, there is a trend in lower P95 signal and number of probes detected (Figure 1-5b). Furthermore, when probe-wise Pearson correlations were calculated between OPD and non-OPD samples, OPD-treated samples had higher correlation (Figure 1-5c). Taken together, this may suggest that OPD improves overall RNA quality, including those RNA present minimally in the sample which allows for more efficient WGA of all RNA species.

Indeed, when I performed pairwise comparison of all the probes in AFB15 of a higher amount of input RNA (1:2 dilution, AFB15_1.2) and that of a lower input (1:4 dilution, AFB15_1.4), I continued to observe strong Pearson correlation between AFB15_OPD_1.2 and AFB15_OPD_1.4. The same comparison in AFB15_OPD_1.4 and AFB15_nOPD_1.4 showed overall lower signal intensities, manifesting as points falling below the red diagonal line (Figure 1-6, upper right). Perhaps unsurprisingly, the same

lower signal intensity was observed between non-OPD treated AFB_1.2 and AFB_1.4 (Figure 1-6, lower right).

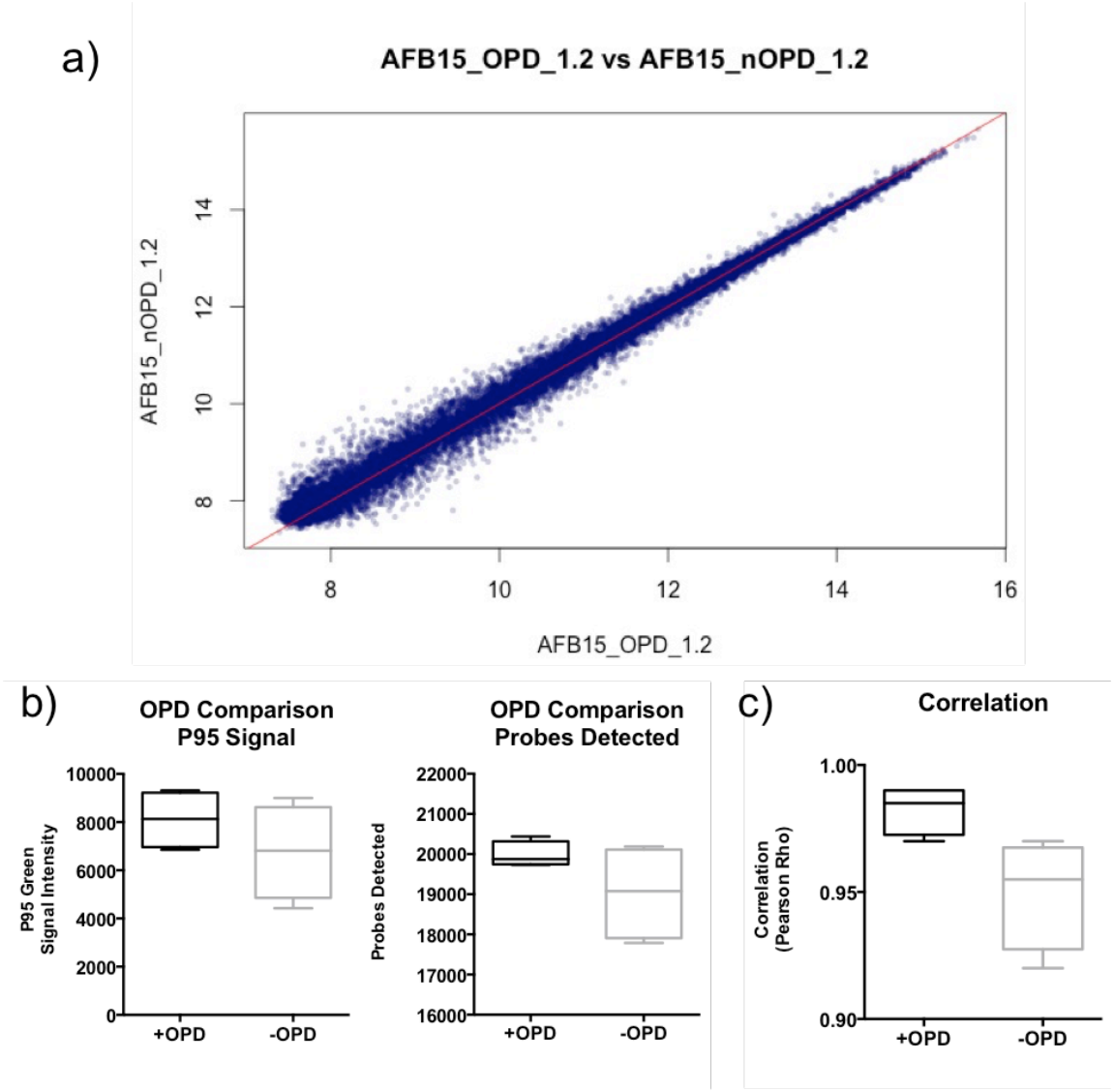


Figure 1–5: Comparison of DASL performance between OPD and non-OPD treated samples

a) Pairwise correlation between all the probes in OPD and non-OPD treated AFB15. Pearson R = 0.99. b) P95 green signal intensity and number of probes detected across all DASL arrays comparing OPD treatment. c) Pairwise Pearson correlation comparing samples treated with or without OPD of equal RNA input into DASL.

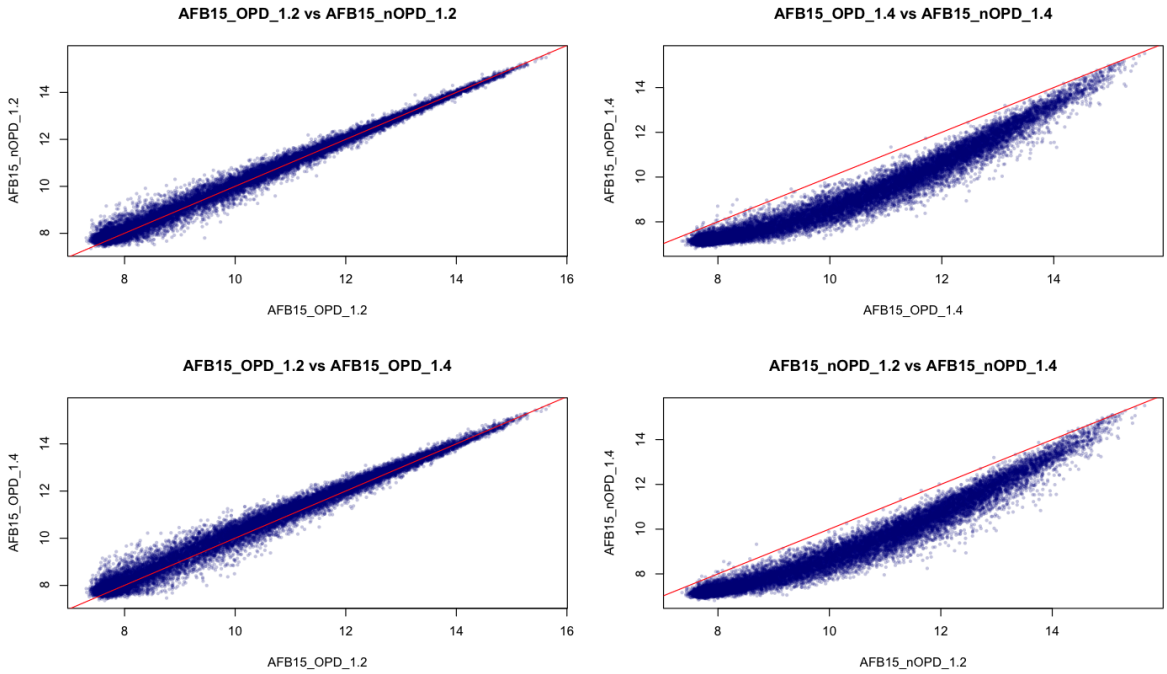


Figure 1–6: OPD does not affect gene expression ranks, but improves signal intensities in DASL of samples with low RNA input

Results of pairwise correlation comparing samples with varying OPD treatment groups and RNA input into DASL microarray. Diagonal red line describes a linear relationship between both axes.

1.2.1.5: Conclusion

Taken together, I concluded that light staining of FFPE samples does not alter profiles or signal intensities of transcripts and is a viable method to visualize the tissue for dissection to enrich for tumor cells, and therefore tumor signals. OPD appears to improve performance of samples with low input into the DASL microarray, and will be useful in situations where RNA yields are limited. Both modifications should be considered in RNA extraction of small and challenging lesions.

1.2.2: Optimization of lab protocols for Illumina Human Methylation 450K microarray

Our lab has extensive experience performing multiplex QPCR-based DNA methylation analysis in FFPE tissue sections and has developed a protocol where the DNA purification and bisulfite treatment are combined into a single process, minimizing losses that arise during DNA purification. This method, called TNES, uses a DNA extraction buffer (10 mM Tris, 150 mM NaCl, 2 mM EDTA, 0.5% SDS) with overnight proteinase K (PK) digestion.

The workflow of analyzing FFPE materials on the 450K methylation chipset includes a QPCR-based QC step, which assesses the amplifiability of the DNA sample. Unfortunately, the presence of SDS in the extraction buffer inhibits PCR reactions, and precludes its use for the QC analysis. Furthermore, bisulfite-treated DNA (NaBi-DNA) samples derived using this method have not yet been assayed using methylation microarrays. There is also lack of well-defined bioinformatics practices for the analysis of FFPE-derived NaBi DNA, such as sample call rates and detection p-value thresholds.

To that end, the following experiment was designed to 1) investigate the incorporation of the TNES protocol into the 450K microarray workflow, 2) assess the performance of NaBi-DNA generated with this protocol on the 450K microarray, and 3) develop bioinformatics protocols for microarray-based analysis of DNA methylation in FFPE material. We performed 450K microarray analysis on a series of 3 FF/FFPE pairs and 36 FFPE breast tissues, divided into three biological groups – 12 patients with non-proliferative fibrocystic benign breast disease (BBD) who did not develop invasive ductal

carcinoma (IDC), and 12 paired BBD and IDC samples from patients who were diagnosed with BBD and eventually developed IDC. Of the 36 sample cohort, a total of 5 samples were arrayed in duplicate and a single sample was used for limited dilution of input NaBi-DNA into the array workflow. Lastly, a series of Illumina control tissues, both FFPE and FF, were arrayed in the same microarray experiment as controls.

1.2.2.1: Tween-based TNET buffer allows for QPCR-based QC

As mentioned, a limitation of using SDS in the extraction buffer is the inability to use the lysate for QPCR reactions without additional purification of the DNA. We hypothesized that replacing SDS with a non-ionic detergent, such as Tween-80, will allow for cell lysis and activation of proteinase K without inhibiting PCR reactions. We modified the TNES buffer to TNET buffer (10 mM Tris, 150 mM NaCl, 2 mM EDTA, 0.5% Tween-80) for DNA extraction of 2 sections from each tissue and dsDNA from the lysates were quantified using a fluorescence-based dye. TNET lysates were used as input DNA for the QPCR-based Illumina FFPE QC kit.

These TNET-based lysates were quantifiable using Picogreen, and 28/36 (77.8%) FFPE samples were amplified by the assay. QC results were reported as delta C_T, and ranged between 4.5 to 17.16 in this cohort of samples. In comparison, the threshold Illumina suggested was at a threshold of lower than 5. We decided to proceed with 450K microarray analysis of these samples to evaluate the success rates of samples with such varied QC values.

1.2.2.2: Illumina FFPE QC results correlate with percent detected probes and threshold above manufacturer recommended level should be considered

Amount of DNA measured from TNET extraction were used to estimate TNES yields, and were combined as necessary to achieve amounts necessary for the target 1ug input into bisulfite conversion reaction. The resulting ssDNA amounts were used to estimate total input DNA from TNES/TNET and where the values exceed 1ug, the bisulfite reaction was repeated with adjustments to input DNA. Following satisfactory bisulfite treatment, 500ng of NaBi DNA was restored using the Illumina FFPE DNA restoration kit and processed for 450K microarray using manufacturer recommended protocols.

Illumina BeadArray technologies include a probe-wise metric called detection p-value (detP), which estimates presence of signal using the distribution of signal intensities from replicate probes on beads distributed across the microarray compared to the signal intensities of negative control probes. Using a threshold of detP<0.05, we estimated the percent probes detected (call rates) for each sample and compared that to the delta C_T from the FFPE QC kit (Figure 1-7).

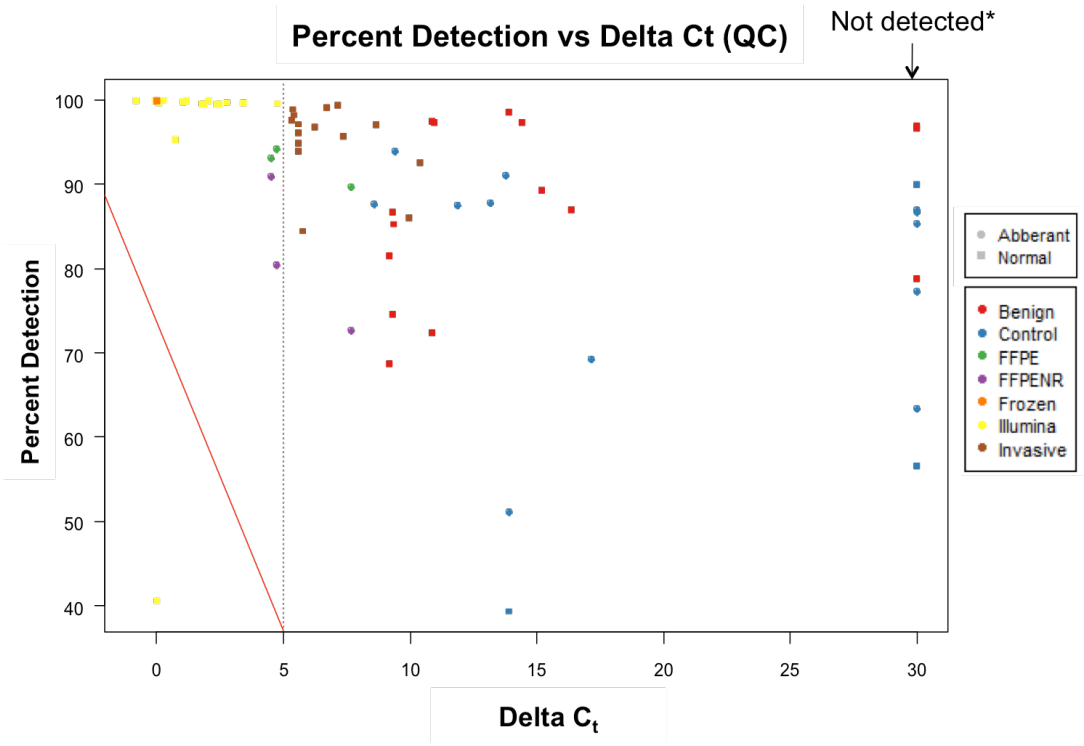


Figure 1–7: Relationship between percent detection (call rate) and delta Ct

Vertical dotted red line represents Illumina recommended threshold of 5. An empiric call rate of 80% was set as the threshold for an FFPE sample with enough detected probes and sufficient quality for downstream analyses.

We observed call rates > 90% for all fresh frozen and restored FFPE samples with delta C_T < 5 (Figure 1-7). Call rates of samples with delta C_T > 5 had more variable call rates, with 23/47 arrays with call rates < 90%. Of the samples with C_T > 5, only 10 arrays had call rates below the empirical threshold of 80%, with most of the loss of acceptable quality samples at delta C_T > 8 Interestingly, of the samples that did not pass QC, half had call rates > 80%.

1.2.2.3: Removal of low performance probes improves reproducibility across duplicates

We assessed the reproducibility between replicates in this experiment and observed a linear relationship between reproducibility and call rate. We hypothesized that this is due to the random nature of signal intensities of poorly performing probes and filtering against these probes will improve reproducibility between replicates. As such, we removed samples with overall poor call rates (< 70%), and filtered against probes that were undetected at detP < 1e-5 in at least 2 samples. The remaining probes were defined as high quality probes.

Of the 5 duplicates, 1 pair (SC30) had a sample with poor call rate (< 40%) and was not used in the identification of high quality probes. To quantify the differences across replicates, we calculated pairwise Pearson correlation and change in rank order of samples by Euclidean distance for duplicates. In the latter, we estimate the similarity of duplicate samples to other samples by measuring the simultaneous change in Euclidean distances. In this scenario, a closer rank order indicates greater increase in similarity compared to other samples in between the initial and final rank.

We observed improved reproducibility across all samples, including SC30, for both Pearson correlation (3% on average) and improvement in rank order of sample similarity. A representative example of increased Pearson correlation is SC33 (Figure 1-8) where we observed increased correlation comparing analysis in high quality probes to all probes.

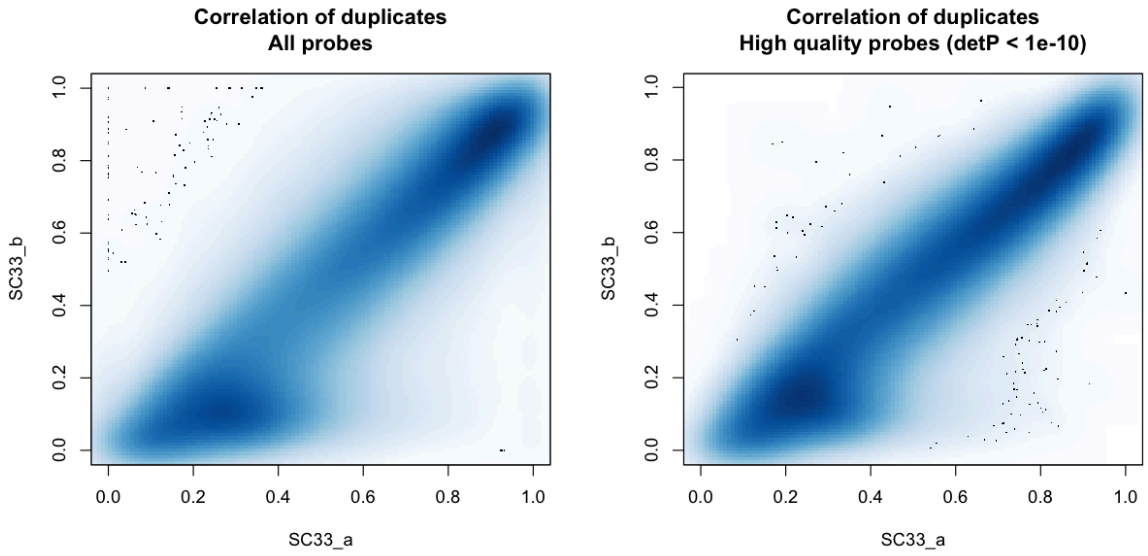


Figure 1–8: Probe-wise comparison in SC33 across all probes or only within high quality probes.

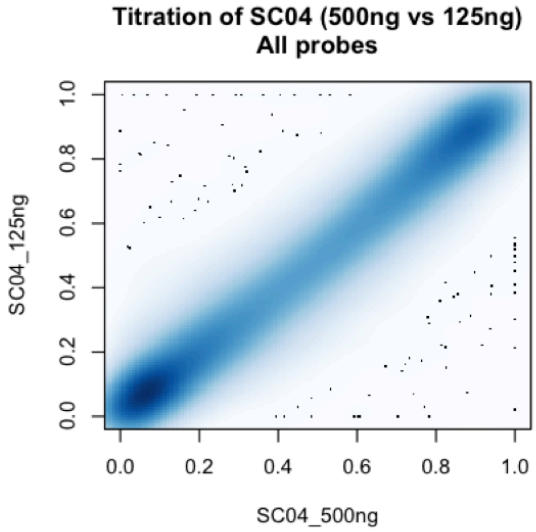
A colored density plot representative example of improved probe-wise comparison when filtering for high quality probes in SC33 duplicates. A darker blue indicates a higher density of probes compared to the surroundings. Less outlier probes on the upper left and lower right are also observed.

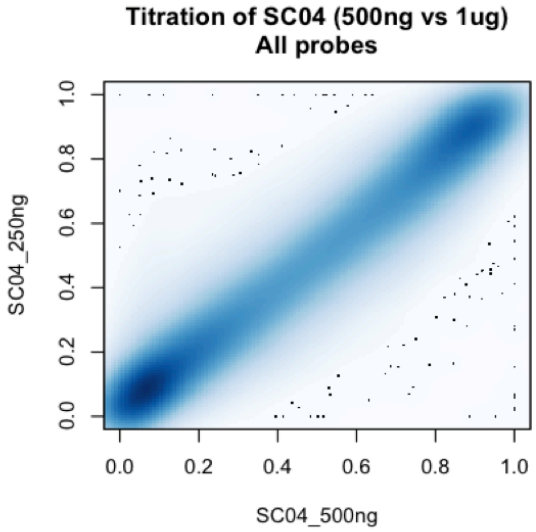
1.2.2.4: Titration experiment revealed good concordance from 125ng to 1ug of input DNA

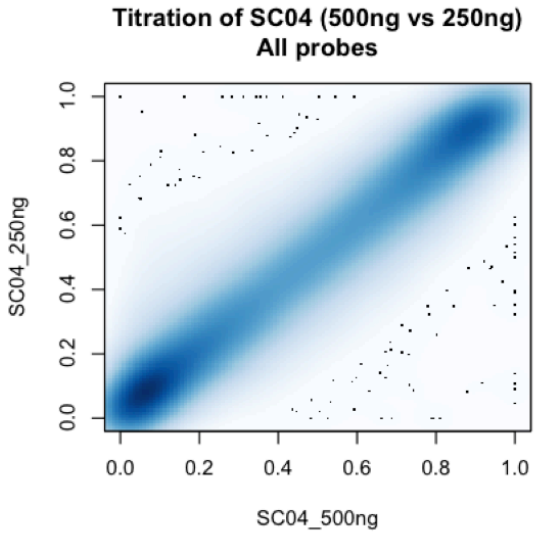
In an effort to assess the upper and lower limits of input material for the FFPE restoration reaction, four different NaBi-DNA inputs of SC04 were used in the FFPE DNA restoration step of the 450K FFPE protocol; 125ng, 250ng, 500ng, and 1ug. SC04 was a sample with relatively high yield and good quality DNA (QC delta $C_T = 5.75$).

We observed strong probe-wise correlations across all four amounts of input NaBi-DNA (Figure 1-9). Perhaps unsurprisingly, there waswas little improvement

observed when this comparison was performed for high quality probes, likely due to the fact that most of the probes in these arrays were detected in all samples (Figure 1-10).





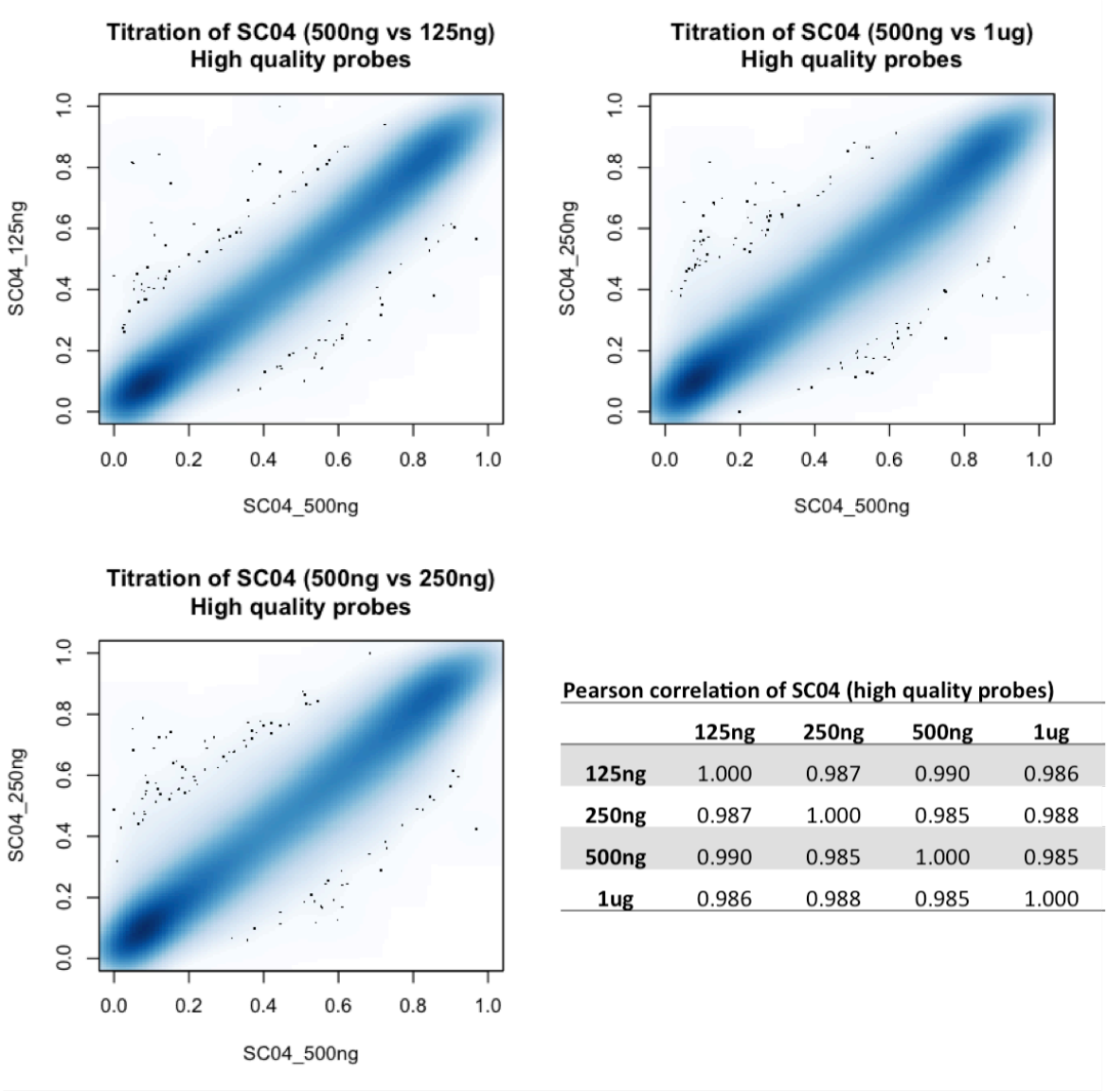


Pearson correlation of SC04 titrations (all probes)

	125ng	250ng	500ng	1ug
125ng	1.000	0.987	0.989	0.986
250ng	0.987	1.000	0.986	0.988
500ng	0.989	0.986	1.000	0.986
1ug	0.986	0.988	0.986	1.000

Figure 1–9: Probe-wise comparison across different NaBi DNA inputs of SC04

Colored density plots comparing reproducibility across different amounts of NaBi input for SC04 across all probes. Pearson correlation shows no statistically significant differences across different inputs. See Figure 1-10 for high quality probe comparison.



Pearson correlation of SC04 (high quality probes)

	125ng	250ng	500ng	1ug
125ng	1.000	0.987	0.990	0.986
250ng	0.987	1.000	0.985	0.988
500ng	0.990	0.985	1.000	0.985
1ug	0.986	0.988	0.985	1.000

Figure 1–10: Probe-wise comparison in high quality probes across different NaBi DNA inputs of SC04

Colored density plots comparing reproducibility across different amounts of NaBi input for SC04 across high quality probes. Pearson correlation shows no statistically significant differences across different inputs. There was also no qualitative or quantitative difference compared to the same analysis across all probes (Figure 1-9).

1.2.2.5: Conclusion

From this pilot experiment, we have established the use of Tween-80 in place of SDS for extraction of a subset of the sections to estimate the amount of DNA extracted and performed QPCR-based Illumina QC reactions. The TNET lysates were successfully combined with TNES lysates to maximize DNA yield for subsequent bisulfite treatment. We also observed call rates > 80% in 37/47 samples with > 5 delta C_T. Restriction of delta C_T to < 9 increased the samples with > 80% call rates to 93.5% (15/16) of the samples with 500ng of input NaBi-DNA. This suggests that with increased input NaBi-DNA, a relaxed threshold can be used for delta C_T in selecting samples for the microarray. Samples with increased delta C_T have rapidly decreased success rates, to as low as 50%.

Furthermore, we explored the use of stringent detP thresholds in defining high quality probes and showed increased concordance between replicates, suggesting improvement in measuring true signal intensities.

Finally, we have determined that in a sample of sufficient quality for good microarray results using 125ng NaBi-DNA, increasing the amount to 1ug did not change the methylation profile of the sample. This suggests that for consistency in experimental design, increasing input NaBi-DNA across all samples to accommodate samples of questionable quality will likely improve overall call rates with no detrimental effect to data quality in samples of high quality.

1.2.3: Efficient co-extraction of RNA and DNA

The focus of this thesis is the multiomic analysis of archival FFPE material across various neoplasms, including early stage disease such as ductal carcinoma in situ (DCIS). Using DCIS as an example, such lesions tend to be smaller compared to their invasive counterparts and efficient retrieval of DNA and RNA is crucial for maximizing the analyses, both on high throughput platforms or otherwise, that could be performed in these samples. Furthermore, in retrospective studies, FFPE blocks of the disease of interest may be limited, and only a small number of sections may be available which may impede the ability to extract enough RNA and DNA from the same sample, forcing the investigator to analyze only one nucleic acid species.

Co-extraction of RNA and DNA from the sample can serve to overcome section limitations, and have the potential of increasing overall RNA and DNA yields. Beyond that, in studies of disease with heterogeneous cell populations, co-extraction will provide matched nucleic acid fractions and allow for better integration of data generated across different molecular platforms. Therefore, we designed a pilot study to compare yield and quality of co-extraction methods to previously optimized RNA or DNA extraction methods, and incorporate it into our laboratory workflow for microarray profiling of FFPE material.

In this study, we compared the optimized individual extraction methods to two co-extraction protocols; 1) Qiagen Allprep FFPE DNA/RNA kit, and 2) a Trizol-based co-extraction method developed by Kotorashvili et al. We assessed this in 7 FFPE breast cancer samples with 18 sections each; where 3 sections were used for Allprep, 3 sections

for the Kotorashvili protocol, 3 sections for TNES, and 3 sections for Roche Highpure FFPET RNA kit, with every other slide distributed into the different assessments.

1.2.3.1: RNA yields were comparable across methods but TNES had higher DNA yields compared to co-extraction methods

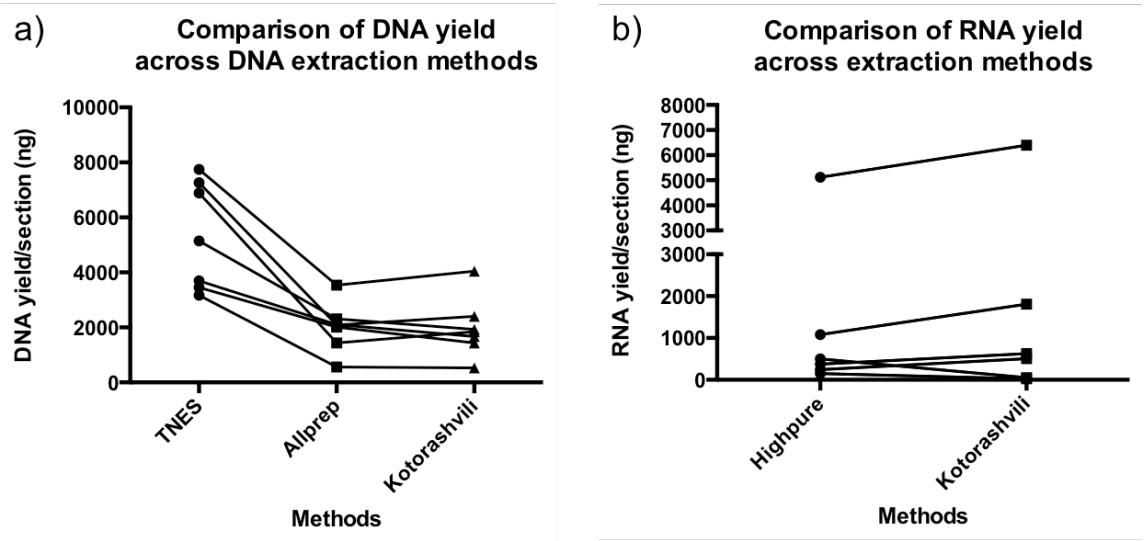


Figure 1–11: DNA and RNA yields from different extraction methods

a) DNA yield comparison across three separate methods; TNES for only DNA extraction, and Allprep and Kotorashvili for RNA/DNA co-extraction. b) RNA yield comparing Highpure RNA-only extraction and the Kotorashvili method.

There was an average of a 3-fold statistically significant increase in the DNA yield per section obtained from the TNES method compared to either co-extraction methods (Figure 1-11a). This is expected, as the TNES protocol does not contain a DNA purification step compared to the other two methods, which contain column-based DNA purification steps. In a previous experiment, we've shown that Allprep RNA and Kotorashvili extraction methods had comparable RNA yields (data not shown), and in this experiment we observed comparable yields between Highpure and the Kotorashvili method (Figure 1-11b).

1.2.3.2: Conclusion and method selection

Taken together, this suggests that while the co-extraction methods are less efficient compared to recovering DNA in lysates for bisulfite conversion reactions, the RNA yields are comparable to RNA-only extraction methods. Note that with the co-extraction method, we will be able to perform DNA/RNA extraction on double the number of slides compared to individual methods, effectively doubling the RNA yield and decreasing the difference between DNA yields of the competing methods. Furthermore, we have yet to compare the efficiency of DNA recovery against protocols with DNA purification, a process which will incur losses in recovery. As such, a co-extraction method will maximize the yield of nucleic acids, should the investigator choose to analyze both RNA and DNA. If only one species is of interest, I would recommend using methods specific to it to minimize cost and processing time.

Between the two co-extraction methods tested, I opted to use the Qiagen Allprep FFPE RNA/DNA kit due to increased reproducibility, more efficient time use, and lower cost. Firstly, the Kotorashvili method includes phase separation steps that may introduce variability across samples and, even more so, across operators. I aim to establish a workflow for FFPE materials that maximizes reproducibility for all FFPE-related high throughput genomic work. Secondly, while both protocols require multi-day processing of FFPE samples, the Allprep kit requires less active processing time, which allows for more efficient staggering in the processing of large batches of samples. Lastly, due to the use of RNaseOut in the Kotorashvili method, the price per sample reaches a little over

$20, while the price per sample of the Qiagen kit is $12.82, at the commercial rate without factoring in institutional discounts.

In conclusion, the general workflow for processing of FFPE materials for high throughput nucleic acid analyses will be a modified Allprep co-extraction protocol (see Methods), supplemented with TNES extraction where necessary (for example where increased DNA yields for methylation analysis are needed).

1.3: Final workflow for high throughput analysis of FFPE-derived nucleic acids

We have developed a comprehensive series workflow for analysis of FFPE tissue (Table 3). The tissue will first be assessed for cellularity and the need for enrichment, either by macro- or microdissection. Light H&E staining can be performed as necessary. Following that, extraction, QC, and processing of the molecular platform should be performed based on Table 3.

Table 3: Workflow for various high throughput –omics analyses of FFPE material

	Nucleic acid extraction	QC	Processing	Verified methods	Bioinformatics protocols
RNA-based assays					
Microarray	Highpure	QPCR-based		DASL	Assess P95 for sample-wise error
Sequencing	Highpure	Bioanalyzer on input and library for high MW fragments	Adjust input into library to high MW fragments	Illumina RNA Access, Total RNA-seq	RNASeQC to assess sample-wise success
DNA-based assays					
Methylome					
Microarray	TNES/TNET followed by NaBi	Illumina FFPE QC kit (delta Ct < 9)	Illumina FFPE restoration kit	Illumina 450K	Filter for high quality probes and samples with > 80% call rate
Genome					
Microarray	Qiagen DNA	Illumina FFPE QC kit (delta Ct < 9)	Illumina FFPE restoration kit		
Sequencing	Qiagen DNA	Bioanalyzer on input and library		Targeted sequencing	Filter for high quality calls with multiple reads
RNA & DNA assays					
Various platforms	Qiagen Allprep RNA/DNA		Refer to individual platform suggestions		

1.4: Materials and Methods

Patient selection and tissue collection

We used patient registries here at Johns Hopkins Hospital and at collaborating institutions to identify cases and controls that matched study criteria and had documented long term follow up. Tissues were obtained with approval of the respective institutional IRBs. Study pathologists reviewed archival H&E sections to select FFPE tissue blocks. Unstained tissue sections were obtained and macro dissected using pathologist annotated H&E sections for orientation and macrodissection for enrichment.

RNA/DNA extraction and quantification

RNA/DNA extractions during comparative studies were performed according to manufacturer recommended protocols (Roche Highpure FFPET RNA kit, Qiagen Allprep FFPE RNA/DNA kit) or published protocol in the case of the Kotorashvili method. Protocols for TNES/TNET extraction are appended. Following that, protocols were optimized for extraction from sections or cores, all of which are also appended.

Quantification of RNA and DNA were performed using Nanodrop2000 and the Qubit fluorometer (Qiagen) using appropriate kits (RNA HS, RNA BR, DNA HS, and DNA BR). The 260/230 and 260/280 ratios were used to assess sample purity and solvent contamination. Qubit derived measurements were ultimately used to calculate nucleic acid input into microarray platforms.

Quality control

RNA: SYBR Green-based QPCR for GAPDH using custom primers designed for short amplicons was used to measure the capacity for amplification of RNA samples. Personal correspondence with the Lowe's Family Genomics Core at Johns Hopkins Bayview Medical Center showed correlation between this assay and performance on the DASL microarray as measured by P95 signal and number of probes detected. At the request of the core facility, the primer sequences will not be reported, but is available in the lab database. In our experiments, we used the iTaq™ Universal SYBR® Green Supermix

(Bio-Rad, Hercules CA). Reverse transcription was performed on 50ng of input RNA using MMLV and random hexamers without second strand synthesis into a final volume of 50uL. 10uL was used in each QPCR reaction performed in triplicate. In this analysis, samples with C_T value $\leq$ 33 should be considered for microarray using standard protocols, and increased input should be consider for samples with values > 33.

DNA: Illumina FFPE QC kit was performed using the iTaq™ Universal SYBR® Green Supermix and was regarded as the main quality control step for 450K and other DNA-based microarrays. Illumina reported that a delta C_T of > 5 compared to control should not be used for restoration and subsequent microarray, but our experience show us that a delta C_T of up to 9 allowed for > 90% of the samples having call rates above 80%. As such, the recommended threshold for delta C_T is 9, prioritizing samples with lower delta C_T.

Bisulfite conversion

Bisulfite conversion was performed using the EZ DNA Methylation-Gold™ Kit (Zymo Research, Irvine CA), with modifications introduced per Appendix I of the manufacturer's recommended protocol. The detailed protocol is appended at the end of the thesis.

DNA restoration and microarray

The DASL microarray was processed at the Lowe's Family Genomics Core Facility at Johns Hopkins Bayview Medical Center. The FFPE DNA restoration and 450K microarray was performed by the Sidney Kimmel Comprehensive Cancer Center (SKCCC) Microarray Core Facility.

Data processing and analysis

Quality control metrics for Illumina-based arrays were estimated using Illumina's GenomeStudio software, and validated in the R Statistical Environment using Bioconductor packages or custom functions to extract control probe signal intensities. Illumina preprocessing was performed on the DASL microarray using GenomeStudio and exported as a data matrix for analysis in R. 450K microarray was read and preprocessed with Illumina algorithm using the minfi package in R and analyzed using Bioconductor packages and custom functions.

Chapter 2: Identification of copy number variation from high density methylation microarrays

2.1: Introduction

2.1.1: Rationale

Both genetic and epigenetic alterations are implicated in the development of cancer [77]. Genetic lesions, such as translocations, amplifications, insertions, deletions, and point mutations, have been implicated in promoting the development of cancer through alteration of expression or activating and in-activating mutations of tumor suppressors and oncogenes [78]. Epigenetic modifications, through DNA methylation or histone modifications, lead to the silencing or reactivation of genes, which can translate into phenotypic changes that also lead to carcinogenesis [79, 80]. DNA methylation of CpG islands in the promoter region and first exon of genes result in changes in gene transcription. CpG island hypermethylation is a phenomenon common across multiple cancer types, and has been shown to lead to silencing of tumor suppressor genes [81].

Integration of genetic and epigenetic data can provide a more complete view of disease processes, including underlying pathogenic mechanisms [82]. However, there are limitations on performing concurrent genetic and epigenetic characterization of tumor samples, including cost and availability of sufficient material. The ability to read out multiple data types from a single platform both minimizes cost and ensures a single source of test material.

2.1.2: Methylation and SNP microarray technologies

At the time of the study, the Illumina Infinium Human Methylation HM450K microarray (HM450K) is the leading methylation microarray containing 485,577 CpG probes that are widely distributed across both intra- and intergenic regions of the genome. The SNP6 microarray platform from Affymetrix has been extensively used to obtain high-resolution genome-wide DNA copy number estimates. The SNP6 microarray contains 1,852,600 probes, with 906,600 SNP probes for variant detection and 946,000 non-polymorphic probes for copy number estimation.

2.1.3: Platform similarities and study setup

Both SNP and methylation microarrays use probes tagged with different fluorophores to identify genetic and epigenetic variants of a given genome locus, and the technical aspects of both array platforms are very similar. Furthermore, the bisulfite treatment of DNA that precedes hybridization on methylation arrays chemically creates the equivalent of induced SNPs at unmethylated CpG dinucleotide sites, allowing a similar analytical approach on the two array platforms.

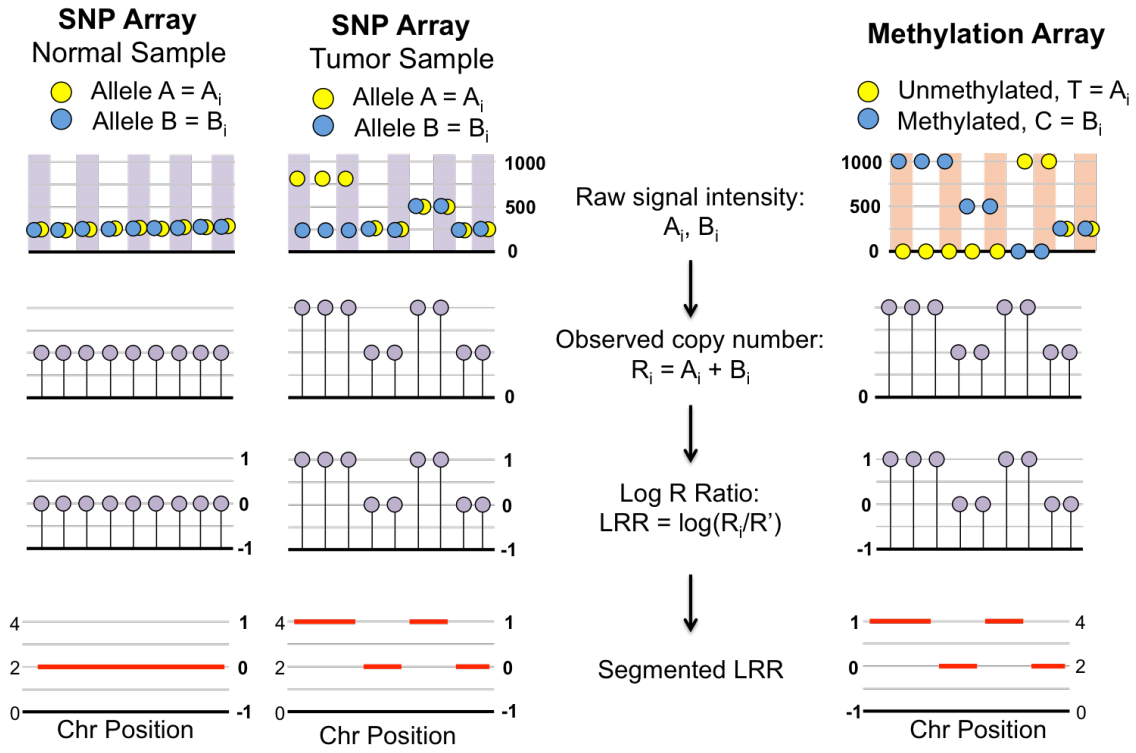


Figure 2–1: Graphical representation of SNP and methylation array similarities.

The two left panels show the graphical representation of the process of obtaining CNV information from a normal sample (left) and tumor sample (right). The rightmost panel shows the same process using a methylation microarray, which can be viewed as an induced SNP array with bisulfite treatment converting the unmethylated C of a CpG dinucleotide to a T. Each alternative colored column represents a locus in the genome interrogated by the array. The filled circles of different colors represent the probes and the height represent the signal intensities. The purple filled circles represent the summed signal intensities and finally the red horizontal bars represent segmented genomes.

These similarities have driven previous efforts to obtain copy number variation (CNV) information from HM450K microarrays [83-85]. The cumulative intensity values from the unmethylated probe, U, and the methylated probe, M, is theoretically a proxy of total DNA copy number at that locus. This is the approach used by Sturm et al. [83] to characterize the CNV in a series of glioblastomas. The "getCN" function from the minfi package on Bioconductor sums the raw intensities of U and M to obtain total intensities [84]. Feber et al. most recently proposed a statistical pipeline that provides copy number

information from HM450K arrays [85]. However, there has been no study to date investigating in detail the circumstances in which it is possible to reliably obtain CNV information from methylation arrays.

Herein, I study the experimental parameters allowing the reliable assessment of CNV using HM450K arrays, suggest an optimized method that provides detailed CNV estimation, and summarize its performance on various Cancer Genome Atlas (TCGA) datasets where both methylation and CNV data are available for several tumor types. A better understanding of when one can efficiently use and how to interpret CNV information from HM450K arrays will help in deciding when reliable CNV calls can be made using just one platform. This in turn would significantly reduce costs and tissue requirements, and ensure that both measures are derived from an identical DNA sample.

2.2: Methods

2.2.1: Data download and analysis

The Illumina Human Methylation HM450K .idat files (TCGA level 1 data) for THCA, BRCA, and LUSC were downloaded from the Broad Institute's Firehose Genome Data Analysis Center (GDAC) server (data freeze 12/2013). Processed Affymetrix SNP6 array data (TCGA Level 3 data) and the accompanying GISTIC results for the same tumors were downloaded from the same server, for comparison. Data was analyzed using the R statistical environment (Version 3.1.1), packages from Bioconductor, and custom functions.

2.2.2: Estimating copy number using Epicopy

Raw methylation data were processed using the functional normalization algorithm adapted from the developer's version of the minfi package [84] to return red-green channel data.

Log2 signal intensities for both the methylated, M_i, and unmethylated, U_i, channels, as calculated in minfi, were summed together to obtain total signal intensity, t_i, of genomic position *i*. Normal samples (n=55) from the THCA dataset were used to represent the diploid genome, and signal intensities for position *i* in normal samples are represented using $T_{i,J,}$ where J = [68]. Specifically, at each genomic position *i* represented on the array, we calculated $\hat{T}_i$ = mode(T_{ik}), as estimated using the naive estimator from the modeest package [86]. These values were then used to calculate the log R ratio (LRR) of the intensities, Δt_{ij}, for genomic position *i* in sample *j*.

$$LRR, \Delta t_{ij} = \log(t_{ij}) - \log(\hat{T}_i)$$

Finally, the mean Δt_{ij} was centered at zero and subjected to circular binary segmentation (CBS), as implemented by the DNAcopy package [87], using default options, to obtain copy number estimates λ_{ij}, which represents log R ratio of a given genomic position i for sample j. GISTIC 2.0 was then used to identify gene-level copy number events as λ_{gj} using parameters defined in the following section.

2.2.3: Selecting model parameters

The CN value from SNP6 Affymetrix array, θ_{gj}, was used as the true copy number of gene *g* for sample *j*. Copy number values at gene *g* for SNP arrays are categorized as deleted, amplified, or copy-neutral, where $\theta_j \geq 0.3$ is considered an amplification and $\theta_j \leq -0.3$ a deletion, based on empiric decisions from previous studies.

Using the THCA dataset, the optimal mean segment CN values for making amplification and deletion calls were selected from a range of values between 0.03 and 0.3 to maximize the accuracy of λ_{gj}, $\frac{\sum_{N=1}^{j}(\sum_{n=1}^{g} B|\theta_{gj}|-B|\lambda_{gj}|)/n}{N}$, where B is a binarizing term. This threshold value is varied for Δt_{jg} to find the optimal cutoffs for making amplification/deletion calls using the Epicopy method by maximizing accuracy in gene-level CNV.

The mean gene level threshold, identified from the step above, was then used to determine the minimum number of probes required for a confident segment call. We hypothesized that normal thyroid tissue samples have copy number neutral (2n) genomes and that any segment that had an absolute CN value ($|\lambda_{jg}|$) above 0.15 was defined as false positive. We aimed to minimize the false positive rate while controlling for probe number threshold, which translates to genomic coverage. With increased probe number threshold requirement, the less amount of the genome is covered.

2.2.4: Calling copy number events

GISTIC2.0 [88] was performed to identify focal and arm level events that are 1) recurrently amplified or deleted in each tumor type and 2) generate gene-level copy number estimates, $\lambda_{gj,}$ for gene *g* in sample *j*. Default parameters were used when applying GISTIC to SNP arrays, including the requirements that segments include at least 5 tags and that log R ratios, $|\theta_{gj}| \geq 0.3$, when calling amplifications and deletions. Parameters were derived for use with CHAMP-CNV, and Epicopy. Specifically, I required that segments contain at least 200 tags and log R ratios, $|\theta_{gj}| \geq 0.15$ for Epicopy and 200 tags and LRR of 0.11 for CHAMP-CNV.

2.2.5: Performance metrics

Several measures were used to compare results obtained from methylation arrays using Epicopy to those derived using other algorithms or platforms. Concordance, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the ROC curve were calculated using the pROC package [89]. These were used to benchmark performance from methylation arrays against standard SNP arrays, as each metric assesses different aspects of the technology. Concordance allows us to measure the degree of accuracy across measures. Sensitivity reports the percent of CNVs detected by SNP6 that is detected by Epicopy, while specificity identifies the proportion of copy number neutral genes identified by Epicopy. The PPV allows us to estimate the

percent of positive calls made by Epicopy that is identified by the SNP6 array. In other words, of all the CNVs detected by Epicopy, how many are true. The same for NPV, except for genes detected as copy number neutral. The AUC estimates overall Epicopy performance, regardless of threshold.

The reproducibility index is a Jaccard similarity coefficient calculated at the gene-level [90], which calculates the ratio of the intersection of CNVs across both sets over the union of both sets. This measure is similar to sensitivity but treats methods symmetrically rather than assuming that one method represents a benchmark. The reproducibility index for sample *j* was calculated as:

$$Reproducibility\ Index_j = \frac{A \cap B}{A \cup B} = \frac{|CNVs\ identified\ by\ both\ methods|}{|CNVs\ identified\ by\ either\ method|}$$

Local regression to highlight trend in figures was performed using the locfit function from the locfit R package [91].

2.3: Results and discussion

2.3.1: Sample selection

Three TCGA datasets, thyroid carcinoma (THCA) [92], breast carcinoma (BRCA) [5], and lung squamous cell carcinoma (LUSC) [93], were chosen for model development and validation because a large number of samples with paired SNP and methylation arrays were available for these datasets. THCA, which has few, but frequently recurrent CNVs [92], was used for model development, while BRCA and

LUSC, representing cancers with many CNVs per sample, were used for testing. Combined, these three datasets are representative of the CNV spectrum in human cancer [94].

2.3.2: Feasibility and probe coverage

The feasibility of obtaining CNV calls from the HM450K array depends on having sufficient probe coverage and the ability to optimally normalize probe signal intensities. Illumina HM450K microarrays have a smaller genomic coverage than the current generation of SNP arrays, having 485,577 probes compared to 1.8 million probes (906,600 SNP probes and 946,000 non-polymorphic copy number probes) in the SNP6 platform from Affymetrix. The latter was chosen by the TCGA consortium to obtain DNA copy number estimates, and the distribution of probes across the genome differs as well. Although the probes in the HM450K array are distributed across both inter- and intragenic regions (Figure 2.2), they are concentrated in intragenic regions and particularly in gene promoters. This may result in different regions across the genome having varying sensitivities for making CNV calls.

Another factor likely to affect performance is that the Illumina Methylation array is designed with two different probe chemistries, with unique distributions of the probe intensities [95]. The probe type is closely correlated to the CG content of the probe sequence, so that different regions of the genome are enriched for each probe type. This issue was addressed in two ways in the Epicopy algorithm. First, the **functional normalization algorithm** (funnorm) [84] used for preprocessing includes steps to

minimize differences between the two probe types. Second, normal samples suffering from the same technical concerns but having little or no copy number variation are used to standardize probe level estimates of abundance.

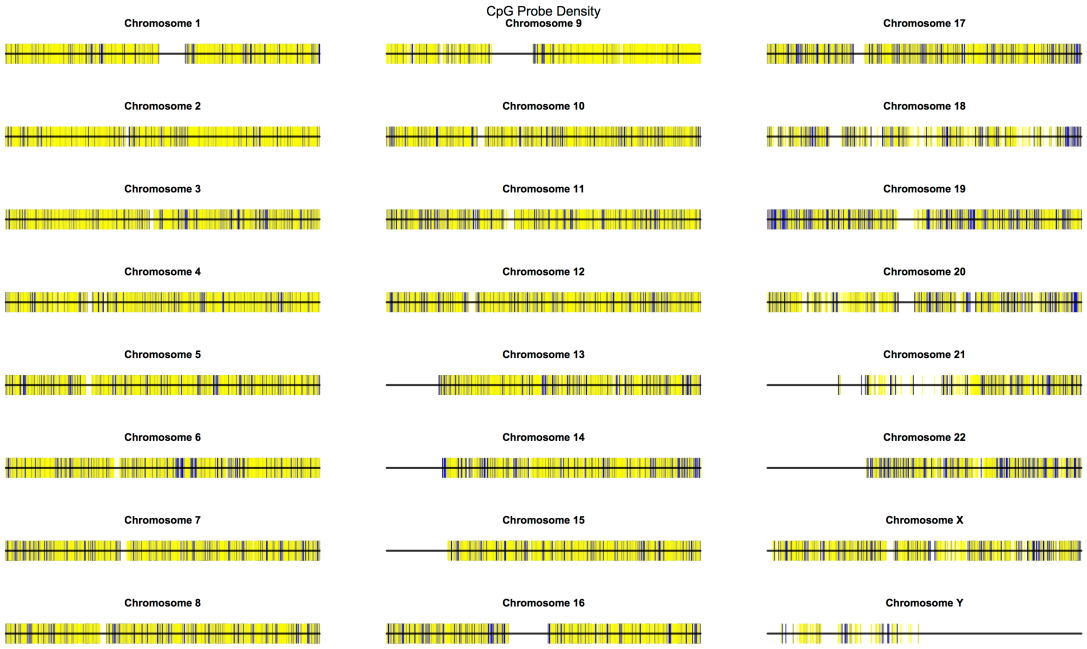


Figure 2–2: HM450K probe coverage

Chromosomes are not to scale compared to each other. Despite having only 485,577 probes, there was good coverage of all but 1 autosome (Chr 21) and Chromosome X. Colors indicate probe chemistry type: blue: Infinium I; yellow: Infinium II

2.3.3: Obtaining copy number calls with Epicopy

The Epicopy pipeline is shown in Figure 2-2. Briefly, after normalization, the log ratios of probe intensities of tumor samples to the mode of normal reference samples were calculated and mean centered before segmentation using the circular binary segmentation (CBS) algorithm [96]. This pipeline was used to obtain segment information in the THCA, BRCA, and LUSC datasets. Following segmentation, GISTIC2.0 [88] was used both to estimate gene-level copy number calls for each sample and identify significant copy number altered regions (SCNA). The resulting data can be compared directly to TCGA SNP6 results, which were processed using Broad Institute's Copy Number Pipeline-analyzed SNP6 data analysis [97]; this includes both CBS and GISTIC 2.0 results.

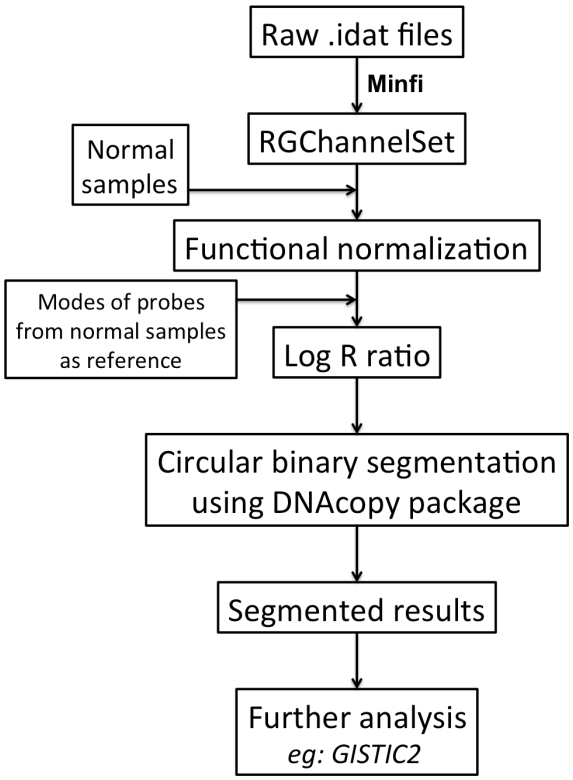


Figure 2–3: Epicopy pipeline

As the methylation microarray is not optimized for obtaining CN information, I observe some differences in the segments derived using SNP6 and Epicopy. The log R ratios (LRR), or magnitude changes in copy number compared to reference, are lower in Epicopy segments than in their SNP6 counterparts, which are approximately twice as

large. The direction of copy number change, however, remains the same. Furthermore, Epicopy-derived segments are more fragmented than SNP6 segments. For example, a single CNV event, identified as a single segment in SNP6, may be represented by multiple adjacent segments in Epicopy. In spite of these differences, Epicopy results closely approximated results obtained by SNP6 CNV analysis. This is illustrated in a representative comparison of SNP and Epicopy CN profiles from a breast tumor sample showing that Epicopy is able to detect chromosomal, arm, and focal copy number changes (Figure 2-4).

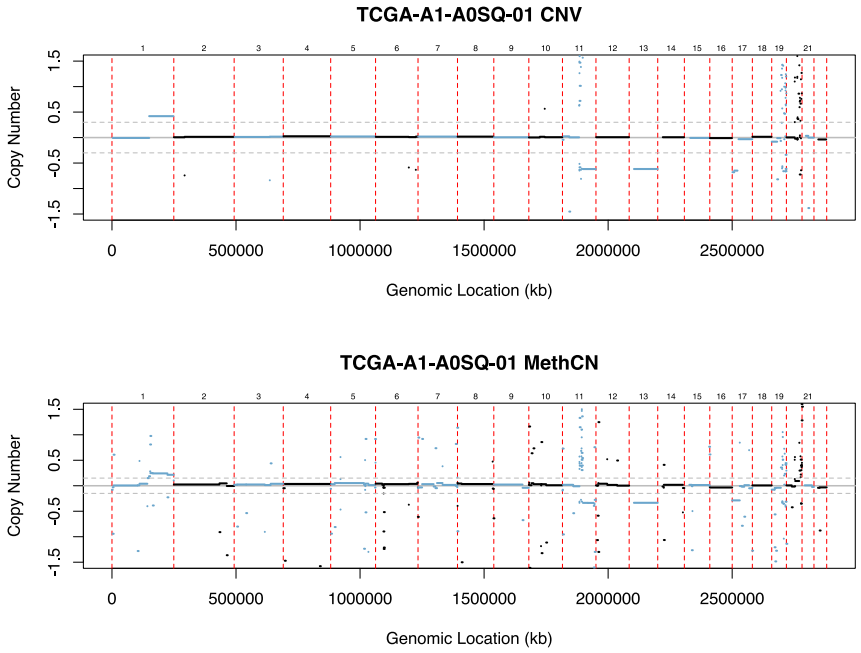


Figure 2–4: Representative example of Epicopy- and SNP-derived copy number profile

Copy number profiles of the same breast cancer sample from SNP array (top) and Epicopy (bottom). Note the lower copy number values on the Epicopy derived CNV information. The y-axis represents copy number (LRR). The x-axis represents genomic location. Dotted red lines signals a transition across chromosomes. The horizontal blue and black bars represent the segments with alternating color signifying chromosome transition. The dotted horizontal line is the threshold of making a CNV call.

2.3.4: Model parameters

On the most basic level, CNV calls can be made as a function of two parameters. The first was the magnitude of change, or log R ratio (LRR), which describes the minimum amount change in signal compared to a reference control needed to call amplifications and deletions with confidence. The second was the minimum probe count needed to define a segment. Inspection of the mean segment LRRs derived by Epicopy revealed that the dynamic range of LRRs was narrower in methylome arrays than in SNP arrays. Thus, values used frequently for SNP arrays are too liberal to be used with methylation arrays.

I derived optimal values for both parameters using the TCGA thyroid data. As detailed in methods, GISTIC 2.0 [88], with the default minimum number of probes per segment filter of n = 5, was applied to segmented data from both SNP and Epicopy-derived CNV profiles to infer gene specific levels for each sample, and make a call for each gene (amplified, deleted, or neutral) to be used as a standard of comparison. The concordance of calls between Epicopy and the SNP array was used as the metric to identify the optimal threshold to detect a CNV.

Increasing the CNV threshold sequentially from 0.03 to 0.3, I identified 0.15 as the value which maximizes the median accuracy across samples in THCA at 0.996 (Figure 2.5).

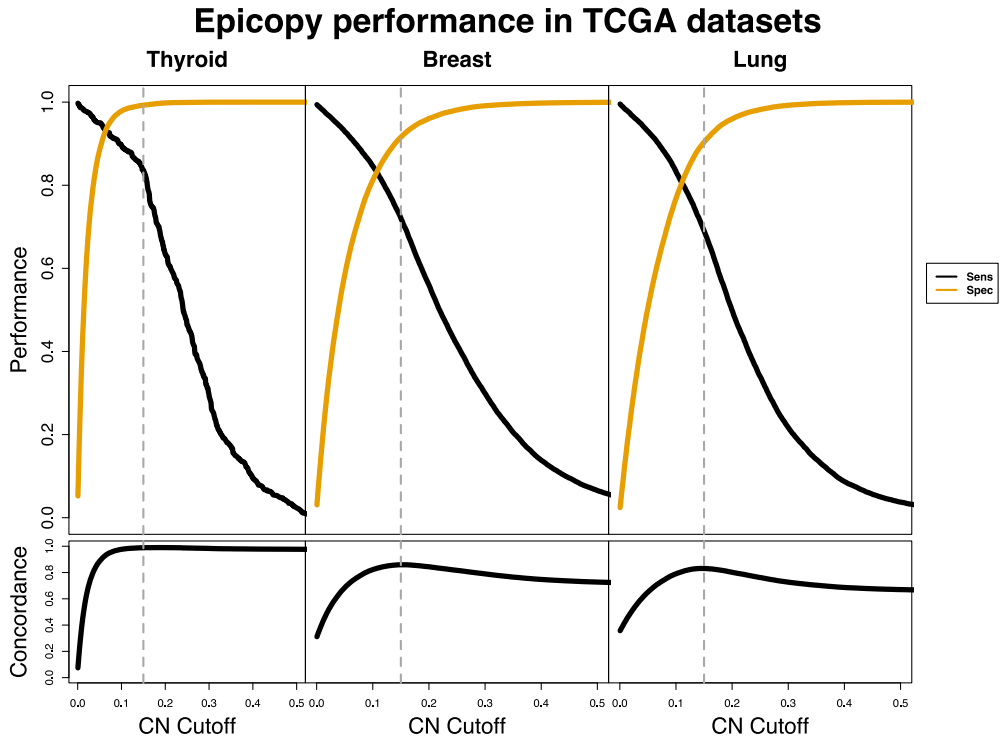


Figure 2–5: Gene-level performance of Epicopy

Performance of gene-level Epicopy calls against SNP analysis in three TCGA datasets; THCA, BRCA, and LUSC using a CN threshold of 0.15 and 200 probes per segment. In the top panels, the tan line represents specificity while the black line represents sensitivity. The bottom panel shows the concordance, or accuracy, of gene-level data.

Next, I used Epicopy segmented data for normal tissue to identify the minimum number of probes per segment needed for confident CNV calls. Here, I expected the optimal probe number threshold to minimize the false positive rate (FPR, the rate of segments with CNV that passes the probe number filter) while retaining the highest coverage of the genome. I observed that both metrics are inversely correlated with probe number and identified that a minimum probe number of 50 reduces the FPR to 0.1%, with a genomic coverage of 95.8% (Figure 2.6). This decrease of FPR continued to 200 probes per segment before the specificity decreased significantly. This suggests that the

optimal threshold for the number of probes per segment lies between 50 – 200 (Figure 2.6). Similar results were obtained for CHAMP-CNV. As a conservative measure, the analysis presented in the rest of this manuscript was performed using the 200-probe cutoff for both CHAMP-CNV and Epicopy.

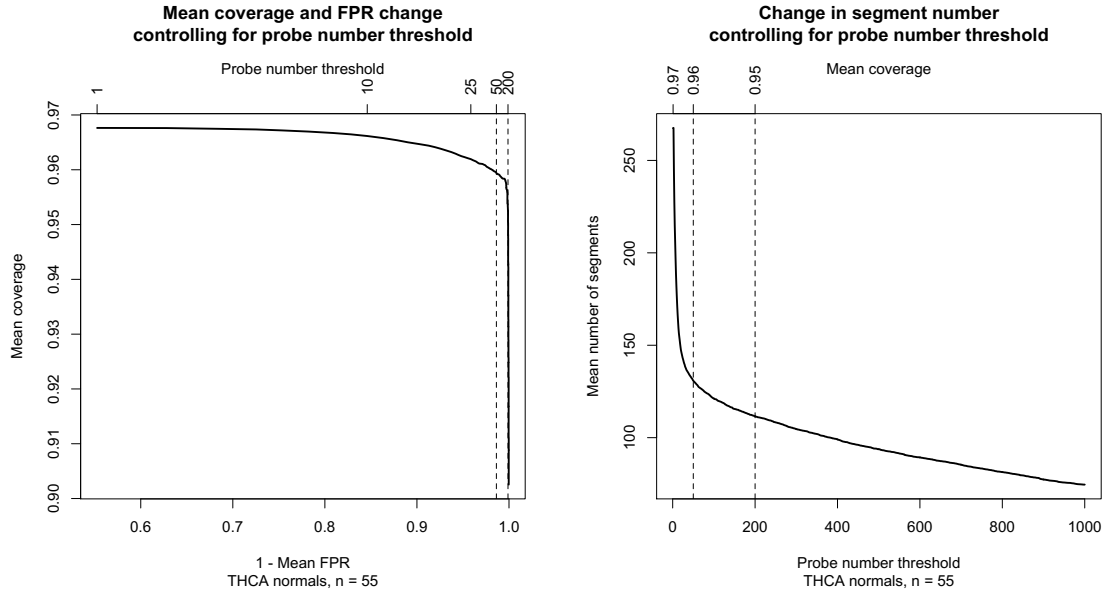


Figure 2–6: Coverage, FPR, and number of segments.

Average of 55 normal thyroid samples. a) Change in mean coverage (y-axis) and 1 – Mean FPR (x-axis) with increasing probe number threshold (top-axis). B) Change in mean number of segments with increasing probe number threshold. Mean coverage is shown on the top-axis.

We further investigated these thresholds using tumors from the same dataset by calculating sensitivity and specificity on a gene level, averaged across all samples as a function of mean segment CN threshold (Figure 2-6, left-most), and found a sensitivity of 84% and specificity of 99%.

To further understand the effect of the segmental LRRs identified by the SNP analysis on the ability of Epicopy to identify a lesion, we analyzed the percent gene-level amplifications detected by Epicopy in the THCA dataset while increasing the LRR of

these amplifications on the SNP microarrays (Figure 2-7). As expected, Epicopy was more likely to detect amplifications that were estimated by the SNP array to have a high copy number.

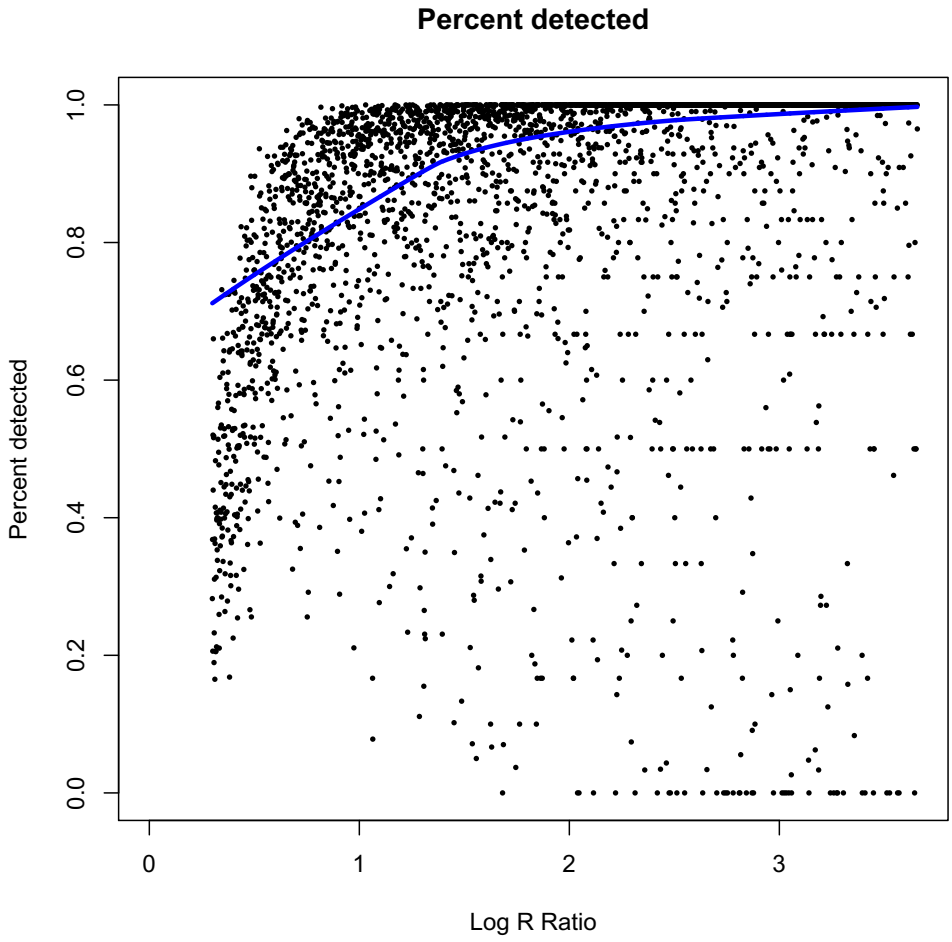


Figure 2–7: Percent alterations detected by LRR.

Each point represents the average of all segment identified by SNP array in the THCA dataset, disregarding the length or probe number within the segment. The x-axis represents the LRR of the segment in the SNP array and the y-axis represents the percent of those segments identified by Epicopy. The blue line is the local regression line fitted using the locfit function in R.

2.3.5: Epicopy performance on gene-wise correlations

We tested Epicopy's performance at the gene level, on the BRCA and LUSC datasets, using the CNV results from the SNP arrays as the standard of comparison. Measures of performance included overall accuracy, sensitivity, and specificity, evaluated at the gene level. Of note, we used thyroid normal tissue from TCGA as the reference diploid samples for both BRCA and LUSC, reflecting the commonly occurring situation where well-matched reference samples are not available.

With the thresholds of log R ratio (LRR), or magnitude of change, set at 0.15 and number of probes per segment set at 200, the accuracy of the method in the THCA, BRCA and LUSC datasets was 99%, 86%, and 83%, respectively. The sensitivity of Epicopy was 84%, 72% and 69%, respectively (Table 1, Figure 2-5), while the specificity was 99%, 92%, and 90%, respectively.

Table 4: Epicopy and CHAMP-CNV AUCs across 3 TCGA datasets

Dataset	Method	AUC
THCA	Epicopy	0.97
THCA	CHAMP-CNV	0.97
BRCA	Epicopy	0.90
BRCA	CHAMP-CNV	0.85
LUSC	Epicopy	0.88
LUSC	CHAMP-CNV	0.76

We further calculated a reproducibility index between SNP and Epicopy gene CNVs (Figure 2=8). This measure, which is based on the Jaccard distance [90], describes

the probability that a copy number alteration identified on either platform is found on both and has the advantage of treating the SNP and Epicopy results symmetrically. We observed an average reproducibility between CN calls from the SNP6 platform and from Epicopy of 27%, 57%, and 51% for THCA, BRCA, and LUSC, respectively.

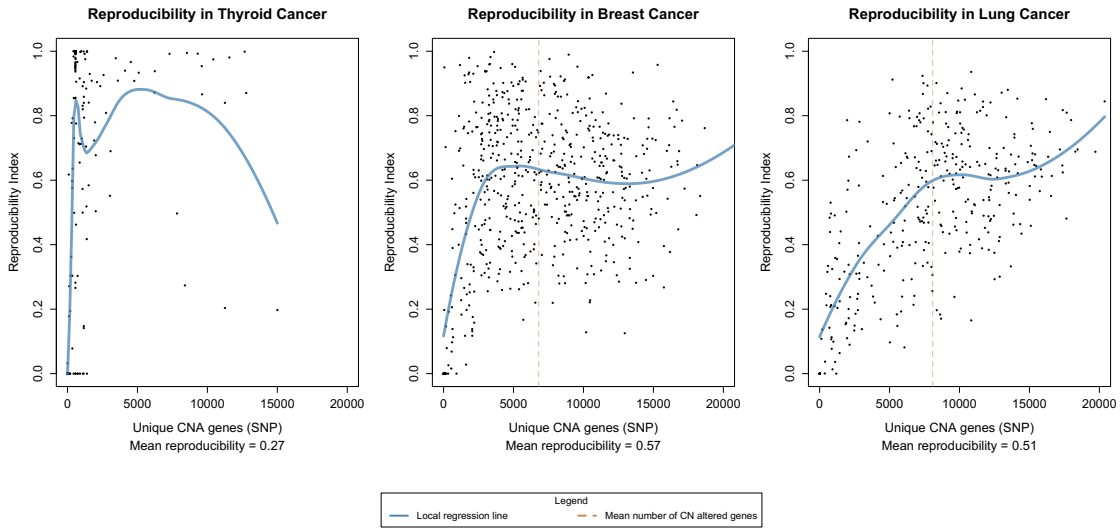


Figure 2–8: Reproducibility index

Each point represents a sample. Reproducibility index was calculated as the number of intersecting CN genes detected by both methods divided by the union of CN genes detected by both methods. The blue lines indicate a fitted average across all the points, calculated using the package locfit.

To put these reproducibility results in perspective, studies evaluating the CN detection reproducibility across SNP array platforms and even across different CNV-calling algorithms on the same platform have shown that reproducibility in replicate experiments ranges between 39% and 79%, even for within platform comparisons, while reproducibility across platforms ranges between 25% to 50% [98-100]. Specifically, the

maximum number of reproducible copy number alterations detected by the SNP6 platform as assessed by Pinto et al. using an identical algorithm was 79% [98].

Thyroid cancer is distinctive among these 3 tumor types because of its very rare copy number changes overall, and the low level of agreement seen in this analysis may be attributable to this [92, 94]. The reproducibility index weighs CNV events being called by either Epicopy or SNP analysis. In samples with no events identified by SNP analysis, as often occurred in the case of the THCA dataset, even a single spurious CNV event identified by Epicopy caused the reproducibility index to be zero.

Based on this, we believe that there is an upper limit to the reproducible CNV detection rate, given the present array technologies, and that Epicopy's performance, as assessed by the agreement between Epicopy and SNP6 measurements of CNV from high density methylation microarrays, is comparable to that seen between different SNP array platforms.

2.3.6: Epicopy performance on recurrent amplifications and deletions

All comparisons described so far were performed at the gene level, but it is common to report more highly summarized versions of these results, focusing on the most frequently recurring events that are likely to be driving the development and progression of disease. To assess Epicopy's ability to recapitulate such analyses, we used GISTIC2.0 [88], which employs a probabilistic method to identify peaks within the genome where recurrent CNV events occur within a set of samples.

TCGA SNP6 datasets were processed using the Broad Institute's Copy Number Pipeline-analyzed SNP6 data [97], which uses CBS to obtain CN segments. TCGA released these data with the GISTIC 2.0 output, which contains both the gene-level CNVs and recurrent CNVs. Since characterization of recurrent CNVs can be used to identify driver events, we compared Epicopy-based calls and SNP6 arm level events identified by GISTIC to assess Epicopy's ability to detect recurrent CNVs.

In BRCA and LUSC, two tumor types that are characterized by a high number of CNVs, Epicopy was able to identify 70% (Figure 2.9, 2.10) of peaks identified in the SNP6 platform.

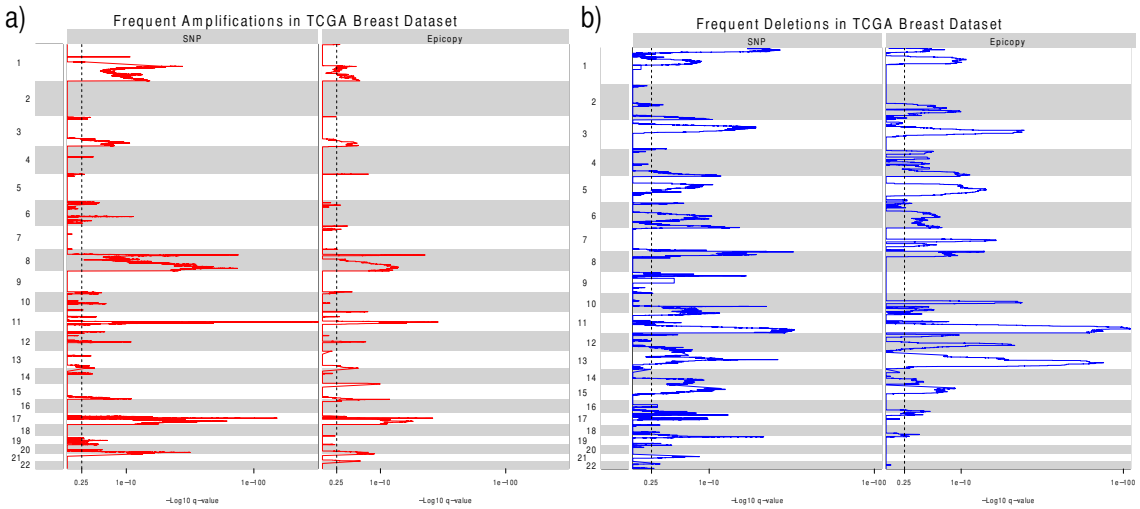


Figure 2–9: GISTIC comparison for BRCA validation dataset.

Comparison of the GISTIC results obtained by SNP analysis and Epicopy-derived values. A) Frequent (recurrent) amplifications identified by SNP- (left) and Epicopy- (right) –derived results. B) Frequent deletions identified by SNP- (left) and Epicopy- (right) –derived CNV results. There was 72% overlap between the recurrently altered peaks identified across both platforms.

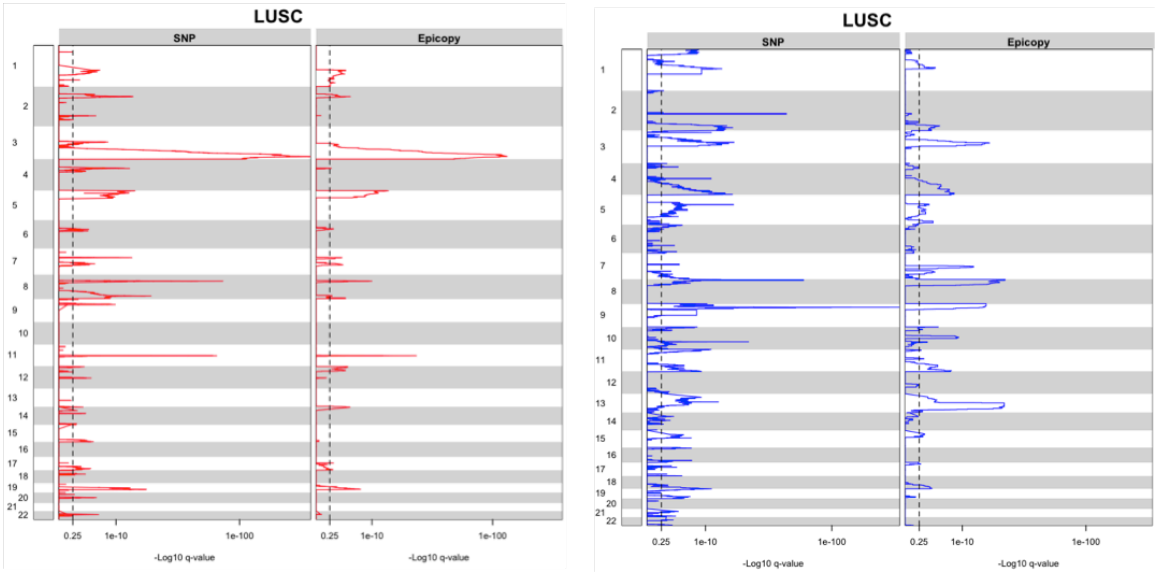


Figure 2–10: GISTIC results for LUSC validation dataset

Interestingly, there were peaks identified in the Epicopy data that were not seen by SNP6 in the THCA dataset (Figure 2-11). These may be false positives, which would reflect limitations in using HM450K for CNV profiling, or true positive calls that are detected by Epicopy and missed by SNP6-based analysis.

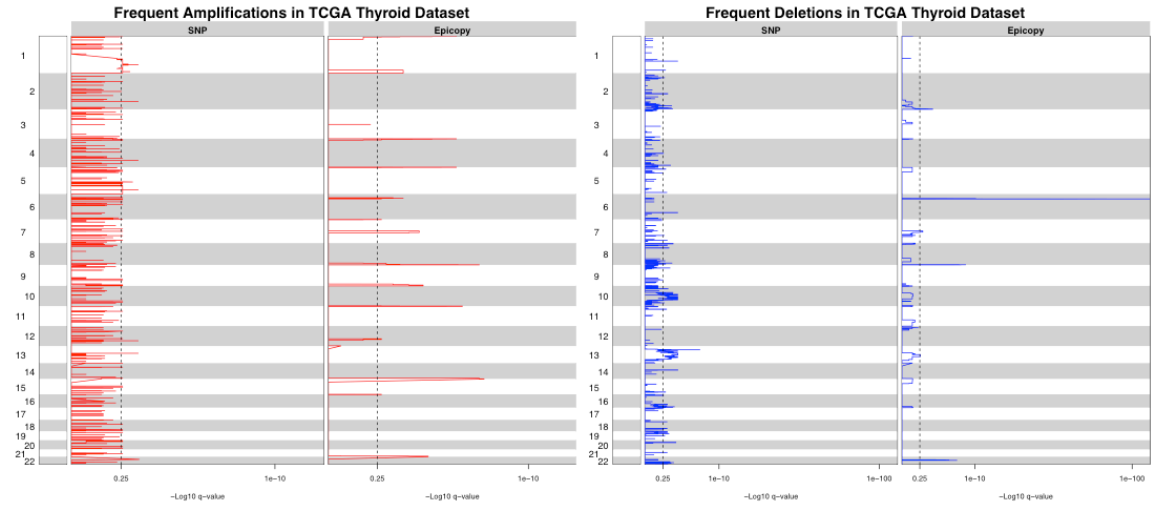


Figure 2–11: GISTIC results for the THCA dataset.

As discussed, there are regions of the genome where HM450K probe coverage is denser than in the SNP6 array and these peaks may fall into such regions. Indeed, when we investigate the probe density in these peaks for SNP6 and HM450K arrays, HM450K had more probes in 9 out of 12 peaks, suggesting that these are regions where HM450K is more sensitive at detecting CNV than SNP6. Furthermore, when we investigated the probe density of Illumina HM450K methylation array compared to the Affymetrix SNP6 array, we were able to show that Illumina CpG probes were enriched around transcriptional start sites (TSS) and exons (Figure 2-12). The enrichment of CpG probes in and around gene bodies suggests that in regions of the genome where the functional consequences of CNV is well understood, Epicopy-derived CNV profiles may be as sensitive or more sensitive than SNP arrays, especially for focal CNVs. In support of that, Feber et al. have shown that the HM450K CpG array was able to identify a PTCH1 focal deletion undetected by the Illumina CytoSNP array [85]. Of note, some biologically relevant thyroid driver genes are present in these peaks; e.g., TERT and AKT1 are amplified while BRCA2 was deleted. TERT amplification has been shown to be significant in familial papillary thyroid cancer (FPTC) patients [101] and genetic alterations in all three genes have been described by TCGA [92].

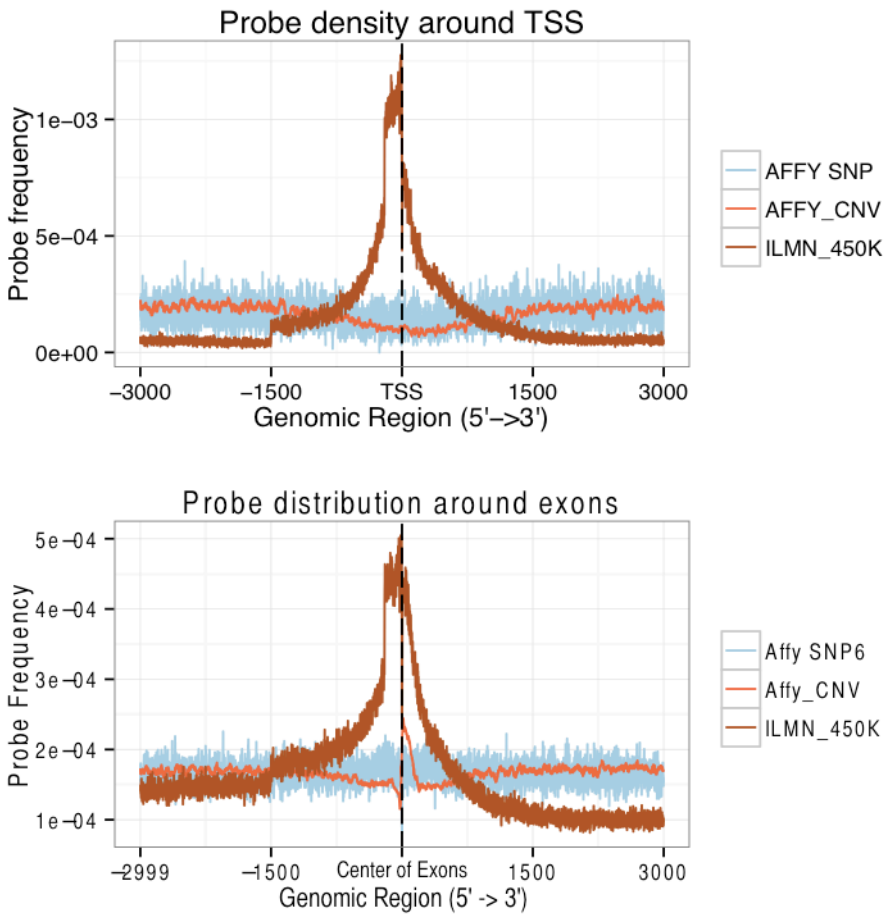


Figure 2–12: Probe density around TSS and exons for HM450K and SNP6.0 arrays

The distribution of probes in 450K and SNP6 microarrays around (a) transcriptional start sites (TSS) and (b) center of exons. Perhaps unexpectedly, 450K probes are concentrated around TSS and exons of coding genes, suggesting that in regions where we understand the direct implications of copy number change, 450K microarray can capture the changes as well as, if not better, compared to SNP6 microarray.

Additionally, there was a distinct peak at chromosome 6p22 detected by Epicopy but not the SNP array in the THCA dataset. The Illumina probes in this peak are situated in HLA genes, a known hypervariable region (Figure 2-13, red boxes). Furthermore, nearby probes outside of this hypervariable region show no CNV, suggesting that the deletion in this peak is indeed unique to probes within HLA genes. This implies a lack of

probe binding from probe mismatch due to genetic variation of this HLA region rather than the actual loss of copy number in this region. Therefore, it is recommended that these probes be removed from analysis when using the HM450K platform to profile CNV, and this is implemented by Epicopy.

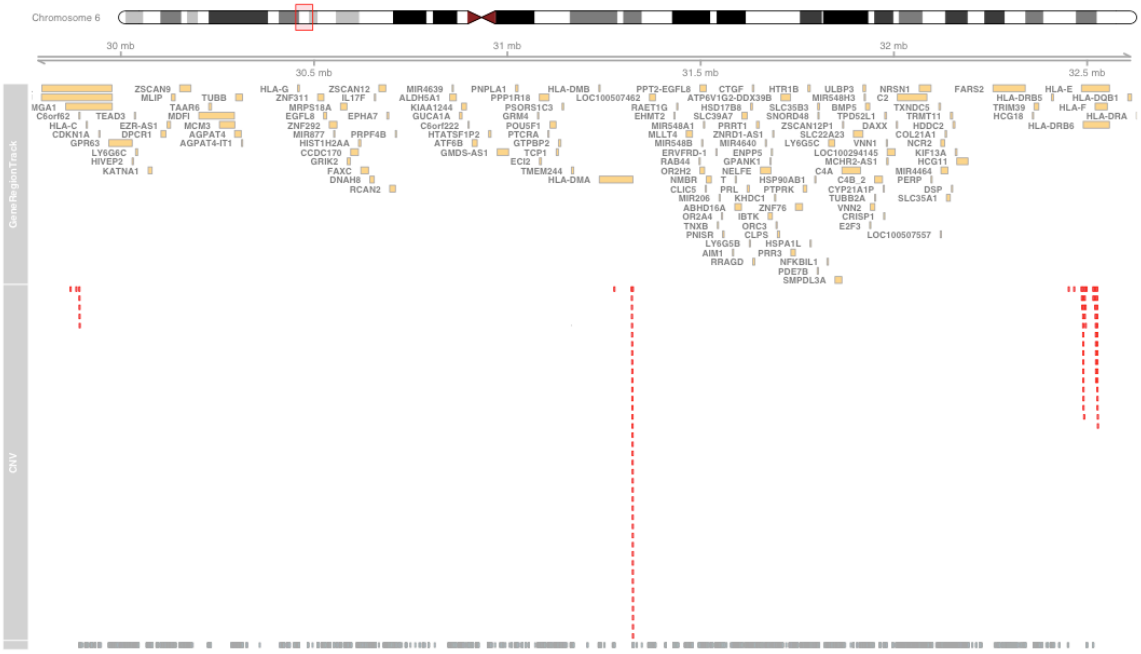


Figure 2–13: Probes for HLA genes are enriched in chr6q22

Region of the DNA with copy number loss in chr6q22. Each tan colored rectangle represents a gene. The red bars and ticks on the middle panel represents probes that contributed to chr6 loss calls in THCA data. The grey bars in the bottom panel represent other probes in the region that did not contribute to the call. Notice that the red ticks align with HLA genes.

2.3.7: Comparison of Epicopy to an existing method

Feber et al. [85] recently published a method to identify CNV from HM450K array data using the ChAMP pipeline [102], which uses a quantile normalization step as opposed to the functional normalization adopted by Epicopy. We compared the

performance of Epicopy with ChAMP-CNV to assess both methods' ability to make accurate individual gene-level calls, as well as recurrent CNV calls using GISTIC2.0. ChAMP-CNV was performed on Level-1 TCGA data using recommended parameters. The same post-segmentation processing of identifying optimal thresholds was performed to obtain comparable datasets between Epicopy and ChAMP-CNV.

We plotted the receiver-operating characteristic (ROC) curves for correct classification of gene level CNV and calculated the area under the curve (AUC) for gene level amplification and deletions in each of the TCGA tumor types (Figure 2-14) to assess the overall performance of both methods against CN value thresholds.

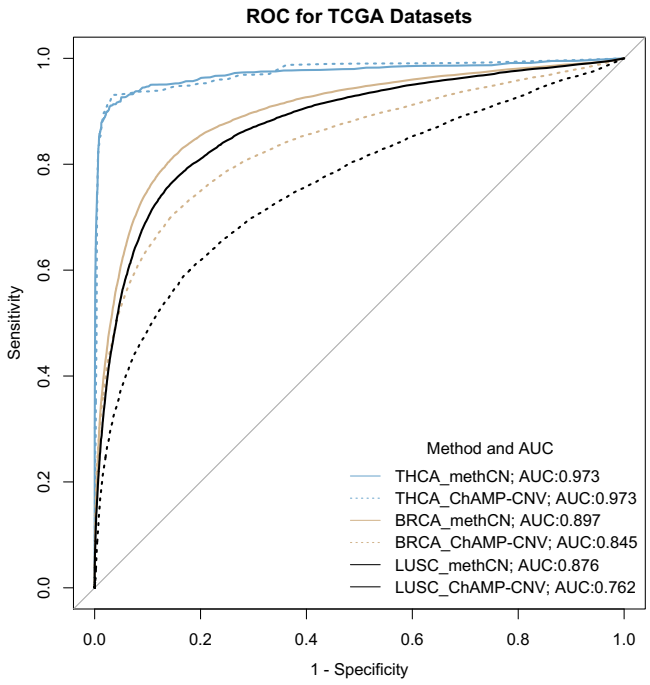


Figure 2–14: ROC analysis comparing Epicopy and CHAMP-CNV performance

Threshold selection was performed for both methods and the results gene-level CNV. ROC analysis was performed with LRR as the predictor and gene-level CNV results (|gene_LRR|>0.3 as an event) from SNP analysis downloaded from the Firehose server as the response. The curves are detailed in the inset legend.

While the performance of these two methods was equal in the THCA training set, Epicopy showed improved performance compared to ChAMP-CNV in the BRCA and LUSC validation datasets. This can be attributed to slight differences in data normalization and in the reference samples used. Epicopy uses functional normalization followed by probe type-specific normalization, while ChAMP-CNV uses quantile normalization. Where Epicopy uses a series of normal samples as reference intensities, ChAMP-CNV uses the median intensities across all tumor samples. This leads to a difference in reference intensities, which manifests itself when the log R ratios are calculated.

2.4: Conclusion

This study assessed the efficiency of using HM450K CpG microarrays to profile CNV in a series of different tumor types. With ample probe coverage across the genome, especially within promoter and exonic regions of genes, HM450K can be used to obtain CNV information in the human genome. We presented a series of tools in the Epicopy pipeline to identify CNV using the Illumina HM450K methylation array with high probe density across the genome. Using publicly available paired SNP and methylation array data from TCGA, we showed that Epicopy has a sensitivity of 70% and PPV of 75% across THCA, BRCA, and LUSC datasets. Epicopy is available at https://github.com/sean-cho/Epicopy.

With the increased interest by the scientific community in understanding the interactions between genetic and epigenetic changes in disease, Epicopy represents a valuable tool, allowing users to obtain CNV information from a DNA methylation

microarray chip. Some of the pressing questions in cancer biology that address patient outcomes and treatment can be answered using multiplatform analyses of clinical samples with long-term follow-up information. Such studies are often limited to archival samples where available tissues are frequently scarce. Being able to analyze both genomic and epigenomic data from a single DNA input will allow for more samples to be analyzed and also allow for better correlation of genomic and epigenomic data, since both analyses are performed on the same sample. As such, we believe that methods such as Epicopy and CHAMP-CNV are crucial in allowing the scientific community to more completely characterize molecular changes across multiple platforms.

Chapter 3: Using Epicopy

Summary: Epicopy is a computational pipeline that allows users to obtain copy number variation (CNV) information from Illumina Human Methylation 450K microarray data. It comes with a companion package, EpicopyData, which contains raw data for normal samples that can be used as substitute normal reference intensities in the instances where normal samples are unavailable. Epicopy can be run as a complete pipeline using a single function with a comprehensive list of arguments, or can alternatively be run as individual sub-routines.

Availability: Epicopy and its supplementary package, EpicopyData, are freely available as R (>=3.0.0) packages hosted on Github and Bioconductor under the Artistic 2.0 license.

Contact: sean.cho@jhmi.edu; lcope1@jhmi.edu

Supplementary Information: https://github.com/sean-cho/Epicopy

3.1. Introduction

Both genetic and epigenetic alterations contribute to oncogenesis. Therefore, a more comprehensive picture of these events in cancer can be obtained by performing multiomic analyses, as evidenced most prominently by studies performed in recent years by The Cancer Genome Atlas (TCGA) consortium. Multiomic studies, while informative, can be challenging to obtain in low resource settings, where funding or sample availability and quality are limited. To that end, we propose Epicopy, a method to derive copy number variation (CNV) data from existing data obtained from high density Illumina Human Methylation 450K microarrays.

Epicopy has a companion package, EpicopyData, which contains raw data from a series of normal tissues (thyroid, breast, and lung) derived from TCGA data that can be used as normal reference for signal intensities, in the event that users have few or no available normal samples.

A standard Epicopy analysis pipeline (Figure 2-3), that can be run with optional parallelization (1), reads raw Illumina idat files from a target directory, (2) filters SNP-adjacent probes, (3) performs functional normalization, (4) compares it to reference normal, (5) segments the data using circular binary segmentation (CBS), and (6) returns the segmentation results and a marker file suitable for GISTIC 2.0 analysis.

Epicopy includes many arguments for finer control of the process within the main function at each step of the process (Figure 2-3). For better control of the program, users can run separate aspects of the program independently.

3.2. Implementation and usage

3.2.1. Implementation and standard parameters

Epicopy is implemented as an R (>=3.0.0) package and is available on the Bioconductor website (http://www.bioconductor.org/packages/release/bioc/html/Epicopy.html). It has dependencies on the R packages *minfi*, *DNAcopy*, and *ParDNAcopy*.

A standard Epicopy run uses functional normalization, followed by dye-specific quantile normalization, to normalize raw data between tumor and normal samples. The default reference intensity calculation is performed using a naïve mode estimation implemented by the *modeest* package. Finally, segmentation is performed using CBS with the *sdundo* argument.

To allow for customizing, many arguments and even sub-routines are built into the Epicopy package. Some of the key arguments include:

target_dir: Directory containing raw idat files and a sample sheet.

output_dir: Output directory into which segment and marker files are deposited.

project_name: Suffix for output files.

Normals: Indicate either column name within sample sheet to look for normal samples (indicated by the case-insensitive character string "normal") or one of the included normal tissue cohorts with one of "all", "thyroid", "breast", or "lung".

Ref: The method of calculating reference intensities from normal samples, which defaults to mode, and can alternatively be set to median.

ncore: Number of cores to be allocated for parallel segmentation.

filterbycount, *minprobes*: Whether to filter final segmentation file for segments with more than the specified minimum number of probes.

3.2.2. Setup & usage

There are two key components necessary for running Epicopy; (1) raw Illumina idat files in a single directory and, in the same directory, (2) an Illumina sample sheet detailing which chips/samples to analyze, with optional columns delineating normal status and sample names. An example dataset is included at https://github.com/sean-cho/EpicopyData.

Running the Epicopy pipeline requires a single line of code. The following argument runs Epicopy on a series of samples with normal status indicated in the column "*sample_status*" with 4 cores.

```
epicopy(target_dir = 'target/', output_dir = 'output/',
Normals = 'sample_status', Ref = 'mode', ncore = 4)
```

Additional arguments are available, as well as individual sub-routines. Finally, both Epicopy segment and marker file outputs are compatible with GISTIC 2.0, to facilitate optional further analyses.

3.2.3. Additional tools

The ability to make confident CNV assessments depends on the probe density of a region and by evaluating a region of interest (of say, a gene). Epicopy includes a function to evaluate regions of interest in the human genome for probe coverage in the 450K microarray, enabling users to decide if Epicopy will provide confident CNV calls for their needs..

Epicopy also includes a function that generates Manhattan plots from segmented data output for quick visualization of CNV calls.

3. Considerations

Segmented CNV data obtained from Epicopy have different log R ratio and minimum probe per segment thresholds and distributions compared to SNP microarray derived

CNV data. Based on an analysis comparing SNP and Epicopy derived CNV lesions, we recommend a minimum number of 50 probes to identify segments with high confidence.

Chapter 4: Multiomic analysis of ductal carcinoma in situ

4.1: Introduction

4.1.1: Breast cancer statistics

Breast cancer is the most diagnosed form of carcinoma and second leading cause of cancer-related deaths in women in United States. According to the age-adjusted SEER data for the year 2009 [103], 155.7 in 100,000 women have received a diagnosis of breast carcinoma and about 20.1% of those diagnoses were reported as carcinoma in situ. In addition, 23 in 100,000 women have succumbed to breast carcinoma in 2009.

Recent advances in detection, prognostication and therapy has improved outcome in breast cancer survival, but there exists a subpopulation of patients who do not benefit from current form of disease management and another population of patients who are over-treated, especially patients with in situ disease [104-107].

4.1.2: Mammography: Risk versus benefit

Studies over the years have been conducted to study the benefits and risks associated with mammography. The benefits of mammography include 1) early detection of a lesion that will eventually evolve into life-threatening disease and 2) the ability to effectively treat the lesion at the time of screen detection, both which directly translate into reduced mortality [107, 108]. However, there exists the risk of over-diagnosis of

benign lesions that will not progress into invasive disease or present with clinical symptoms during the lifetime of the patient [109].

Two comprehensive meta-analyses published in 2012 using SEER data (US) [105] and data from UK [110] studies highlight this risk-benefit relationship and review many observations and trials conducted to study the risk-benefit ratio of mammography screening. The reduction in mortality is calculated as the risk of mortality of screened women subtracted from the mortality of unscreened women. Both studies, in agreement with numbers obtained by previous investigations [104, 106-109], estimate a mortality risk reduction of 20 – 30% in all women screened.

Observational studies have noted that the increased incidence of early stage breast cancer (DCIS/local) do not have an equivalent effect in reducing the incidence of late stage breast cancer (regional/metastatic), indicating potential over-diagnosis. In particular, SEER data obtained over the last three decades [105] indicate an increase in the number of diagnosed early stage breast cancer, with a only a slight reduction of late stage breast cancer incidences in the group of women over the age of 40, where mammography is recommended, and no reduction in women under the age of 40 where mammography is uncommon (Figure 4-1).

The risk of over-diagnosis can be estimated using various methods [105, 109, 110], but can be loosely defined as the number of screen-detected cases in the screening arm exceeding the number of clinically detected cases within the control arm after the lead time, or the time required for a pre-cancerous lesion to develop into invasive carcinoma. The lead time for early breast cancers is ill-defined and studies have generally accepted a follow-up time equal to that of the period of screening as a reasonable

estimate of lead time. Various studies, including the UK panel and Bleyer et al., have found varying rates of over-diagnosis, ranging from 11% to 60%, with a median of 28%. Even if the numbers do not agree, there is a significant possibility that routine screening mammography may lead to over-diagnosis of early stage breast cancer and may contribute to over-treatment of the disease.

4.1.3: Ductal Carcinoma In Situ (DCIS) incidences from mammography screens

Breast cancer can be histopathologically classified into a few general subtypes depending on the site of origin and presentation, but the most prevalent forms of breast disease remain the ductal subtypes, invasive ductal carcinoma (IDC) and ductal carcinoma in situ (DCIS), which originate from the epithelium of mammary ducts.

DCIS represents the non-invasive local breast disease, where malignant cells proliferate within the confines of the duct and no invasion through the basement layer or myoepithelial cells are observed. Historically, DCIS was diagnosed after the excision of an IDC or a clinical event, such as nipple discharge or palpable mass. However, with the advent of mammographic screening, an increased number of DCIS cases are observed before symptoms occur, as highlighted in Figure 4-2 (SEER data).

Putting the SEER data into perspective, there has been an increase in the incidence of DCIS of 50 per 100,000 women in US, now reaching 56.5 cases of DCIS per 100,000 women, a 10-fold increase. DCIS now represents 17% of all newly diagnosed breast cancers over a 20-year period. [103, 105] The UK panel found screen-detected DCIS rates among newly diagnosed breast cancers at 20% versus clinically detected

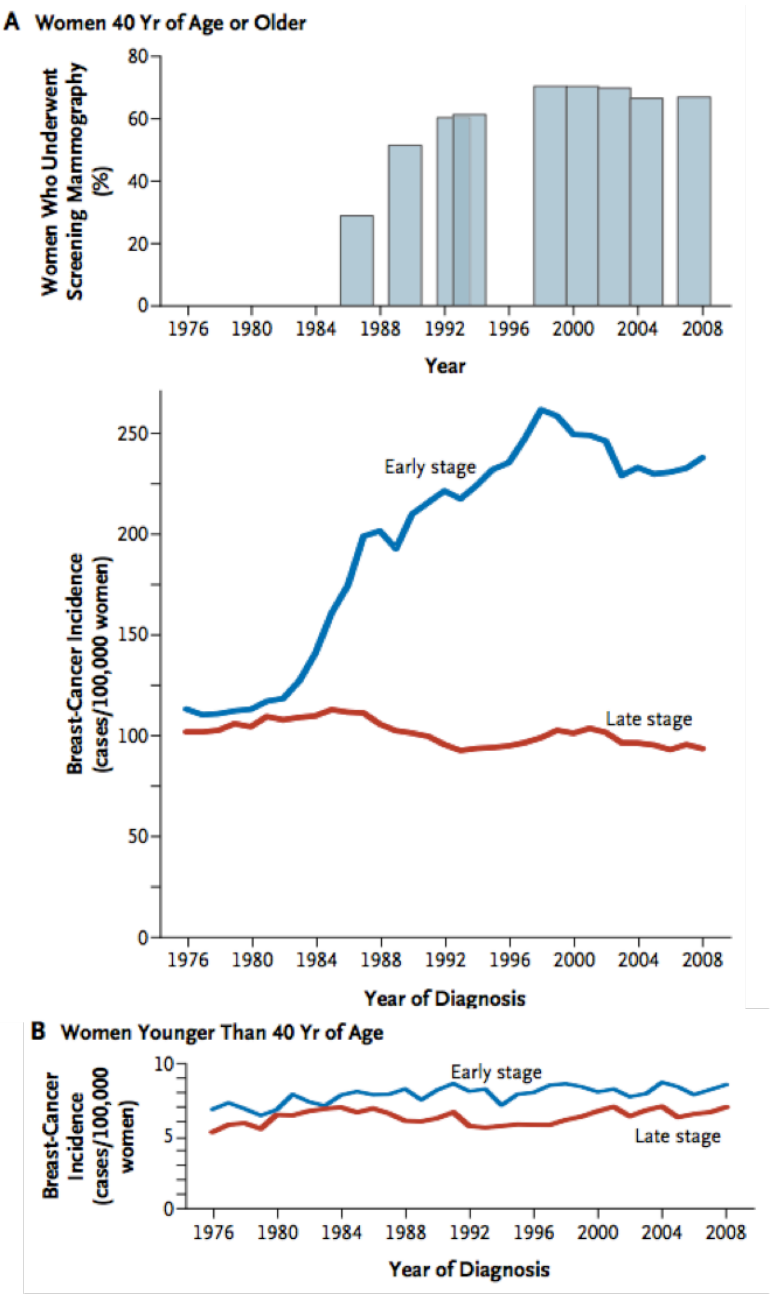


Figure 4–1: Age and HRT-adjusted SEER data for breast cancer incidences

There was an increase in early stage breast cancer in the a) screened group (women >40 yrs of age) vs b) unscreened at the advent of mammography.

DCIS rates at 5% [110]. Other studies have put the rate of DCIS among newly diagnosed breast carcinomas at 15 – 20% [106, 111, 112].

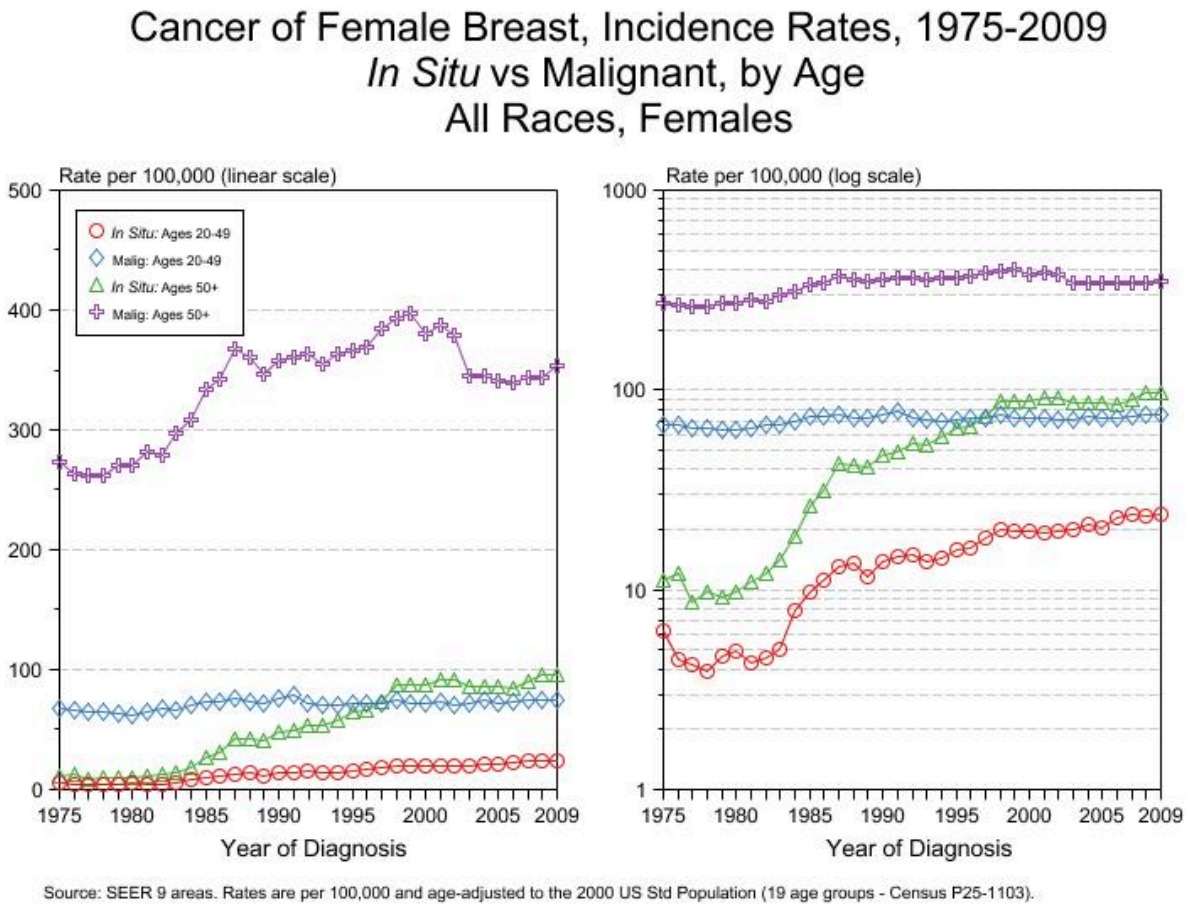


Figure 4–2: In situ versus malignant of female breast cancer by age

From 1980s to the 2009, since the introduction of routine screening mammography, an increase of malignant breast cancer was observed. During the same time period, an increase of in situ disease by 7 to 8 fold was observed in both younger and older age groups since the 1980s.

4.1.4: Natural history of DCIS and its clinical implications

Although DCIS does not present as a life-threatening disease, it is widely believed that DCIS lesions are non-obligate precursors that can progress into invasive ductal

carcinomas (IDC). Therefore, the treatment goal for DCIS is the prevention of progression in the form of IDC. The extent of progression usually cannot be directly observed because the standard treatment involves surgical removal of the DCIS. Retrospective studies aimed at estimating the number of progressive DCIS have studied the rate of IDC formation from DCIS misdiagnosed as benign breast disease (BBD). These cases of DCIS were treated with a simple biopsy, which allowed the lesion to progress at a natural rate. [113-116] These studies have estimated between 14% to 53% progression at a median follow-up of 10 years. These estimates are likely to be negatively biased, as DCIS that are misdiagnosed as BBD are generally small and lower in grade.

Another method of estimating the natural rate of DCIS progression is to investigate the incidence of IDC recurrence from DCIS lesions treated only with breast conserving surgery. Numerous studies have recorded ipsilateral recurrence rates ranging from 10% to 25% [112]. Among these studies, about a half of the cases, ranging from 46% - 76%, recurred as IDCs, reinforcing the hypothesis that not all DCIS will progress to IDC.

4.1.5: Current therapy, clinical risk stratification, and the problem of over-treatment

4.1.5.1: Mastectomy versus breast conserving surgery

Traditionally, DCIS has been treated with radical mastectomy. [117] However, trials in localized IDC showing comparable survival between mastectomy and

lumpectomy/breast conserving surgery (BCS) have prompted a shift in treatment towards BCS for patients diagnosed with DCIS [118]. Briefly, the NSABP B-06 trial randomized 2163 patients with IDCs <4cm in diameter into 3 arms; total mastectomy, BCS followed by radiotherapy (BCS+RT) and BCS only. At 20 years of follow-up, there was a significant difference between local recurrences in the mastectomy vs BCS+RT trial arms, but no difference in overall survival or deaths caused by breast disease. [118] The Milan trial compared 701 women with IDCs measuring <2cm that were randomized into radical mastectomy or BCS + RT [119]. Similarly, after 20 years of follow-up, while there was a higher risk of local recurrence (2%, mastectomy vs 9%, BCS+RT), the overall survival and deaths caused by breast disease remained similar between both arms. While there have been no trials comparing mastectomy to BCS+RT in DCIS, it is reasonable to extrapolate these findings to DCIS given that it is a premalignant lesion deemed less aggressive.

4.1.5.2: Radiotherapy and adjuvant therapy

Two pivotal DCIS trials performed by NSABP evaluated the effect of radiotherapy and adjuvant hormonal therapy (tamoxifen) on DCIS patients treated with BCS. NSABP B-17 enrolled 817 DCIS patients into two randomized arms; 1) BCS and 2) BCS+RT. At 12-year follow-up, the local disease free recurrence for the group receiving radiotherapy was significantly reduced (8%) compared to the BCS only group (15%) [120]. NSABP-24 [121], on top of corroborating the effectiveness of radiotherapy, established the role of tamoxifen in treating DCIS. 1804 patients were randomized into

two arms; 1) BCS+RT+placebo and 2) BSC+RT+tamoxifen. The 7-year cumulative incidence for invasive local recurrence in the group receiving placebo was 11% versus 7.7% in the tamoxifen treated group (p=0.07). Furthermore, the ER-status for 628 patients were available and the effectiveness of tamoxifen was clear in these patients (relative risk, RR =0.41, p=0.0002) [121]. Figure 4-3 summarizes these findings.

Based on these trials, patients presenting with DCIS, be it by clinical detection or mammography-screened, are generally treated with BCS+RT and given adjuvant tamoxifen therapy depending on their ER status.

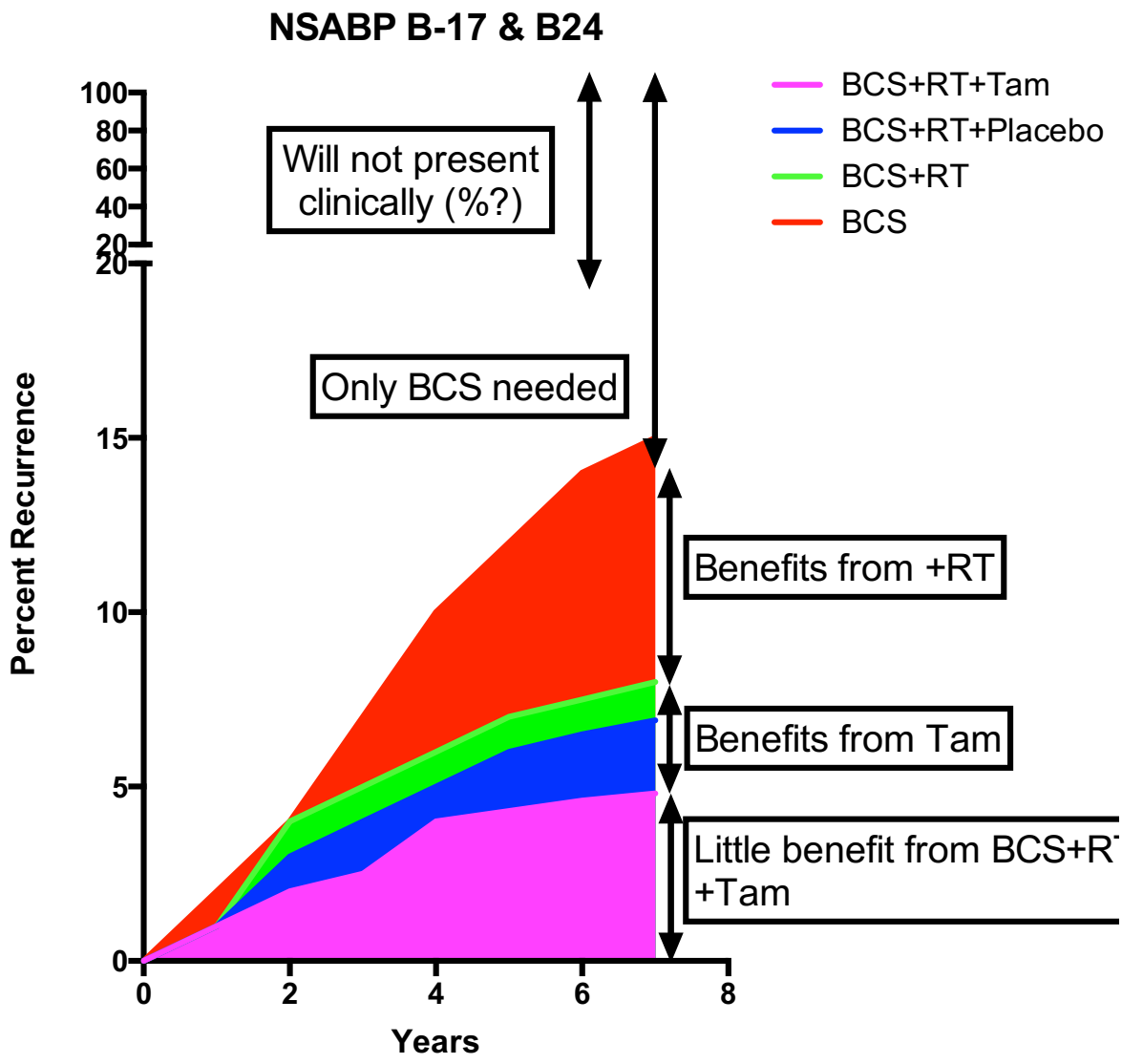


Figure 4–3: IDC recurrence rate of clinically relevant subgroups estimated from the results of NSABP B-17 and B-24.

Estimation of IDC recurrence rates from NSABP-17 and -24[122] in five clinically relevant subgroups with different clinical management strategies. An estimated 5% of patients with DCIS will experience recurrence as IDC even with BCS, radiation, and tamoxifen, and represents a subset of the patients who will need more aggressive therapy. On the other end of the spectrum, 85% of the patients will need at most BCS for a recurrence free survival of 7 years. A proportion of these patients will not need therapy, as the lesion will not present as a clinically relevant disease in the lifetime of the woman.

Even with a 90% cure rate for DCIS using current therapy, I hypothesize from these meta-analyses and trials that there exist 6 groups of clinically relevant patients subgroups:

1) Patients who have screening-detected DCIS who will not present breast disease clinically in their lifetime, representing a group of patients who are at the extreme end of over-treatment
2) Patients who have DCIS that will eventually present clinically and benefit from BCS, but not radiotherapy, representing another group of patients who are commonly over-treated
3) Patients who will benefit from BCS+RT, but not Tam
4) Patients who will benefit from BCS+Tam, but not RT
5) Patients who will benefit from BCS+RT+Tam
6) Patients with progressive disease in spite of BCS+RT+Tam, representing a group of patients who are under-treated

Given these subgroups of patients, there exists a need to stratify them into clinically relevant groups for optimal disease management.

4.1.6: Current DCIS classifications, treatment modalities, and prognostic potential

DCIS is a heterogeneous disease, both in its etiology and clinical presentation. There are many clinicopathological characteristics that have been used to classify DCIS lesions, albeit with limited prognostic success.

4.1.6.1: Clinicopathological features

Clinically, the patient's age, race, family history, tumor size, and surgical margins have been evaluated as risk factors of breast carcinoma and DCIS recurrence, with conflicting reports regarding the association of these factors with IDC recurrence. However, most studies have found that family history of breast cancer, surgical margins, and mode of detection are associated with higher risk of recurrence.

Histologically, the architectural appearance of DCIS lesions allow them to be classified into micropapillary, papillary, solid, or cribiform subtypes. Most often, DCIS lesions exhibit more than one architectural subtype. On top of that, tumor size, nuclear grade, degree of differentiation, extent of comedo-necrosis, and presence of calcifications are also used to describe DCIS lesions. Furthermore, the surgical margins after BCS of < 1mm have been shown to increase the risk of developing locoregional IDC recurrence.

In addition to these features, expression of certain proteins, such as the expression of hormonal receptors (estrogen and progesterone receptors) and Her2/neu have been used in the clinic to classify IDCs and may have value in predicting DCIS risk.

Clinically, steroid receptor status in particular has been used to prescribe adjuvant hormonal therapy [121].

Kerlikowske et al. [123] systematically evaluated the clinicopathological factors associated with DCIS and IDC recurrence and showed that mode of detection, margins, and nuclear grade are associated with increased risk of IDC in a multivariate model. More recently, Zhang et al. [124] performed a meta-analysis to identify predictors of locoregional IDC progression and identified that, in observational studies, the presence comedo-necrosis and positive margins results in higher likelihood of IDC progression. Interestingly, and relevant to the discussion of over-treatment, is that the same meta-analysis of randomized clinical trials (RCTs) identified that mode of detection was related to risk of IDC progression, specifically that non-screen detected cancers were more likely to progress to IDC.

4.1.6.2: Proposed scoring systems

Following these studies, various routinely available clinicopathological features discussed earlier were used in the effort to develop standardized scoring systems proposed for clinical use for risk stratification in DCIS. Many histological classification systems were designed based on the Bloom-Richardson classification derived from invasive breast cancer (IBC), which depends on tubule formation, nuclear polymorphism, and mitotic rate. One such index is the updated Van Nuys Prognostic Index (VNPI), which incorporates 1) Bloom-Richardson grading of the tumor, 2) tumor size, 3) presence of comedo-necrosis, and, most recently, 4) age, to prognosticate patients into three risk

groups; low risk (VNPI 4-6), intermediate risk (VNPI 7-9), and high risk (VNPI 10-12). In the retrospective study where the updated VNPI was developed and first reported, 706 patients underwent BCS, and Silverstein [125] showed that in the VNPI risk groups of patients with up to 12 years of follow-up, low risk patients had a 1% recurrence rate, intermediate risk patients had a 20% recurrence rate, of which 44% were IDC, and the high risk patients had a 50% recurrence rate, of which 39% were IDC. Other studies attempting to validate the performance of VNPI found either less significant log-odds ratios between risk groups [126] or no difference in risk [127] for disease free survival. Of note is that none of these studies, including the original, found a statistically significant difference in the probability of breast cancer related deaths across risk groups.

Other scoring systems have been proposed, but none have been adopted in universal guidelines, as experts disagree on which system is most reproducible [128]. Furthermore, many classifications systems do not distinguish between the prognostication of DCIS recurrence *vs*. IDC progression, and show little to no association with breast cancer related mortality. Therefore, there is a clinical need and opportunity for the design of a molecular test that may complement the currently established classification systems that allows clinicians to distinguish progressive versus non-progressive DCIS with a high negative predictive value (NPV).

4.1.7: Molecular properties of DCIS and markers of progression

Gene candidate approach work in our lab and others has shown that DNA methylation changes occur in early stages of cancer development. For example,

hypermethylation and transcriptional silencing of the gene 14-3-3σ (stratifin, SFN), a tumor suppressor involved in breast oncogenesis, was observed in stages of breast hyperplasia as early as atypical hyperplasia [129, 130]. In support of that, transcriptomic, genomic, and epigenomic profiling of DCIS, either by itself or in comparison with IDC, paired or otherwise, have revealed that many of the hallmarks of IDC are present in DCIS [131], including chromosome 1q and 8q gains, chromosome 16p loss, TP53 mutation, and HOX-family gene promoter methylation [132] .

In studies of limited sample size (between 26 to 74 samples), investigators have shown successful and high confidence PAM50 classification of DCIS samples [132-134]. Interestingly, the distributions of PAM50 subtypes across DCIS do not reflect distributions observed in IDC, with the DCIS cohorts presenting with more HER2-enriched (29% - 38% in DCIS compared to 15% – 20% in IDC) but less luminal subtypes, with consistent IHC/FISH histological classification. While small sample size and low power, as well as selection bias for high grade DCIS, may explain this disparity, it is also important to note that previous studies have shown highly variable HER2 expression in DCIS, ranging from 28% to 65% [135-137]. Furthermore, a recent NSABP prospective study evaluating the use of Trastuzumab in DCIS identified that 34.9% of a cohort of 5861 patients were HER2+ [138].

Analyses of DCIS with concurrent IDC identified similar oncogenic genomic and transcriptomic changes in both DCIS and IDC lesions of the same patient [139]. More surprising is the discovery that in studies comparing mixed DCIS and IDC samples across different grades, hierarchical clustering of genes differentially expressed across lesion groups (normal vs. DCIS vs. IDC, with ER-status being equal) showed clustering

first by individuals, followed by grade, with little support for clustering of DCIS vs. IDC [134, 139].

At the time of writing, none of the published high-throughput molecular studies were performed in DCIS samples with long-term follow-up information, and evaluating molecular changes of aggressiveness (high versus low grade, IDC versus DCIS in synchronous lesions, etc.) was analyzed as a proxy for progression.

In studies comparing genetic alterations in synchronous DCIS with IDC, amplifications of MYC, HER2, FGFR1, and CCND1 have been identified to be distinct between DCIS and synchronous IDC events [140]. Beyond that, a higher degree of genomic instability reported as the percent of the genome altered (via amplification or deletion) is associated with increased DCIS grade [133]. Gene expression classifiers have been identified that distinguish between DCIS and IDC [141] and high and low-grade DCIS [133, 134, 142]. Unfortunately, given the limited data available on DCIS samples, none of these were validated with an independent external dataset.

To my knowledge, at the time of writing, there is one study that evaluated the use of several IHC markers beyond the canonical ER, PR, HER2, and Ki67, and systematically assessed their relationship with IDC recurrence. In a gene candidate approach using IHC, Kerlikowske et al. evaluated the relationship between a series of protein markers and IDC recurrence in a case-control study with a cohort of 329 DCIS patients with long-term follow-up. The authors showed that in a multivariable model correcting for clinicopathological factors, DCIS lesions expressing all three p16, CPX-2 and Ki67 markers were at increase risk of IDC recurrence [143].

4.2: Study design and methods

4.2.1: Motivation

We hypothesize that molecular markers of progression are present in early stages of the disease and identification of these markers will allow us to stratify patients into different risk groups with appropriate clinical management protocols.

As reported, high-throughput molecular studies of DCIS have been limited to DCIS lesions comparing pure DCIS with samples harboring synchronous DCIS and IDC in the same section. In the published literature, all lesions were collected at the same time, with short-term follow up, which are study designs with limited capacity to address the question of progression because of the lack of true non-progressive controls, which are DCIS samples that do not develop IDC over a long period of observation [144], *e.g.*, > 10 years.

The available literature suggests extensive concordance between the molecular alterations in DCIS and IDC, especially as assessed by genetic and transcriptomic platforms, which suggests the need to study and profile the epigenomic changes in DCIS as well. Furthermore, these molecular similarities were observed in pure DCIS across all grades, and may indicate small effect sizes between progressive and non-progressive DCIS, speaking to the need of a large study cohort and high resolution methodologies to identify such differences.

4.2.2: Study design

This study was designed as multicenter, nested case-control study of DCIS. We identified a cohort of 100 patients with DCIS and no evidence of invasive disease at presentation that progressed to invasive breast cancer within 10 years, after a minimal interval of 12 months, recruited from four leading breast cancer treatment centers. Given the difficulty in identifying cases matching these criteria, the racial composition was limited to Caucasians, the largest cohort available, in order to minimize confounding factors at this early stage, where our priority was in confirming our hypothesis that a clinically relevant prognostic molecular signature can be identified in DCIS. 100 DCIS control cases were matched to the same selection criteria, and were selected to reflect the clinical subsets of the cases, including histological nuclear grade, margin status, and adjuvant treatments, as well as approximate age and year of diagnosis (both within a 5 year window). All controls had a disease-free follow up of a minimum of 10 years. Cases and matched controls were treated and followed at the same institutions, and had similar rates of mastectomies (35%), radiation treatments (30%) and hormone therapy (10%). All relevant clinical information has been captured in an anonymized research database, and 20 unstained sections with matching H&E stained control slides were obtained from the relevant tissue blocks with coded identifiers.

The three study pathologists, Dr. Gabrielson (JHU), Dr. van Diest (Utrecht), and Dr. Hawes (USC) reviewed all cases and controls in order to ensure diagnostic consistency through the Aperio digital imaging system (Aperio Inc., Vista, CA), which

allows very efficient remote visualization and interactive annotation of the entire histological slide at high (20x and 40x) resolution.

In the initial study design, transcriptomic analysis was analyzed using the Illumina DASL microarray. Unfortunately, during sample accrual and nucleic acid extraction, the DASL assay was discontinued and the platform was switched to the Affymetrix pipeline of WGA by WT Pico followed by HTA2 microarray profiling. DNA methylation analysis was performed using Illumina FFPE restoration followed by 450K methylation microarray. Lastly, due to DNA yield limitations, genetic analysis was performed using copy number data derived from 450K microarray data by Epicopy instead of the initial plan of using Illumina 1M SNP Beadchip microarray.

4.2.3: Patient identification and sample collection

We used patient registries here at Johns Hopkins Hospital and at collaborating institutions to identify cases and controls that matched study criteria and had documented long term follow up. Tissues were obtained with approval of the respective institutional IRBs. Study pathologists reviewed archival H&E sections to select FFPE tissue blocks. A total of 98 progressive DCIS cases, 98 non-progressive DCIS controls, 12 DCIS adjacent normal tissue, and 5 reduction mammoplasty samples were profiled using gene expression (Affymetrix HTA2) and methylation platforms (Illumina 450K).

4.2.4: DNA/RNA extraction and quality control

Pathologist annotated adjacent H&E sections were used as guide for tissue orientation and macrodissection of unstained tissue sections to enrich for >70% DCIS epithelial cells. Following that, DNA and RNA were extracted using Allprep FFPE RNA/DNA kits (Qiagen) with modifications to deparaffinization, digestion, and wash steps. The modified protocol is appended.

Quantification of RNA and DNA was performed using a Nanodrop2000 and a Qubit fluorometer (Qiagen) using appropriate kits (RNA HS, RNA BR, DNA HS, and DNA BR). The 260/230 and 260/280 ratios were used to assess sample purity and solvent contamination. Qubit derived measurements were ultimately used to calculate nucleic acid input for the microarray platforms.

4.2.5: Quality control and microarray

Quality control was performed using the Illumina FFPE QC kit with the iTaq™ Universal SYBR® Green Supermix and was regarded as the main quality control step for 450K and other DNA-based microarrays. Samples with delta C_T <9 were used in the study and case-control pairs with lower delta C_T were prioritized. Bisulfite conversion was performed using the EZ DNA Methylation-Gold™ Kit (Zymo Research, Irvine CA), with modifications introduced per Appendix I of the manufacturer's recommended protocol. The detailed protocol is appended at the end of the thesis. NaBi-converted DNA

was submitted to the SKCCC Microarray Core Facility for FFPE DNA restoration and profiling using the Illumina 450K microarray.

4.2.6: Data pre-processing and QC

Unless otherwise stated, data analysis was performed in R Statistical Environment using base, Bioconductor, and custom packages. P-values were corrected using Benjamini-Hochberg's method for false discovery rate estimation.

Illumina 450K Methylation Quality control metrics for Illumina-based arrays were estimated using Illumina's GenomeStudio software, and validated through control probe signal intensities extracted through the minfi software in R. GenomeStudio-derived detection p-values (detP) with a threshold of $p < 0.01$ were used to calculate sample-wise call rates, and samples with call rates of less than 80% were removed from the analysis. Raw beta-value density plots were plotted and samples with aberrant beta-value density plots (without a bimodal distribution with means around 0.1 for unmethylated regions and 0.9 for methylated regions) were removed from the analysis. Probe-wise detP were estimated and probes with > 95% coverage across remaining samples were retained for analyses. Probes with interrogated CpGs 2bp from a known SNP with a population minor allele frequency (MAF) >5% were removed. Functional normalization was performed on the final set of high quality samples and probes to obtain final methylation dataset.

Epicopy-derived CNV High quality samples and probes from the methylation pre-processing were used as input into Epicopy to generate CNV information for DCIS samples. Default Epicopy parameters were used with reduction mammoplasty normal samples serving as reference samples. CNV profiles were assessed for typical segmentation parameters, and samples with aberrant parameters were discarded from downstream analyses.

TCGA Data Processed TCGA data were downloaded from Broad Institute's Firehose server.

4.2.7: Methylome data analysis

Exploratory data analysis was performed using principal component analysis (PCA) and unsupervised clustering using Euclidean distance was performed on variable probes across all the samples (standard deviation, SD, above 3 interquartile ranges (IQR) of the standard deviation, n = 2963). Probe-wise differential methylation analyses across various groups were performed using limma on probes with SD above 1.5 IQR (n = 132,174). *DMRcate* was used to identify differentially methylated regions with a Gaussian kernel bandwidth of 1000 and a scaling factor of 2, resulting in a sigma of 500. Methylation scores for various molecular classes were calculated as transformed mean beta-value. Briefly, the beta-values for each probe was multiplied using the sign of the moderated t-statistic from limma, and a mean transformed beta-value for each sample

was calculated. A larger value represents a molecular phenotype closer to the positive contrast from the limma analysis. For N differentially methylated probes,

$$Methylation\ score_j = \frac{\sum_i^N \left(\beta_{ij} \times sgn(t_i)\right)}{N}$$

Consensus clustering was performed to identify stable epigenetic clusters and probe clusters. To assign functional groups to these series of probes, we used the "Compute Genesets Overlap" tool hosted on the Molecular Signatures Database (MSigDB) against C2:CGP (chemical and genetic perturbations gene sets).

4.2.8: Copy number data analysis

GISTIC 2.0 was used to identify regions of recurrent amplifications across tumor adjacent normal, case, and control samples. Default input parameters were used. Recurrent copy number results were extracted for custom plots and analyses. Gene-wise copy number information was used to identify recurrent gene-wise copy number changes. Cytogenetic band copy number changes were estimated using gene-wise copy number information and were used in clustering analyses.

4.3: Results and Discussion

4.3.1: Methylome analysis reveals distinct methylation patterns in normal tissue consistent with oncogenic development that is validated in the TCGA breast cancer dataset

Genomic data are often represented by sparse matrices and the same is true for 450K data with 485,512 probes. To reduce dimensionality, data was first subsetted into phenotype-naïve probes with SD above 1.5 IQR of standard deviation across all samples. Exploratory analysis was performed on another subset of probes with SD > 3 IQR. PCA revealed clustering of normal and tumor-adjacent normal tissues on the first and second component of the PCA, which collectively explained 39% of all the variation observed in the dataset (Figure 4-4a). Unsupervised hierarchical clustering was performed on the same set of probes to validate and visualize PCA results, and identify clinicopathological features that may contribute to further clusters. The majority of the most variable probes are located within CpG islands, and a general signature of hypermethylation was observed in a subset of the DCIS samples (Figure 4-4b). DCIS samples did not cluster by progression status, which may suggest that progression status was not the largest contributor to molecular differences, and that biologically, these two classes are very similar. Of note, a few non-progressive DCIS samples clustered near the normal samples.

Differential methylation analysis using limma was performed comparing all DCIS and reduction mammoplasty derived normal tissues (D-N comparison, Figure 4-4c and Table 5) to identify DCIS-specific probes. Hypermethylation of CpG islands in promoter regions of genes was observed in DCIS, consistent with observations from reported

studies in cancer, including breast cancer. This effect was observed predominantly in CpG islands (Figure 4-4d, 9-fold), and to a smaller extent, shores (2.2-fold). Comparatively, there does not seem to be a difference between the proportion of hypermethylated and hypomethylated probes in shelves and open sea regions.

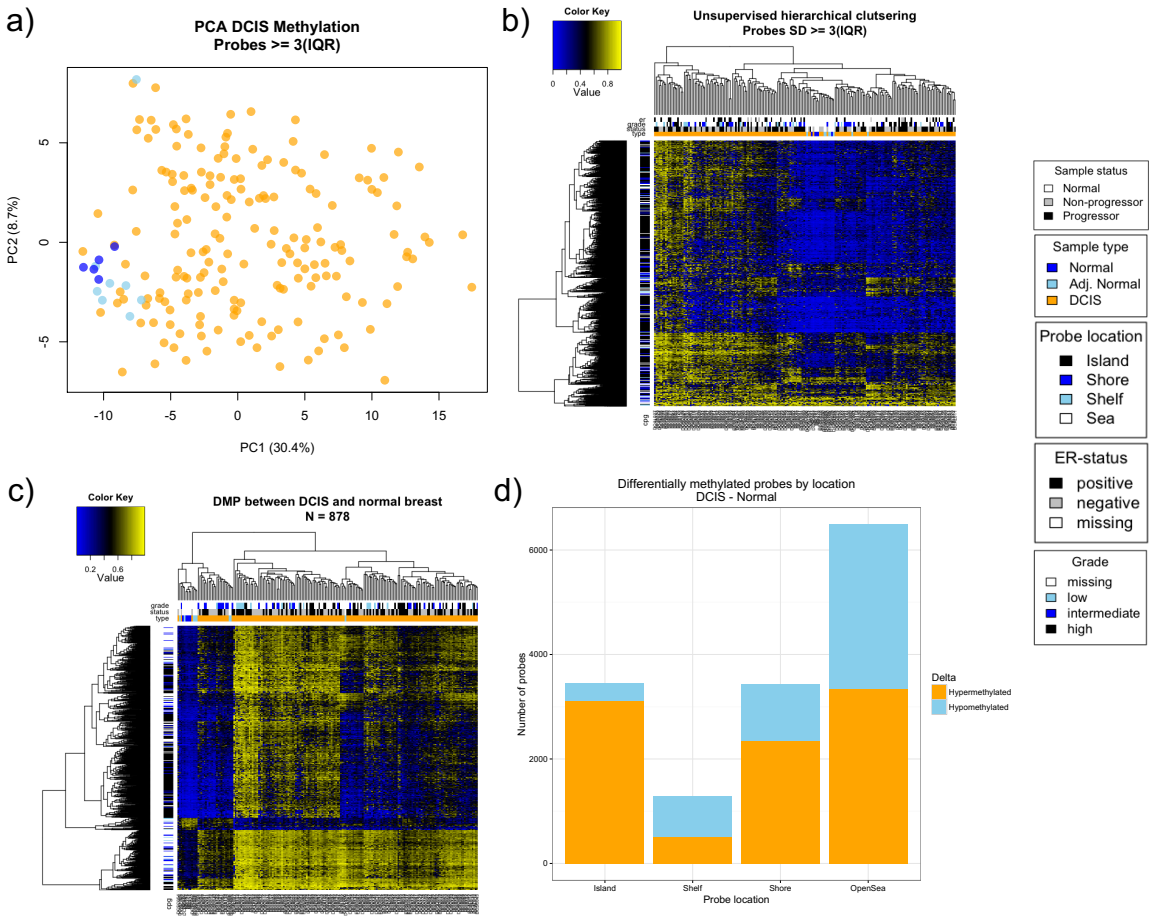


Figure 4–4: Distinct methylation profiles between normal and DCIS tissues.

a) PCA analysis on probes with SD > 3 IQR showed clustering of normal samples on both PC1 and PC2. b) Unsupervised hierarchical clustering using complete linkage revealed that normal tissues are unmethylated in most of these CpG island probes compared to DCIS. c) Differentially methylated probes (DMPs) identified using limma (FDR < 0.05, delta beta > 0.3) revealed general hypermethylation in DCIS. d) DMPs (FDR < 0.05) located in CpG islands tend to be hypermethylated (9-fold) compared to DMPs in other regions (between 0.7- to 2.2-fold).

Table 5: Top 50 DMPs between DCIS and reduction mammoplasty normal samples (D-N)

probeID	delta_beta	ave_beta	t	pval	fdr	b	gene_symbol	Relation_to_Island
cg23908638	0.34	0.78	9.25	2.76E-17	3.65E-12	28.15	GABBR1	OpenSea
cg10017626	0.39	0.75	8.45	5.03E-15	3.33E-10	23.25		Shore
cg17387577	0.48	0.65	8.34	1.02E-14	4.49E-10	22.59	NCOR2	OpenSea
cg16924776	0.34	0.76	8.24	1.85E-14	6.10E-10	22.03		OpenSea
cg23922724	0.42	0.70	8.11	4.14E-14	1.09E-09	21.27		OpenSea
cg19321887	0.34	0.71	8.06	5.83E-14	1.28E-09	20.95		OpenSea
cg25656762	0.38	0.71	8.01	8.08E-14	1.52E-09	20.64		OpenSea
cg22954906	0.35	0.73	7.92	1.37E-13	2.26E-09	20.15	LOC121952	OpenSea
cg10531355	0.38	0.68	7.86	2.01E-13	2.95E-09	19.78	SERINC5	OpenSea
cg20589096	0.31	0.76	7.79	3.14E-13	4.15E-09	19.36	MOBKL1A	OpenSea
cg20547777	0.39	0.69	7.69	5.66E-13	6.80E-09	18.81	EXT1	OpenSea
cg18082788	-0.33	0.18	-7.68	6.20E-13	6.83E-09	18.72	ZC3H12D	OpenSea
cg08092318	0.36	0.64	7.62	8.90E-13	8.91E-09	18.38		OpenSea
cg13799919	0.36	0.69	7.61	9.44E-13	8.91E-09	18.33	POLR1D	OpenSea
cg07694621	0.36	0.70	7.57	1.18E-12	9.88E-09	18.12		OpenSea
cg08325813	-0.34	0.32	-7.57	1.20E-12	9.88E-09	18.10	LFNG	OpenSea
cg21949305	0.29	0.75	7.56	1.27E-12	9.88E-09	18.05	C22orf45;ADORA2A	OpenSea
cg06527213	0.33	0.73	7.54	1.44E-12	1.06E-08	17.93	KLHL25;MIR1276	Shelf
cg01629007	0.33	0.77	7.47	2.22E-12	1.54E-08	17.52	PXDN	OpenSea
cg10717189	0.35	0.68	7.40	3.32E-12	2.19E-08	17.14		Shore
cg01298514	0.29	0.77	7.38	3.77E-12	2.34E-08	17.02	VEGFA	Shore
cg18188653	0.38	0.61	7.37	3.90E-12	2.34E-08	16.99	LPP	OpenSea
cg03048488	0.37	0.70	7.30	5.89E-12	3.14E-08	16.61		Shelf
cg07211212	-0.30	0.17	-7.30	6.02E-12	3.14E-08	16.58		Shore
cg12198841	0.41	0.68	7.29	6.25E-12	3.14E-08	16.55		OpenSea
cg06012347	0.30	0.72	7.28	6.55E-12	3.14E-08	16.51		OpenSea
cg18356785	0.50	0.77	7.28	6.60E-12	3.14E-08	16.50	C1QTNF4	Island
cg21527078	0.42	0.76	7.28	6.64E-12	3.14E-08	16.49	VGLL4	Island
cg04315771	0.38	0.68	7.25	8.01E-12	3.65E-08	16.32	NUP62;IL4I1;ATF5	Shore
cg16324121	-0.31	0.36	-7.24	8.31E-12	3.66E-08	16.28	IL17RE	Shelf
cg22109795	0.32	0.66	7.23	8.85E-12	3.77E-08	16.22		Shelf
cg02573551	0.30	0.74	7.22	9.37E-12	3.87E-08	16.17	EPHB3	Shore
cg15384589	0.35	0.67	7.21	1.03E-11	4.14E-08	16.08		OpenSea
cg09627520	0.33	0.63	7.16	1.33E-11	5.18E-08	15.84	PXK	OpenSea
cg01393234	-0.33	0.26	-7.15	1.47E-11	5.55E-08	15.75		OpenSea
cg22740835	0.37	0.74	7.13	1.64E-11	6.03E-08	15.64	DDR2	OpenSea
cg21880888	0.38	0.70	7.12	1.73E-11	6.18E-08	15.59	DDR2	OpenSea
cg10960266	0.30	0.78	7.11	1.81E-11	6.31E-08	15.55	VGLL4	Island
cg18084609	0.37	0.73	7.10	1.93E-11	6.53E-08	15.49	COCH	Shore
cg06721601	0.31	0.86	7.09	2.09E-11	6.78E-08	15.42	CUX1	OpenSea
cg01446571	0.28	0.74	7.08	2.10E-11	6.78E-08	15.41		OpenSea
cg20946037	0.34	0.68	7.06	2.39E-11	7.36E-08	15.29		OpenSea
cg07642822	0.32	0.75	7.06	2.40E-11	7.36E-08	15.29	ZNF787	Island
cg10824259	0.29	0.77	7.06	2.48E-11	7.44E-08	15.26	NEDD9	OpenSea
cg08845721	0.32	0.69	7.02	3.02E-11	8.86E-08	15.07	NR3C1	Shore
cg06556244	0.35	0.61	6.98	3.84E-11	1.10E-07	14.84	EPHA1	OpenSea
cg08414108	0.35	0.74	6.97	4.10E-11	1.15E-07	14.78	SYNJ2	OpenSea
cg20899625	0.35	0.73	6.96	4.32E-11	1.17E-07	14.73	PLXNA1	OpenSea
cg25827524	0.29	0.74	6.96	4.34E-11	1.17E-07	14.73	WWP1	OpenSea
cg16217794	0.35	0.64	6.94	4.81E-11	1.27E-07	14.63		OpenSea

DMPs identified from this analysis (Table 5) include probes located in promoter regions or gene bodies of genes implicated in breast cancer development [80], including RASSF1A, TP73, CDKN2A (p16), GSTP1, MGMT, APC, and HOX family genes. I also observed DMPs in genes related to estrogen receptor (ER) signaling; ESR1, RUNX3, and NCOR2.

I next tested the hypothesis that cancer-related methylation events occur in DCIS by analyzing the methylation profiles of the identified DMPs in the TCGA breast cancer (BRCA) dataset. In support of that hypothesis, these differential methylation events were also observed comparing IDC and tumor adjacent normal tissue (Figure 4-5). Taken together, this suggests that global oncogenic methylation changes occur in DCIS, before the development of IDC.

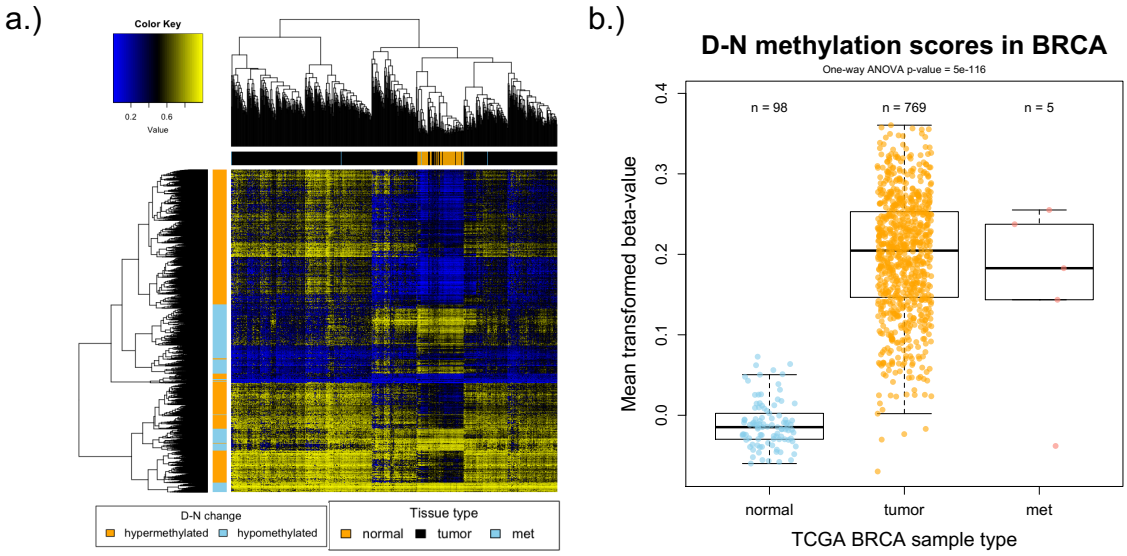


Figure 4–5: DMPs in D-N show consistent change in invasive breast cancer.

a). 14,652 DMPs between DCIS and normal tissue in the TCGA BRCA dataset. Note that the methylation profiles in the tumors are concordant with the methylation profiles in DCIS when compared to normal breast tissue. b) D-N methylation scores calculated for tumor adjacent normal, tumor, and metastasis samples.

Studies have shown that DNA methylation changes occur in blocks across the human genome and clusters of neighboring CpGs known as differentially methylated regions (DMRs) [80, 145, 146]. DMRs act as functional units that affect change in gene expression, where hypermethylation in promoter region or first exons leads to gene silencing, and concerted hypermethylation in the gene body was often correlated to increased gene expression. While DMRs are a natural consequence of statistical smoothing of DNA methylation sequencing data, microarray data are represented as probes that span specific genetic loci and are not as readily quantifiable into methylation blocks. To address the need to identify functionally relevant changes in DNA methylation, computational algorithms have been developed to perform smoothing of microarray data [147, 148]. Consistent with this theory, I observed concerted differentially methylated neighboring probes in many genes and performed DMR analysis using *DMRcate* [147]. Differential methylation in the promoter region of RASSF1A is shown as a representative example (Figure 4-6).

A total of 678 DMRs were identified comparing DCIS and normal tissues, including in regions where aberrant methylation have been observed in breast cancer (Table 6) [80, 149]. Promoter hypermethylation of the HOX family genes of master regulators were commonly observed. HOX genes have been implicated in the oncogenesis and aggressive phenotypes in breast cancer [150-153]. Recently, work in our lab has shown that HOX genes regulate cell fate transition [154], invasiveness [155], and endocrine therapy resistance [156]. Furthermore, tumor suppressor genes involved in DNA repair, such as TP73, APC, CCND2, and MGMT, were also differentially

methylated. I also identified DMRs in the estrogen responsive genes, ESR1, RUNX3, and FOXA2, suggesting that at least a subset of the DCIS have dysregulated ER signaling.

Figure 4–6: Differentially methylated region identified in a CpG island in the promoter region of RASSF1A in D-N analysis.

RASSF1A promoter hypermethylation have been implicated in breast cancer development and this was observed in a D-N comparison. The chromosomal ideogram shows the chromosome the gene is located in and the red vertical bar highlights the locus displayed in the panels below. The genomic axis for hg19 is displayed. The blue block arrows identify genes and the transcription direction. Green bars represent CpG sites and the purple box highlights identified DMRs. The line charts on the bottom show beta-value change in DCIS (orange) and normal tissue (blue).

Table 6: Differentially methylated regions in DCIS (D-N)

Chromosome	Start	End	Width	CpG N	Min FDR	Stouffer	Max delta	Mean beta	Overlapping promoter(s)
chr1	3566950	3568245	1296	21	6.21E-12	1.24E-07	0.25	0.15	WRAP73, TP73
chr1	3605979	3607425	1447	11	6.52E-09	3.10E-05	0.19	0.09	TP73
chr1	25255838	25258679	2842	24	1.21E-15	1.66E-13	0.26	0.14	RUNX3
chr10	131264786	131265073	288	4	6.45E-05	7.77E-03	0.27	0.11	MGMT
chr11	32421514	32421845	332	4	1.11E-07	1.82E-04	0.39	0.22	WT1
chr11	32454718	32456340	1623	12	1.06E-06	3.29E-05	0.23	0.14	WT1, WT1-AS
chr11	32459760	32459954	195	2	7.25E-04	2.60E-02	0.13	0.12	WT1-AS
chr11	67351271	67352041	771	6	1.03E-04	8.88E-04	0.20	0.13	GSTP1
chr12	4380586	4382188	1603	16	1.26E-09	5.25E-08	0.24	0.13	CCND2
chr12	4383281	4384751	1471	7	5.83E-07	7.39E-05	0.19	0.12	CCND2, CCND2-AS2, CCND2-AS1
chr12	54345056	54346784	1729	8	1.66E-08	2.96E-06	0.42	0.16	HOXC12
chr12	54349169	54349349	181	2	2.35E-05	1.64E-03	0.21	0.20	HOXC12, AC012531.23
chr12	54412565	54413384	820	6	3.43E-08	9.63E-05	0.31	0.19	HOXC4, RP11C11.14, AC012531.25
chr12	54446944	54448913	1970	14	8.20E-06	4.09E-05	0.20	0.14	HOXC4
chr16	82660206	82660873	668	8	1.07E-07	3.76E-05	0.21	0.11	CDH13
chr17	46655164	46656093	930	17	2.60E-07	9.23E-04	0.21	0.11	HOXB4, MIR10A, HOXB3
chr17	46667683	46667812	130	2	1.79E-04	3.52E-03	-0.15	-0.13	HOXB-AS3, HOXB3
chr17	46690336	46692248	1913	7	9.90E-05	5.29E-03	0.19	0.10	HOXB8, HOXB7
chr17	46703646	46704004	359	6	3.55E-04	4.17E-03	0.13	0.10	HOXB9
chr17	46711017	46711446	430	5	2.30E-04	1.21E-03	0.22	0.18	MIR196A1, HOXB7
chr17	46806445	46806935	491	5	9.54E-05	2.74E-03	0.20	0.12	HOXB13
chr18	49866065	49868552	2488	14	1.69E-07	2.14E-07	0.23	0.12	DCC, RP11-25O3.1
chr18	60984485	60984656	172	2	8.58E-06	7.10E-04	-0.27	-0.20	BCL2
chr19	42408316	42408464	149	3	1.01E-04	1.71E-03	-0.19	-0.19	ARHGEF1
chr2	176957304	176958174	871	5	7.20E-06	4.05E-04	0.27	0.15	HOXD13
chr2	176963583	176964720	1138	11	5.99E-13	3.22E-08	0.30	0.18	HOXD12
chr2	176971304	176973275	1972	14	3.42E-14	1.68E-07	0.30	0.15	HOXD11, AC009336.1, HOXD10
chr2	176980837	176982230	1394	9	2.32E-16	1.81E-06	0.42	0.19	HOXD10
chr2	176986460	176988505	2046	16	4.30E-16	1.24E-07	0.33	0.16	HOXD9
chr2	176993017	176995556	2540	17	3.99E-10	5.46E-08	0.25	0.15	HOXD8
chr2	177001256	177001909	654	5	3.45E-06	1.82E-04	0.29	0.18	HOXD-AS2, HOXD3
chr2	177052903	177054306	1404	11	7.40E-06	2.01E-04	0.17	0.09	HOXD1, HOXD-AS1
chr20	22565881	22567920	2040	12	3.03E-11	1.19E-08	0.27	0.17	FOXA2
chr3	25469914	25469925	12	3	2.03E-07	2.59E-05	0.16	0.14	RARB
chr3	79815639	79817278	1640	10	1.52E-07	2.17E-08	0.25	0.15	ROBO1
chr5	112073348	112074043	696	14	1.57E-13	6.22E-08	0.30	0.17	APC
chr6	152127812	152128426	615	7	9.03E-06	2.95E-03	-0.25	-0.12	ESR1
chr7	19157263	19158134	872	13	1.37E-05	6.73E-03	0.16	0.09	TWIST1, AC003986.7
chr7	27135147	27136424	1278	10	2.06E-15	2.07E-08	0.24	0.16	HOXA1, HOTAIRM1
chr7	27145972	27146445	474	4	2.48E-05	2.83E-04	0.27	0.21	HOXA-AS2
chr7	27168688	27171401	2714	26	2.07E-31	9.40E-19	0.32	0.18	HOXA4, HOXA-AS3
chr7	27182493	27184375	1883	31	1.29E-23	2.73E-07	0.28	0.14	HOXA5
chr7	27190431	27191564	1134	7	4.09E-05	2.67E-03	0.14	0.11	RP1O19.22, RP1O19.23, HOXA3, HOXA6, HOXA-AS3
chr7	27195036	27198189	3154	18	4.01E-08	2.01E-07	0.23	-0.01	HOXA7, RP1O19.21
chr7	27204349	27205658	1310	15	3.18E-09	1.47E-05	0.24	0.13	HOXA9
chr7	27213610	27214201	592	10	9.80E-07	7.13E-04	0.27	0.15	HOXA10, RP1O19.20
chr7	27238910	27239763	854	4	2.92E-04	1.79E-03	0.21	0.15	HOXA13, HOTTIP
chr7	116166408	116166824	417	4	1.32E-05	7.30E-04	0.14	0.10	CAV1
chr8	41165699	41167278	1580	8	4.97E-05	5.03E-05	0.24	0.14	SFRP1

4.3.2: Tumor-adjacent normal tissues display intermediate hallmarks of DCIS

Results from exploratory analysis using PCA, unsupervised hierarchical clustering, and D-N differential methylation analysis show that DCIS-adjacent normal tissue cluster closer to normal breast tissue than to DCIS (Figure 4-4). Interestingly, within a subset of D-N DMPs, I observed methylation profiles intermediate to that between DCIS and normal tissue in the series of tumor-adjacent normal tissue. Indeed, when I calculated a methylation score for the JHU DCIS cohort, the DCIS adjacent

normal tissues had methylation scores intermediate to that of normal breast tissue and DCIS (Figure 4-7a), and this result was statistically significant (Bonferroni adjusted pairwise Wilcox test, $p < 0.05$ across all comparisons). A subset of DCIS-specific probes was differentially methylated between adjacent normal and normal tissue, with the adjacent normal tissue showing intermediate methylation profiles (Figure 4-7b). To identify if there are oncogenic differences between probes differentially methylated between DCIS-adjacent normal and normal breast tissue (A-N), I identified a series of 1214 DMPs with p-values < 0.01, and used this series to calculated a methylation score in the TCGA dataset. This analysis showed a statistically significant difference between methylation scores of tumor and tumor adjacent normal samples (Figure 4-7c). This signature also distinguishes normal tissue from DCIS (Figure 4-7d).

This observation was not due to contamination of DCIS cells during macrodissection because a contamination derived profile will show intermediate methylation across most probes, and this was not evident in the clustering of all DCIS-specific probes (Figure 4-6c). In addition, this effect has been reported in both DCIS and IDC in the literature. Concordant oncogenic gene expression changes have been observed in DCIS- and IDC- adjacent stroma [157]. Furthermore, other studies have shown field cancerization [158] in breast cancer, where cancer-associated genetic [159], epigenetic [130, 160], transcriptomic [161, 162], and telomeric [163] changes were observed in neighboring normal breast epithelium. To our knowledge, this is the first study which identified global methylation changes in DCIS-adjacent normal tissue.

Our limited sample size suggests that this effect happens on a subset of DCIS specific methylation changes, and further studies evaluating the change in prognostic

markers in these DCIS adjacent tissues are important. From a clinical testing point of view, the presence of such field effects would eliminate the need for extensive micro- or macrodissection, increasing ease-of-use and technical reproducibility.

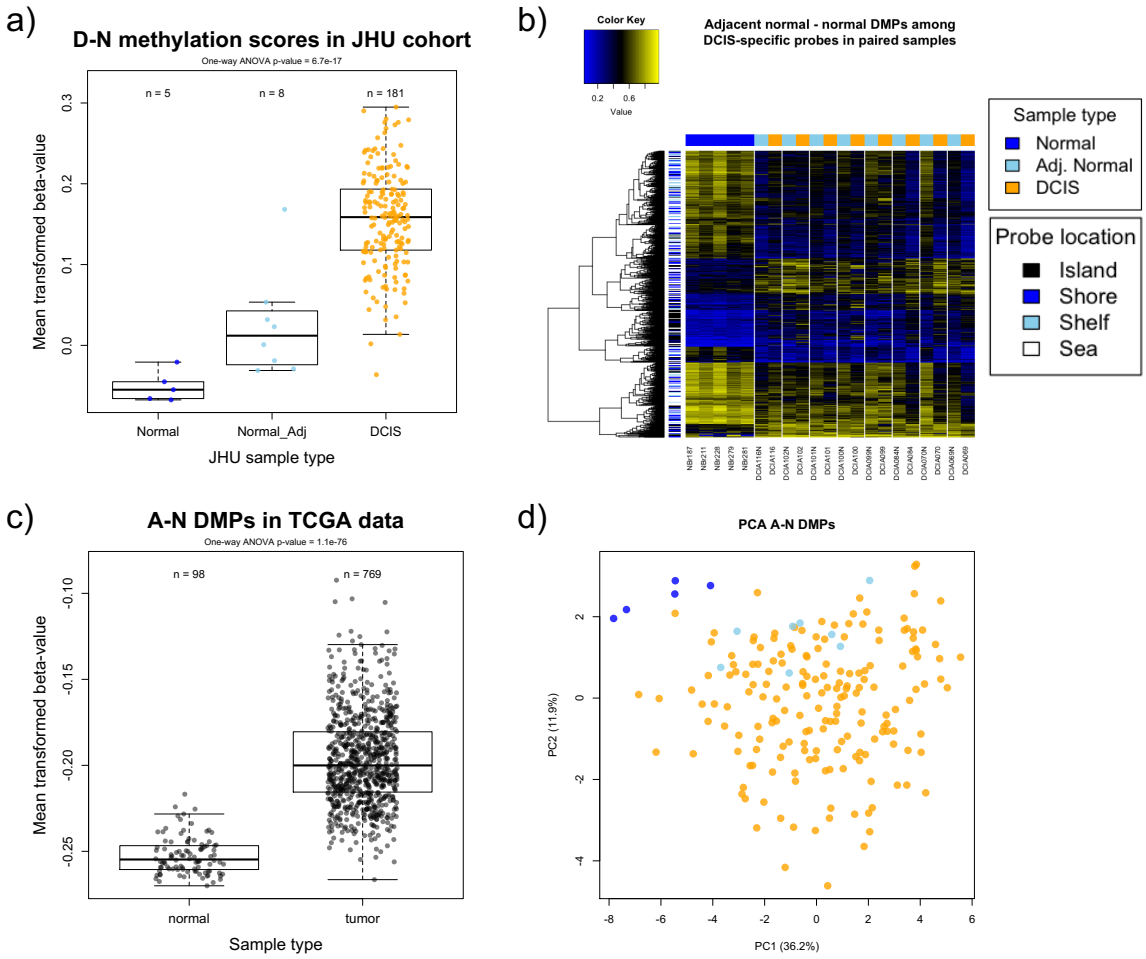


Figure 4–7: Hallmarks of DCIS and oncogenic methylation observed in DCIS adjacent normal

a). D-N methylation scores across normal, DCIS-adjacent normal, and DCIS show that DCIS-adjacent normal has intermediate methylation scores, suggesting that a subset of D-N probes are altered in these samples. b) DMPs comparing DCIS-adjacent normal and normal (A-N) profiles in normal, DCIS adjacent normal, and paired DCIS reveal that these methylation profiles are intermediate in the DCIS-adjacent normal. c) Higher methylation scores observed in tumors compared to normal in A-N probes suggests that A-N probes are associated with tumorigenicity. d) PCA of A-N probes show clustering of DCIS samples with DCIS-adjacent normal away from reduction mammoplasty normal.

4.3.3: Unsupervised clustering identified four methylation clusters

Consensus clustering was performed using the ConsensusClusterPlus package in R in an effort to identify methylation subtypes in DCIS-specific probes (D-N probes, FDR < 0.05) in these DCIS samples. Probes were restricted to 14,652 DCIS-specific set to enrich for probes with functional relevance in disease. Consensus clustering is a resampling based method, which allows us to assess the stability of discovered clusters and address the question of over-fitting, prominent in these high dimensional datasets. Using the *partitioning around medoid* (PAM) algorithm, using Pearson correlation with average linkage and 80% resampling, we found that 4 clusters of samples, that we have named epitypes, are most stable in this set of probes (Figure 4-8 and 4-9). Hierarchical clustering with complete linkage was used to identify clusters of probes, and this revealed 3 major clusters, which were functionally classified using a hypergeometric test against the C2:CGP (chemical and genetic perturbations) gene sets.

Progressor status was not associated with any of the epitypes, but epitype 1 was enriched for high nuclear grade DCIS (Table 7, Figure 4-9). Interestingly, epitype 4 showed highly methylated CpG islands. To assess if this was a global event, we calculated average beta-values for all CpG island probes, and observed that epitype 4 had higher mean hypermethylation in promoter-specific probes compared to the other epitypes ($p < 0.0001$, pairwise Wilcoxon test with Bonferroni adjustment).

All of the probes were enriched for Polycomb protein (PRC2, SUZ12, and EED) targets, suggesting dysregulation of DNA methylation machinery, a phenomenon observed across multiple cancer types [164]. Clusters 1 and 4 were enriched for probes

located within or near genes involved in epithelial-mesenchymal transition (EMT), while probes in clusters 2 and 3 were enriched for estrogen responsive genes. Interestingly, cluster 4 is the only cluster where probes were predominantly located within CpG islands and there probes were hypermethylated in a subset of DCIS sample, a phenotype known as CIMP, which is common in cancer [79].

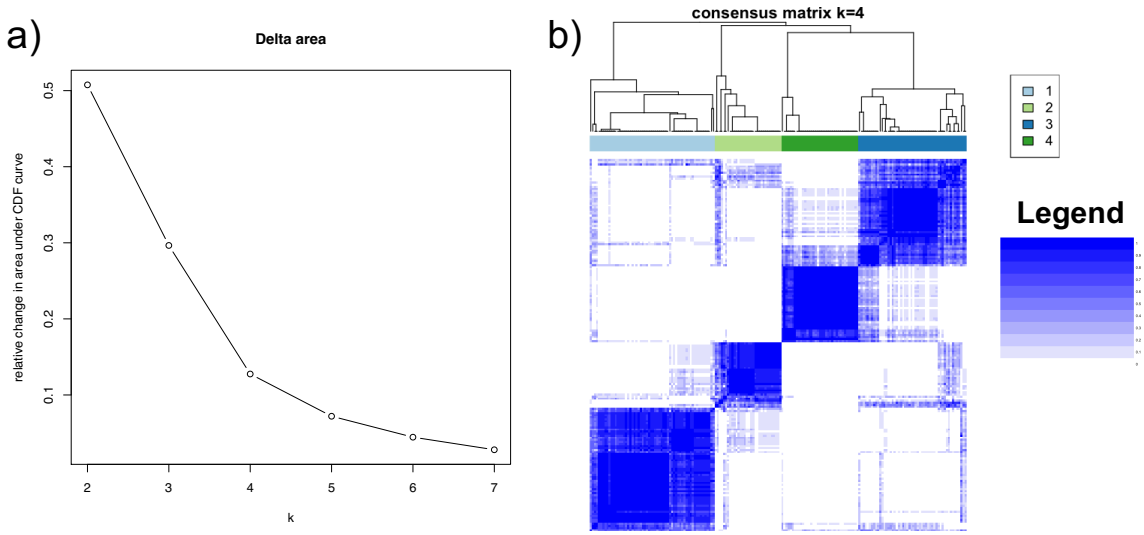


Figure 4–8: Consensus cluster metrics for selection of optimal K

a) Change in area under the CDF curve identifies k = 4 as the inflection point. b) Consensus matrix for k = 4 shows good consensus and stability across four clusters.

Table 7: Association of high grade DCIS with DCIS methylation epitype 1

Grade	Methylation cluster, n (%)			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
High	18 (75.0)	4 (21.1)	10 (40.0)	7 (38.9)
Intermediate	4 (16.7)	14 (73.9)	8 (32.0)	4 (22.2)
Low	2 (8.3)	1 (5.3)	7 (28.0)	7 (38.9)

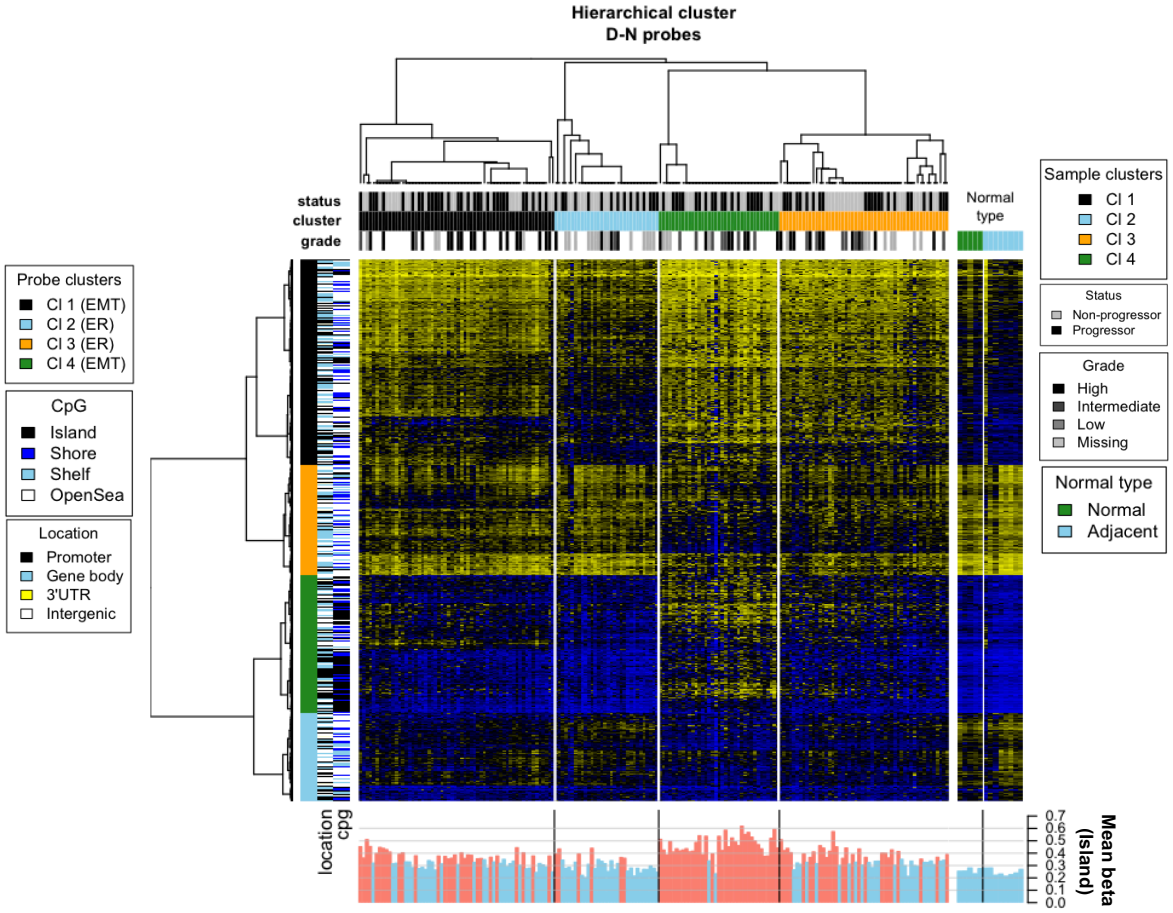


Figure 4–9: Clustering of DCIS samples

Clustering of DCIS samples of 14,529 probes differentially expressed between DCIS and normal samples. Samples were clustered using consensus clustering and four stable clusters were identified. Probes were clustered using hierarchical clustering with the Ward clustering algorithm. Methylation profiles of 5 pure normal samples and 8 DCIS-adjacent normal are displayed on the right. While progressor status did not correlate with any cluster, epitype 1 showed enrichment for high grade DCIS. Barplots on the bottom show the island methylation score, average beta-value (methylation levels) of all variable probes (SD > 1.5 IQR). Red colored barplots highlight samples with IMS above the median of all DCIS samples. Note that most samples in epitype 4 had IMS above the median.

4.3.4: Differential methylation analysis on DCIS-specific genes between progressive and non-progressive DCIS shows no DMPs

We performed differential methylation analysis using limma to identify methylation associated with progression within DCIS-specific probes identified in the D-N analysis. This analysis revealed no statistically significant probes differentially methylated between cases and controls (Table 8). The same analysis was repeated with most variable probes > 1.5 IQR, but the results remained similar (data not shown).

Furthermore, a Cox regression analysis controlling for age and radiotherapy did not show improvement in individual probes predicting risk of progression (data not shown). A supervised principal component analysis was performed using the *superpc* package in the R Statistical Environment to explore and assess the possibility of more complex interactions between probes that may contribute to progression status. Probes that were significantly differentially methylated between progressive and non-progressive DCIS were more closely associated with epitype than progression status (Table 9, Figure 4-10).

These results suggest that there no clear-cut methylation differences between progressive and non-progressive DCIS in this cohort, at least without further subclassification using additional molecular and/or clinical features. It has been shown that DCIS, like IDC, can be separated into different molecular subtypes, and complex molecular interactions between subtypes and progression may impede our ability to identify robust progression markers with the currently available data.

Table 8: DMPs comparing case and control in DCIS-specific probes

ProbeID	Delta_Beta	Ave_Beta	t	pval	FDR	B	Gene_Symbol	Relation_to_Island
cg02532672	0.07	0.70	4.60	7.55E-06	1.11E-01	3.45	HEATR2	Island
cg17439800	-0.07	0.44	-3.93	1.17E-04	8.61E-01	0.92		OpenSea
cg15206981	0.06	0.53	3.82	1.81E-04	8.82E-01	0.53	SPRY1	Shelf
cg03128029	-0.05	0.36	-3.43	7.45E-04	1.00E+00	-0.76	NOP58	OpenSea
cg11214507	-0.07	0.41	-3.40	8.27E-04	1.00E+00	-0.85		OpenSea
cg21923959	-0.05	0.43	-3.34	1.02E-03	1.00E+00	-1.04	POU2AF1	OpenSea
cg15093997	0.06	0.44	3.33	1.02E-03	1.00E+00	-1.05	CHST3	Shore
cg12081643	-0.06	0.55	-3.32	1.07E-03	1.00E+00	-1.09	COL4A1	OpenSea
cg19304088	-0.06	0.32	-3.21	1.56E-03	1.00E+00	-1.43	PITX2	Shelf
cg08858272	-0.05	0.38	-3.20	1.60E-03	1.00E+00	-1.45	NALCN	OpenSea
cg09025324	-0.10	0.39	-3.20	1.60E-03	1.00E+00	-1.45	DSE	Shore
cg24201034	0.05	0.25	3.20	1.63E-03	1.00E+00	-1.46	SHROOM3	Island
cg06099431	0.05	0.13	3.16	1.85E-03	1.00E+00	-1.58		Island
cg18475969	-0.06	0.31	-3.15	1.90E-03	1.00E+00	-1.60		OpenSea
cg17425818	-0.05	0.43	-3.13	2.02E-03	1.00E+00	-1.65	NRP1	OpenSea
cg02928365	-0.05	0.36	-3.13	2.04E-03	1.00E+00	-1.66	HLX	Shore
cg20140333	-0.06	0.52	-3.12	2.12E-03	1.00E+00	-1.70		OpenSea
cg23014549	-0.06	0.53	-3.10	2.26E-03	1.00E+00	-1.76	KIF26B	OpenSea
cg12789884	-0.06	0.56	-3.07	2.42E-03	1.00E+00	-1.82		OpenSea
cg13787850	-0.05	0.31	-3.07	2.43E-03	1.00E+00	-1.82		OpenSea
cg13027727	-0.06	0.62	-3.07	2.44E-03	1.00E+00	-1.82	SYT7	OpenSea
cg05095252	-0.06	0.45	-3.05	2.58E-03	1.00E+00	-1.87	LYPD6	OpenSea
cg26926765	-0.05	0.49	-3.05	2.59E-03	1.00E+00	-1.88	C6orf142	OpenSea
cg08750510	-0.06	0.47	-3.03	2.77E-03	1.00E+00	-1.94		OpenSea
cg07960083	-0.05	0.60	-3.03	2.79E-03	1.00E+00	-1.94		Shelf

Table 9: Probes associated with progression status as identified by supervised PCA

ProbeID	Gene_Symbol	Relation_to_Island
cg19044229	MAP3K11	Island
cg13787850		OpenSea
cg02928365	HLX	Shore
cg08858272	NALCN	OpenSea
cg06099431		Island
cg15093997	CHST3	Shore
cg02532672	HEATR2	Island
cg12081643	COL4A1	OpenSea
cg15206981	SPRY1	Shelf
cg21923959	POU2AF1	OpenSea
cg04129308	TCF21	Shore
cg17439800		OpenSea
cg07960083		Shelf
cg03128029	NOP58	OpenSea
cg24201034	SHROOM3	Island

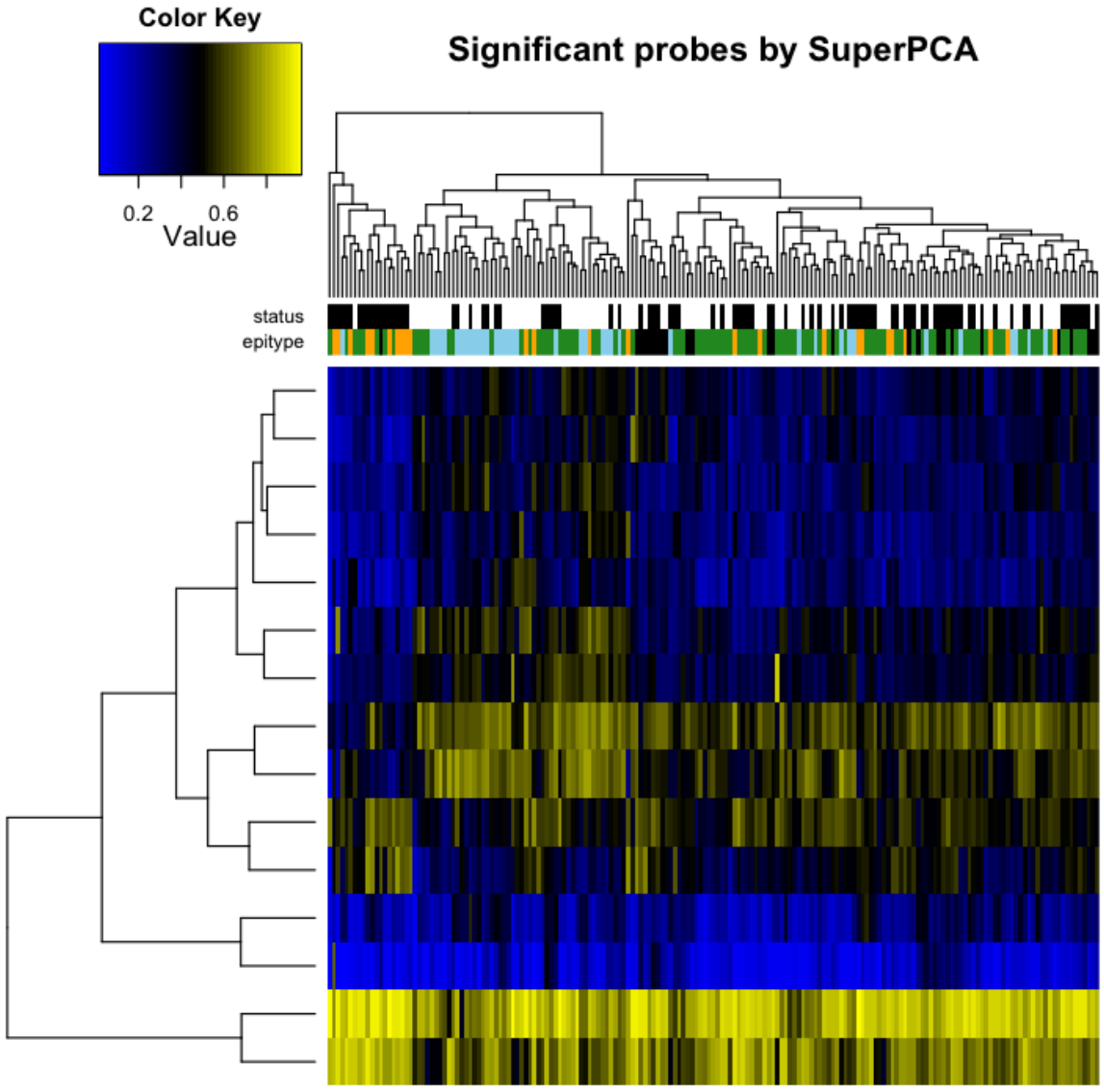


Figure 4–10: Supervised principal component analysis

Super PCA identified 15 probes associated with progression status across 3 principal coponents. Interestingly, these probes had greater association with epitype than with progression status.

4.3.5: CNV data recapitulate previously identified recurrent CNVs in DCIS

The CNV profiles of all DCIS samples were tabulated by genetic location into proportions with a given alteration and compared it to proportions estimated by a meta-analysis performed by Rane et al. on previously published DCIS CNV studies. We observed good concordance between previously published DCIS profiles and profiles from the JHU cohort. Interestingly, we observed increase incidences of CNV in parts of the genome, which may be due to the enrichment for progression cases in our cohort compared to the average DCIS population (Figure 4-11).

4.3.6: Differences in proportions of CNVs in progressive and non-progressive DCIS suggest molecular lesions of interest

Furthermore, we compared the CNV proportion between progressive and non-progressive DCIS and identified regions which tend to be altered according to progression status (Figure 4-12). The most prominent of these include CNV in chromosome 8, where we observed increased copy number loss in progressors and high copy number gain in non-progressors. This may speak to the presence of a tumor, or "progression", suppressor gene in this region.

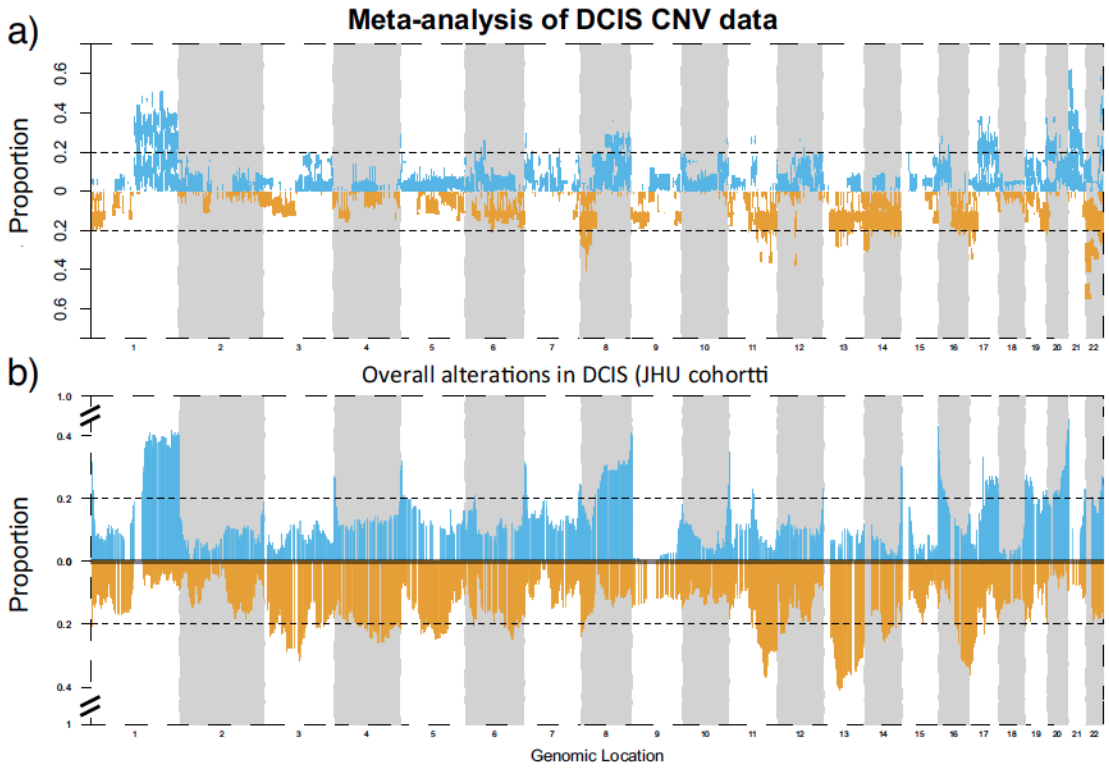


Figure 4–11: CNV events by incidence in previously published studies and JHU cohort

a) Proportion of previously published DCIS samples from the meta-analysis of Rane et al. which showed CNV alterations at specific regions of the genome. b) The same information for the proportion of DCIS samples in our study. Many of the events identified in the meta-analysis were observed in our dataset. Furthermore, there are some regions in the JHU cohort that were not observed in the meta-analysis, e.g., the copy number loss in chromosome 3, which may be a result of enrichment for progressive DCIS compared to previous studies.

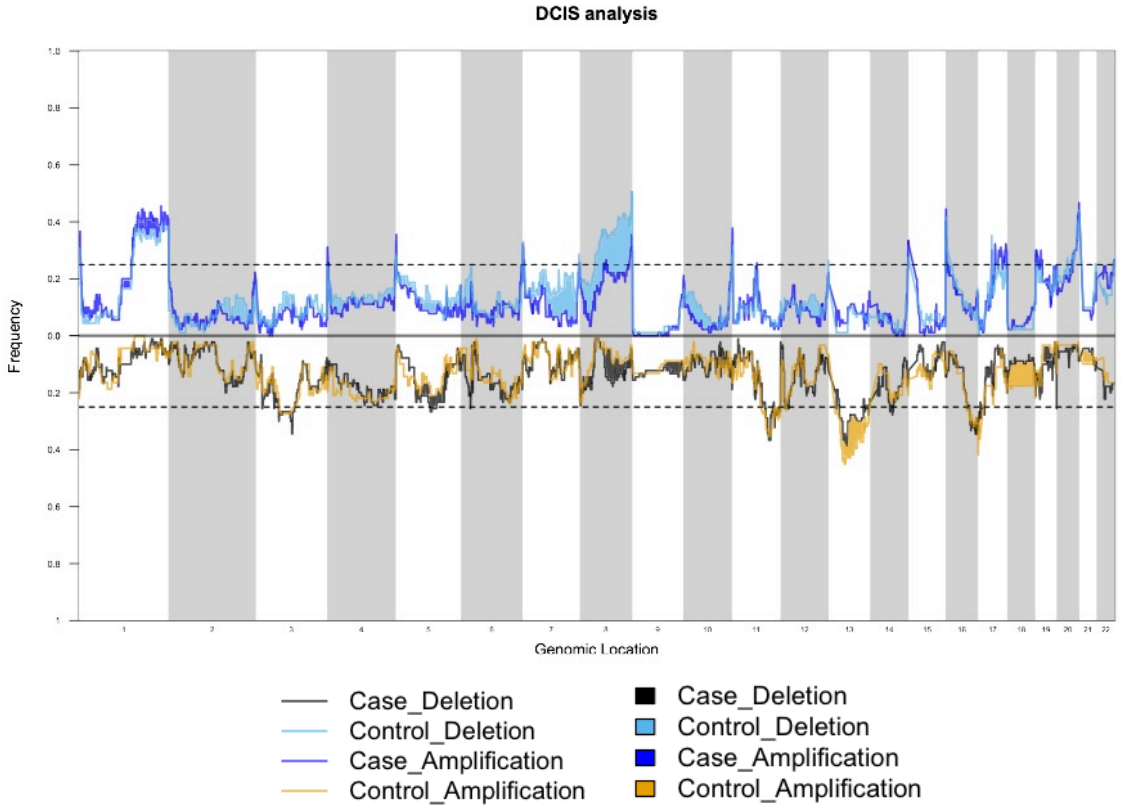


Figure 4–12: Comparison of CNV incidences across case and controls in JHU DCIS cohort

4.4: Conclusion

We detected methylation changes between breast reduction mammoplasty normal and DCIS that confirm previously published findings, suggesting that biologically relevant data were obtained in this low resource setting limited by tumor size and FFPE-derived DNA & RNA. Furthermore, we observed global methylation field effects associated with malignancy in DCIS adjacent normal tissue, extending the candidate gene observations from previous studies. We also identified four stable methylation epitypes of DCIS that showed associations to tumor grade. Furthermore, one epitype exhibited

patterns similar to breast tumors with the CpG island hypermethylation phenotype (CIMP), where we observed overall hypermethylation of CpG islands. Differential methylation analysis and supervised PCA to identify progression-related features revealed no statistically significant probes or differentially methylated regions, and could be a result of the molecularly complex phenotypes in breast cancers and DCIS.

A CNV analysis revealed CNV incidences largely similar to previously published studies, suggesting that there are biologically relevant CNV data obtained using Epicopy. A comparison of amplified and deleted regions in progressors and non-progressors revealed several regions where incidences differ significantly. Interestingly, chromosome 8 exhibited a relationship where copy number loss was prevalent in progressors while copy number gain was enriched in the non-progressors.

Chapter 5: Multiomic analysis and prognostic biomarker discovery in ER-negative breast cancer of patients who did not receive chemotherapy

5.1: Introduction

In 2015, there were a total of 231,840 newly diagnosed cases of breast cancer and 40,290 breast cancer-related deaths in women in United States. Of these, an estimated 15 – 20% were of the triple negative breast cancer (TNBC) subtype, which test negatively for three clinical markers used in treatment decisions; estrogen receptor (ER), progesterone receptor (PR), and Her2-amplification (HER2). Unlike their ER/PR and HER2 counterparts which are treated with hormonal therapy and HER2-targetted therapies respectively, TNBCs currently have no targetable driver alterations. Beyond the context of treatment, TNBCs are also naturally associated with early onset of disease and more rapidly progressive disease.

Due to the lack of targetable therapies, a number of seminal clinical trials were performed in the early 1990s by multicenter study groups to assess the impact of adjuvant chemotherapy on disease free survival in ER-negative breast cancer patients. One such group was the International Breast Cancer Surgical Group (IBCSG) which performed two key trials, Trial VIII and Trial IX [165], that randomized early stage, operable, node negative breast cancer patients into two arms, one where patients were treated only with surgery and local radiation, and the other where patients received surgery, radiation, and a CMF (cyclophosphamide, methotrexate, and 5-fluorouracil) chemotherapy regiment (Figure 5-1). At 12-years of follow-up, there was a statistically significant survival benefit of 15% in the CMF-treated arm (70% disease free) compared to those who did not

receive chemotherapy (55% disease free). Other studies reflect disease free survival benefits similar to this study [166]. We can identify three subgroups of patients in terms of benefit to chemotherapy from these studies; 1) 15% of TNBC patients benefit from chemotherapy, 2) 30% of patients need better therapy than CMF, and 3) 55% of patients with operable, node-negative TNBC disease do not need chemotherapy of at least up to 12 years after initial diagnosis.

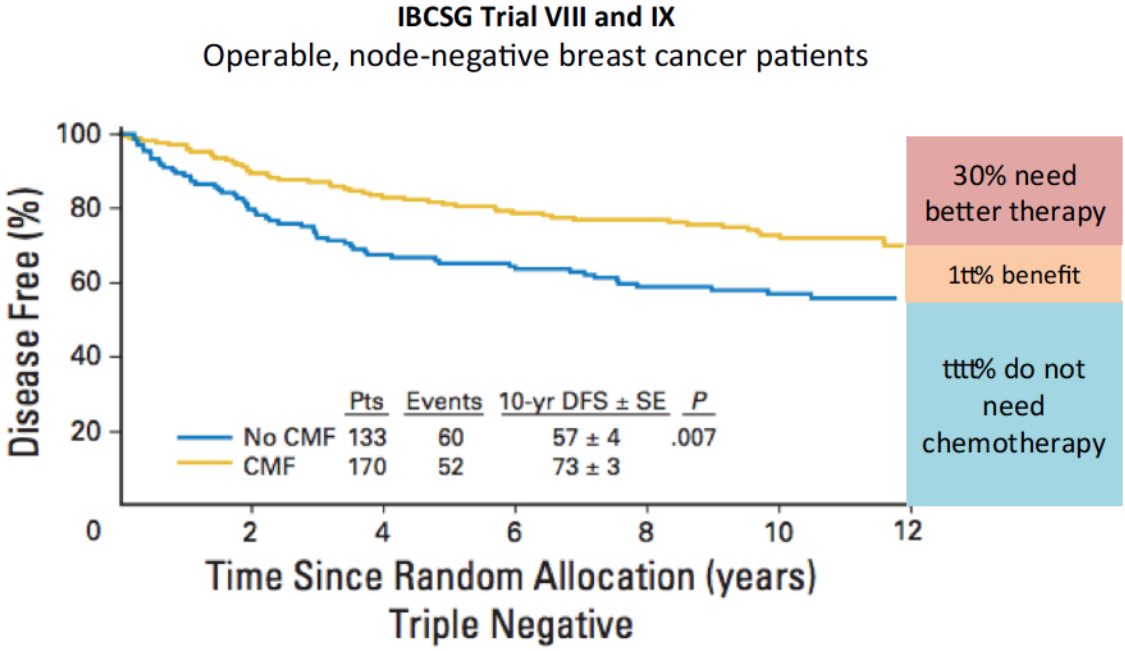


Figure 5–1: International Breast Cancer Surgical Group – Trial VIII and IX TNBC 12 year follow up

12-year follow-up results from IBCSG trials VIII and IX of the TNBC patients in the study [165]. At 12-years, there was a statistically significant survival advantage of 15% in patients receiving CMF chemotherapy compared to patients who did not. Looking at this Kaplan-Meier curve, we also observe that 30% of the patients had a recurrent event within 12-years, suggesting the need for better therapy. More importantly, 55% of the patients in the no CMF arm were disease free 12-years after the initial therapy, suggesting that there exist a group of TNBC patients who do not need chemotherapy and the current modality of treating all TNBC patients with adjuvant chemotherapy represents a problem of over-treatment for a substantial subset of patients.

Given the relatively aggressive nature of the disease and the lack of clinicopathological or molecular features that allow clinicians to stratify patients into different risk groups, it is understandable, and even logical, that clinical management of the disease to use the maximal therapeutic options at our disposal to minimize the risk of metastatic disease, even at the risk of over-treating half of TNBC patients. Based on the most conservative estimates, approximately 35,000 women are diagnosed with TNBC, with half, or 17,500, of them receiving unnecessary chemotherapy, which exposes the patients to harmful side effects and potential comorbidities. Therefore, there exists an unmet clinical need to identify markers of stratifying patients into different risk groups that will allow the field to apply more targeted therapies to this patient population.

To that end, molecular markers serve as attractive targets. Recent technological advances led to the identification of several gene expression molecular subtypes of breast cancer, which are associated with known clinical markers used in therapeutic decisions; ER, PR, HER2, and Ki67, the last of which is a marker of proliferation [167]. In fact, these intrinsic molecular subtype terminologies have been widely adopted in clinical practice, suggesting a change in the mindset of thinking about breast cancer in molecular terms [168].

Such observations have led to the idea that further molecular profiling of TNBCs will reveal additional functional subgroups with new therapeutic targets or identify markers of aggressiveness and risk. In a study analyzing 587 TNBC samples from publically available datasets, Lehmann et al. [169] identified subtypes within TNBCs, including luminal androgen receptor (LAR), immunomodulatory, basal-like, mesenchymal, and mesenchymal stem-like subtypes. Since then, multiple studies have

validated the existence of at least two of these subtypes, namely the LAR [170] and immunomodulatory subtypes. Such studies have shown the possibility of stratifying TNBCs into functional subtypes with different risks of progression, but the fundamental clinical question about chemotherapy benefit remains to be addressed.

Current clinical molecular marker test panels designed to address this question fall short for TNBCs. The most commonly used molecular panel is the 21-gene Oncotype DX panel [57], which stratifies patients into three risk groups of low, intermediate, and high risk, for the purpose of determining the need for chemotherapy in ER-positive disease, and offers no stratification in ER-negative disease. The 70-gene assay Mammaprint assigns 95% of all ER-negative disease to its high-risk category, offering little additional information for the treatment decisions in TNBC [171]. Taken together, this speaks to the need of a robust molecular marker panel, which will allow us to identify patients with low risk of recurrence in the absence of chemotherapy and spare patients from unnecessary side effects and comorbidities of treatment.

5.2: Study design and methods

5.2.1: Motivation

The motivation of this study was thus to identify a set of molecular biomarkers of early disease with high negative predictive value (NPV) to prognosticate patients treated with only surgery and radiotherapy. We estimate an NPV of 0.95 would be required to significantly impact clinical practice. To achieve this, we aim to use high-density array

technologies for gene expression and methylation to perform molecular profiling to identify probes associated with recurrence in TNBC patients who did not receive adjuvant chemotherapy.

5.2.2: Study design

This study was designed as multicenter, nested case-control study of early stage, node negative, ER-negative invasive ductal carcinomas (IDC) of patients who never received adjuvant or neo-adjuvant chemotherapy of their primary tumors. We identified a series of 75 cases, which are patients with a recurrence event with at least a 6-month lead time after the initial treatment, with the lead time implemented to filter against patients whose recurrence were due to incomplete local therapy. An equal number of controls were identified with at least 5 years of follow-up with no indication of recurrence. In assembling this cohort, we controlled for clinicopathological features that may influence recurrence, including histological grade, margin status, and adjuvant treatments, as well as approximate age and year of diagnosis (both within a 5 year window). All relevant clinical information has been captured in an anonymized research database, and 20 unstained sections with matching H&E stained control slides are being obtained from the relevant tissue blocks with coded identifiers.

In order to ensure diagnostic consistency, our study pathologist, Dr. Gabrielson (JHU), reviewed all cases and controls using the Aperio digital imaging system (Aperio Inc., Vista, CA), which allows very efficient remote visualization and interactive annotation of the entire histological slide at high (20x and 40x) resolution. For this study,

we were able to obtain three sets of molecular information. Gene expression profiling was performed using the Illumina DASL microarray, while methylation profiling was obtained using the Illumina 450K Human Methylation microarray (IL450K). CNV data were estimated using Epicopy on data from IL450K.

5.2.3: Patient identification and sample collection

We used patient registries here at Johns Hopkins Hospital and at collaborating institutions to identify cases and controls that matched study criteria and had documented long term follow up. Tissues were obtained with approval of the respective institutional IRBs. Unstained tissue sections were obtained and macro dissected using pathologist annotated H&E sections for orientation and macrodissection for enrichment. A total of 75 recurrent TNBC cases, 77 non-recurrent TNBC controls, and 5 reduction mammoplasty samples were profiled using gene expression (Illumina DASL) and DNA methylation platforms (Illumina 450K).

5.2.4: DNA/RNA extraction and quality control

H&E sections were macrodissected to enrich for >70% DCIS epithelial cells. Following that, DNA and RNA were extracted using Allprep FFPE RNA/DNA kit (Qiagen) with modifications to the deparaffinization, digestion, and wash steps. The modified protocol is appended.

Quantification of RNA and DNA was performed using Nanodrop2000 and the Qubit fluorometer (Qiagen) using appropriate kits (RNA HS, RNA BR, DNA HS, and DNA BR). The 260/230 and 260/280 ratios were used to assess sample purity and solvent contamination. Qubit-derived measurements were used to calculate nucleic acid input into microarray platforms.

5.2.5: Quality control and microarray

Illumina FFPE QC kit was performed using the iTaq™ Universal SYBR® Green Supermix and was regarded as the main quality control step for 450K and other DNA-based microarrays. Samples with delta C_T <9 were used in the study and case-control pairs with lower delta C_T were prioritized. Bisulfite conversion was performed using the EZ DNA Methylation-Gold™ Kit (Zymo Research, Irvine CA), with modifications introduced per Appendix I of the manufacturer's recommended protocol. The detailed protocol is appended at the end of the thesis. NaBi-converted DNA was submitted to the SKCCC Microarray Core Facility for FFPE DNA restoration and profiling using the Illumina 450K microarray. RNA QC was performed using QPCR with primers targeting the GAPDH gene and samples with delta C_T larger than 32 were not profiled using the DASL microarray.

5.2.6: Data pre-processing and QC

Unless otherwise stated, data analysis was performed in R Statistical Environment using base, Bioconductor, and custom packages. P-values were corrected using Benjamini-Hochberg's method for false discovery rate estimation.

Illumina DASL Microarray P95 green signal intensities were used as a measure to assess overall gene expression for a given sample. Outliers with low P95s were excluded. Normalization was performed using median absolute deviation (MAD). This decision was performed based on the idea that WGA in DASL randomly amplifies RNA present in lower amounts, leading to the artifactual generation of bimodal distributions of genes present in marginal copies and quantile normalization-based methods are not appropriate in such a setting.

Illumina 450K Methylation Quality control metrics for Illumina-based arrays were estimated using Illumina's GenomeStudio software, and validated through control probe signal intensities extracted through the *minfi* software in R. GenomeStudio-derived detection p-values (detP) with a threshold of $p < 0.01$ were used to calculate sample-wise call rates and samples with call rates of less than 80% were removed from downstream analyses. Raw beta-value density plots were plotted and samples with aberrant beta-value density plots (without a bimodal distribution with means around 0.1 for unmethylated regions and 0.9 for methylated regions) were removed from analysis. Probe-wise detP were estimated and probes with > 95% coverage across remaining samples were retained

for analyses. Probes with interrogated CpGs 2bp from a known SNP with a population minor allele frequency (MAF) of > 5% were removed. Functional normalization was performed on the final set of high quality samples and probes to obtain the final methylation dataset.

Epicopy-derived CNV High quality samples and probes from the methylation pre-processing were used as input into Epicopy to generate CNV information for ER-negative tumor samples. Default Epicopy parameters were used with reduction mammoplasty normal samples serving as reference samples. Samples with aberrant profiles were discarded from downstream analyses. For clustering purposes, the CNV data from GISTIC 2.0 were summarized into cytobands by taking the mean of all genes in a given cytoband.

TCGA Data Processed TCGA data were downloaded from the Broad Institute's Firehose server.

5.2.7: Integrative data analysis

Exploratory analysis was performed using principal component analysis (PCA) and unsupervised hierarchical clustering, using Euclidean distances and Ward's algorithm, to identify outliers that can be removed from downstream analysis and identify high level clustering of the data. Final clustering analysis was performed using hierarchical clustering with Ward's algorithm on the 500 most variable genes from

transcriptomic data, and the same sample-wise dendrogram was used to cluster molecular data from the other platforms.

Gene-wise clusters in the transcriptome data were identified and a hypergeometric test-based analysis was performed on clusters of genes in Molecular Signatures Database (MSigDB) to assign functional information to these gene clusters. This was done using the "Compute Genesets Overlap" tool hosted on the Molecular Signatures Database. Analysis was performed against Hallmark and C2 gene sets, and gene sets in the top 10 were empirically summarized by both test significance and relevance to breast cancer (e.g. breast cancer related chemical and genetic perturbations were ranked higher than pancreatic cancer gene sets).

Differential expression analysis was performed between the three stable clusters to identify functional processes enriched within each cluster. *Limma* analysis was performed comparing each group to its counterparts with Benjamini-Hochberg's false discovery rate (FDR) method for p-value adjustment, and results were used in a GSEA-like, rank-based gene set analysis. Briefly, mean moderated t-statistics were used to rank genes. Unlike GSEA, the gene universe is first restricted to the 4,386 genes annotated in the Hallmark gene sets, which protects against over-enrichment bias for well-studied genes. The *wilcoxGST* function from the *limma* package was used to estimate enrichment significance, with the direction of alternative hypothesis specified to calculate enrichment for a specific phenotype. Finally, results were visualized using the Java implementation of GSEA using the pre-ranked method on moderated t-statistics.

5.2.8: PAM50 classification and leukocyte infiltration estimation

PAM50 classification was performed using the *intrinsic.cluster.predict* function from the *genefu* package against the pam50.scale model. Briefly, genes from transcriptomic data were mapped to EntrezID identifiers, which are matched to EntrezID identifiers in the PAM50 model. This prediction first calculates gene-wise Z-scores and estimates the Spearman rank-based correlation score to each of 5 molecular subtypes and assigns a sample with the PAM50 class with the strongest Spearman correlation.

Leukocyte infiltration was estimated using methylation data. Promoter methylation of lineage specific genes remain as one of the most stable molecular marks in cells of different lineages, and was the concept used in this estimation method. Leukocyte and breast specific markers, with 1000 markers for each tissue type, were identified using leukocyte data from GSE35069 [172] and normal reduction mammoplasty samples from our study. Final leukocyte proportions were estimated by identifying the mode of 2000 leukocyte to tumor ratios for each probe.

5.2.9: Gene expression and probe methylation scores by gene voting

Gene expression scores for different molecular phenotypes were estimated using a "gene voting" method, calculated as the mean of transformed, normalized log2 signal intensities. Briefly, the gene-wise Z-scores were estimated across all samples and multiplied using the sign of the moderated t-statistic from limma, or other statistical tests that indicate direction of change. The mean of this value was then calculated as the gene

expression score for a given sample, where a larger value represents a molecular profile closer to the positive contrast from the statistical test. For N differentially methylated probes in n samples;

$$Gene\ expression\ score_j = \frac{\sum_i^N \left(\frac{\beta_{ij} - \hat{\beta}_i}{n} \times sgn(t_i) \right)}{N}$$

Methylation scores for various molecular classes were calculated as transformed mean beta-value. Briefly, the beta-values for each probe was multiplied using the sign of the moderated t-statistic from limma and mean transformed beta-value for each sample was calculated. A larger value represents a molecular phenotype closer to the positive contrast from the limma analysis. For N differentially methylated probes,

$$Methylation\ score_j = \frac{\sum_i^N \left(\beta_{ij} \times sgn(t_i) \right)}{N}$$

5.2.10: Estimating proportion of altered genome from Epicopy-derived CNV data

Segmented data derived using Epicopy was used to estimate the proportion of altered genome (deleted or amplified). An absolute log R ratio (LRR) threshold of 0.15 was used to identify CNV regions across all samples and the fraction of that compared to regions of the genome with probe coverage in Epicopy was calculated as the proportion of genome altered.

5.2.11: Genes associated with recurrence status

Differentially expressed genes associated with recurrence status were identified using a two-step algorithm. First, a Wilcoxon test was performed to pre-rank and identify the top 10% of genes that were differentially expressed between cases and controls. Next, a Cox regression was performed on these genes to identify the genes with best association with recurrence status, controlling for age and radiotherapy described in the following model:

$$Recurrence = \beta + Age + Therapy$$

To identify the best p-value cut off for picking genes, a 10-fold bootstrap analysis was performed for p-value cut offs of 0.001, 0.01, 0.05, and 0.1. In this bootstrap analysis, prediction accuracy for recurrence was assessed using an ROC analysis, with the AUC metric adjusted by dividing it with the number of genes identified with each cut off.

From this analysis, genes with p-value $\leq$ 0.01 were defined as recurrence-associated, and a recurrence score (RS) was calculated for every sample using the gene voting method. An ROC analysis for predicting recurrence was performed using the RS as predictors to assess overall performance of these probes and a Kaplan-Meier analysis was performed separating tumors by median RS into two groups.

Independent external validation was performed using data from GSE31519, where a series of 264 adjuvant chemotherapy-free TNBC samples were curated and normalized from various public datasets. Genes from both DASL and GSE31519 were mapped to EntrezIDs and overlapping genes were used to calculate RS for each sample. Kaplan-

Meier analysis was performed on these samples with two groups identified from the median RS.

5.3: Results and Discussion

5.3.1: Unsupervised clustering identified three stable clusters associated with PAM50 subtypes

Unsupervised clustering on the transcriptomic data identified three stable clusters, with clusters associated with PAM50 subtypes and clinical parameters (Figure 5-2). Interestingly, we observed classification of some of these ER-negative tumors as luminal A (LumA) and luminal B (LumB) subtypes. Recall that this cohort of patients were ER-negative by immunohistochemistry (IHC), and should have few, if any, ER-positive samples. We also observed 6 Her2-enriched (HER2) samples, which was not unexpected since the selection of these patients occurred without Her2-amplification information, as it was not available in most cases. Moreover, differences in methylation and CNV profiles were observed across these clusters (Figure 5-3).

Cluster 1 was enriched for samples with the LumA and HER2 PAM50 subtypes, and has high expression of ER-responsive and HER2-related genes as observed in the expression data (Figure 5-3). It is also the cluster where the patients present with the highest age of diagnosis (Figure 5-2), which is in line with previous observations that ER-positive breast cancer tend to occur in older women [173], suggesting that this subset of tumors recapitulates the presentation of ER-positive IDC. Furthermore, all of the ER-

negative samples classified as LumA were controls, which showed no recurrence, agreeing with previously published findings that the LumA subtype is the least aggressive subtype of breast cancer.

Basal-like and normal-like PAM50 subtypes were distributed across clusters 2 and 3. In Cluster 2, we observed up-regulation of immune-related genes, and relatively unperturbed CNV profiles (Figure 5-3). This is also the cluster with unmethylated CpG islands compared to the other clusters. This collective observation could be explained by the presence of immune infiltrates, which results in over-expression of immune-related genes and dilute out cancer-specific CNV alterations and promoter hypermethylation signals.

Finally, cluster 3 appears heterogeneous for both the expression and methylation datasets. A subset of cluster 3 showed upregulation of differentiation- and basal breast cancer-related genes, but more importantly, most samples in this cluster display genomic instability in the form of high degrees of CNV. Previous studies have shown the existence of a group of breast cancer tumors with complex genomic alterations, which are also enriched in ER-negative disease.

To identify molecular processes that were altered in each of these clusters, we performed sequential *limma* analysis comparing one of the clusters against the remaining two. The moderated t-statistics were then used for a custom GSEA-like approach against Hallmark gene sets to identify functional differences (see methods). A GSEA-like approach was used as it does not require empirically determined cut-offs for differential expression and allow us to use information from all analyzed genes in a setting that recapitulates biological systems where gene expression changes are often concordant but

not always increased in the same magnitude [174]. The results were visualized using the Java implementation of GSEA [175].

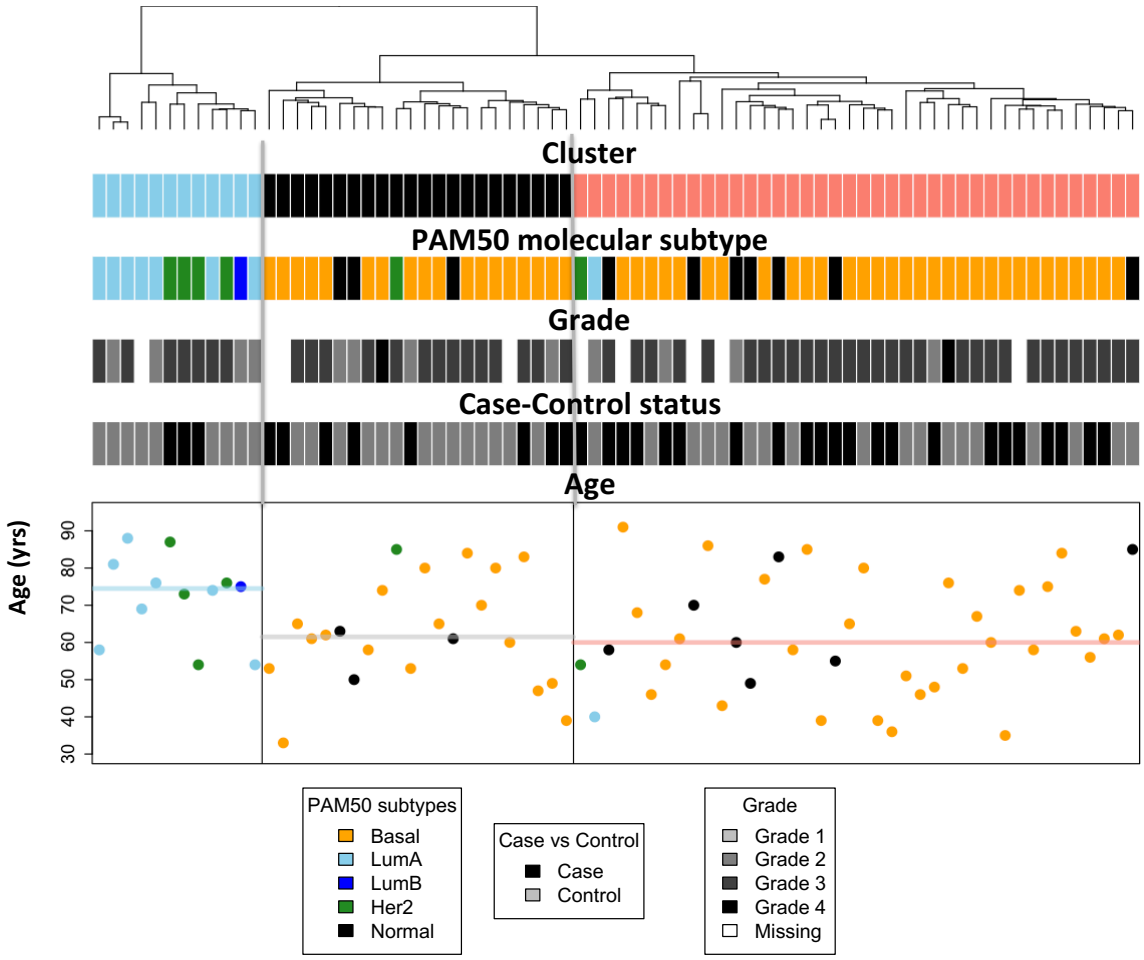


Figure 5–2: Unsupervised clustering analysis on JHU ER-negative cohort identifies 3 stable clusters associated with PAM50 status and clinical features

Cluster 1 (blue) is enriched for LumA and Her2 subtypes and is the cluster with the patients diagnosed at the highest age. Interestingly, all of the LumA subtype samples were controls. Basal and normal-like samples were distributed across clusters 2 and 3. There are no significant differences in grade or case-control status across all 3 clusters.

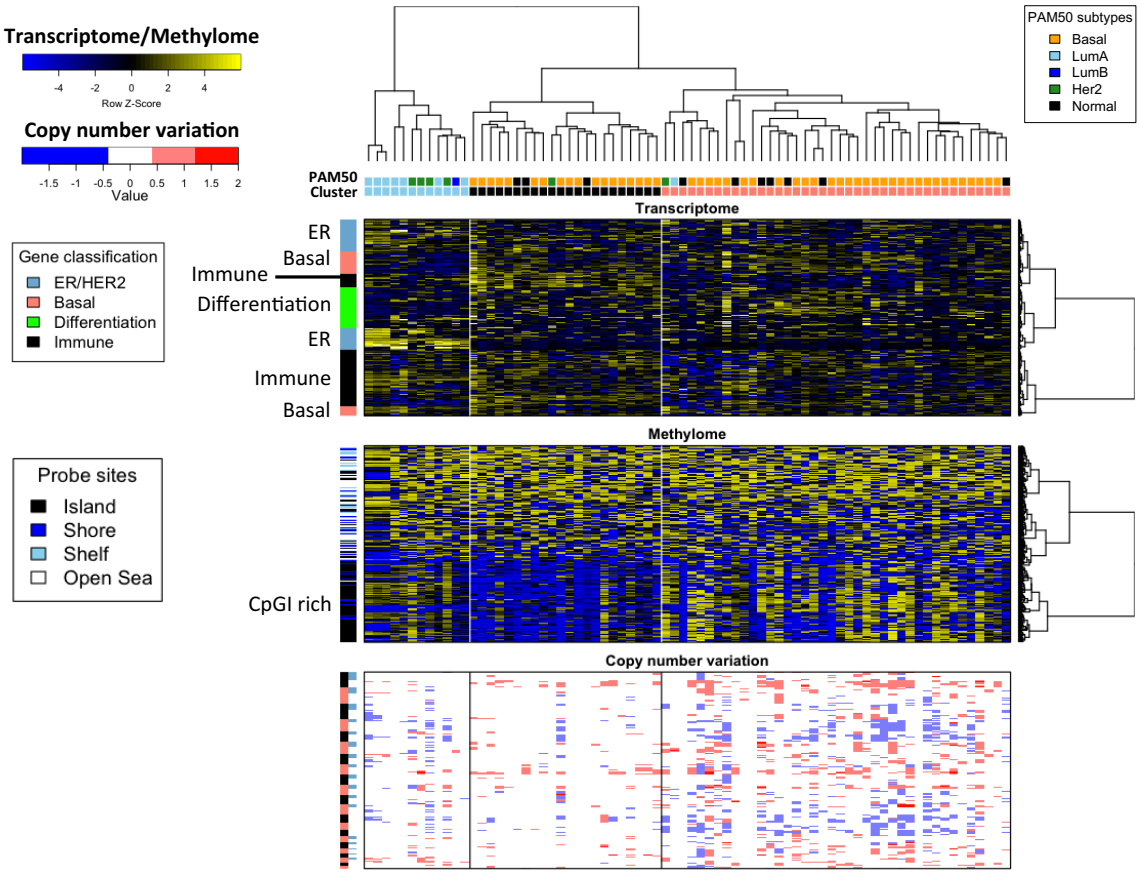


Figure 5–3: Molecular profiles of unsupervised clustering of JHU ER-negative cohort identifies distinct molecular differences across expression, methylation, and copy number platforms

Heatmaps of all 3 molecular genomic datasets reveal distinct alterations across the 3 clusters. Cluster 1 shows high expression of ER-responsive and HER2-related genes, in line with observations in LumA subtypes. Cluster 2 shows increased expression of immune-related genes, and hypomethylated CpG islands compared to the other two clusters. Interestingly, cluster 2 was also the most CN quiet cluster. This can be explained by infiltration of immune cells, which leads to higher immune gene expression and dilution of tumor-specific CNV and methylation changes. Cluster 3 shows heterogeneity in both gene expression and methylation data, with a slightly higher increase in expression of basal and differentiation related genes. The most evident molecular alteration in these samples is the high degree of CNV changes in these samples, suggesting a high degree of genomic instability and recapitulates tumors with complex CNV-profiles identified by previous studies.

5.3.2: Enrichment for hormonal receptor gene sets are observed in cluster 1 and is driven by androgen receptor expression

Gene set enrichment analysis identified three Hallmark gene sets enriched in cluster 1, which are the androgen response, estrogen early response, and estrogen late response gene sets, suggesting that there are hormonal pathways active in these samples (Figure 5-4). When we analyzed the expression of the different hormone receptors in these samples, we observed equivalent, low estrogen (ESR1) and progesterone (PGR) expression across all three clusters. On the other hand, androgen receptor (AR) expression was increased in samples in cluster 1 (Figure 5-4, $p < 0.001$), suggesting that the enrichment of hormonal response pathways was due to the activation of AR. ER-negative breast cancers overexpressing AR have been reported as luminal AR disease, and patients with luminal AR breast cancer have been shown to have good prognosis [169].

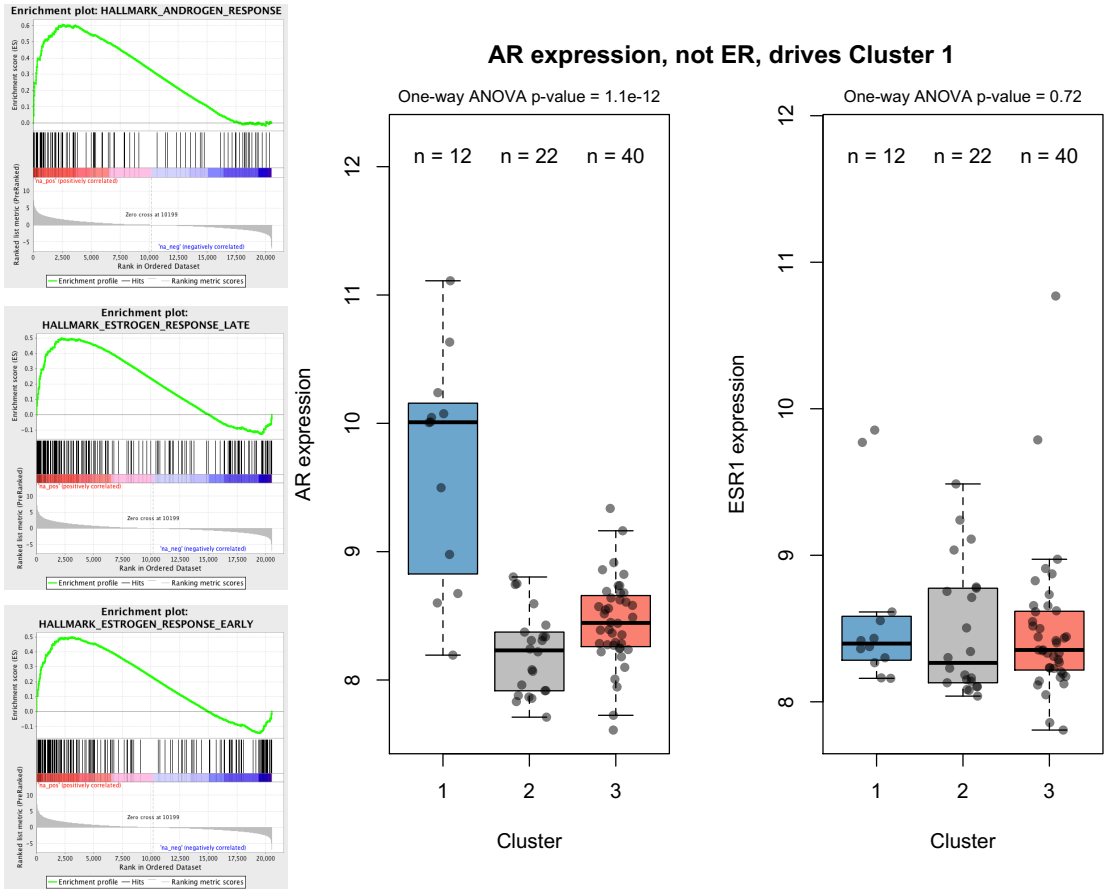


Figure 5–4: Cluster 1 is enriched for hormonal receptor pathways and is driven by androgen receptor (AR) expression

Gene set enrichment analysis shows positive enrichment for 3 Hallmark hormone receptor response gene sets. AR is upregulated in cluster 1, while ER expression was comparably low across all three subtypes. The same was observed for PR (data not shown).

5.3.3: Cluster 2 exhibits immune-related signatures, leukocyte infiltration, and upregulation of immune checkpoint genes

Gene set enrichment analysis revealed strong positive enrichment for immune related gene sets, including interferon alpha and gamma, IL2-STAT5 signaling, and inflammatory response gene sets (Figure 5-5). We characterized the expression for three markers of cytotoxic T-cells and cytolytic activity; CD8A, granzyme (GZMA), and perforin (PRF1). CD8A encodes for the alpha chain of the CD8 receptor, which is expressed on CD8 T-cells, which are key players in the antitumor immune response. GZMA and PRF1 are cytolytic effectors upregulated upon CD8 T-cell activation. All three markers were upregulated in cluster 2 (Figure 5-6a). Furthermore, this observation was validated when we predicted the proportion of leukocyte infiltrates in the tumor using methylation data, and showed that there was a higher degree of leukocyte infiltration within cluster 2 (Figure 5-6b), supporting the data observed in the gene expression dataset.

The presence of these cytotoxic T-cell markers in these tumors suggests that the tumors have developed methods for immune evasion preventing immune-mediated tumor elimination. To assess the possibility of immune evasion, we investigated the expression of LAG3, a T-cell exhaustion marker and saw upregulation of LAG3 within this group of samples. Furthermore, we assessed the expression of two immune evasion genes, PD1 and CTLA4, and observed upregulation of CTLA4, but not PD1, suggesting that the immune evasion pathway used by this subset of tumors was CTLA4-mediated.

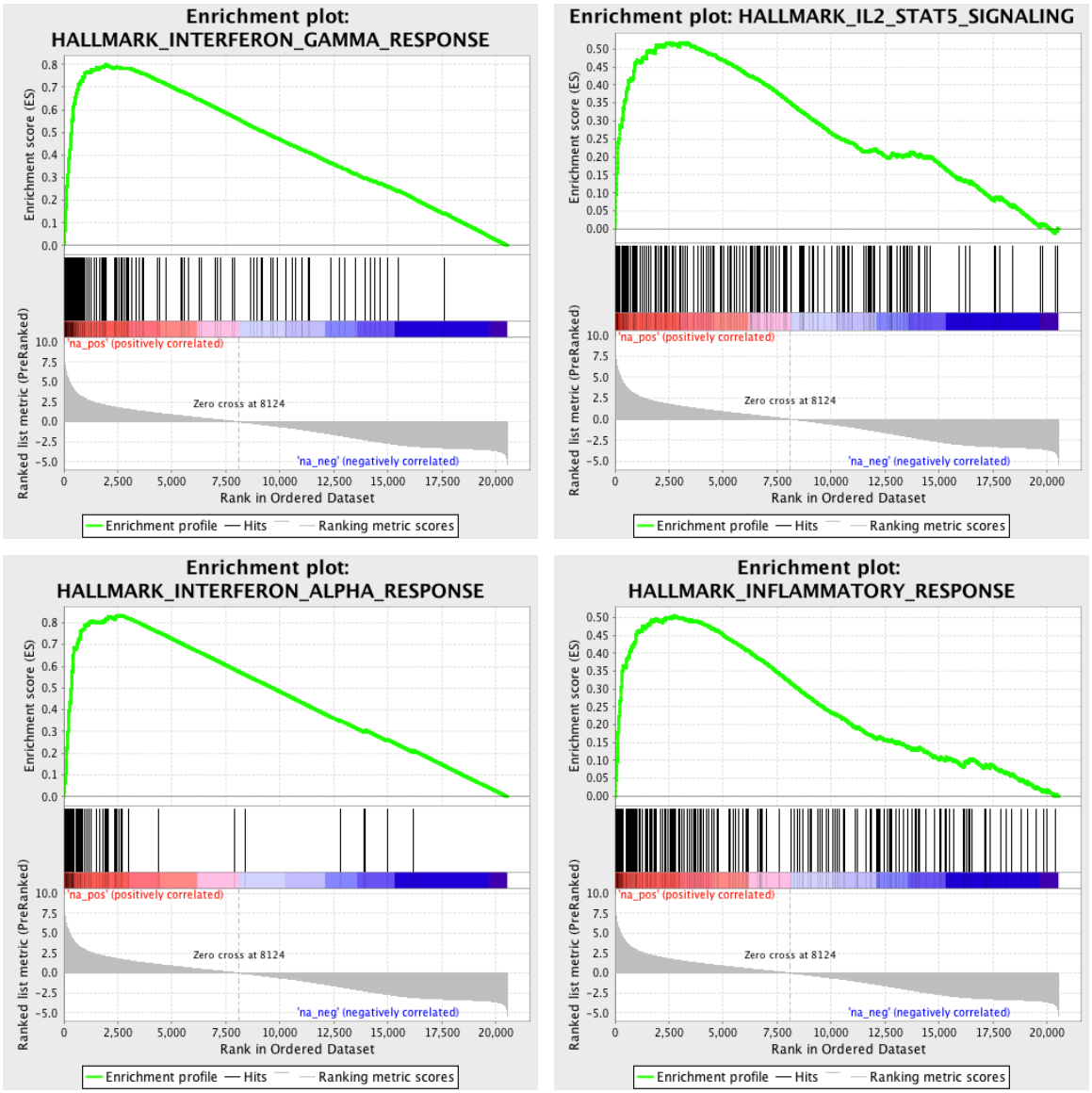


Figure 5–5: Positive enrichment for Hallmark immune gene sets in Cluster 2

Strong positive enrichment for 4/4 Hallmark immune gene sets was observed in cluster 2, suggesting an immune component in these tumors. Of note is the tight up-regulation of genes in the IFN-alpha response, where most genes were enriched for in the top 75th percentile.

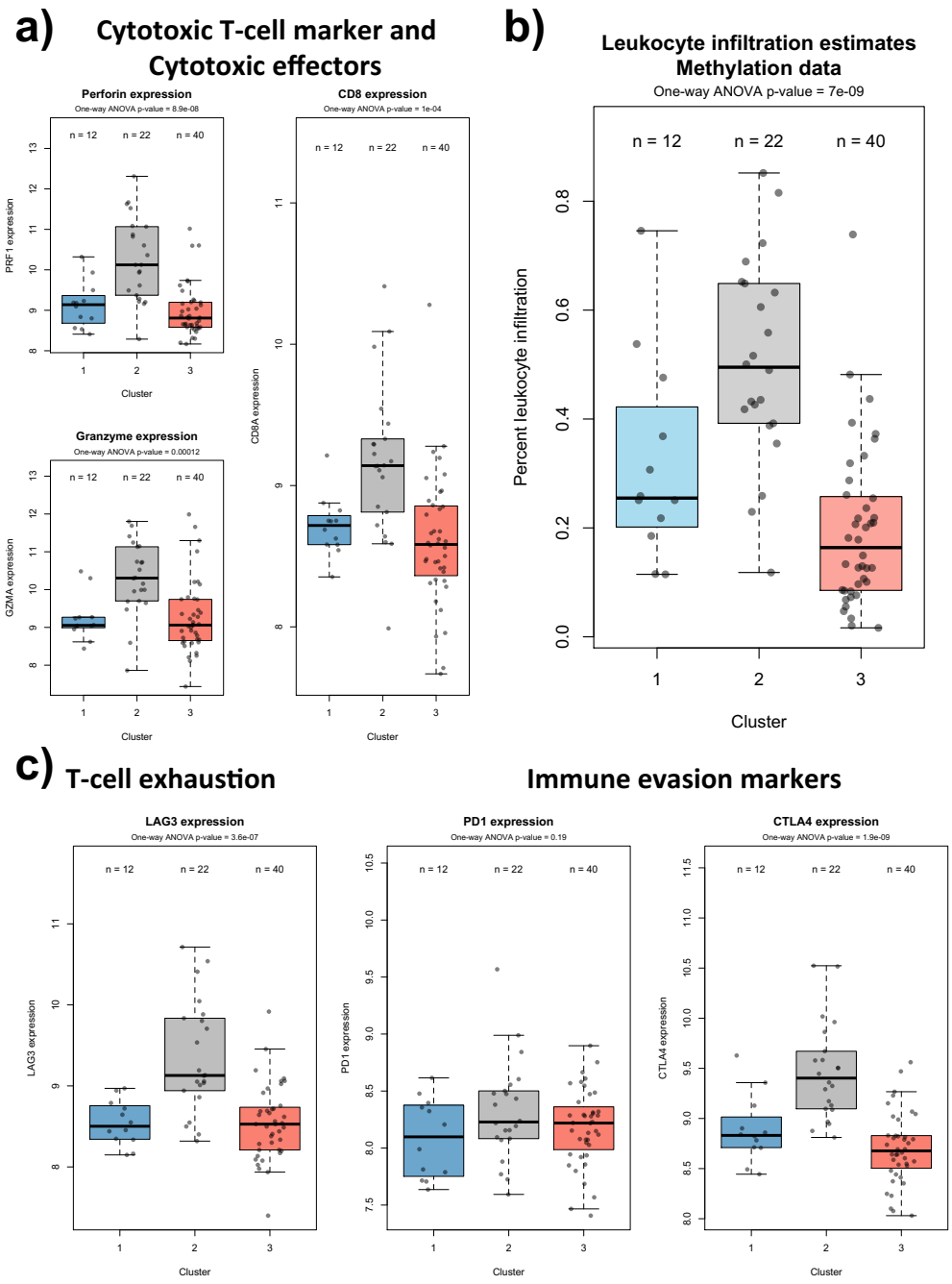


Figure 5–6: Immune markers up-regulated in cluster 2

.a) Up-regulation of cytotoxic T-cell and cytolytic markers in cluster 2. b) Increased leukocyte infiltration predicted from methylation data observed in cluster 2. c) The presence of immune evasion phenotypes in cluster 2 is driven by CTLA4 expression.

5.3.4: Copy number variation high cluster 3 show negative enrichment for DNA repair

Samples in cluster 3 show the largest proportion of genome alteration among all the samples ($p < 0.001$, Figure 5-7a), suggesting that there was a high degree of genomic instability in these tumors. Gene set analysis revealed negative enrichment for DNA repair (Figure 5-7b), suggesting that there may be defects in the DNA repair pathway that manifest in complex CNV profiles.

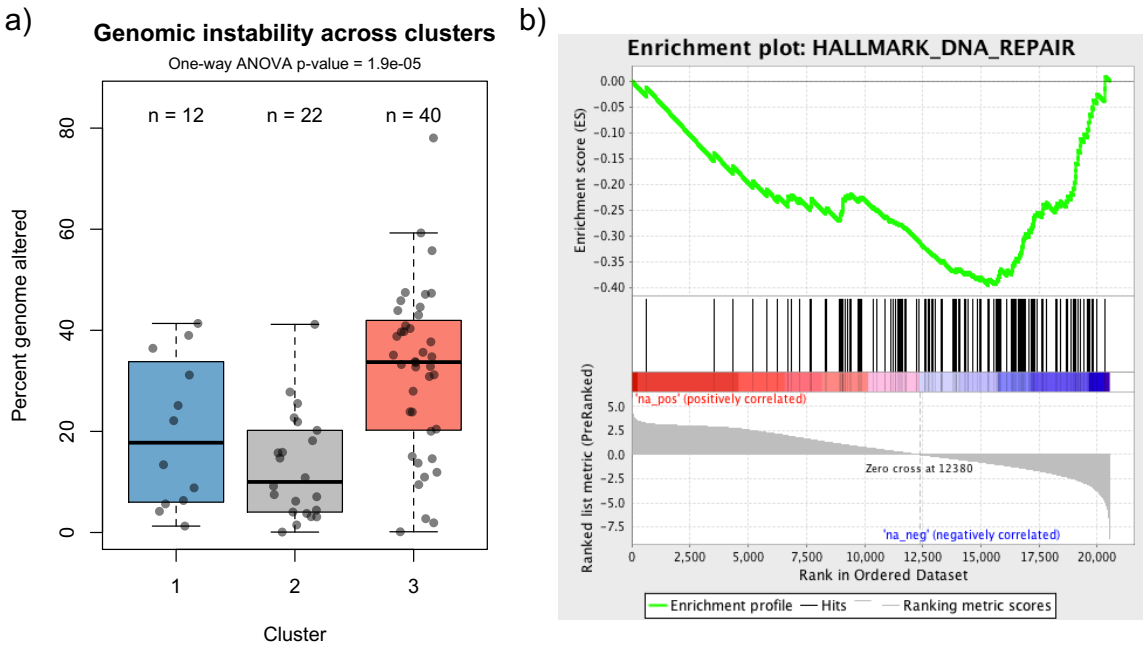


Figure 5–7: High degree of CNV observed in cluster 3 with negative enrichment of DNA repair gene set

a) Percent genome altered, calculated as total base changed divided by total base in the genome measurable by Epicopy, is highest in cluster 3. b) Negative enrichment for Hallmark DNA Repair gene set was observed.

5.3.5: Differences in survival observed across 3 clusters

We next assessed the survival of patients in each of the three clusters by Kaplan-Meier analysis (Figure 5-8). Patients in the AR-driven cluster have the best disease free survival, while patients in the CNV-high cluster had the worst prognosis ($p = 0.11$). This is in line with previous observations that patients with luminal A disease had good outcomes, and holds true as well for a subset of luminal ER-negative tumors driven by AR.

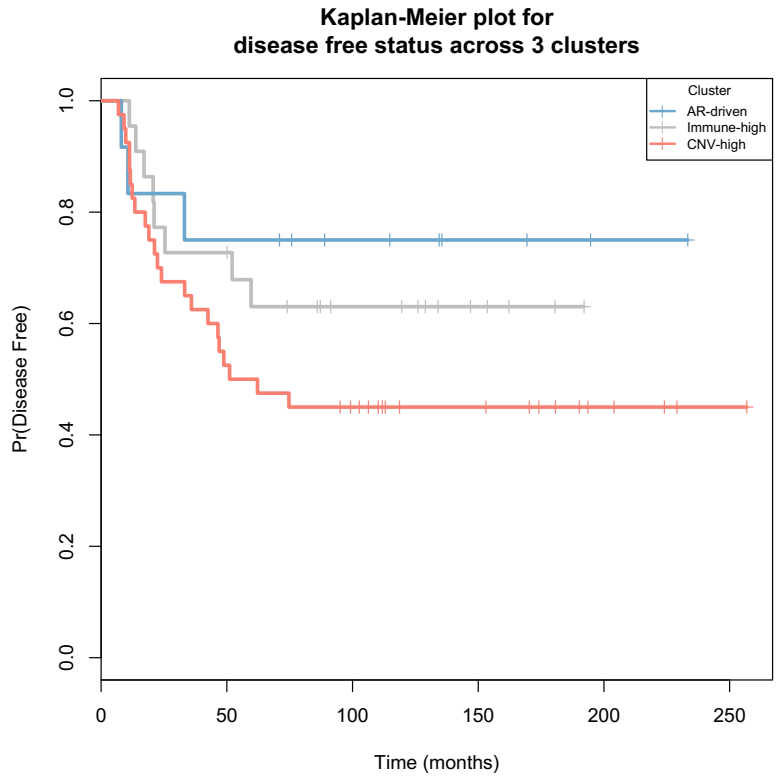


Figure 5–8: Kaplan-Meier analysis for survival across three ER-negative clusters

KM analysis identified that AR-driven disease had the best prognosis while patients with high, complex CNV alterations had the poorest survival.

5.3.6: Identification of transcriptome markers associated with recurrence and independent external validation

We identified 130 genes that are associated with recurrence in the JHU ER-negative cohort and calculated a recurrence score (RS) using a gene voting algorithm (see Methods), where a high RS represents a higher risk of recurrence. These genes collectively predicted recurrence status with an AUC of 0.914 (Figure 5-9a). A Kaplan-Meier analysis separating patients into two groups by median RS revealed that patients with higher RS have poorer disease free survival (DFS) with a hazard ratio of 10.2 (Figure 5-9b). Following this, we calculated RS for TNBCs patients who did not receive adjuvant chemotherapy from an independent external dataset, GSE31519. Bifurcating the patients into two groups by median RS, we observed poorer DFS in the patients with higher RS as well (5-9c), suggesting that some of the genes identified by this analysis are indeed associated with recurrence.

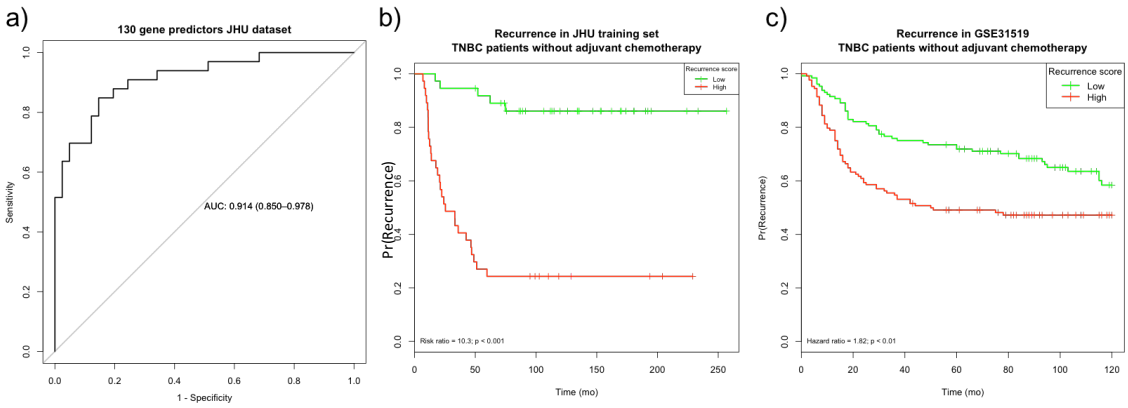


Figure 5–9: Gene expression markers associated with recurrence

a) A 130-gene predictor identified from a Cox regression analysis adjusting for age and radiotherapy was used to calculate recurrence scores (RS). ROC analysis on the ability of RS to predict recurrence status. An AUC of 0.914 was observed. b) KM analysis for DFS in samples in the JHU ER-negative cohort divided by median RS identified better DFS in samples with low RS with a hazard ratio of 10.3 ($p < 0.001$). c) KM plot showing DFS differences between samples with high and low RS in an independent external dataset, GSE31519. Patients used in this analysis were the subset of patients who did not receive adjuvant chemotherapy.

5.4: Conclusion and clinical implications

Herein, we have demonstrated the ability to perform genomic analysis in a series of archival FFPE ER-negative breast cancers. Using PAM50 subtype prediction, we identified a subset of LumA-like tumors that are ER-negative by immunohistochemistry. We defined three stable clusters by unsupervised hierarchical clustering of gene expression data, and functional analyses revealed that these clusters were driven by different biological mechanisms. Furthermore, we identified a series of 130 genes that predict recurrence status and were validated in an external dataset.

The LumA rich cluster was driven by high AR expression and enriched for hormone receptor response gene sets. These luminal-like tumors behave clinically similar to previously identified luminal AR tumors, with relatively good prognosis and higher

patient age at diagnosis. Clinically, we can explore the possibility of treating these patients with androgen-targeted therapies such as the use of anti-androgens and androgen synthesis inhibitors.

A series of samples with high levels of immune infiltrates were observed, with molecular profiles suggestive of an immune evasion phenotype driven by upregulation of the immune checkpoint gene CTLA4. CTLA4 inhibitors have been developed and are well tolerated [176] in patients. The use of these inhibitors may be considered in the treatment of this cohort of patients.

Finally, we identified a cluster of samples with heterogeneous gene expression and DNA methylation profiles, which were characterized by a large degree of genetic change manifesting as CNV events. Gene set enrichment analysis suggests DNA repair defects in these tumors, and may be optimally targeted by chemotherapy or PARP1 inhibitors.

Looking forward, we have defined three groups of ER-negative breast cancer tumors, with functionally relevant pathways and potential therapeutic targets, some which are better tolerated than chemotherapy. Further refinement of the gene set associated with recurrence across these three groups may allow us to identify features to not only optimally decide on therapeutic options, but also to decide on patients who will not require additional therapy.

Chapter 6: FFPE RNA-seq analysis of follicular thyroid cancer reveals transcriptomic landscape and identifies markers of metastasis

6.1: Introduction

Thyroid cancer statistics show an alarming increase in incidence, almost quadrupling from 3.6 per 100,000 in 1973 to 15 per 100,000 in 2012, as illustrated by the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program [54]. Mortality from thyroid cancer, however, remains unchanged, with over 95% of patients experiencing excellent outcomes. This phenomenon is likely to reflect a trend of "cancer over diagnosis", where a silent disease reservoir that would have remained undetected and asymptomatic is detected during routine clinical assessment driven by increased access to healthcare screening, as exemplified by screening for prostate cancer in the elderly [177, 178]. Indeed, thyroid cancer is one of the cancers where routine clinical screening has increased substantially due to improvements in detection methods, including ultrasonography and fine-needle aspiration biopsy (FNAB) [179]. There are, however, issues complicating the optimal management of thyroid cancer [180, 181], including diagnostic ambiguity in FNAB diagnoses and inadequate prognostic markers, the latter being the focus of this study.

Follicular carcinomas (FC) represent 10% - 15% of all newly diagnosed thyroid cancers, and over 90% of FCs are indolent and have a probability of metastasis under 5%. The subset of FCs that does metastasize, however, tends to have a more aggressive course of disease than their papillary counterparts, given their tendency to metastasize through the hematogenous route leading directly to distant organ metastasis [182]. Despite having

a low metastatic rate, the lack of robust risk stratification has resulted in thyroidectomy followed by radioiodine remaining the mainstay of treatment for most screening-detected FCs [180, 181]. The lack of appropriate prognostication leads to overtreatment of FCs, which exposes patients to complications from total thyroidectomy and requires lifelong thyroid hormone supplementation [177]. While a significant body of work has been done on trying to identify markers that stratify risk in follicular thyroid tumors, most of the work so far has been focused on distinguishing follicular adenomas (FAs) from FCs. The former is believed to be a benign, encapsulated thyroid nodule which does not pose any risk of metastasis, although it is generally thought of as precursor to thyroid cancer, since, in contrast to hyperplastic nodules, these are clonal lesions.

FAs are indistinguishable from FCs based on cytologic, sonographic, or clinical features alone, and the pathognomonic feature of FCs is that they show capsular and/or vascular invasion. FAs and minimally invasive FCs with only microscopic penetration of the tumor capsule without vascular invasion can be treated through simple resection of the tumor or hemithyroidectomy, which usually allows the remnant thyroid to maintain a euthyroid state without the need for lifelong hormone substitution.

Markers distinguishing FCs from FAs include genetic alterations and mRNA expression changes. RAS mutations are prevalent in FCs, at a rate of 40-50%, and have been explored as a potential molecular marker for prognosis. Unfortunately, RAS mutations are also present in 20-40% of FAs, leading to poor positive predictive value (PPV) for this molecular marker [182]. Similarly, PAX8-PPARG fusion occurs in 30-40% of FCs and about 10% in FAs [182, 183]. Studies have explored using mRNA

expression as detection markers for FC compared to FAs, using transcripts such as PCSK, CCND2, and LGALS3, and others [184].

In contrast, no study has investigated molecular risk factors for metastasis in FCs to date. Due to the low rate of metastasis in FCs, merely distinguishing FCs from FAs will continue to result in the overtreatment of large numbers of indolent FCs [185]. There is a pressing need for a diagnostic test with high negative predictive value for metastatic disease, available preoperatively, which will have a large impact in controlling the current rates of total thyroidectomy and adjuvant radiation for a usually indolent disease.

Due to the rarity of metastatic FC, the most realistic manner of identifying metastatic markers in FCs is to retrospectively identify cohorts of FCs with available long-term follow-up information. Most archival samples are stored as formalin-fixed, paraffin-embedded (FFPE) tissue blocks, a process that significantly degrades nucleic acids. While there have been some success [186, 187] in performing RNA-seq on FFPE materials, to our knowledge this has not be done in FCs, or thyroid samples, to date.

Herein, we performed a proof-of-principle study examining the feasibility of performing RNA-seq on 4 metastatic and 4 non-metastatic primary FC FFPE samples, and assessed the ability to detect differential expression and perform splice variant analysis on such material, and further determined the molecular profiles of these 8 FC samples as compared to PTCs studied by The Cancer Genome Atlas (TCGA)-thyroid consortium.

6.2: Methods

6.2.1: Patient sample collection

A retrospective pilot study was conducted under Institutional Review Board approval using archival FFPE samples of patients diagnosed with FC and treated at Johns Hopkins Hospital. This study was designed to identify and compare transcriptomic profiles of primary FC lesions of four patients presenting with stage IV metastatic FC and four patients presenting with FC confined to the thyroid and who had no metastatic disease at 5 years of follow-up with comparable clinical variables, including age and treatment regimen (Table 10).

6.2.2: RNA Extraction and Quality Assessment

Unstained histological slides were macro-dissected to enrich for tumor cells (>75%) using a consecutive H&E section annotated by the study pathologist as reference. RNA was extracted from the samples and DNase treated using the Maxwell(r) 16 LEV RNA FFPE Purification Kit (Promega, Madison WI) following the manufacturers protocol. The resulting RNA was analyzed for UV absorbance wavelength ratios (Nanodrop; 260/230, 260/280) to determine purity and concentration. The amount of RNA was normalized to the DV200 value obtained from the Agilent RNA Tapestation, representing the fraction of RNA >200bp in that sample. Where necessary, samples were

concentrated using sodium acetate/ethanol precipitation to have a DV200-normalized input of 1ug RNA in 10uL.

RNA fragment distribution was analyzed by the Tapestation and found to be highly degraded, as was expected for FFPE samples, eliminating the need for fragmentation before library preparation.

6.2.3: Library preparation

Ribosomal RNA (rRNA) depletion was performed using the Ribozero Gold rRNA Depletion Kit (Illumina, San Diego). TruSeq Stranded Total RNA Library Prep Kit (Illumina, San Diego) was used for library preparation following manufacturer's protocols, and performed using an Agilent Bravo A automated workstation (Agilent, Wilmington DE). Final libraries were analyzed by Tapestation to determine average fragment size. A normalized pool of all 8 samples was sequenced on Illumina MiSeq sequencer as a final QC measure.

6.2.4: RNA-sequencing and data processing

Sequencing was performed using Illumina HiSeq 2500 at 2 samples per lane using an 8-lane flowcell to produce approximately 150 million paired-ended sequencing reads of 48 base pairs per sample. Reads were aligned to the Human Genome Reference Consortium build 38 (GRCh38) using Tophat2 [188] and assembled into transcripts using CLASS2 [189]. CLASS2 implements a rigorous statistical model to recognize and filter

intronic ‘noise’ due to reads from unspliced RNA, which is abundant in ribosomal RNA depleted libraries, and therefore is particularly well suited to the analysis of FFPE samples. The collection of transcripts merged across the samples was used to identify differentially expressed genes between metastatic and non-metastatic primary tumors with Cuffdiff2 [190], and to determine differential alternative splicing events (exon skipping, mutually exclusive exons, alternative exon ends) with rMATS [191]. Results were inspected and visually validated using the Integrative Genomics Viewer (IGV) [192].

6.2.5: Data analysis

Gene set analyses were performed using the wilcoxGST function from the limma Bioconductor package [193] using gene set collections obtained from Molecular Signatures Database (MSigDB) [194, 195]. The gene universe was restricted to intersecting mapped genes and genes present in gene set collections. Additional analyses were performed using custom functions as indicated.

Metastasis scores for TCGA primary thyroid tumors were derived using 140 genes that were differentially expressed between metastatic and non-metastatic primary FCs in our study. First, we performed gene-wise scaling of log2 RSEM values into Z-scores within each cohort. This transforms weights of relative differences in expression between genes into unweighted terms.

$\beta'_{gi} = \frac{\beta_{gi} - \overline{\beta_g}}{\sigma_g}$, where β_{gi} is the log2 RSEM of gene g of sample i

Next, we multiplied the Z-scores of a given gene with the sign (+1/-1) of the CuffDiff2 test-statistic of the same gene to obtain positive relationships between metastasis and expression in each of the genes.

Let N be the number of genes in the marker set and n be the number of samples, and B'' be the matrix after test-statistic correction,

$$B'' = \begin{pmatrix} \beta'_{gi} & \cdots & \beta'_{gn} \\ \cdots & \cdots & \cdots \\ \beta'_{Ni} & \cdots & \beta'_{gi} \end{pmatrix} \times (\alpha_1 \quad \cdots \quad \alpha_g)$$, where α is the Cuffdiff2 test statistic of gene g

Finally, the mean of all the genes for each sample were calculated as the metastasis scores M.

$$M = \frac{\sum_{g=1}^{N} \beta_{gi}}{N}$$

Receiver operating characteristic (ROC) analyses were performed using the pROC package [89] in R Statistical Environment with confidence intervals calculated using the Delong method [196]. In both the FVPTC and PTC TCGA cohorts, the responses for the ROC analysis were distant metastasis and the predictors were the metastatic scores.

6.3.6: Comparison with TCGA data

The Picard tools [197] workflow was used to obtain RSEM data from RNA-seq data of our FC cohort mapped against RefSeq release 74 for comparison with the TCGA THCA dataset. Processed TCGA RNA-seq data was downloaded from the Firehose GDAC hosted by the Broad Institute [70].

Follicular variant papillary thyroid cancer (FVPTC) specific genes were identified using Boruta [198] with a p-value cutoff of 0.05 and maxRuns of 1,000 to minimize tentative genes, and assessing 10,000 trees against FVPTC and non-FVPTC sample classifications. Spearman distances were calculated using 1 – Spearman correlation of FVPTC-specific genes between this study's FC cohort and individual samples from TCGA.

6.2.7: Mutational analysis

Samtools/bcftools [199] and GATK [200] were used to make variant calls in a subset of commonly genetically altered genes in thyroid cancer including BRAF, HRAS, NRAS, KRAS, and EIF1AX. Mutational calls were inspected and manually validated using IGV.

6.2.8: Pyrosequencing

Pyrosequencing for the RAS genes was performed as described previously [201, 202]. The limit of detection for pyrosequencing is approximately 5% mutant alleles (or 10% tumor cells in a tissue sample). A signal of 4% to 5% mutant alleles is considered indeterminate and a signal of 3% mutant alleles or less is reported as a negative result if corrected for the histologically estimated tumor cell percentage in the sample. An indeterminate result (4%–5%) or a positive result with low mutant allele level (6%– 10%) triggered a review of the H&E slide and reevaluation of the estimated tumor cell

percentage. Specimens with no mutation detected and with less than 70% tumor cellularity were also reevaluated and reported as tumor cellularity below the limit of detection of the assay. The interpretation of complex pyrogram patterns due to the mutation of 2 or more nucleotides on the same allele was resolved by the software program Pyromaker (http://pyromaker.pathology.jhmi.edu; accessed March 21, 2014) [203].

6.3: Results

6.3.1: RNA-sequencing of FFPE tissue samples is a viable method for whole transcriptome analysis of FCs.

In this pilot experiment on 8 FC FFPE tumors, 7 samples yielded satisfactory RNA-seq results, with a median of 136 million reads of which 87.5% (range 31.2% to 91.35%) uniquely-mapped to the genome (Figure 6-1a). The single failed sample, M02, yielded only 104 million reads, of which 4% mapped to the genome, and was removed from further analysis. For the remaining samples, 13% to 20% of the genomic alignments mapped to exonic regions in the genome (Figure 6-1b). Finally, our use of an rRNA depletion method, which was made necessary by the inability to use poly-A selection of transcripts due to FFPE-induced fragmentation of nucleic acids (Ribo-Zero Gold rRNA Removal Kit, Illumina, San Diego) did not result in any transcript position biases (Figure 6-1c). We determined that the sample with the lowest number of mapped reads, M01, was suitable for further analysis based on a principal component analysis (PCA) of a

subset of the 500 most variable genes expressed across all samples, which showed that it did not behave as an outlier (Figure 6-1d).

6.3.2: Initial CuffDiff2 analysis reveals differentially expressed genes and identified a sample with late metastasis

Analysis of the 3 metastatic FC samples and 4 non-metastatic FC samples, revealed 93 differentially expressed loci, mapping to 86 genes (Supplementary Table 1). Two clusters were detected in a hierarchical clustering analysis of these 7 samples using the 85 differentially expressed genes, separating the metastatic samples from their non-metastatic counterparts (Figure 6-2). Interestingly, one of the non-metastatic samples, I04, showed a gene expression profile intermediate between the metastatic tumors and non-metastatic tumors, residing within the metastatic cluster. After updating available clinical follow-up information 2 years after the initial case selection, this sample was revealed to be from a patient who eventually developed a distant metastatic event at 10.5 years, much later compared to the other 3 metastatic cases that had presented with stage IV disease, suggesting that some of molecular changes associated with metastasis can occur early in the evolution of FC's.

6.3.3: Differential gene expression analysis on reclassified sample phenotype identifies 140 differentially expressed genes

Sample I04, the sample from a patient with late metastasis, was then reclassified as a late metastatic primary tumor. Differential expression analysis using this new classification revealed 161 statistically significant (FDR < 0.05), differentially expressed loci, mapping to 140 genes (Figure 6-3, Table 11), including 114 genes that overlapped with the original analysis.

6.3.4: Gene set enrichment analysis reveals enrichment for epithelial-mesenchymal transition (EMT) and oncogenic pathways

Gene set enrichment analysis performed using a competitive, mean-rank gene set enrichment test and the "hallmark" gene set curated by Molecular Signatures Database (MSigDB), revealed enrichment with an FDR < 0.05 for gene sets related to epithelial mesenchymal transition (EMT), steroid hormones, p53 signaling, KRAS signaling, and hypoxia (Table 12, Figure 6-4).

6.3.5: Genes significantly differentially expressed between metastatic and non-metastatic primary tumors show similar trends in TCGA thyroid cancer dataset

Significantly differentially-expressed thyroid-relevant genes identified from our dataset as either previously described to play a functional role in thyroid cells or having

been used as candidate prognostic markers in previous studies, were compared across tumor classes in the TCGA THCA dataset. PCSK2 and MFGE8 had increased expression in metastatic vs non-metastatic FC in our cohort, and showed increased expression in thyroid cancer compared to normal thyroid tissue in the TCGA data. Conversely, we observed down-regulation of DIO1 and CHI3L1 in our dataset, which was also seen in TCGA comparing thyroid cancer to normal tissue (Figure 6-5a). Furthermore, of the 140 differentially expressed genes identified by this study, 100 had concordant trends in TCGA data (Figure 6-5b).

6.3.6: FCs are molecularly more similar to FVPTCs than classical PTCs

We identified a set of genes that best separate FVPTCs from non-FVPTCs in the TCGA dataset using Boruta, a random-forest based machine-learning model (Figure 6-6). Using the selected genes, Spearman distances were calculated between JHU FC samples and TCGA FVPTC and PTC samples. The distances were summarized as the median distance between each FC and TCGA FVPTC or PTC samples grouped by follicular fraction. All 7 FC cases were significantly more likely to be closer in distance to FVPTCs than PTCs, with a stepwise increase in distance between PTCs of high follicular content to classical PTCs ($p < 0.0001$, Figure 6-7a).

6.3.7: FC Metastasis markers identify metastatic FVPTCs but not metastatic PTCs

Metastasis scores for TCGA thyroid cancer samples were calculated as a mean Z-score-normalized expression of 140 DE genes distinguishing metastatic from non-metastatic primary FCs (see Methods). ROC analyses were performed for FVPTC and PTC subsets of the TCGA data using distant metastasis as the response and metastasis scores as predictors (Figure 6-7b to 6-7e). In the FVPTCs, we observed a statistically significant AUC of 0.946 [95% confidence interval: 0.893, 1.00] in the metastasis scores predicting distant metastasis (Figure 6-7b) while the scores did not predict distant metastasis in classic PTCs 0.562 [0.303, 0.82]. Furthermore, these metastasis scores did not predict lymph node (LN) metastasis in FVPTCs (Figure 6-7d), consistent with a signature specific to hematogenous spread, and, while significant, do not prove to be a clinically useful set of markers within the PTC LN cohort (Figure 6-7e). This could be due to the functional differences required for lymphatic metastasis and differences in molecular progression between PTCs and FTCs.

6.3.8: Splice variant analysis using rMATs identifies differentially skipped exon events in genes relevant to thyroid cancer

Splice variant analysis was performed using rMATs for skipped exon (SE) and mutually exclusive exon (MXE) events, starting from the combined gene and transcript sets assembled from the RNA-seq samples. While there were no differential MXE events,

the analysis identified 7 differentially expressed SE events (Table 13), in the following genes: NPC2, TG, MACF1, ACSL3, ARID1B, UTRN (Figure 6-8a), and RMST.

6.3.9: Identification of RAS and EIF1AX mutations

Hotspot mutations in the BRAF and RAS genes, as well as genes that were identified as mutated in FVPTCs from TCGA's thyroid cancer analysis, were assessed using MutSig and Integrated Genomics Viewer (IGV) with the queried locations summarized in Supplementary Table 1. Of the 7 samples, 3 of the 4 metastatic cases were found to have RAS mutations, with 2 in KRAS and 1 in NRAS. M03 had an NRAS-Q61R mutation, M04 had a KRAS-Q61R mutation, and LM01, the case originally classified as non-metastatic, had a KRAS-G12S mutation (Figure 6-8b, Table 14). Furthermore, two indolent samples showed splice junction mutations in EIF1AX at A113, a mutation previously found in FVPTCs by TCGA (Figure 6-8b, Table 14).

6.3.10: Validation of RAS mutations by pyrosequencing

We performed pyrosequencing to validate the mutations identified from our RNA-seq analysis (Table 5). One sample with NRAS Q61R mutations were validated in our pyrosequencing analysis. Consistent with our RNA-seq data, none of the samples showed HRAS mutations.

6.4: Discussion

We performed RNA-seq on FFPE material derived from 8 primary FCs, 4 of which were non-metastatic and 4 were metastatic at the time of diagnosis. To our knowledge, this is the first study assessing the use of RNA-seq to study the transcriptome of thyroid FFPE samples. Our results, with a success rate of 7/8 samples, indicate that RNA-seq is a feasible platform to analyze thyroid FFPE samples. Furthermore, we observed no transcript position bias for library preparation from the rRNA depletion protocol used.

A principal component analysis (PCA) revealed separation based on metastasis status on the first component and further suggested an intermediate sample between non-metastatic and metastatic FCs. Hierarchical cluster analysis and PCA using genes identified from an initial CuffDiff2 analysis showed that one of the tumors classified as non-metastatic clustered with the metastatic tumors. After updating all available clinical follow-up information 2 years after initial case selection, that sample was found to be from a patient who developed distant metastatic disease at 10.5 years post-treatment. We reclassified that sample as a metastatic sample, and reanalyzed the data. While we cannot exclude the possibility that additional cases classified as indolent could show similar late disease progression (current follow periods range from 66-204 months), this is unlikely given the very low rate of such late events in initially indolent FC. Following this second analysis, we identified 140 differentially expressed genes with 114 overlapping the first analysis.

Genes over-expressed in metastatic samples include SLC34A2, LRP4, PBX3, STK32A, PCSK2, and MFGE8. SLC34A2 is a sodium/phosphate co-transporter, which has been shown to be up-regulated in PTCs compared to normal thyroid tissue [204-206]

and is also the target of a preclinical antibody-drug conjugate (ADC) for ovarian, lung, and thyroid tumors [207]. LRP4 up-regulation in thyroid cancer has been shown in multiple studies, the both classic PTCs [208-213] as well as FVPTCs compared to FAs [214]. Likewise, STK32A up-regulation has been observed in PTCs [213]. The homeobox transcription factor, PBX3, is a proto-oncogene [215] that has been shown to be a target of metastasis-suppressor microRNA let-7c and plays a role in promoting cell proliferation and metastasis in colon cancer [216, 217]. PCSK2 was used in a three-gene model to distinguish FCs from FAs [218], and found to be over-expressed in FCs by microarray and qRT-PCR methods [219]. MFGE8 is a pro-angiogenic factor in several tumors [220-222], and has been shown to be up-regulated in thyroid cancer [208, 210, 223, 224].

Genes down-regulated in metastatic samples include DIO1, TFCP2L1, SLC5A8, PPARG, and RHOB. DIO1 is a Type I Thyroxin Deiodinase that activates thyroid hormone by converting the prohormone thyroxin (T4) to the bioactive 3,3',5-triiodothyronine (T3) [225]. This is consistent with previous studies that showed DIO1 to be under-expressed across all subtypes of thyroid carcinomas compared to normal tissue [226-229]. Kim et al. observed [205] TFCP2L1 down-regulation in PTCs compared to normal tissue by microarray and QPCR. Furthermore, TFCP2L1 is down-regulated in highly malignant anaplastic thyroid cancers (ATCs) compared to benign goiters and contributes to silencing of CRYAB, a potential tumor suppressor [230]. Finally, TFCP2L1 was found to be down-regulated in FVPTCs compared to FAs in a microarray analysis [214]. SLC5A8 is a tumor suppressor that is commonly hypermethylated and down-regulated in cancer [231-234], including PTCs [235]. PPARG is also down-

regulated in metastatic FCs compared to indolent FCs. The PAX8/PPARG fusion may be present in about a third of FCs, although not detected in this small series, and acts as a dominant negative inhibitor of wild-type PPARG [236, 237]. Beyond PPARG/PAX8 fusions, PPARG down-regulation has been observed in various cancers and is associated with poorly differentiated tumors [238], including FCs [239, 240]. RHOB is a small GTPase with a tumor suppressor role [241], and re-expression of RHOB have been shown to be a requirement in cell-cycle arrest through its activity in up-regulating p21 [242, 243]. Furthermore, RHOB had been shown to play an anti-metastatic role in cancer and is inactivated by the RAS pathway [243, 244], a key pathway activated in FCs.

Our gene set analysis revealed that, remarkably, EMT (epithelial-to-mesenchymal transition) was found to be the most significantly enriched gene set using a mean-rank gene set test on differential expression statistics comparing metastatic and non-metastatic primary FCs on the MSigDB hallmark gene set (Table 3, Figure 6-4). EMT is a developmental process during which epithelial cells acquire mesenchymal traits that allow them to migrate, invade, and disseminate. EMT is co-opted by cancer cells to initiate invasion and metastasis [245]. Among the top 5% of genes differentially expressed between metastatic and non-metastatic primary tumors, six of those were a subset of the EMT gene set; FMOD, GJA1, RHOB, SGCD, DPYSL3, and IGFBP3 (Figure 6-4). Among the down-regulated genes, GJA1 [246], FMOD [247], and RHOB [244] have been implicated to play a role in preventing metastasis or to be down-regulated in metastatic solid tumors, while SGCD is a component of the sarcoglycan complex, which acts as a mechanosignalling connection from the cytoskeleton to the extracellular matrix and has been observed to be downregulated in breast and prostate

cancer [248, 249]. DPYSL3 and IGFBP3 appear to have a context-dependent dual inhibitory and stimulatory role in cancer. DPYSL3, or CRMP4, is a cell adhesion molecule that has been shown to promote metastasis in pancreatic and gastric cancers, but is a metastasis suppressor in breast. We observed a stepwise increase in DPYSL3 in more malignant phenotypes in TCGA data (Figure 6-5), including metastasis, suggesting that DPYSL3 contributes to a metastatic phenotype in thyroid cancer [250-252]. Interestingly, VEGF promotes the upregulation of DPYSL3 in gastric cancer [253], and FCs have been known to metastasize through the hematologic route [182], a process tied to neoangiogenesis [254]. While IGFBP3 is correlated with good prognosis in gastric cancer, it promotes transendothelial migration in oral squamous cell carcinoma [255] and TGF-beta-mediated EMT in esophageal cancer [256]. However, similar to DPYSL3, it showed a stepwise increase in expression in more malignant thyroid cancer phenotype in TCGA data (Figure 6-5b), which may suggest its role in promoting an aggressive phenotype in thyroid cancer.

We also identified concurrent enrichment for the hallmark hypoxia gene set, a pathway shown in an FC cell line model to induce EMT [257]. The KRAS signaling pathway was also implicated. RAS is the most commonly mutated oncogene family in FCs, with the most commonly mutated member being NRAS, followed by HRAS and KRAS. Unfortunately, KRAS is the only gene in the RAS family that was featured in the KRAS signaling gene set collection, making it difficult to identify the specific RAS gene driving this observation. Enrichment for the p53 signaling pathway was also observed, and p53 mutation plays a key role in anaplastic thyroid cancers, a more aggressive, poorly-differentiated subtype of epithelial thyroid cancer [258, 259].

While estrogen- and androgen-related factors have been implicated in thyroid cancer [260], and such pathways were enriched in this analysis, this might be contributed by the increased in proportion of male-to-female patients in our dataset. We are unable to distinguish this confounding factor given the small number of samples in this pilot experiment.

As an external validation of our metastasis markers, we investigated the expression of the DE genes in the TCGA thyroid cancer RNA-seq dataset and were able to detect changes between tumor versus normal and, in some cases, metastatic versus non-metastatic primary tumors that are consistent with our findings (Figure 6-5, Supplementary Table 1). To be exact, 100/140 mapped and validated genes from our analysis show concordant, statistically significant differences between aggressive and non-aggressive tumors. Interestingly, we observe a stepwise change in gene expression from normal thyroid tissue to FVPTC to PTC for many of the genes (see Figure 6-5, Supplementary Table 1). This may suggest that there exists a parallel biological process in metastatic disease that mimics differences between FVPTCs and PTCs.

By definition, FVPTCs consist to 99% of tumor tissue showing follicular-like morphology, similar to that of FCs, and we hypothesize that subsets of FVPTCs, such as the majority that lacks BRAF mutations, are molecularly more similar to FCs than their classical PTC counterparts. To our knowledge, there are currently no genome-wide molecular data exploring the relationship between FVPTCs in relation to PTCs and FCs. To minimize confounding by batch effect, we employed a non-parametric method of calculating Spearman distances between our 7 FC samples and all the TCGA thyroid FVPTC and PTC tumors using 92 FVPTC-specific genes identified using Boruta (Table

16). FCs are molecularly more similar to FVPTCs than PTCs in the context of FVPTC-specific genes (Figure 6-7). Furthermore, FCs are also molecularly closer to PTCs with high morphologically follicular fractions (but below the 99% threshold for FVPTC diagnosis), compared to PTCs with low follicular fraction. Metastasis scores, calculated as a mean Z-score-normalized expression value, were used as predictors in an ROC analysis comparing primary tumors with or without distant metastasis in the TCGA FVPTC and PTC cohorts respectively. The metastasis scores were prognosticative for FVPTCs but not PTCs (Figure 6-7, b to e) by ROC analysis.

Splice variant analysis using rMATS identified 7 genes with differential SE events, including the thyroid relevant gene thyroglobulin (TG). Other genes of interest include UTRN and MACF1. In this pilot study, exon 66 (ENSE00001084869.1) of the canonical isoform of UTRN is skipped with higher frequency in metastatic FCs (FDR = 0.02). UTRN deletion, truncation, and frameshift mutations have been observed in breast carcinoma, neuroblastoma, and melanomas [261, 262]. Interestingly, UTRN is a target of mir-206, which had been shown to inhibit metastasis-relevant traits in ATCs [263]. Given that a putative binding site of mir-206 is in the skipped exon, this splice variant may offer advantages in evading mir-206 regulation. MACF1 (or ACF7) regulates cytoskeletal-focal adhesion dynamics and plays an important role in epidermal migration [264]. Furthermore, MACF1 was recently found to be differentially spliced in breast cancer compared to patient-matched normal tissue [265].

Finally, we investigated the presence of the highly recurrent mutations in RAS, BRAF, and EIF1AX identified by TCGA's thyroid cancer analysis in our FC cohort. We identified 3 canonical RAS mutations in metastatic primary samples and 2 splice-site

mutations in the EIF1AX gene in non-metastatic primary samples (Table 13). Pyrosequencing of the HRAS and NRAS canonical mutations validated the mutational profiles observed in these samples using RNA-seq.

6.5: Conclusion

This proof-of-concept study showed that RNA-seq is a viable platform to assess transcriptomic and, to a certain extent, genetic changes in FFPE thyroid tissues. We have described a high success rate in identifying metastatic samples and identified both novel and previously recognized gene expression changes related to tumor metastasis, albeit in a small sample set. Our identification of a transcriptional pattern characteristic of metastatic FC in a tumor originally classified as non-metastatic that eventually developed metastasis at 10.5 years suggest that some of these molecular changes occur early in the developmental timeline of the tumor, and may speak to the ability to better identify patients whose tumors will require clinical intervention.

Using the TCGA thyroid dataset, we further discovered concordant gene expression changes in aggressive versus indolent tissue (tumor versus normal tissue, and metastatic versus non-metastatic primary tumor) and identified enrichment for gene sets of cellular processes related to increased aggressiveness and invasiveness such as EMT, hypoxia, and oncogenic signaling pathways.

A distance-based analysis using genes distinguishing FVPTCs and PTCs revealed that FCs are molecularly closer to FVPTCs than classical PTCs, and might suggest a reconsideration of the relationship between FCs, FVPTCs, and PTCs. Finally, we showed

the ability to detect mutations in FFPE material from RNA-seq, and documented mutations previously identified in FVPTCs in this small cohort of FCs.

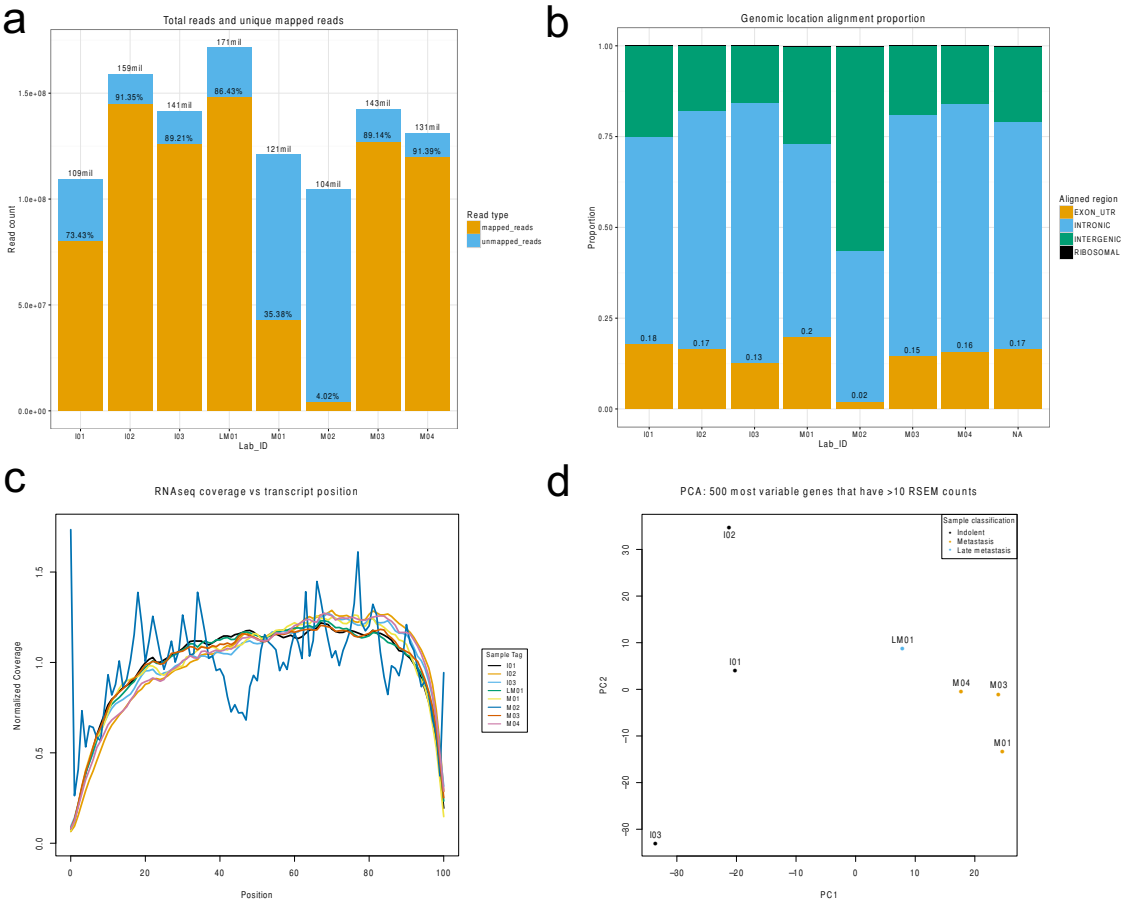


Figure 6–1: QC metrics for RNA-seq of FFPE FTC

a) Total read counts separated into mapped (orange) and unmapped (light blue) reads for each of eight samples. Two samples had poor mapping, M01 and M02. b) Genomic mapping of all eight FTC samples reveal mapping comparable to total RNA-seq in FF tissues, except for M02. c) RNAseq coverage vs. transcript position plot assesses a 3' bias effect that correlates with RNA quality. Other than M02, other samples reveal good coverage. Following this, M02 was dropped from further analyses. d) Principal component analysis (PCA) to identify outliers in datasets and provide unsupervised exploratory analysis of the data. No outliers were observed, including M01 that had poorer than expected mapping, suggesting that the sample can be used for downstream analyses. Interestingly, separation in metastatic status of these primary FTCs was observed on PC1. More importantly, LM01, which developed metastasis 10 years after the initial diagnosis, had a PC1 position intermediate of metastatic and non-metastatic FTCs, suggesting that transcriptomic changes associated with metastatic potential was present early in disease.

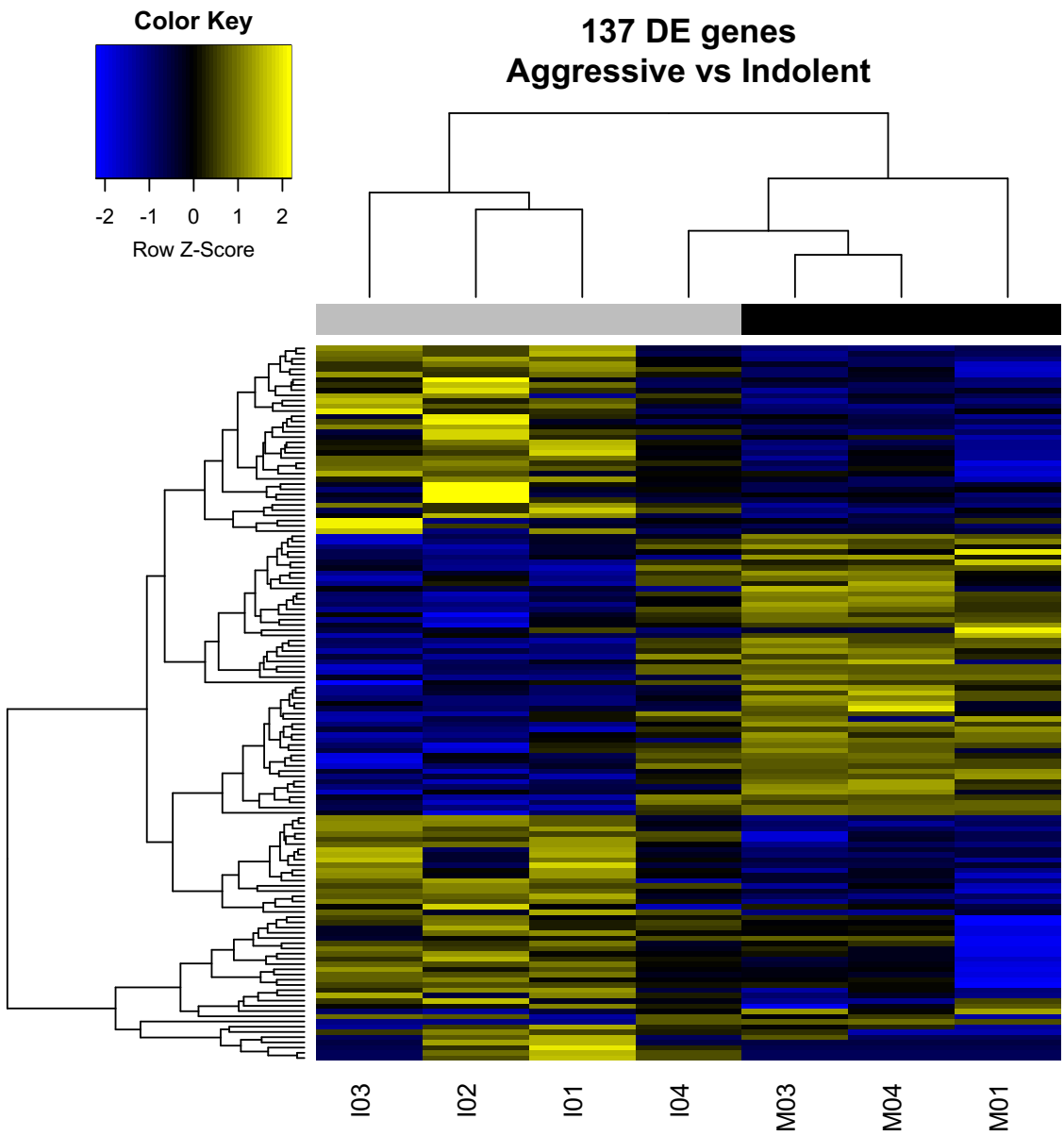


Figure 6–2: DE genes between non-metastatic and metastatic samples, without reclassification of late metastatic sample

Hierarchical clustering results for 137 DE probes identified via Cuffdiff analysis comparing non-metastatic (indolent) and metastatic samples (I, grey bars, vs M, black bars). Interestingly, one of the samples, I04, had gene expression profiles of these genes intermediate between non-metastatic and metastatic samples. Upon updating clinical follow up information for these series of FTCs, I04 was revealed to have a metastatic event at >10 years after the initial diagnosis. I04 was reclassified as LM01. This observation suggests that markers of metastasis can be identified years before the actual metastasis event.

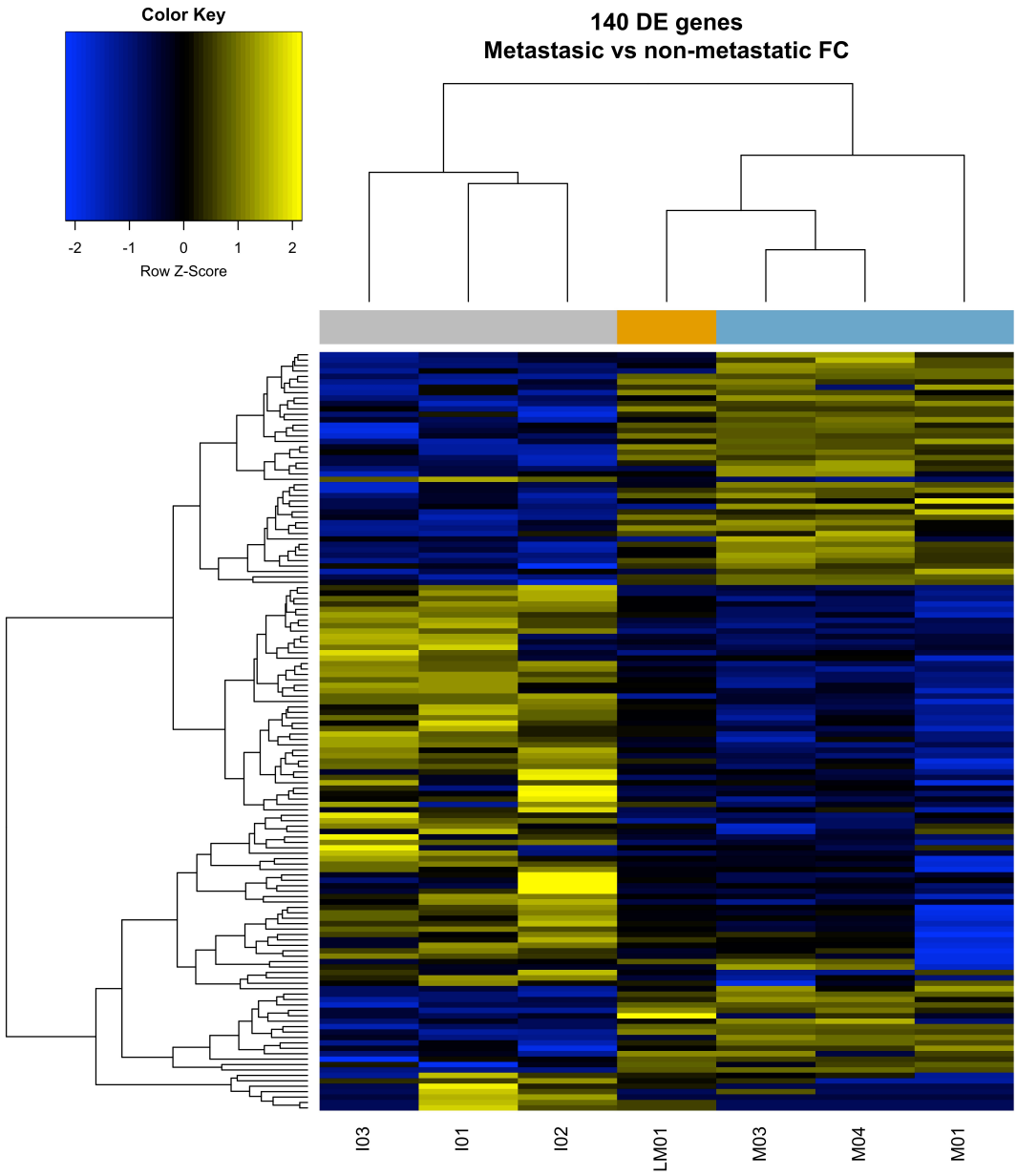


Figure 6–3: DE genes between non-metastatic and metastatic sample, with LM classified as metastatic

Hierarchical clustering of 140 DE genes identified using Cuffdiff comparing metastatic and non-metastatic samples, with LM01 classified as a metastatic sample. This was performed to capture a set of genes with a purer metastatic signal. Perhaps unsurprisingly, LM01 had a gene expression profile intermediate between metastatic and non-metastatic samples. A total of 114 genes identified by this new analysis overlapped with the genes discovered when LM01 was considered an indolent disease.

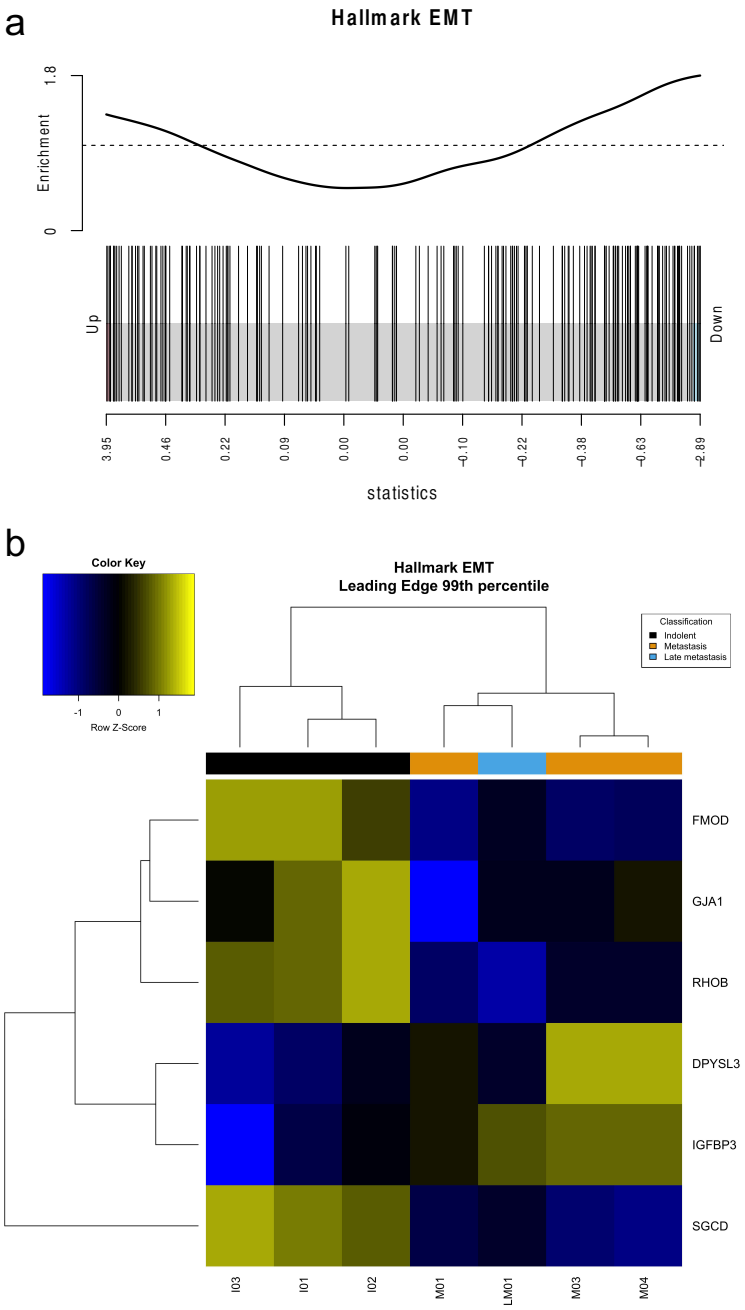


Figure 6–4: Hallmark EMT gene set enrichment results between metastatic and non-metastatic FTC

a) Enrichment for Hallmark EMT was observed in genes differentially expressed between metastatic and non-metastatic FTCs. b) Heatmap showing RSEM values of leading edge genes (99^{th} percentile) identified in the EMT geneset.

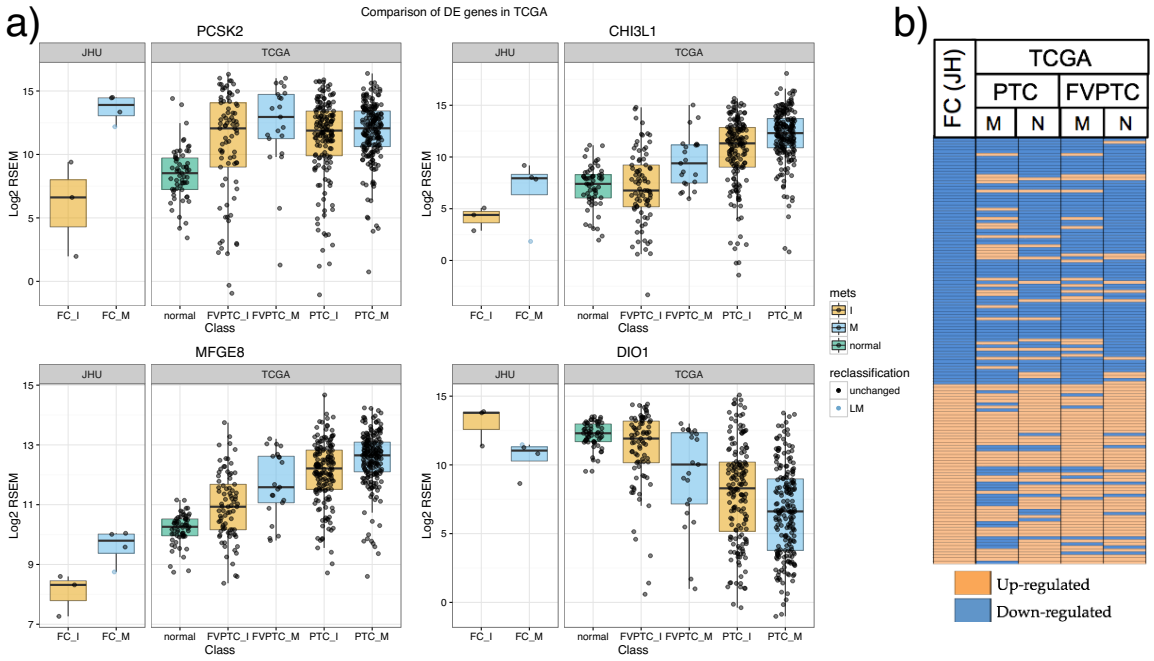


Figure 6–5: Expression of genes DE between metastatic and non-metastatic FTC in different subgroups of TCGA thyroid cancer dataset

a) The expression of four genes, PCSK2, CHI3L1, MFGE8, and DIO1, in TCGA thyroid cancer dataset. Tumor adjacent normal and primary thyroid cancer samples were used in this analysis. The primary cancer samples were separated into their histological subtypes, papillary thyroid cancer (PTC) and follicular variant of papillary thyroid cancer (FVPTC). Beyond that, the primary samples were also divided into two classes I, for indolent, and M, for samples with any metastasis event – to the lymph node or distant site. From the DE genes identified by the JHU FTC cohort, trends of up- and down-regulation were preserved when comparing an aggressive phenotype to a less aggressive phenotype. b) Simplified heatmap comparing direction of change of 140 genes DE between metastatic and non-metastatic FCs in four different TCGA comparisons. PTC-M: Metastatic primary PTC vs. non-metastatic primary PTC; PTC-N: Primary PTC vs. tumor adjacent normal tissue. FVPTC-M and FVPTC-N, similar to PTC except in the FVPTC subset of primary tumors. Orange shows upregulation and blue downregulation. The trend is consistent when comparing a more aggressive to a less aggressive phenotype.

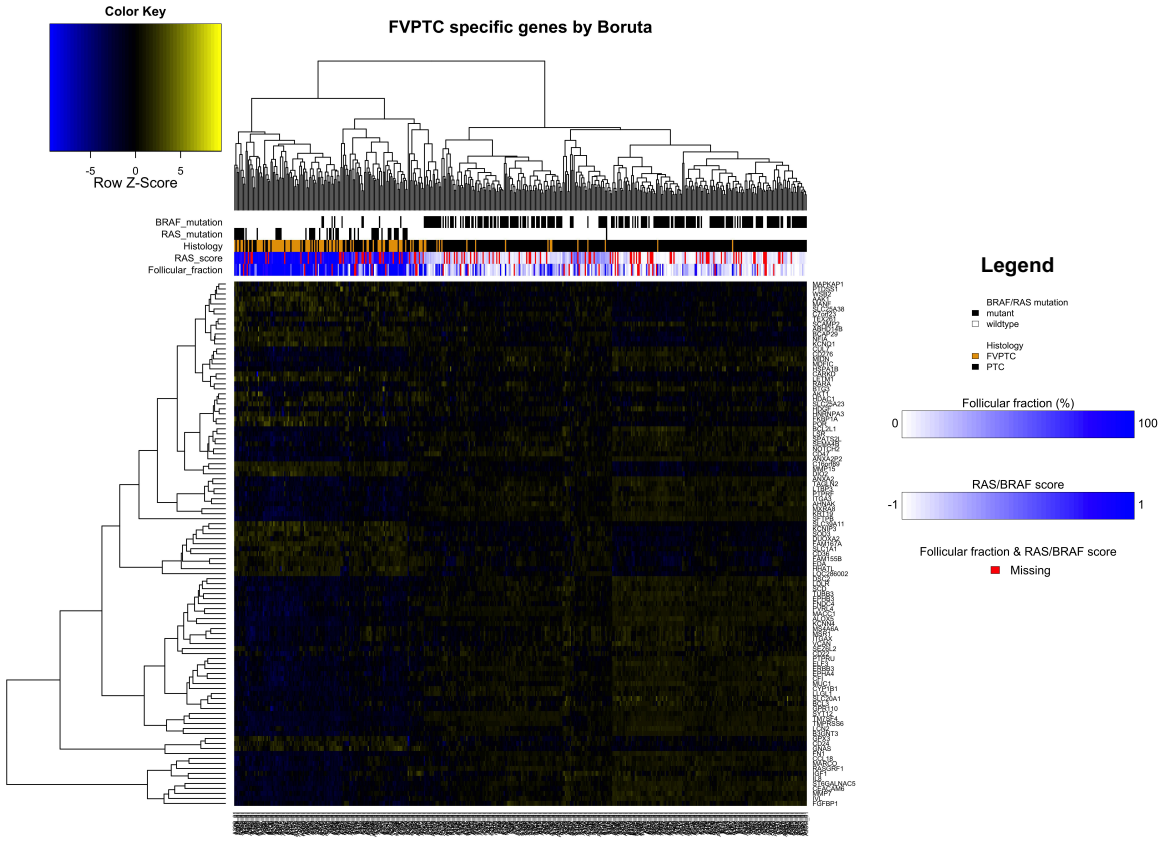


Figure 6–6: FVPTC-specific genes identified using Boruta comparing classical PTC and FVPTC in the TCGA thyroid cancer dataset

Using Boruta, we identified a series of genes that were able to distinguish FVPTCs from classical PTCs, and results are summarized in this heatmap. PTC classification, histological, and molecular features are highlighted on the color bars in the top; BRAF and RAS mutation, histological subtype, RAS/BRAF score as measured by TCGA, and follicular fraction estimated by TCGA pathologists. Missing information is highlighted in red. We observed separation of FVPTCs and PTCs into two separate clusters. More interestingly, we observed the clustering of some PTC samples in the FVPTC cluster, with these PTCs having high follicular fractions and RAS scores, suggesting that they are molecularly driven by FVPTC-related pathways.

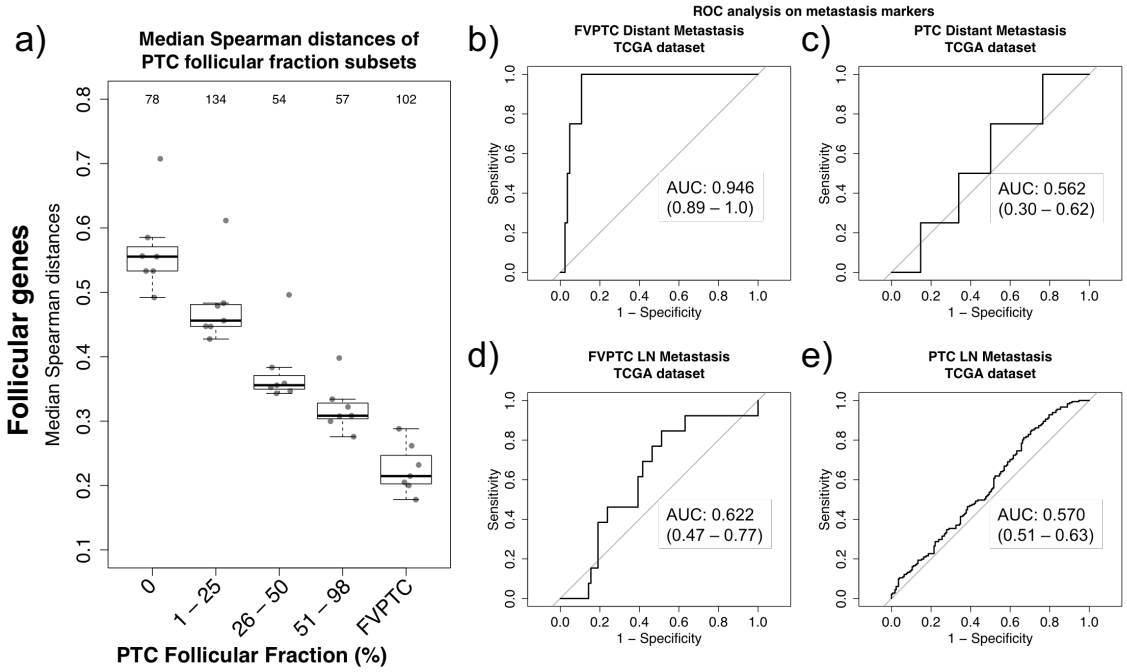


Figure 6–7: FTCs are molecularly similar to FVPTCs and markers of distant metastasis in FTCs predicts distant metastasis in FVPTCs

a) Median Spearman distances for each JHU FTC sample was calculated against primary PTCs in the TCGA THCA dataset using follicular genes differentially expressed between FVTPC and PTC in the TCGA dataset. The median distance between each FTC sample and groups of TCGA separated into different follicular fraction groups. FTC samples are molecularly most similar to FVTPCs, with a stepwise increased distance from PTCs with decreasing follicular fraction. b-e) ROC analysis on the ability to predict metastasis in TCGA thyroid cancer samples. Briefly, metastasis scores were calculated using a gene voting method and were used as predictors against different metastatic outcomes in different PTC subtype; b) distant metastasis in FVPTC, c) distant metastasis in PTC, d) LN metastasis in FVPTC, and e) LN metastasis in PTC. FTC-metastasis markers were only predictive of distant metastasis in FVPTCs (AUC 0.946, 95% CI: 0.89 – 1.0).

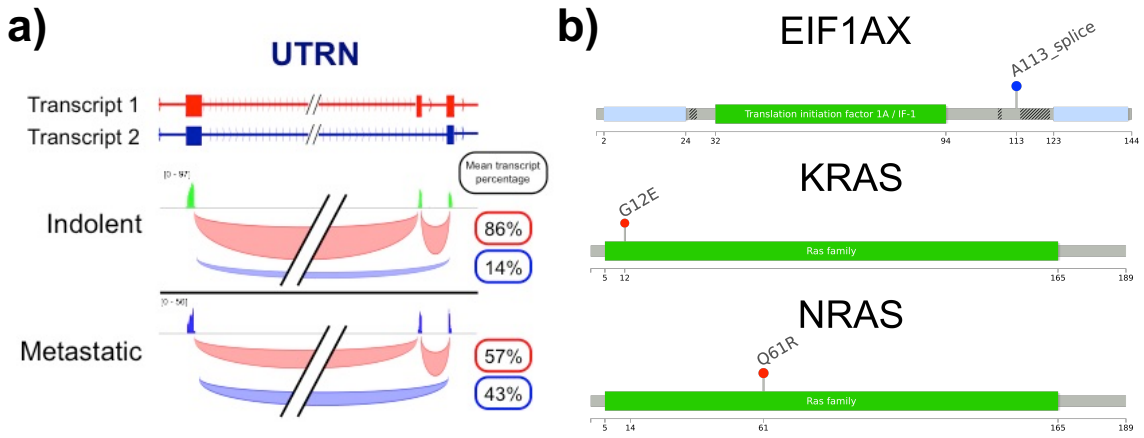


Figure 6–8: Differential splicing event observed in UTRN and mutations in three known FTC and FVPTC driver genes

a) Representative example of splice variant result identifying a preferential exon skipping event in UTRN in metastatic FTCs. b) Three genes previously identified as mutated with high frequency in FTC and FVPTCs were identified in 5/7 FTC samples, including EIF1AX, KRAS, and NRAS. Interestingly, RAS mutations were only present in metastatic samples and EIF1AX mutations were only present in non-metastatic samples.

Table 10: Patient demographics for FVPTC metastasis study

Sample_Name	Classification	Gender	Age	Race	Size
I01	Indolent	F	36	H	7.2
I02	Indolent	M	20	W	8
I03	Indolent	F	48	W	3.5
I04	Indolent	M	77	M	8
A01	Aggressive	F	57	B	4.5
A02	Aggressive	F	56	W	4.1
A03	Aggressive	M	59	W	2.2
A04	Aggressive	M	74	W	5

Table 11: Top 50 differentially expressed genes between metastatic and non-metastatic FTC

gene	Mean_met	Mean_nonmet	log2_FC	test_stat	p_value	q_value
DIO1	33.81	178.08	2.40	2.73	5.00E-05	0.01
SGIP1	46.86	8.25	-2.51	-1.59	5.00E-05	0.01
AC113949.1,LPHN2	3.31	17.44	2.40	1.54	5.00E-05	0.01
VAV3	42.47	11.11	-1.94	-1.29	5.00E-05	0.01
F5	0.85	48.89	5.85	2.53	5.00E-05	0.01
FMOD	8.72	52.60	2.59	2.89	5.00E-05	0.01
ARID5B	4.07	18.21	2.16	2.30	5.00E-05	0.01
C10orf131,CC2D2B,ENTPD1,RP11-248J23.6,RP11-248J23.7,RP11-429G19.3,RP11-690P14.4	181.29	35.58	-2.35	-1.11	5.00E-05	0.01
RP1-59M18.2,SERGEF,TPH1	71.83	13.62	-2.40	-1.60	5.00E-05	0.01
MUC15	48.84	6.05	-3.01	-3.30	5.00E-05	0.01
LRP4	38.11	2.34	-4.02	-2.90	5.00E-05	0.01
RASGRP2	0.85	3.65	2.10	0.76	5.00E-05	0.01
LRRK2	141.32	2.50	-5.82	-2.95	5.00E-05	0.01
HOXC10,HOXC4,HOXC5,HOXC6,HOXC9,RP11-834C11.12,RP11-834C11.14	0.30	2.68	3.17	0.94	5.00E-05	0.01
PIWIL1,RP11-117L5.1	91.36	9.76	-3.23	-0.81	5.00E-05	0.01
AVPR1A,RP11-1022B3.1	2.76	112.17	5.35	3.69	5.00E-05	0.01
RP11-230G5.2	18.93	0.32	-5.89	-0.98	5.00E-05	0.01
MIR4495,RP11-1016B18.1	126.86	3.39	-5.23	-1.37	5.00E-05	0.01
NRXN3	1.52	14.89	3.30	1.40	5.00E-05	0.01
ASPG	1.25	14.26	3.51	0.88	5.00E-05	0.01
SLC7A8	7.16	33.37	2.22	1.94	5.00E-05	0.01
LTBP2	27.40	8.50	-1.69	-1.79	5.00E-05	0.01
PGF	9.94	54.09	2.44	2.21	5.00E-05	0.01
IGHA1,IGHD3-10,IGHG1,IGHG2,IGHG3,IGHG4,IGHGP,IGHJ1,IGHJ2,IGHJ3,IGHJ3P,IGHJ4,IGHJ5,IGHJ6,IGHM,IGHV1-18,IGHV1-2,IGHV1-3,IGHV2-5,IGHV3-11,IGHV3-20,IGHV3-21,IGHV3-48,IGHV3-6,IGHV3-7,IGHV4-4,IGHV6-1,IGHV7-34-1,RP11-731F5.2	185.54	6425.41	5.11	14.93	5.00E-05	0.01
GABRG3	6.15	0.01	-9.95	-0.13	5.00E-05	0.01
BNIP3P5,CAPN3,GANC,RP11-164J13.1	49.74	14.01	-1.83	-1.29	5.00E-05	0.01
ALDH1A2	0.35	2.36	2.74	0.44	5.00E-05	0.01
DAPK2	118.72	24.68	-2.27	-0.99	5.00E-05	0.01
AC245033.1,GOLGA2P10	10.18	0.15	-6.10	-0.64	5.00E-05	0.01
NTRK3	7.68	1.12	-2.77	-0.82	5.00E-05	0.01
SLC47A1	26.31	2.98	-3.14	-1.84	5.00E-05	0.01
HOXB-AS1,HOXB-AS3	0.16	1.49	3.18	0.30	5.00E-05	0.01
CACNA1G	0.13	2.16	4.06	0.98	5.00E-05	0.01
SRCIN1	18.54	2.23	-3.05	-1.63	5.00E-05	0.01
TMC6,TNRC6C-AS1	30.42	8.35	-1.87	-0.94	5.00E-05	0.01
ARHGAP28	1.48	17.11	3.53	1.72	5.00E-05	0.01
CTD-2527I21.4,FXYD1,FXYD7	0.18	1.13	2.63	0.36	5.00E-05	0.01
CTC-339O9.1	0.00	21.70	inf	nan	5.00E-05	0.01
CTD-3252C9.4,MIR24-2	16.15	56.83	1.81	2.40	5.00E-05	0.01
DMKN	64.40	183.06	1.51	1.39	5.00E-05	0.01
RHOB	22.43	97.96	2.13	2.89	5.00E-05	0.01
MBOAT2	19.04	3.46	-2.46	-1.11	5.00E-05	0.01
EFEMP1	4.90	28.04	2.52	1.66	5.00E-05	0.01
ST6GAL2	110.21	19.15	-2.53	-2.09	5.00E-05	0.01
TFCP2L1	2.49	21.45	3.11	2.59	5.00E-05	0.01
PCSK2	153.81	2.00	-6.27	-3.95	5.00E-05	0.01
SOGA1	11.69	54.49	2.22	1.29	5.00E-05	0.01
PRAME	0.01	2.92	8.95	0.11	5.00E-05	0.01
SCUBE1	0.20	17.00	6.42	1.44	5.00E-05	0.01
PPARG	3.93	107.71	4.78	2.43	5.00E-05	0.01

Table 12: Gene set enrichment analysis results of FTC metastasis dataset

Gene_set	P_value	FDR	N	Leading_edge_genes_95th_percentile
EPITHELIAL_MESENCHYMAL_TRANSITION	0	0	176	IGFBP3, RHOB, SGCD, FMOD, GJA1, DPYSL3
ESTROGEN_RESPONSE_EARLY	0	0	175	CA12, GJA1, FRK, FOXC1, DEPTOR, TOB1, SLC7A2
NOTCH_SIGNALING	1.00E-05	7.00E-05	26	FZD1
ADIPOGENESIS	1.00E-05	7.00E-05	164	PPARG, TOB1
P53_PATHWAY	2.00E-05	0.00019	172	TOB1
TNFA_SIGNALING_VIA_NFKB	5.00E-05	0.00038	179	DUSP1, RCAN1, RHOB, DUSP4
ESTROGEN_RESPONSE_LATE	0.00013	0.00092	180	CA12, FRK, FOXC1, TOB1, SERPINA1
UV_RESPONSE_DN	0.00015	0.00095	131	MET, PPARG, GJA1, DUSP1, ID1
OXIDATIVE_PHOSPHORYLATION	2.00E-04	0.00114	163	
HYPOXIA	0.00115	0.00577	182	IGFBP3, PGF, DUSP1, CA12
MTORC1_SIGNALING	0.0015	0.00684	179	
INTERFERON_GAMMA_RESPONSE	0.00178	0.00741	171	NOD1, ARID5B
ANDROGEN_RESPONSE	0.00202	0.00775	88	ARID5B
IL2_STAT5_SIGNALING	0.00325	0.01161	174	RHOB
KRAS_SIGNALING_UP	0.01118	0.03725	173	IGFBP3, KIF5C
GLYCOLYSIS	0.01592	0.04974	173	IGFBP3, B3GAT1, MET

Table 13: CLASS splice variant analysis results

Gene_symbol	chr	strand	exonStart	exonEnd	upstreamES	upstreamEE	downstreamES	downstreamEE	FDR	Delta_inclusion
NPC2	chr14	-	74480701	74480779	74480003	74480288	74484414	74484587	7.39E-08	0.026
TG	chr8	+	132869728	132869826	132868114	132868223	132871347	132871551	2.98E-05	0.033
MACF1	chr1	+	39480919	39481030	39479797	39480009	39484600	39485195	0.00077644	0.321
ACSL3	chr2	+	222900673	222900780	222887829	222887888	222908732	222908781	0.00337043	0.589
ARID1B	chr6	+	157150724	157150853	157148623	157148951	157167039	157167185	0.01897395	0.325
UTRN	chr6	+	144827347	144827386	144820881	144821018	144827610	144827659	0.02312159	-0.292
RMST	chr12	+	97496493	97496554	97495930	97496056	97530654	97530859	0.02312159	-0.029

Table 14: Mutations in thyroid driver genes identified in JHU FTC cohort

Sample_ID	Reclassify	Status	EIF1AX	KRAS	NRAS
I02	I02	Indolent	p.A113_splice	0	0
I03	I03	Indolent	p.A113_splice	0	0
I04	M05	Late metastatic	0	p.G12S	0
M03	M03	Metastatic	0	0	p.Q61R
M04	M04	Metastatic	0	p.Q61R	p.Q61R

Table 15: FVPTC-specific genes when compared to PTC identified by Boruta

SYMBOL	ENTREZID	GENENAME	MAP
AHNAK2	113146	AHNAK nucleoprotein 2	14q32.33
AKT1	207	v-akt murine thymoma viral oncogene homolog 1	14q32.32
ALOX5	240	arachidonate 5-lipoxygenase	10q11.2
ANXA2P2	304	annexin A2 pseudogene 2	9p13
ANXA2	302	annexin A2	15q22.2
B3GNT3	10331	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 3	19p13.1
BCAP29	55973	B-cell receptor-associated protein 29	7q22.3
BCL2L1	598	BCL2-like 1	20q11.21
BCL3	602	B-cell CLL/lymphoma 3	19q13.1-q13.2
C16orf89	146556	chromosome 16 open reading frame 89	16p13.3
C1orf130	NA	NA	NA
C7orf23	NA	NA	NA
CARKD	55739	carbohydrate kinase domain containing	13q34
CCL18	6362	chemokine (C-C motif) ligand 18 (pulmonary and activation-regulated)	17q12
CD24	100133941	CD24 molecule	6q21
CD276	80381	CD276 molecule	15q23-q24
CD36	948	CD36 molecule (thrombospondin receptor)	7q11.2
CD47	961	CD47 molecule	3q13.1-q13.2
CDC42EP3	10602	CDC42 effector protein (Rho GTPase binding) 3	2p21
CEACAM6	4680	carcinoembryonic antigen-related cell adhesion molecule 6 (non-specific cross reacting antigen)	19q13.2
CETN2	1069	centrin, EF-hand protein, 2	Xq28
CFI	3426	complement factor I	4q25
CNTNAP2	26047	contactin associated protein-like 2	7q35
CYP1B1	1545	cytochrome P450, family 1, subfamily B, polypeptide 1	2p22.2
DDX60	55601	DEAD (Asp-Glu-Ala-Asp) box polypeptide 60	4q32.3
DSC2	1824	desmocollin 2	18q12.1
DSC3	1825	desmocollin 3	18q12.1
DUOXA2	405753	dual oxidase maturation factor 2	15q15.1
EDA	1896	ectodysplasin A	Xq12-q13.1
EHBP1L1	254102	EH domain binding protein 1-like 1	11q13.1
ELF3	1999	E74-like factor 3 (ets domain transcription factor, epithelial-specific)	1q32.2
EPHA4	2043	EPH receptor A4	2q36.1
EPHB3	2049	EPH receptor B3	3q27.1
ERBB3	2065	erb-b2 receptor tyrosine kinase 3	12q13
F5	2153	coagulation factor V (proaccelerin, labile factor)	1q23
FAM155B	27112	family with sequence similarity 155, member B	Xq13.1
FGFBP1	9982	fibroblast growth factor binding protein 1	4p15.32
FN1	2335	fibronectin 1	2q34
FNDC4	64838	fibronectin type III domain containing 4	2p23.3
FOSL2	2355	FOS-like antigen 2	2p23.3
GABRB2	2561	gamma-aminobutyric acid (GABA) A receptor, beta 2	5q34
GNAS	2778	GNAS complex locus	20q13.3
GPR110	NA	NA	NA
GPX1	2876	glutathione peroxidase 1	3p21.3
GTF3C1	2975	general transcription factor IIIC, polypeptide 1, alpha 220kDa	16p12
HDGF	3068	hepatoma-derived growth factor	1q23.1
HHATL	57467	hedgehog acyltransferase-like	3p22.1
HSPA1B	3304	heat shock 70kDa protein 1B	6p21.3
IGF1	3479	insulin-like growth factor 1 (somatomedin C)	12q23.2
IL10RA	3587	interleukin 10 receptor, alpha	11q23
IL8	NA	NA	NA
ITGA3	3675	integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)	17q21.33
ITGAX	3687	integrin, alpha X (complement component 3 receptor 4 subunit)	16p11.2
IVL	3713	involucrin	1q21
KCNIP3	30818	Kv channel interacting protein 3, calsenilin	2q21.1
KCNN4	3783	potassium channel, calcium activated intermediate/small conductance subfamily N alpha, member	19q13.2
KCNQ1	3784	potassium channel, voltage gated KQT-like subfamily Q, member 1	11p15.5
KIAA1217	56243	KIAA1217	10p12.31
KRT19	3880	keratin 19, type I	17q21.2
LCN2	3934	lipocalin 2	9q34

(cont)

(Table 15, cont)

SYMBOL	ENTREZID	GENENAME	MAP
LDLR	3949	low density lipoprotein receptor	19p13.2
LETM1	3954	leucine zipper-EF-hand containing transmembrane protein 1	4p16.3
LGR6	59352	leucine-rich repeat containing G protein-coupled receptor 6	1q32.1
LOC286002	NA	NA	NA
LRP1	4035	low density lipoprotein receptor-related protein 1	12q13.3
LSR	51599	lipolysis stimulated lipoprotein receptor	19q13.12
LTBP3	4054	latent transforming growth factor beta binding protein 3	11q13.1
MACC1	346389	metastasis associated in colon cancer 1	7p21.1
MARCO	8685	macrophage receptor with collagenous structure	2q14.2
MIDN	90007	midnolin	19p13.3
MMP15	4324	matrix metallopeptidase 15 (membrane-inserted)	16q13
MMP7	4316	matrix metallopeptidase 7	11q22.2
MSR1	4481	macrophage scavenger receptor 1	8p22
MUC1	4582	mucin 1, cell surface associated	1q21
MXRA8	54587	matrix-remodelling associated 8	1p36.33
MYH10	4628	myosin, heavy chain 10, non-muscle	17p13
NFE2L3	9603	nuclear factor, erythroid 2-like 3	7p15.2
NOTCH2	4853	notch 2	1p13-p11
POR	5447	P450 (cytochrome) oxidoreductase	7q11.2
PPL	5493	periplakin	16p13.3
PTDSS1	9791	phosphatidylserine synthase 1	8q22
PTPRF	5792	protein tyrosine phosphatase, receptor type, F	1p34
PVRL4	81607	poliovirus receptor-related 4	1q23.3
QSOX1	5768	quiescin Q6 sulfhydryl oxidase 1	1q24
RASGRF1	5923	Ras protein-specific guanine nucleotide-releasing factor 1	15q24.2
REEP5	7905	receptor accessory protein 5	5q22-q23
RGN	9104	regucalcin	Xp11.3
RPS6KA2	6196	ribosomal protein S6 kinase, 90kDa, polypeptide 2	6q27
RUNX1	861	runt-related transcription factor 1	21q22.3
S100B	6285	S100 calcium binding protein B	21q22.3
SCD	6319	stearoyl-CoA desaturase (delta-9-desaturase)	10q24.31
SEMA4B	10509	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic d	15q25
SEZ6L2	26470	seizure related 6 homolog (mouse)-like 2	16p11.2
SFTPB	6439	surfactant protein B	2p12-p11.2
SLC20A1	6574	solute carrier family 20 (phosphate transporter), member 1	2q13
SLC25A23	79085	solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 23	19p13.3
SLC25A38	54977	solute carrier family 25, member 38	3p22.1
SLC39A11	201266	solute carrier family 39, member 11	17q24.3-q25.1
SLC5A8	160728	solute carrier family 5 (sodium/monocarboxylate cotransporter), member 8	12q23.1
SNRPB	6628	small nuclear ribonucleoprotein polypeptides B and B1	20p13
SPATS2L	26010	spermatogenesis associated, serine-rich 2-like	2q33.1
ST6GALNAC5	81849	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltr	1p31.1
SYT12	91683	synaptotagmin XII	11q13.2
TAGLN2	8407	transgelin 2	1q21-q25
TCTA	6988	T-cell leukemia translocation altered	3p21
TEX261	113419	testis expressed 261	2p13.3
TM7SF4	NA	NA	NA
TMEM63B	55362	transmembrane protein 63B	6p21.1
TMPRSS6	164656	transmembrane protease, serine 6	22q12.3
TSPAN7	7102	tetraspanin 7	Xp11.4
TUBB3	10381	tubulin, beta 3 class III	16q24.3
WASF3	10810	WAS protein family, member 3	13q12
YWHAE	7531	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon	17p13.3

Chapter 7: Concluding remarks and recommendations

This body of work describes and addresses important fundamental questions about the feasibility of performing genomic analysis on formalin fixed paraffin embedded (FFPE) tissue using high-throughput genome-wide technologies such as sequencing and microarrays. We define this as a problem of performing sparse, high dimensional data analysis in a low resource setting given the relatively low nucleic acid yield and quality, often due to limited availability of tissue.

In the process of answering these questions, we have optimized extraction methods for maximizing yield and quality of RNA and DNA from FFPE materials, developed FFPE-specific quality control metrics and workflows pre- and post-microarray and sequencing experiments, and identified operating limits of sequencing and microarray technologies. Ultimately, we demonstrated the ability to obtain high quality datasets from FFPE-derived RNA and DNA, and showed increased reproducibility by implementing FFPE-specific approaches or modifications to existing protocols.

Beyond that, we developed Epicopy, which is a computational method that allows users to obtain copy number variation (CNV) information from methylation microarrays, extending the information measured by a single microarray technology. We developed Epicopy using the relatively CNV neutral thyroid carcinoma dataset and validated it on the CNV rich breast and lung squamous carcinoma datasets, allowing us to assess Epicopy's performance through the whole spectrum of CNV change across a gamut of human tumors. We showed good concordance between Epicopy- and SNP-derived CNV

profiles and that reproducibility rates between Epicopy and SNP microarrays are comparable to rates between different SNP microarray platforms.

Finally, we used the tools developed in the first part of this thesis to profile the molecular landscape of ductal carcinoma in situ (DCIS) in the context of disease progression, ER-negative breast cancer of patients who did not receive adjuvant chemotherapy in the context of disease recurrence, and follicular thyroid cancer (FTC) in the context of distant metastasis.

In the DCIS study, we analyzed the methylation profile and copy number alterations in a retrospective case-control study of DCIS that progressed to IDC and those that did not. We observed a global methylation field effect in DCIS-adjacent normal tissue and classified DCIS into four stable methylation phenotypes or *epitypes* that show associations with tumor nuclear grade and a CIMP-like phenotype. While differential methylation analyses revealed few differences between progressors and non-progressors, copy number analysis identified regions of the genome with differential CNV events between these groups.

Our multiomic analysis of ER-negative breast cancer was motivated by the clinical need to identify patients who will do well without adjuvant chemotherapy. We aimed to identify subtypes within ER-negative disease, the molecular processes that drive these tumors, and discover biomarkers of recurrence in a cohort of patients with long-term clinical follow-up that did not receive chemotherapy. We observed three stable clusters, that we defined as AR-driven, immune-high, and CNV-high. The AR-driven tumors displayed hallmarks of luminal breast cancers, with high expression of hormone receptor response genes. In these histologically ER-negative tumors, ER expression was

low, and androgen receptor (AR) was upregulated, suggesting that the AR is driving hormonal response. The immune-high tumors had high levels of expression of cytotoxic markers by expression data, and were estimated to have high degree of leukocytic infiltrates by methylation data. The immune exhaustion marker LAG3 and immune checkpoint gene CTLA4 were upregulated, while PD-1 and PDL-1 were not, suggesting that CTLA4-driven immune evasion occurred. The CNV high tumors displayed genomic instability manifesting as a high incidence of amplifications and deletions across the genome. These tumors showed down-regulation of genes related to the DNA damage repair pathway, consistent with the observed molecular phenotype. We proposed the use of more targeted therapies for each subtype of tumor, with androgen-targeted therapy for the AR-driven tumors, CTLA4 inhibitors for the immune high tumors, and chemotherapy or PARP1 inhibitors for the CNV high tumors. The analysis was also extended to identify gene expression markers of recurrence, and identified 130 genes with the ability to do so. A recurrence score (RS) calculated from these genes showed the ability to predict recurrence in the JHU ER-negative cohort and an independent external TNBC dataset of patients who did not receive adjuvant chemotherapy.

Lastly, we demonstrated the ability to perform total RNA-seq analysis of a series of primary FTC tumors. These tumors were obtained from patients who either presented with distant metastasis (stage IV) or were metastasis-free for more than 6 years. The clinical question asked was the ability to identify tumors with the capacity to form distant metastasis using molecular markers at the time of diagnosis. We identified a series of 140 genes, enriched for epithelial-mesenchymal transition (EMT) genes, that were differentially expressed between metastatic and non-metastatic disease. This gene set also

predicted the metastasis of a single FTC tumor 10 years before its clinical manifestation. Using TCGA data, we further demonstrated that FTCs and FVPTCs are molecularly similar and showed the ability of our markers to predict distant metastasis in FVPTCs, suggesting that we are capturing the signature of a biological process with these markers. This speaks to the potential of not only using these markers for predicting metastasis, but also using them to discover more appropriate druggable targets.

Collectively, this body of work demonstrated our ability to recover nucleic acids from FFPE tissues, obtain high quality data from high-throughput microarray & NGS molecular platforms, and maximize the data obtained from the generated datasets. The identification of biologically relevant molecular landscapes in three different tumors types suggests the broad applicability of these methods. Clinically relevant biomarkers were discovered in the ER-negative breast cancer and FTC studies, and we are hopeful that with additional molecular information on the DCIS cohort that we will be able to obtain subtype-specific markers of progression.

Cell culture models can be used to validate biological findings of the ER-negative cohorts, and experiments can be designed to identify both available and novel drugs to more appropriately treat these patients. Genes that best distinguish these classes can be used, either in the form of IHC markers or molecular panels, as companion diagnostics for appropriate treatment.

In the ER-negative and FTC studies, the discovered prognostic biomarkers can be explored as candidates for feature selection and used in the development of a bench-based molecular assay. With appropriate technical and independent external validation,

these can serve as clinical markers in addition to routinely available clinicopathological information for better stratification of patient treatment groups.

Looking forward, this work confirms the ability to perform retrospective studies with well-controlled clinical parameters in FFPE material, granting us access to the treasure trove of information previously locked in archival disease tissue repositories. We are optimistic about the promise of these methodologies, and their applicability in discovering ways to improve clinical outcomes in the most resource efficient manner.

Bibliography

1. Tomczak K, Czerwinska P, Wiznerowicz M: **The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.** *Contemp Oncol (Pozn)* 2015, **19:**A68-77.
2. Abramovitz M, Leyland-Jones B: **Application of array-based genomic and epigenomic technologies to unraveling the heterogeneous nature of breast tumors: on the road to individualized treatment.** *Cancer Genomics Proteomics* 2007, **4:**135-145.
3. Kottaridis PD, Gale RE, Frew ME, Harrison G, Langabeer SE, Belton AA, Walker H, Wheatley K, Bowen DT, Burnett AK, et al: **The presence of a FLT3 internal tandem duplication in patients with acute myeloid leukemia (AML) adds important prognostic information to cytogenetic risk group and response to the first cycle of chemotherapy: analysis of 854 patients from the United Kingdom Medical Research Council AML 10 and 12 trials.** *Blood* 2001, **98:**1752-1759.
4. Robinson DR, Wu YM, Vats P, Su F, Lonigro RJ, Cao X, Kalyana-Sundaram S, Wang R, Ning Y, Hodges L, et al: **Activating ESR1 mutations in hormone-resistant metastatic breast cancer.** *Nat Genet* 2013, **45:**1446-1451.
5. Cancer Genome Atlas N: **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490:**61-70.
6. Mamounas EP, Wickerham DL, Fisher B, Geyer CE, Julian TB, Wolmark N: **Chapter 42. The NSABP Experience.** In *Kuerer's Breast Surgical Oncology.* Edited by Kuerer HM. New York, NY: The McGraw-Hill Companies; 2010
7. McCready DR, Miller NA, Youngson BJ: **Chapter 20. Invasive Breast Carcinoma.** In *Kuerer's Breast Surgical Oncology.* Edited by Kuerer HM. New York, NY: The McGraw-Hill Companies; 2010
8. Veronesi U, Zurrida S: **Chapter 43. The Milan Cancer Institute's Landmark Clinical Trials.** In *Kuerer's Breast Surgical Oncology.* Edited by Kuerer HM. New York, NY: The McGraw-Hill Companies; 2010
9. Kokkat TJ, Patel MS, McGarvey D, LiVolsi VA, Baloch ZW: **Archived Formalin-Fixed Paraffin-Embedded (FFPE) Blocks: A Valuable Underexploited Resource for Extraction of DNA, RNA, and Protein.** *Biopreservation and Biobanking* 2013, **11:**101-106.
10. Group EBCTC: **Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. Early Breast Cancer Trialists' Collaborative Group.** *Lancet* 1992, **339:**71-85.
11. Bancroft JD, Gamble M, Jensen K: **Theory and Practice of Histological Techniques (6th Edition).** *Journal of neuropathology and experimental neurology* 2008, **67:**633.
12. Eltoum I, Fredenburgh J, Myers RB, Grizzle WE: **Introduction to the theory and practice of fixation of tissues.** *Journal of Histotechnology* 2001, **24:**173-190.
13. Fox CH, Johnson FB, Whiting J: **Formaldehyde fixation.** *J histochem ...* 1985.

14. Eltoum I, Fredenburgh J, Grizzle WE: **Advanced concepts in fixation: 1. Effects of fixation on immunohistochemistry, reversibility of fixation and recovery of proteins, nucleic acids, and other molecules from fixed and processed tissues. 2. Developmental methods of fixation.** *Journal of Histotechnology* 2001, **24:**201-210.
15. Fraenkel-Conrat H, Brandon BA, Olcott HS: **The reaction of formaldehyde with proteins; participation of indole groups; gramicidin.** *The Journal of biological chemistry* 1947, **168:**99-118.
16. Fraenkel-Conrat H, Cooper M: **The Reaction of Formaldehyde with Proteins - Journal of the American Chemical Society (ACS Publications).** *Journal of the American ...* 1945.
17. Fraenkel-Conrat H, Mecham DK: **THE REACTION OF FORMALDEHYDE WITH PROTEINS.** 1949.
18. Fraenkel-Conrat H, Olcott HS: **Reaction of formaldehyde with proteins; participation of the guanidyl groups and evidence of crosslinking.** *Journal of the American Chemical Society* 1946, **68:**34-37.
19. Fraenkel-Conrat H, Olcott HS: **The reaction of formaldehyde with proteins; cross-linking between amino and primary amide or guanidyl groups.** *Journal of the American Chemical Society* 1948, **70:**2673-2684.
20. Metz B, Kersten GF, Hoogerhout P, Brugghe HF, Timmermans HA, de Jong A, Meiring H, ten Hove J, Hennink WE, Crommelin DJ, Jiskoot W: **Identification of formaldehyde-induced modifications in proteins: reactions with model peptides.** *J Biol Chem* 2004, **279:**6235-6243.
21. McGhee JD, Von Hippel PH: **Formaldehyde as a probe of DNA structure. I. Reaction with exocyclic amino groups of DNA bases.** *Biochemistry* 1975, **14:**1281-1296.
22. McGhee JD, Von Hippel PH: **Formaldehyde as a probe of DNA structure. II. Reaction with endocyclic imino groups of DNA bases.** *Biochemistry* 1975, **14:**1297-1303.
23. McGhee JD, Von Hippel PH: **Formaldehyde as a probe of DNA structure. 3. Equilibrium denaturation of DNA and synthetic polynucleotides.** *Biochemistry* 1977, **16:**3267-3276.
24. McGhee JD, Von Hippel PH: **Formaldehyde as a probe of DNA structure. 4. Mechanism of the initial reaction of formaldehyde with DNA.** *Biochemistry* 1977, **16:**3276-3293.
25. Jackson V: **Studies on histone organization in the nucleosome using formaldehyde as a reversible cross-linking agent.** *Cell* 1978, **15:**945-954.
26. Jackson V: **Formaldehyde cross-linking for studying nucleosomal dynamics.** *Methods (San Diego, Calif)* 1999, **17:**125-139.
27. Casanova-Schmitz M, Heck HDA: **Effects of formaldehyde exposure on the extractability of DNA from proteins in the rat nasal mucosa.** *Toxicology and Applied Pharmacology* 1983, **70:**121-132.
28. Chaw YFM, Crane LE, Lange P, Shapiro R: **Isolation and identification of cross-links from formaldehyde-treated nucleic acids.** *Biochemistry* 1980, **19:**5525-5531.

29. Chomczynski P, Sacchi N: **Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction.** *Anal Biochem* 1987, **162:**156-159.
30. Abramovitz M, Ordanic-Kodani M, Wang Y, Li Z, Catzavelos C, Bouzyk M, Sledge GW, Jr., Moreno C, Leyland-Jones B: **Optimization of RNA extraction from FFPE tissues for expression profiling in the DASL assay.** *BioTechniques* 2008, **44:**417.
31. Bonin S, Hlubek F, Benhattar J, Denkert C, Dietel M, Fernandez PL, Höfler G, Kothmaier H, Kruslin B, Mazzanti CM, et al: **Multicentre validation study of nucleic acids extraction from FFPE tissues.** *Virchows Archiv* 2010, **457:**309-317.
32. Bonin S, Stanta G: **Nucleic acid extraction methods from fixed and paraffin-embedded tissues in cancer diagnostics.** *Expert Review of Molecular Diagnostics* 2013, **13:**271-282.
33. Chung J-Y, Braunschweig T, Hewitt SM: **Optimization of Recovery of RNA From Formalin-fixed, Paraffin-embedded Tissue.** *Diagnostic Molecular Pathology* 2006, **15:**229-236.
34. Doleshal M, Magotra AA, Choudhury B: **Evaluation and Validation of Total RNA Extraction Methods for MicroRNA Expression Analyses in Formalin-Fixed, Paraffin-Embedded Tissues.** *The Journal of Molecular ...* 2008.
35. Funabashi KS, Barcelos D, Visona I, Silva MSe: **DNA extraction and molecular analysis of non-tumoral liver, spleen, and brain from autopsy samples: The effect of formalin fixation and paraffin embedding.** *... -Research and Practice* 2012.
36. Kizys MML, Cardoso MG, Lindsey SC, Harada MY, Soares FA, Melo MCC, Montoya MZ, Kasamatsu TS, Kunii IS, Giannocco G, et al: **Optimizing nucleic acid extraction from thyroid fine-needle aspiration cells in stained slides, formalin-fixed/paraffin-embedded tissues, and long-term stored blood samples.** *Arquivos Brasileiros de Endocrinologia & Metabologia* 2012, **56:**618-626.
37. Kotorashvili A, Ramnauth A, Liu C, Lin J, Ye K, Kim R, Hazan R, Rohan T, Fineberg S, Loudig O: **Effective DNA/RNA Co-Extraction for Analysis of MicroRNAs, mRNAs, and Genomic DNA from Formalin-Fixed Paraffin-Embedded Specimens.** *PloS one* 2012, **7:**e34683.
38. Linton KM, Hey Y, Dibben S, Miller CJ, Freemont AJ, Radford JA, Pepper SD: **BioTechniques - Methods comparison for high-resolution transcriptional analysis of archival material on Affymetrix Plus 2.0 and Exon 1.0 microarrays.** *BioTechniquescom* 2009, **47:**587-596.
39. Ludyga N, Grünwald B, Azimzadeh O, Englert S, Höfler H, Tapio S, Aubele M: **Nucleic acids from long-term preserved FFPE tissues are suitable for downstream analyses.** *Virchows Archiv* 2012, **460:**131-140.
40. Mark Abramovitz MO-KYWZLCCMBGWSJCSMBL-J: **Optimization of RNA extraction from FFPE tissues for expression profiling in the DASL assay.** *BioTechniques* 2008, **44:**417.
41. Masuda N, Ohnishi T, Kawamoto S, Monden M, Okubo K: **Analysis of chemical modification of RNA from formalin-fixed samples and optimization of**

molecular biology applications for such samples. *Nucleic Acids Research* 1999, **27:**4436-4443.

42. Muñoz-Cadavid C, Rudd S, Zaki SR, Patel M, Moser SA, Brandt ME, Gómez BL: **Improving Molecular Detection of Fungal DNA in Formalin-Fixed Paraffin-Embedded Tissues: Comparison of Five Tissue DNA Extraction Methods Using Panfungal PCR.** *Journal of clinical ...* 2010.
43. Potluri K, Mahas A, Kent MN, Naik S, Markey M: **Genomic DNA extraction methods using formalin-fixed paraffin-embedded tissue.** *Analytical Biochemistry* 2015, **486:**17-23.
44. Roberts L, Bowers J, Sensinger K, Lisowski A, Getts R: **Identification of methods for use of formalin-fixed, paraffin-embedded tissue samples in RNA expression profiling.** *Genomics* 2009.
45. Ton CC, Vartanian N, Chai X, Lin MG, Yuan X, Malone KE, Li CI, Dawson A, Sather C, Delrow J, et al: **Gene expression array testing of FFPE archival breast tumor samples: an optimized protocol for WG-DASL® sample preparation.** *Breast cancer research and treatment* 2010, **125:**879-883.
46. Turashvili G, Yang W, McKinney S, Kalloger S: **Nucleic acid quantity and quality from paraffin blocks: Defining optimal fixation, processing and DNA/RNA extraction techniques.** *... and molecular pathology* 2012.
47. Abdueva D, Wing M, Schaub B, Triche T: **Quantitative Expression Profiling in Formalin-Fixed Paraffin-Embedded Samples by Affymetrix Microarrays.** *The Journal of Molecular ...* 2010.
48. Srinivasan M, Sedmak D, Jewell S: **Effect of Fixatives and Tissue Processing on the Content and Integrity of Nucleic Acids.** *The American journal of pathology* 2002.
49. Williams C, Ponten F, Moberg C, Söderkvist P, Uhlén M, Pontén J, Sitbon G, Lundeberg J: **A High Frequency of Sequence Alterations Is Due to Formalin Fixation of Archival Specimens.** *The American journal of pathology* 1999, **155:**1467-1471.
50. Bibikova M, Talantov D, Chudin E, Yeakley JM, Chen J, Doucet D, Wickham E, Atkins D, Barker D, Chee M, et al: **Quantitative Gene Expression Profiling in Formalin-Fixed, Paraffin-Embedded Tissues Using Universal Bead Arrays.** *The American journal of pathology* 2004, **165:**1799-1807.
51. **Whole-Genome DASL® HT Assay for Expression Profiling in FFPE Samples** [http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_whole_genome_dasl.pdf]
52. Ravo M, Mutarelli M, Ferraro L, Grober OMV, Paris O, Tarallo R, Vigilante A, Cimino D, De Bortoli M, Nola E, et al: **Quantitative expression profiling of highly degraded RNA from formalin-fixed, paraffin-embedded breast tumor biopsies by oligonucleotide microarrays.** *Laboratory Investigation* 2008, **88:**430-440.
53. Reinholz MM, Eckel-Passow JE, Anderson SK, Asmann YW, Zschunke MA, Oberg AL, McCullough AE, Dueck AC, Chen B, April CS, et al: **Expression profiling of formalin-fixed paraffin-embedded primary breast tumors using cancer-specific and whole genome gene panels on the DASL® platform.** *BMC Medical Genomics* 2010, **3:**60.

54. **GeneChip® WT Pico Kit (WT Pico Kit): Datasheet** [http://media.affymetrix.com/support/technical/datasheets/wt_pico_kit_datasheet.pdf]
55. **GeneChip® Human Gene 2.0 ST Array: Datasheet** [http://media.affymetrix.com/support/technical/datasheets/hugene_2_st_datasheet.pdf]
56. Sinicropi D, Qu K, Collin F, Crager M, Liu M-L, Pelham RJ, Pho M, Rossi AD, Jeong J, Scott A, et al: **Whole Transcriptome RNA-Seq Analysis of Breast Cancer Recurrence Risk Using Formalin-Fixed Paraffin-Embedded Tumor Tissue.** *PloS one* 2012, **7:**e40092.
57. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, et al: **A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.** *N Engl J Med* 2004, **351:**2817-2826.
58. Norton N, Sun Z, Asmann YW, Serie DJ, Necela BM, Bhagwate A, Jen J, Eckloff BW, Kalari KR, Thompson KJ, et al: **Gene Expression, Single Nucleotide Variant and Fusion Transcript Discovery in Archival Material from Breast Tumors.** *PloS one* 2013, **8:**e81925.
59. Cieslik M, Chugh R, Wu Y-M, Wu M, Brennan C, Lonigro R, Su F, Wang R, Siddiqui J, Mehra R, et al: **The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing.** *Genome research* 2015, **25:**1372-1381.
60. Lips EH: **Reliable High-Throughput Genotyping and Loss-of-Heterozygosity Detection in Formalin-Fixed, Paraffin-Embedded Tumors Using Single Nucleotide Polymorphism Arrays.** *Cancer research* 2005, **65:**10188-10191.
61. Oosting J, Lips EH, van Eijk R, Eilers PHC, Szuhai K, Wijmenga C, Morreau H, van Wezel T: **High-resolution copy number analysis of paraffin-embedded archival tissue using SNP BeadArrays.** *Genome research* 2007, **17:**368-376.
62. Krijgsman O, Israeli D, Haan JC, van Essen HF, Smeets SJ, Eijk PP, Steenbergen RDM, Kok K, Tejpar S, Meijer GA, Ylstra B: **CGH arrays compared for DNA isolated from formalin-fixed, paraffin-embedded material.** *Genes Chromosomes & Cancer* 2012, **51:**344-352.
63. Sikora MJ, Thibert JN, Salter J, Dowsett M, Johnson MD, Rae JM: **High-efficiency genotype analysis from formalin-fixed, paraffin-embedded tumor tissues.** *The pharmacogenomics journal* 2011, **11:**348-358.
64. Tuefferd M, De Bondt A, Van Den Wyngaert I, Talloen W, Verbeke T, Carvalho B, Clevert D-A, Alifano M, Raghavan N, Amaratunga D, et al: **Genome-wide copy number alterations detection in fresh frozen and matched FFPE samples using SNP 6.0 arrays.** *Genes Chromosomes & Cancer* 2008, **47:**957-964.
65. Harada S, Henderson LB, Eshleman JR, Gocke CD, Burger P, Griffin CA, Batista DAS: **Genomic changes in gliomas detected using single nucleotide polymorphism array in formalin-fixed, paraffin-embedded tissue: superior results compared with microsatellite analysis.** *The Journal of molecular diagnostics : JMD* 2011, **13:**541-548.

66. Thompson ER, Herbert SC, Forrest SM, Campbell IG: **Whole genome SNP arrays using DNA derived from formalin-fixed, paraffin-embedded ovarian tumor tissue.** *Human Mutation* 2005, **26:**384-389.
67. Schweiger MR, Kerick M, Timmermann B, Albrecht MW, Borodina T, Parkhomchuk D, Zatloukal K, Lehrach H: **Genome-Wide Massively Parallel Sequencing of Formaldehyde Fixed-Paraffin Embedded (FFPE) Tumor Tissues for Copy-Number- and Mutation-Analysis.** *PloS one* 2009, **4:**e5548.
68. Yost SE, Smith EN, Schwab RB, Bao L, Jung H, Wang X, Voest E, Pierce JP, Messer K, Parker BA, et al: **Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens.** *Nucleic Acids Research* 2012, **40:**e107-e107.
69. Kerick M, Isau M, Timmermann B, Sültmann H, Herwig R, Krobitsch S, Schaefer G, Verdorfer I, Bartsch G, Klocker H, et al: **Targeted high throughput sequencing in clinical cancer Settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity.** *BMC Medical Genomics* 2011, **4:**68.
70. Wagle N, Berger MF, Davis MJ, Blumenstiel B, DeFelice M, Pochanard P, Ducar M, Van Hummelen P, Macconaill LE, Hahn WC, et al: **High-Throughput Detection of Actionable Genomic Alterations in Clinical Tumor Samples by Targeted, Massively Parallel Sequencing.** *Cancer discovery* 2012.
71. Oh E, Choi Y-L, Kwon MJ, Kim RN, Kim YJ, Song J-Y, Jung KS, Shin YK: **Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples.** *PloS one* 2015, **10:**e0144162.
72. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, Jane-Valbuena J, Friedrich DC, Kryukov G, Carter SL, et al: **Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine.** *Nature medicine* 2014, **20:**682-688.
73. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM: **Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis.** *BMC Bioinformatics* 2010, **11:**587.
74. Thirlwell C, Eymard M, Feber A, Teschendorff A, Pearce K, Lechner M, Widschwendter M, Beck S: **Genome-wide DNA methylation analysis of archival formalin-fixed paraffin-embedded tissue using the Illumina Infinium HumanMethylation27 BeadChip.** *Methods* 2010, **52:**248-254.
75. Dumenil TD, Wockner LF, Bettington M, McKeone DM, Klein K, Bowdler LM, Montgomery GW, Leggett BA, Whitehall VL: **Genome-wide DNA methylation analysis of formalin-fixed paraffin embedded colorectal cancer tissue.** *Genes Chromosomes Cancer* 2014, **53:**537-548.
76. de Ruijter TC, de Hoon JP, Slaats J, de Vries B, Janssen MJ, van Wezel T, Aarts MJ, van Engeland M, Tjan-Heijnen VC, Van Neste L, Veeck J: **Formalin-fixed, paraffin-embedded (FFPE) tissue epigenomics using Infinium HumanMethylation450 BeadChip assays.** *Lab Invest* 2015, **95:**833-842.
77. You JS, Jones PA: **Cancer genetics and epigenetics: two sides of the same coin?** *Cancer Cell* 2012, **22:**9-20.

78. Stratton MR, Campbell PJ, Futreal PA: **The cancer genome.** *Nature* 2009, **458:**719-724.
79. Esteller M: **Epigenetics in cancer.** *N Engl J Med* 2008, **358:**1148-1159.
80. Jones PA, Baylin SB: **The fundamental role of epigenetic events in cancer.** *Nat Rev Genet* 2002, **3:**415-428.
81. Herman JG, Baylin SB: **Gene silencing in cancer in association with promoter hypermethylation.** *N Engl J Med* 2003, **349:**2042-2054.
82. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, et al: **Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.** *Cancer Cell* 2010, **17:**98-110.
83. Sturm D, Witt H, Hovestadt V, Khuong-Quang DA, Jones DT, Konermann C, Pfaff E, Tonjes M, Sill M, Bender S, et al: **Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma.** *Cancer Cell* 2012, **22:**425-437.
84. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA: **Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.** *Bioinformatics* 2014, **30:**1363-1369.
85. Feber A, Guilhamon P, Lechner M, Fenton T, Wilson GA, Thirlwell C, Morris TJ, Flanagan AM, Teschendorff AE, Kelly JD, Beck S: **Using high-density DNA methylation arrays to profile copy number alterations.** *Genome Biol* 2014, **15:**R30.
86. Poncet P: **modeest: Mode Estimation. R package version 2.1.** 2012.
87. Seshan VE, Olshen A: **DNAcopy: DNA copy number data analysis. R package version 1.40.0.** 2012.
88. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G: **GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.** *Genome Biol* 2011, **12:**R41.
89. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M: **pROC: an open-source package for R and S+ to analyze and compare ROC curves.** *BMC Bioinformatics* 2011, **12:**77.
90. Shi G: **Multivariate data analysis in palaeoecology and palaeobiology -- a review.** *Palaeogeography, Palaeoclimatology Palaeoecology* 1993, **105:**199 - 234.
91. **locfit: Local Regression, Likelihood and Density Estimation** [http://CRAN.R-project.org/package=locfit]
92. Cancer Genome Atlas Research N: **Integrated genomic characterization of papillary thyroid carcinoma.** *Cell* 2014, **159:**676-690.
93. Cancer Genome Atlas Research N: **Comprehensive genomic characterization of squamous cell lung cancers.** *Nature* 2012, **489:**519-525.
94. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C: **Emerging landscape of oncogenic signatures across human cancers.** *Nat Genet* 2013, **45:**1127-1133.

95. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F: **Evaluation of the Infinium Methylation 450K technology.** *Epigenomics* 2011, **3:**771-784.
96. Olshen AB, Venkatraman ES, Lucito R, Wigler M: **Circular binary segmentation for the analysis of array-based DNA copy number data.** *Biostatistics* 2004, **5:**557-572.
97. G. S, B. T, J. G, Institute B: **Copy Number Inference Pipeline Documentation.** 2012.
98. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, et al: **Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants.** *Nat Biotechnol* 2011, **29:**512-520.
99. Baumbusch LO, Aaroe J, Johansen FE, Hicks J, Sun H, Bruhn L, Gunderson K, Naume B, Kristensen VN, Liestol K, et al: **Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors.** *BMC Genomics* 2008, **9:**379.
100. Curtis C, Lynch AG, Dunning MJ, Spiteri I, Marioni JC, Hadfield J, Chin SF, Brenton JD, Tavare S, Caldas C: **The pitfalls of platform comparison: DNA copy number array technologies assessed.** *BMC Genomics* 2009, **10:**588.
101. Capezzone M, Cantara S, Marchisotta S, Filetti S, De Santi MM, Rossi B, Ronga G, Durante C, Pacini F: **Short telomeres, telomerase reverse transcriptase gene amplification, and increased telomerase activity in the blood of familial papillary thyroid cancer patients.** *J Clin Endocrinol Metab* 2008, **93:**3950-3957.
102. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S: **ChAMP: 450k Chip Analysis Methylation Pipeline.** *Bioinformatics* 2014, **30:**428-430.
103. Howlader N, Noone A, Krapcho M, Neyman N, Aminou R, Altekruse S, Kosary C, Ruhl J, Tatalovich Z, Cho H: **SEER Cancer Statistics Review, 1975–2009 (Vintage 2009 Populations), National Cancer Institute. Bethesda, MD. Based on November 2011 SEER data submission, posted to the SEER web site, April 2012.** 2012.
104. Ernster VL, Ballard-Barbash R, Barlow WE, Zheng Y, Weaver DL, Cutter G, Yankaskas BC, Rosenberg R, Carney PA, Kerlikowske K, et al: **Detection of ductal carcinoma in situ in women undergoing screening mammography.** *J Natl Cancer Inst* 2002, **94:**1546-1554.
105. Bleyer A, Welch HG: **Effect of three decades of screening mammography on breast-cancer incidence.** *N Engl J Med* 2012, **367:**1998-2005.
106. Bijker N, Donker M, Wesseling J, den Heeten GJ, Rutgers EJ: **Is DCIS breast cancer, and how do I treat it?** *Curr Treat Options Oncol* 2013, **14:**75-87.
107. Allegra CJ, Aberle DR, Ganschow P, Hahn SM, Lee CN, Millon-Underwood S, Pike MC, Reed SD, Saftlas AF, Scarvalone SA, et al: **National Institutes of Health State-of-the-Science Conference statement: Diagnosis and Management of Ductal Carcinoma In Situ September 22-24, 2009.** *J Natl Cancer Inst* 2010, **102:**161-169.

108. Schwartz GF, Solin LJ, Olivotto IA, Ernster VL, Pressman PI: **Consensus Conference on the Treatment of In Situ Ductal Carcinoma of the Breast, April 22-25, 1999.** *Cancer* 2000, **88:**946-954.
109. Erbas B, Provenzano E, Armes J, Gertig D: **The natural history of ductal carcinoma in situ of the breast: a review.** *Breast Cancer Res Treat* 2006, **97:**135-144.
110. Independent UKPoBCS: **The benefits and harms of breast cancer screening: an independent review.** *Lancet* 2012, **380:**1778-1786.
111. Virnig BA, Tuttle TM, Shamliyan T, Kane RL: **Ductal carcinoma in situ of the breast: a systematic review of incidence, treatment, and outcomes.** *J Natl Cancer Inst* 2010, **102:**170-178.
112. Freedman GM: **Risk stratification in ductal carcinoma in situ: the role of genomic testing.** *Curr Oncol Rep* 2013, **15:**7-13.
113. Eusebi V: **Duct carcinoma in situ of the breast: an overview.** *Arkh Patol* 2011, **73:**26-29.
114. Sanders ME, Schuyler PA, Dupont WD, Page DL: **The natural history of low-grade ductal carcinoma in situ of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up.** *Cancer* 2005, **103:**2481-2484.
115. Page DL, Dupont WD, Rogers LW, Jensen RA, Schuyler PA: **Continued local recurrence of carcinoma 15-25 years after a diagnosis of low grade ductal carcinoma in situ of the breast treated only by biopsy.** *Cancer* 1995, **76:**1197-1200.
116. Collins LC, Tamimi RM, Baer HJ, Connolly JL, Colditz GA, Schnitt SJ: **Outcome of patients with ductal carcinoma in situ untreated after diagnostic biopsy: results from the Nurses' Health Study.** *Cancer* 2005, **103:**1778-1784.
117. Fisher B, Land S, Mamounas E, Dignam J, Fisher ER, Wolmark N: **Prevention of invasive breast cancer in women with ductal carcinoma in situ: an update of the National Surgical Adjuvant Breast and Bowel Project experience.** *Semin Oncol* 2001, **28:**400-418.
118. Fisher B, Anderson S, Bryant J, Margolese RG, Deutsch M, Fisher ER, Jeong JH, Wolmark N: **Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer.** *N Engl J Med* 2002, **347:**1233-1241.
119. Veronesi U, Cascinelli N, Mariani L, Greco M, Saccozzi R, Luini A, Aguilar M, Marubini E: **Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer.** *N Engl J Med* 2002, **347:**1227-1232.
120. Wapnir IL, Dignam JJ, Fisher B, Mamounas EP, Anderson SJ, Julian TB, Land SR, Margolese RG, Swain SM, Costantino JP, Wolmark N: **Long-term outcomes of invasive ipsilateral breast tumor recurrences after lumpectomy in NSABP B-17 and B-24 randomized clinical trials for DCIS.** *J Natl Cancer Inst* 2011, **103:**478-488.
121. Allred DC, Anderson SJ, Paik S, Wickerham DL, Nagtegaal ID, Swain SM, Mamounas EP, Julian TB, Geyer CE, Jr., Costantino JP, et al: **Adjuvant tamoxifen reduces subsequent breast cancer in women with estrogen**

receptor-positive ductal carcinoma in situ: a study based on NSABP protocol B-24. *J Clin Oncol* 2012, **30:**1268-1273.
122. Burstein HJ, Polyak K, Wong JS, Lester SC, Kaelin CM: **Ductal carcinoma in situ of the breast.** *N Engl J Med* 2004, **350:**1430-1441.
123. Kerlikowske K, Molinaro A, Cha I, Ljung BM, Ernster VL, Stewart K, Chew K, Moore DH, 2nd, Waldman F: **Characteristics associated with recurrence among women with ductal carcinoma in situ treated by lumpectomy.** *J Natl Cancer Inst* 2003, **95:**1692-1702.
124. Zhang X, Dai H, Liu B, Song F, Chen K: **Predictors for local invasive recurrence of ductal carcinoma in situ of the breast: a meta-analysis.** *Eur J Cancer Prev* 2016, **25:**19-28.
125. Silverstein MJ: **The University of Southern California/Van Nuys prognostic index for ductal carcinoma in situ of the breast.** *Am J Surg* 2003, **186:**337-343.
126. Di Saverio S, Catena F, Santini D, Ansaloni L, Fogacci T, Mignani S, Leone A, Gazzotti F, Gagliardi S, De Cataldis A, Taffurelli M: **259 Patients with DCIS of the breast applying USC/Van Nuys prognostic index: a retrospective review with long term follow up.** *Breast Cancer Res Treat* 2008, **109:**405-416.
127. MacAusland SG, Hepel JT, Chong FK, Galper SL, Gass JS, Ruthazer R, Wazer DE: **An attempt to independently verify the utility of the Van Nuys Prognostic Index for ductal carcinoma in situ.** *Cancer* 2007, **110:**2648-2653.
128. Leonard GD, Swain SM: **Ductal carcinoma in situ, complexities and challenges.** *J Natl Cancer Inst* 2004, **96:**906-920.
129. Ferguson AT, Evron E, Umbricht CB, Pandita TK, Chan TA, Hermeking H, Marks JR, Lambers AR, Futreal PA, Stampfer MR, Sukumar S: **High frequency of hypermethylation at the 14-3-3 sigma locus leads to gene silencing in breast cancer.** *Proc Natl Acad Sci U S A* 2000, **97:**6049-6054.
130. Umbricht CB, Evron E, Gabrielson E, Ferguson A, Marks J, Sukumar S: **Hypermethylation of 14-3-3 sigma (stratifin) is an early event in breast cancer.** *Oncogene* 2001, **20:**3348-3353.
131. Allred DC: **Ductal carcinoma in situ: terminology, classification, and natural history.** *J Natl Cancer Inst Monogr* 2010, **2010:**134-138.
132. Abba MC, Gong T, Lu Y, Lee J, Zhong Y, Lacunza E, Butti M, Takata Y, Gaddis S, Shen J, et al: **A Molecular Portrait of High-Grade Ductal Carcinoma In Situ.** *Cancer Res* 2015, **75:**3980-3990.
133. Vincent-Salomon A, Lucchesi C, Gruel N, Raynal V, Pierron G, Goudefroye R, Reyal F, Radvanyi F, Salmon R, Thiery JP, et al: **Integrated genomic and transcriptomic analysis of ductal carcinoma in situ of the breast.** *Clin Cancer Res* 2008, **14:**1956-1965.
134. Muggerud AA, Hallett M, Johnsen H, Kleivi K, Zhou W, Tahmasebpoor S, Amini RM, Botling J, Borresen-Dale AL, Sorlie T, Warnberg F: **Molecular diversity in ductal carcinoma in situ (DCIS) and early invasive breast cancer.** *Mol Oncol* 2010, **4:**357-368.
135. Collins LC, Schnitt SJ: **HER2 protein overexpression in estrogen receptor-positive ductal carcinoma in situ of the breast: frequency and implications for tamoxifen therapy.** *Mod Pathol* 2005, **18:**615-620.

136. Cornfield DB, Palazzo JP, Schwartz GF, Goonewardene SA, Kovatich AJ, Chervoneva I, Hyslop T, Schwarting R: **The prognostic significance of multiple morphologic features and biologic markers in ductal carcinoma in situ of the breast: a study of a large cohort of patients treated with surgery alone.** *Cancer* 2004, **100:**2317-2327.
137. Somerville JE, Clarke LA, Biggart JD: **c-erbB-2 overexpression and histological type of in situ and invasive breast carcinoma.** *J Clin Pathol* 1992, **45:**16-20.
138. Siziopikou KP, Anderson SJ, Cobleigh MA, Julian TB, Arthur DW, Zheng P, Mamounas EP, Pajon ER, Behrens RJ, Eakle JF, et al: **Preliminary results of centralized HER2 testing in ductal carcinoma in situ (DCIS): NSABP B-43.** *Breast Cancer Res Treat* 2013, **142:**415-421.
139. Ma XJ, Salunga R, Tuggle JT, Gaudet J, Enright E, McQuary P, Payette T, Pistone M, Stecker K, Zhang BM, et al: **Gene expression profiles of human breast cancer progression.** *Proc Natl Acad Sci U S A* 2003, **100:**5974-5979.
140. Cowell CF, Weigelt B, Sakr RA, Ng CK, Hicks J, King TA, Reis-Filho JS: **Progression from ductal carcinoma in situ to invasive breast cancer: revisited.** *Mol Oncol* 2013, **7:**859-869.
141. Hannemann J, Velds A, Halfwerk JB, Kreike B, Peterse JL, van de Vijver MJ: **Classification of ductal carcinoma in situ by gene expression profiling.** *Breast Cancer Res* 2006, **8:**R61.
142. Elias EV, de Castro NP, Pineda PH, Abuazar CS, Bueno de Toledo Osorio CA, Pinilla MG, da Silva SD, Camargo AA, Silva WA, Jr., EN EF, et al: **Epithelial cells captured from ductal carcinoma in situ reveal a gene expression signature associated with progression to invasive breast cancer.** *Oncotarget* 2016.
143. Kerlikowske K, Molinaro AM, Gauthier ML, Berman HK, Waldman F, Bennington J, Sanchez H, Jimenez C, Stewart K, Chew K, et al: **Biomarker expression and risk of subsequent tumors after initial ductal carcinoma in situ diagnosis.** *J Natl Cancer Inst* 2010, **102:**627-637.
144. Polyak K: **Molecular markers for the diagnosis and management of ductal carcinoma in situ.** *J Natl Cancer Inst Monogr* 2010, **2010:**210-213.
145. Timp W, Bravo HC, McDonald OG, Goggins M, Umbricht C, Zeiger M, Feinberg AP, Irizarry RA: **Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors.** *Genome Med* 2014, **6:**61.
146. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, et al: **Increased methylation variation in epigenetic domains across cancer types.** *Nat Genet* 2011, **43:**768-775.
147. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, R VL, Clark SJ, Molloy PL: **De novo identification of differentially methylated regions in the human genome.** *Epigenetics Chromatin* 2015, **8:**6.
148. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA: **Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies.** *Int J Epidemiol* 2012, **41:**200-209.
149. Jovanovic J, Ronneberg JA, Tost J, Kristensen V: **The epigenetics of breast cancer.** *Mol Oncol* 2010, **4:**242-254.

150. Shah N, Sukumar S: **The Hox genes and their roles in oncogenesis.** *Nat Rev Cancer* 2010, **10:**361-371.
151. Jin K, Sukumar S: **HOX genes: Major actors in resistance to selective endocrine response modifiers.** *Biochim Biophys Acta* 2016, **1865:**105-110.
152. Jin K, Sukumar S: **BRCA1: linking HOX to breast cancer suppression.** *Breast Cancer Res* 2010, **12:**306.
153. Chen H, Sukumar S: **HOX genes: emerging stars in cancer.** *Cancer Biol Ther* 2003, **2:**524-525.
154. Teo WW, Merino VF, Cho S, Korangath P, Liang X, Wu RC, Neumann NM, Ewald AJ, Sukumar S: **HOXA5 determines cell fate transition and impedes tumor initiation and progression in breast cancer through regulation of E-cadherin and CD24.** *Oncogene* 2016.
155. Shah N, Jin K, Cruz LA, Park S, Sadik H, Cho S, Goswami CP, Nakshatri H, Gupta R, Chang HY, et al: **HOXB13 mediates tamoxifen resistance and invasiveness in human breast cancer by suppressing ERalpha and inducing IL-6 expression.** *Cancer Res* 2013, **73:**5449-5458.
156. Jin K, Park S, Teo WW, Korangath P, Cho SS, Yoshida T, Gyorffy B, Goswami CP, Nakshatri H, Cruz LA, et al: **HOXB7 Is an ERalpha Cofactor in the Activation of HER2 and Multiple ER Target Genes Leading to Endocrine Resistance.** *Cancer Discov* 2015, **5:**944-959.
157. Ma XJ, Dahiya S, Richardson E, Erlander M, Sgroi DC: **Gene expression profiling of the tumor microenvironment during breast cancer progression.** *Breast Cancer Res* 2009, **11:**R7.
158. Dotto GP: **Multifocal epithelial tumors and field cancerization: stroma as a primary determinant.** *J Clin Invest* 2014, **124:**1446-1453.
159. Ellsworth DL, Ellsworth RE, Love B, Deyarmin B, Lubert SM, Mittal V, Shriver CD: **Genomic patterns of allelic imbalance in disease free tissue adjacent to primary breast carcinomas.** *Breast Cancer Res Treat* 2004, **88:**131-139.
160. Teschendorff AE, Gao Y, Jones A, Ruebner M, Beckmann MW, Wachter DL, Fasching PA, Widschwendter M: **DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer.** *Nat Commun* 2016, **7:**10478.
161. Holst CR, Nuovo GJ, Esteller M, Chew K, Baylin SB, Herman JG, Tlsty TD: **Methylation of p16(INK4a) promoters occurs in vivo in histologically normal human mammary epithelia.** *Cancer Res* 2003, **63:**1596-1601.
162. Trujillo KA, Heaphy CM, Mai M, Vargas KM, Jones AC, Vo P, Butler KS, Joste NE, Bisoffi M, Griffith JK: **Markers of fibrosis and epithelial to mesenchymal transition demonstrate field cancerization in histologically normal tissue adjacent to breast tumors.** *Int J Cancer* 2011, **129:**1310-1321.
163. Heaphy CM, Bisoffi M, Fordyce CA, Haaland CM, Hines WC, Joste NE, Griffith JK: **Telomere DNA content and allelic imbalance demonstrate field cancerization in histologically normal tissue adjacent to breast tumors.** *Int J Cancer* 2006, **119:**108-116.
164. Reddington JP, Sproul D, Meehan RR: **DNA methylation reprogramming in cancer: does it act by re-configuring the binding landscape of Polycomb repressive complexes?** *Bioessays* 2014, **36:**134-140.

165. Colleoni M, Cole BF, Viale G, Regan MM, Price KN, Maiorano E, Mastropasqua MG, Crivellari D, Gelber RD, Goldhirsch A, et al: **Classical cyclophosphamide, methotrexate, and fluorouracil chemotherapy is more effective in triple-negative, node-negative breast cancer: results from two randomized trials of adjuvant chemoendocrine therapy for node-negative breast cancer.** *J Clin Oncol* 2010, **28:**2966-2973.
166. MacDonald D: **Kuerer's Breast Surgical Oncology.** 2010.
167. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747-752.
168. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thurlimann B, Senn HJ, Panel m: **Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013.** *Ann Oncol* 2013, **24:**2206-2223.
169. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA: **Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies.** *J Clin Invest* 2011, **121:**2750-2767.
170. Pietri E, Conteduca V, Andreis D, Massa I, Melegari E, Sarti S, Cecconetto L, Schirone A, Bravaccini S, Serra P, et al: **Androgen receptor signaling pathways as a target for breast cancer treatment.** *Endocr Relat Cancer* 2016, **23:**R485-498.
171. Bueno-de-Mesquita JM, Sonke GS, van de Vijver MJ, Linn SC: **Additional value and potential use of the 70-gene prognosis signature in node-negative breast cancer in daily clinical practice.** *Ann Oncol* 2011, **22:**2021-2030.
172. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlen SE, Greco D, Soderhall C, Scheynius A, Kere J: **Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility.** *PLoS One* 2012, **7:**e41361.
173. Anderson WF, Chatterjee N, Ershler WB, Brawley OW: **Estrogen receptor breast cancer phenotypes in the Surveillance, Epidemiology, and End Results database.** *Breast Cancer Res Treat* 2002, **76:**27-36.
174. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.
175. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP: **GSEA-P: a desktop application for Gene Set Enrichment Analysis.** *Bioinformatics* 2007, **23:**3251-3253.
176. Boutros C, Tarhini A, Routier E, Lambotte O, Ladurie FL, Carbonnel F, Izzeddine H, Marabelle A, Champiat S, Berdelou A, et al: **Safety profiles of anti-CTLA-4 and anti-PD-1 antibodies alone and in combination.** *Nat Rev Clin Oncol* 2016, **13:**473-486.

177. O'Grady TJ, Gates MA, Boscoe FP: **Thyroid cancer incidence attributable to overdiagnosis in the United States 1981-2011.** *Int J Cancer* 2015, **137:**2664-2673.
178. Welch HG, Black WC: **Overdiagnosis in cancer.** *J Natl Cancer Inst* 2010, **102:**605-613.
179. Davies L, Welch HG: **Increasing incidence of thyroid cancer in the United States, 1973-2002.** *JAMA* 2006, **295:**2164-2167.
180. McLeod DS, Sawka AM, Cooper DS: **Controversies in primary treatment of low-risk papillary thyroid cancer.** *Lancet* 2013, **381:**1046-1057.
181. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, et al: **2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer.** *Thyroid* 2016, **26:**1-133.
182. Nikiforov YE, Nikiforova MN: **Molecular genetics and diagnosis of thyroid cancer.** *Nat Rev Endocrinol* 2011, **7:**569-580.
183. Eberhardt NL, Grebe SK, McIver B, Reddi HV: **The role of the PAX8/PPARgamma fusion oncogene in the pathogenesis of follicular thyroid cancer.** *Mol Cell Endocrinol* 2010, **321:**50-56.
184. Cerutti JM: **Employing genetic markers to improve diagnosis of thyroid tumor fine needle biopsy.** *Curr Genomics* 2011, **12:**589-596.
185. Durante C, Haddy N, Baudin E, Leboulleux S, Hartl D, Travagli JP, Caillou B, Ricard M, Lumbroso JD, De Vathaire F, Schlumberger M: **Long-term outcome of 444 patients with distant metastases from papillary and follicular thyroid carcinoma: benefits and limits of radioiodine therapy.** *J Clin Endocrinol Metab* 2006, **91:**2892-2899.
186. Sinicropi D, Qu K, Collin F, Crager M, Liu ML, Pelham RJ, Pho M, Dei Rossi A, Jeong J, Scott A, et al: **Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue.** *PLoS One* 2012, **7:**e40092.
187. Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM: **Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling.** *BMC Genomics* 2014, **15:**419.
188. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol* 2013, **14:**R36.
189. Song L, Sabunciyan S, Florea L: **CLASS2: accurate and efficient splice variant annotation from RNA-seq reads.** *Nucleic Acids Res* 2016.
190. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq.** *Nat Biotechnol* 2013, **31:**46-53.
191. Shen S, Park JW, Huang J, Dittmar KA, Lu ZX, Zhou Q, Carstens RP, Xing Y: **MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data.** *Nucleic Acids Res* 2012, **40:**e61.

192. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer.** *Nat Biotechnol* 2011, **29:**24-26.
193. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res* 2015, **43:**e47.
194. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P: **The Molecular Signatures Database (MSigDB) hallmark gene set collection.** *Cell Syst* 2015, **1:**417-425.
195. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0.** *Bioinformatics* 2011, **27:**1739-1740.
196. DeLong ER, DeLong DM, Clarke-Pearson DL: **Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.** *Biometrics* 1988, **44:**837-845.
197. **Broad Institute: Picard tools** [http://broadinstitute.github.io/picard/]
198. Miron B. Kursa WRR: **Feature Selection with the Boruta Package.** *Journal of Statistical Software* 2010, **36:**1 - 13.
199. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
200. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20:**1297-1303.
201. Li C, Aragon Han P, Lee KC, Lee LC, Fox AC, Beninato T, Thiess M, Dy BM, Sebo TJ, Thompson GB, et al: **Does BRAF V600E mutation predict aggressive features in papillary thyroid cancer? Results from four endocrine surgery centers.** *J Clin Endocrinol Metab* 2013, **98:**3702-3712.
202. An JH, Song KH, Kim SK, Park KS, Yoo YB, Yang JH, Hwang TS, Kim DL: **RAS mutations in indeterminate thyroid nodules are predictive of the follicular variant of papillary thyroid carcinoma.** *Clin Endocrinol (Oxf)* 2014.
203. Chen G, Olson MT, O'Neill A, Norris A, Beierl K, Harada S, Debeljak M, Rivera-Roman K, Finley S, Stafford A, et al: **A virtual pyrogram generator to resolve complex pyrosequencing results.** *J Mol Diagn* 2012, **14:**149-159.
204. Jarzab B, Wiench M, Fujarewicz K, Simek K, Jarzab M, Oczko-Wojciechowska M, Wloch J, Czarniecka A, Chmielik E, Lange D, et al: **Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications.** *Cancer Res* 2005, **65:**1587-1597.
205. Kim HS, Kim do H, Kim JY, Jeoung NH, Lee IK, Bong JG, Jung ED: **Microarray analysis of papillary thyroid cancers in Korean.** *Korean J Intern Med* 2010, **25:**399-407.
206. Galeza-Kulik M, Zebracka J, Szpak-Ulczok S, Czarniecka AK, Kukulska A, Gubala E, Stojcev Z, Wiench M: **[Expression of selected genes involved in transport of ions in papillary thyroid carcinoma].** *Endokrynol Pol* 2006, **57 Suppl A:**26-31.

207. Lin K, Rubinfeld B, Zhang C, Firestein R, Harstad E, Roth L, Tsai SP, Schutten M, Xu K, Hristopoulos M, Polakis P: **Preclinical Development of an Anti-NaPi2b (SLC34A2) Antibody-Drug Conjugate as a Therapeutic for Non-Small Cell Lung and Ovarian Cancers.** *Clin Cancer Res* 2015, **21:**5139-5150.
208. Nikolova DN, Zembutsu H, Sechanov T, Vidinov K, Kee LS, Ivanova R, Becheva E, Kocova M, Toncheva D, Nakamura Y: **Genome-wide gene expression profiles of thyroid carcinoma: Identification of molecular targets for treatment of thyroid carcinoma.** *Oncol Rep* 2008, **20:**105-121.
209. Prasad NB, Somervell H, Tufano RP, Dackiw AP, Marohn MR, Califano JA, Wang Y, Westra WH, Clark DP, Umbricht CB, et al: **Identification of genes differentially expressed in benign versus malignant thyroid tumors.** *Clin Cancer Res* 2008, **14:**3327-3337.
210. Vierlinger K, Mansfeld MH, Koperek O, Nohammer C, Kaserer K, Leisch F: **Identification of SERPINA1 as single marker for papillary thyroid carcinoma through microarray meta analysis and quantification of its discriminatory power in independent validation.** *BMC Med Genomics* 2011, **4:**30.
211. Arora N, Scognamiglio T, Lubitz CC, Moo TA, Kato MA, Zhu B, Zarnegar R, Chen YT, Fahey TJ, 3rd: **Identification of borderline thyroid tumors by gene expression array analysis.** *Cancer* 2009, **115:**5421-5431.
212. Lubitz CC, Ugras SK, Kazam JJ, Zhu B, Scognamiglio T, Chen YT, Fahey TJ, 3rd: **Microarray analysis of thyroid nodule fine-needle aspirates accurately classifies benign and malignant lesions.** *J Mol Diagn* 2006, **8:**490-498; quiz 528.
213. Barros-Filho MC, Marchi FA, Pinto CA, Rogatto SR, Kowalski LP: **High Diagnostic Accuracy Based on CLDN10, HMGA2, and LAMB3 Transcripts in Papillary Thyroid Carcinoma.** *J Clin Endocrinol Metab* 2015, **100:**E890-899.
214. Schulten HJ, Al-Mansouri Z, Baghallab I, Bagatian N, Subhi O, Karim S, Al-Aradati H, Al-Mutawa A, Johary A, Meccawy AA, et al: **Comparison of microarray expression profiles between follicular variant of papillary thyroid carcinomas and follicular adenomas of the thyroid.** *BMC Genomics* 2015, **16 Suppl 1:**S7.
215. Monica K, Galili N, Nourse J, Saltman D, Cleary ML: **PBX2 and PBX3, new homeobox genes with extensive homology to the human proto-oncogene PBX1.** *Mol Cell Biol* 1991, **11:**6149-6157.
216. Han HB, Gu J, Ji DB, Li ZW, Zhang Y, Zhao W, Wang LM, Zhang ZQ: **PBX3 promotes migration and invasion of colorectal cancer cells via activation of MAPK/ERK signaling pathway.** *World J Gastroenterol* 2014, **20:**18260-18270.
217. Han HB, Gu J, Zuo HJ, Chen ZG, Zhao W, Li M, Ji DB, Lu YY, Zhang ZQ: **Let-7c functions as a metastasis suppressor by targeting MMP11 and PBX3 in colorectal cancer.** *J Pathol* 2012, **226:**544-555.
218. Weber F, Shen L, Aldred MA, Morrison CD, Frilling A, Saji M, Schuppert F, Broelsch CE, Ringel MD, Eng C: **Genetic classification of benign and malignant thyroid follicular neoplasia based on a three-gene combination.** *J Clin Endocrinol Metab* 2005, **90:**2512-2521.

219. Puskas LG, Juhasz F, Zarva A, Hackler L, Jr., Farid NR: **Gene profiling identifies genes specific for well-differentiated epithelial thyroid tumors.** *Cell Mol Biol (Noisy-le-grand)* 2005, **51:**177-186.
220. Faibish M, Francescone R, Bentley B, Yan W, Shao R: **A YKL-40-neutralizing antibody blocks tumor angiogenesis and progression: a potential therapeutic agent in cancers.** *Mol Cancer Ther* 2011, **10:**742-751.
221. Francescone RA, Scully S, Faibish M, Taylor SL, Oh D, Moral L, Yan W, Bentley B, Shao R: **Role of YKL-40 in the angiogenesis, radioresistance, and progression of glioblastoma.** *J Biol Chem* 2011, **286:**15332-15343.
222. Silvestre JS, Thery C, Hamard G, Boddaert J, Aguilar B, Delcayre A, Houbron C, Tamarat R, Blanc-Brude O, Heeneman S, et al: **Lactadherin promotes VEGF-dependent neovascularization.** *Nat Med* 2005, **11:**499-506.
223. Finn SP, Smyth P, Cahill S, Streck C, O'Regan EM, Flavin R, Sherlock J, Howells D, Henfrey R, Cullen M, et al: **Expression microarray analysis of papillary thyroid carcinoma and benign thyroid tissue: emphasis on the follicular variant and potential markers of malignancy.** *Virchows Arch* 2007, **450:**249-260.
224. Neutzner M, Lopez T, Feng X, Bergmann-Leitner ES, Leitner WW, Udey MC: **MFG-E8/lactadherin promotes tumor growth in an angiogenesis-dependent transgenic mouse model of multistage carcinogenesis.** *Cancer Res* 2007, **67:**6777-6785.
225. Bianco AC, Kim BW: **Deiodinases: implications of the local control of thyroid hormone action.** *J Clin Invest* 2006, **116:**2571-2579.
226. Aldred MA, Huang Y, Liyanarachchi S, Pellegata NS, Gimm O, Jhiang S, Davuluri RV, de la Chapelle A, Eng C: **Papillary and follicular thyroid carcinomas show distinctly different microarray expression profiles and can be distinguished by a minimum of five genes.** *J Clin Oncol* 2004, **22:**3531-3539.
227. Arnaldi LA, Borra RC, Maciel RM, Cerutti JM: **Gene expression profiles reveal that DCN, DIO1, and DIO2 are underexpressed in benign and malignant thyroid tumors.** *Thyroid* 2005, **15:**210-221.
228. Fryknas M, Wickenberg-Bolin U, Goransson H, Gustafsson MG, Foukakis T, Lee JJ, Landegren U, Hoog A, Larsson C, Grimelius L, et al: **Molecular markers for discrimination of benign and malignant follicular thyroid tumors.** *Tumour Biol* 2006, **27:**211-220.
229. Huang Y, Prasad M, Lemon WJ, Hampel H, Wright FA, Kornacker K, LiVolsi V, Frankel W, Kloos RT, Eng C, et al: **Gene expression in papillary thyroid carcinoma reveals highly consistent profiles.** *Proc Natl Acad Sci U S A* 2001, **98:**15044-15049.
230. Mineva I, Gartner W, Hauser P, Kainz A, Loffler M, Wolf G, Oberbauer R, Weissel M, Wagner L: **Differential expression of alphaB-crystallin and Hsp27-1 in anaplastic thyroid carcinomas because of tumor-specific alphaB-crystallin gene (CRYAB) silencing.** *Cell Stress Chaperones* 2005, **10:**171-184.
231. Babu E, Ramachandran S, CoothanKandaswamy V, Elangovan S, Prasad PD, Ganapathy V, Thangaraju M: **Role of SLC5A8, a plasma membrane**

transporter and a tumor suppressor, in the antitumor activity of dichloroacetate. *Oncogene* 2011, **30:**4026-4037.

232. Porra V, Ferraro-Peyret C, Durand C, Selmi-Ruby S, Giroud H, Berger-Dutrieux N, Decaussin M, Peix JL, Bournaud C, Orgiazzi J, et al: **Silencing of the tumor suppressor gene SLC5A8 is associated with BRAF mutations in classical papillary thyroid carcinomas.** *J Clin Endocrinol Metab* 2005, **90:**3028-3035.
233. Hu S, Liu D, Tufano RP, Carson KA, Rosenbaum E, Cohen Y, Holt EH, Kiseljak-Vassiliades K, Rhoden KJ, Tolaney S, et al: **Association of aberrant methylation of tumor suppressor genes with tumor aggressiveness and BRAF mutation in papillary thyroid cancer.** *Int J Cancer* 2006, **119:**2322-2329.
234. Ganapathy V, Gopal E, Miyauchi S, Prasad PD: **Biological functions of SLC5A8, a candidate tumour suppressor.** *Biochem Soc Trans* 2005, **33:**237-240.
235. Zane M, Agostini M, Enzo MV, Casal Ide E, Del Bianco P, Torresan F, Merante Boschin I, Pennelli G, Saccani A, Rubello D, et al: **Circulating cell-free DNA, SLC5A8 and SLC26A4 hypermethylation, BRAF(V600E): A non-invasive tool panel for early detection of thyroid cancer.** *Biomed Pharmacother* 2013, **67:**723-730.
236. Giordano TJ, Au AY, Kuick R, Thomas DG, Rhodes DR, Wilhelm KG, Jr., Vinco M, Misek DE, Sanders D, Zhu Z, et al: **Delineation, functional validation, and bioinformatic evaluation of gene expression in thyroid follicular carcinomas with the PAX8-PPARG translocation.** *Clin Cancer Res* 2006, **12:**1983-1993.
237. Raman P, Koenig RJ: **Pax-8-PPAR-gamma fusion protein in thyroid carcinoma.** *Nat Rev Endocrinol* 2014, **10:**616-623.
238. Abbosh PH, Nephew KP: **Multiple signaling pathways converge on beta-catenin in thyroid cancer.** *Thyroid* 2005, **15:**551-561.
239. Marques AR, Espadinha C, Frias MJ, Roque L, Catarino AL, Sobrinho LG, Leite V: **Underexpression of peroxisome proliferator-activated receptor (PPAR)gamma in PAX8/PPARgamma-negative thyroid tumours.** *Br J Cancer* 2004, **91:**732-738.
240. Espadinha C, Pinto AE, Leite V: **Underexpression of PPARgamma is associated with aneuploidy and lower differentiation of thyroid tumours of follicular origin.** *Oncol Rep* 2009, **22:**907-913.
241. Huang M, Prendergast GC: **RhoB in cancer suppression.** *Histol Histopathol* 2006, **21:**213-218.
242. Ichijo S, Furuya F, Shimura H, Hayashi Y, Takahashi K, Ohta K, Kobayashi T, Kitamura K: **Activation of the RhoB signaling pathway by thyroid hormone receptor beta in thyroid cancer cells.** *PLoS One* 2014, **9:**e116252.
243. Marlow LA, Reynolds LA, Cleland AS, Cooper SJ, Gumz ML, Kurakata S, Fujiwara K, Zhang Y, Sebo T, Grant C, et al: **Reactivation of suppressed RhoB is a critical step for the inhibition of anaplastic thyroid cancer growth.** *Cancer Res* 2009, **69:**1536-1544.
244. Jiang K, Sun J, Cheng J, Djeu JY, Wei S, Sebti S: **Akt mediates Ras downregulation of RhoB, a suppressor of transformation, invasion, and metastasis.** *Mol Cell Biol* 2004, **24:**5565-5576.

245. Lamouille S, Xu J, Derynck R: **Molecular mechanisms of epithelial-mesenchymal transition.** *Nat Rev Mol Cell Biol* 2014, **15:**178-196.
246. Plante I, Stewart MK, Barr K, Allan AL, Laird DW: **Cx43 suppresses mammary tumor metastasis to the lung in a Cx43 mutant mouse model of human disease.** *Oncogene* 2011, **30:**1681-1692.
247. Lee YH, Schiemann WP: **Fibromodulin suppresses nuclear factor-kappaB activity by inducing the delayed degradation of IKBA via a JNK-dependent pathway coupled to fibroblast apoptosis.** *J Biol Chem* 2011, **286:**6414-6422.
248. Anastasi G, Cutroneo G, Rizzo G, Favaloro A: **Sarcoglycan subcomplex in normal and pathological human muscle fibers.** *Eur J Histochem* 2007, **51 Suppl 1:**29-33.
249. Cutroneo G, Bramanti P, Favaloro A, Anastasi G, Trimarchi F, Di Mauro D, Rinaldi C, Speciale F, Inferrera A, Santoro G, et al: **Sarcoglycan complex in human normal and pathological prostatic tissue: an immunohistochemical and RT-PCR study.** *Anat Rec (Hoboken)* 2014, **297:**327-336.
250. Hiroshima Y, Nakamura F, Miyamoto H, Mori R, Taniguchi K, Matsuyama R, Akiyama H, Tanaka K, Ichikawa Y, Kato S, et al: **Collapsin response mediator protein 4 expression is associated with liver metastasis and poor survival in pancreatic cancer.** *Ann Surg Oncol* 2013, **20 Suppl 3:**S369-378.
251. Kanda M, Nomoto S, Oya H, Shimizu D, Takami H, Hibino S, Hashimoto R, Kobayashi D, Tanaka C, Yamada S, et al: **Dihydropyrimidinase-like 3 facilitates malignant behavior of gastric cancer.** *J Exp Clin Cancer Res* 2014, **33:**66.
252. Kawahara T, Hotta N, Ozawa Y, Kato S, Kano K, Yokoyama Y, Nagino M, Takahashi T, Yanagisawa K: **Quantitative proteomic profiling identifies DPYSL3 as pancreatic ductal adenocarcinoma-associated molecule that regulates cell adhesion and migration by stabilization of focal adhesion complex.** *PLoS One* 2013, **8:**e79654.
253. Chen S, Zhang X, Peng J, Zhai E, He Y, Wu H, Chen C, Ma J, Wang Z, Cai S: **VEGF promotes gastric cancer development by upregulating CRMP4.** *Oncotarget* 2016.
254. Chambers AF, Groom AC, MacDonald IC: **Dissemination and growth of cancer cells in metastatic sites.** *Nat Rev Cancer* 2002, **2:**563-572.
255. Yen YC, Hsiao JR, Jiang SS, Chang JS, Wang SH, Shen YY, Chen CH, Chang IS, Chang JY, Chen YW: **Insulin-like growth factor-independent insulin-like growth factor binding protein 3 promotes cell migration and lymph node metastasis of oral squamous cell carcinoma cells by requirement of integrin beta1.** *Oncotarget* 2015, **6:**41837-41855.
256. Natsuizaka M, Ohashi S, Wong GS, Ahmadi A, Kalman RA, Budo D, Klein-Szanto AJ, Herlyn M, Diehl JA, Nakagawa H: **Insulin-like growth factor-binding protein-3 promotes transforming growth factor-{beta}1-mediated epithelial-to-mesenchymal transition and motility in transformed human esophageal cells.** *Carcinogenesis* 2010, **31:**1344-1353.
257. Yang YJ, Na HJ, Suh MJ, Ban MJ, Byeon HK, Kim WS, Kim JW, Choi EC, Kwon HJ, Chang JW, Koh YW: **Hypoxia Induces Epithelial-Mesenchymal**

Transition in Follicular Thyroid Cancer: Involvement of Regulation of Twist by Hypoxia Inducible Factor-1alpha. *Yonsei Med J* 2015, **56:**1503-1514.
258. Malaguarnera R, Vella V, Vigneri R, Frasca F: **p53 family proteins in thyroid cancer.** *Endocr Relat Cancer* 2007, **14:**43-60.
259. Fagin JA, Matsuo K, Karmakar A, Chen DL, Tang SH, Koeffler HP: **High prevalence of mutations of the p53 gene in poorly differentiated human thyroid carcinomas.** *J Clin Invest* 1993, **91:**179-184.
260. Magri F, Capelli V, Rotondi M, Leporati P, La Manna L, Ruggiero R, Malovini A, Bellazzi R, Villani L, Chiovato L: **Expression of estrogen and androgen receptors in differentiated thyroid cancer: an additional criterion to assess the patient's risk.** *Endocr Relat Cancer* 2012, **19:**463-471.
261. Li Y, Huang J, Zhao YL, He J, Wang W, Davies KE, Nose V, Xiao S: **UTRN on chromosome 6q24 is mutated in multiple tumors.** *Oncogene* 2007, **26:**6220-6228.
262. Cornen S, Guille A, Adelaide J, Addou-Klouche L, Finetti P, Saade MR, Manai M, Carbuccia N, Bekhouche I, Letessier A, et al: **Candidate luminal B breast cancer genes identified by genome, gene expression and DNA methylation profiling.** *PLoS One* 2014, **9:**e81843.
263. Zhang WL, Lv W, Sun SZ, Wu XZ, Zhang JH: **miR-206 inhibits metastasis-relevant traits by degrading MRTF-A in anaplastic thyroid cancer.** *Int J Oncol* 2015, **47:**133-142.
264. Wu X, Kodama A, Fuchs E: **ACF7 regulates cytoskeletal-focal adhesion dynamics and migration and has ATPase activity.** *Cell* 2008, **135:**137-148.
265. Misquitta-Ali CM, Cheng E, O'Hanlon D, Liu N, McGlade CJ, Tsao MS, Blencowe BJ: **Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer.** *Mol Cell Biol* 2011, **31:**138-150.

Appendix

I: Protocols

Optimized Protocol for processing FFPE tissue for RNA/DNA extraction

This protocol uses the AllPrep-DNA/RNA FFPE kit (#80234). For deparaffinization and initial tissue dissection, refer to part (A) for cores and (B) for tissue sections. Continue with RNA purification once completed.

A) FFPE Cores (2 cores per tube, ~ 0.6mm X 2 mm)
Deparaffinize

1. Fill each 1.5 ml microtube with tissue cores (transfer them dry with a tweezers or in some xylene) and approximately 1 ml of PCR-clean xylene total. Pulse vortex 15 sec on high. Incubate 10 min. Centrifuge 1 min. at full speed, room temp. Pipette off xylene. Leave about 20 microliters and don't suck up the sample!
2. Repeat twice for a total of 3x 10 min incubations, totaling 30 min. Attempt to pipette off all xylene.
3. Add 1 ml 100% ethanol. Pulse vortex 15 sec., centrifuge 1 min full speed. Pipette off last couple microliters of ethanol. Open cap and air dry 10 min.

Digest the core tissue

4. Add to the samples 150 µl PKD (supplied) + 10 µl proteinase K (20 mg/ml; supplied), flick to mix. Incubate at 56^{o}C for 3 hr total. When the incubation time is complete, place the samples on ice for 3 minutes.
5. Centrifuge at 20,000 x g for 15 min at 4^{o}C.
6. Transfer the RNA-containing supernatant to a 2 ml low binding Eppendorf microcentrifuge tube (can freeze at ≤ -80^{o}C for up to 1 week). Can also freeze the tissue pellet at ≤ -80^{o}C for later DNA extraction.

B) FFPE Sections (5 sections, 10um each, max 100mm^2)
Deparaffinize

1. Using a clean surgical blade, macro-dissect tissue to enrich for lesion of interest (> 70%) and transfer tissue into 1.7mL Eppendorf tube. Add 1mL of xylene. Pulse vortex 15 sec., and incubate at RT for 10 minutes. Centrifuge 5 min full speed. Pipette off xylene.
2. Repeat twice.
3. Add 1 ml 100% ethanol. Pulse vortex 15 sec., centrifuge 5 min full speed. Pipette off last couple microliters of ethanol. Open cap and air dry 10 min.

Digest the tissue

4. Add to the samples 150 µl PKD (supplied) + 10 µl proteinase K (20 mg/ml; supplied), flick to mix. Incubate at 56^{o}C for 1 hr. When the incubation time is complete, place the samples on ice for 3 minutes.
5. Centrifuge at 20,000 x g for 15 min at 4^{o}C.

6. Transfer the RNA-containing supernatant to a 2 ml low binding Eppendorf microcentrifuge tube (can freeze at ≤ -80°C for up to 1 week). Can also freeze the tissue pellet at ≤ -80°C for later DNA extraction.

A&B)
Purification of RNA

7. Equilibrate the RNA supernatant to room temperature. Preheat a block to 80°C. Transfer the RNA tube to 80°C for exactly 15 min in the preheated heating block.
8. Immediately chill on ice.
9. Processing all tubes at once at room temperature, add 320 µl Buffer RLT (binding buffer, supplied), mix by pipetting gently.
10. Add 1120 µl absolute ethanol (100%), mix by pipetting.
11. Transfer 700 µl to RNeasy MinElute spin (supplied) column placed in a 2 ml collection tube (supplied). Centrifuge ≥ 8,000 x g (≥ 10,000 rpm) for 15s. Discard flow through. Repeat, reusing the column until all the RNA has passed through the column.
12. Wash the column with 350 µl Buffer FRN (supplied), centrifuge 15s ≥ 8,000 x g. Discard the flow through. Change to new collection tube.
13. Add 80 µl of mix containing 10 µl DNaseI stock solution (supplied) + 70µl Buffer RDD (supplied) directly to the membrane. Incubate at room temperature 30 min.
    NOTE: DNaseI stock is prepared, aliquoted, then frozen at -20°C. It can only be F/T once. Handle DNaseI with care. Do not vortex DNaseI mixture, mix by inversion.
14. Add 500 µl Buffer FRN (supplied) to wash out the DNase, centrifuge 15s and SAVE the flow through. Transfer the column to a new collection tube. Mix well and pass the flow through over the column again, centrifuge and DISCARD the new flow through.
15. Wash the column with 500 µl Buffer RPE (supplied), centrifuge 15s, discard the flow through. Repeat.
16. Place the column in a new collection tube. Centrifuge "empty" column at full speed for 5 min.
17. Transfer the column to a new collection tube. Open the cap and air dry 5 min.
18. Elute RNA by adding 20 µl of RNase-free water (supplied) applied directly to the membrane, incubating the column 10 min at room temperature, and centrifuging at full speed for 1 min. Transfer the eluted RNA to a 500 µl microcentrifuge tube and snap freeze on dry ice. Store at -80°C.

Purification of DNA

19. Resuspend the pellet in 180 µl Buffer ATL (supplied), add 40 µl proteinase K (20 mg/ml stock; 3.6 mg/ml final; supplied), and pulse vortex. (total volume = 220µl). Incubate at 56°C overnight (~16 hr).
20. Day 2: Add 10µl PK. Incubate 5 hr at 56°C or until tissue is fully digested.
21. Preheat a heating block to 90°C. Incubate the DNA at 90°C for 2 hr without agitation (to partially reverse formaldehyde modification of nucleic acids).
22. Briefly centrifuge the DNA at room temperature. Add 4 µl RNase (100 mg/ml; supplied) and incubate 2 min at room temperature. (total volume = 224µl)
23. Add 400 µl of a 1:1 mix of Buffer AL (supplied) and absolute ethanol, mix with the sample by pulse vortexing. (total volume = 624µl). Transfer all to QIAmp MinElute Column.

24. Centrifuge column 1 min at ≥ 8,000 x g at room temperature. Discard the flow through.
25. Place the column in a new collection tube, add 700 µl Buffer AW1(supplied), centrifuge 15s at ≥ 8,000 x g. Discard the flow through.
26. Add 700 µl Buffer AW2 (supplied), centrifuge 15s at ≥ 8,000 x g. Discard the flow through.
27. Add 700 µl 100% ethanol, centrifuge 15s at ≥ 8,000 x g. Discard the flow through.
28. Place the column in a new collection tube. Centrifuge full speed 5 min. Discard the collection tube.
29. Air dry column for 5 minutes to remove residual ethanol.
30. Place the QIAamp MinElute (supplied) spin column in a new **1.5 ml** collection tube. Elute with 22µl Buffer ATE (supplied) directly to the center of the spin column membrane. Incubate for 10 min at room temp. Centrifuge at full speed for 1 min to elute the DNA. Repeat once with an additional 22 µl of buffer (no incubation), pooling eluates. Store at -80^{o}C.

DNA Bisulfite Conversion

Using EZ DNA Methylation kit from Zymo Research #D5001, D5002
Fackler Modifications 7-18-11

1. Mix DNA (up to 2.0 µg) + ddH_2O to make a final volume of 42.5 µl in a 500 µl eppendorf tube. Add 7.5 µl M-dilution buffer (supplied in the Zymo kit). Heat at 42° C 30 min in PCR machine to denature the DNA. Up to 3 µg input DNA can be used.
2. During the 30 min incubation in step #1, prepare the CT reagent:
   On a per sample basis combine 71.4 µl water + 17.6 µl M-dilution buffer and 54 mg of conversion reagent.
   On a per vial basis (sufficient for 10.5 -11 samples) add 750 µl ddH_2O + 185 µl M-dilution buffer to the 1.7 ml brown vial containing 567 mg of CT reagent. Rotate in the dark at room temp for 10 min to dissolve. Use immediately.
3. To each sample, add 97.5 µl CT conversion reagent. Mix well with pipette tip. Final volume is 150 µl.
4. Incubate by cycling overnight in PCR machine: 95° C 30 sec, 50-55° C 1 hr for 16 cycles. Hold at 4° C.
5. Clean up: EZ DNA Clean-up (Zymo Research)
   a. Add 4 volumes of M-Binding Buffer to the Zymo-Spin 1C column in a collection tube (e.g. if bisulfite reaction is 150 µl, add 600 µl M-Binding Buffer). Add the sodium bisulfite/DNA reaction to the M-binding buffer in the column. Close the cap and invert tube at least 20 times to mix completely.
   b. Centrifuge at full speed 30 sec and discard the flow-through.
   c. Add 100 µl M-Wash Buffer to the column. Centrifuge at full speed 30 sec.
   d. Add 200 µl M-Desulfonation Buffer to the column. Incubate 15 min at room temperature. Centrifuge full speed for 30 sec.
   e. Add 200 µl M-Wash Buffer to the column. Centrifuge at full speed for 30 sec. Empty the collection tube. Add 200 µl M-Wash Buffer to the column and invert the column several times. Centrifuge full speed for 30 sec. Discard the flow through.
   f. Centrifuge the column 1 min empty to remove all remaining wash buffer.
   g. Transfer the column to a new collection tube. Add 15 µl of either water or elution buffer, preheated to 70°. Allow the water/buffer to sit for 5 min on the column. Centrifuge the column 1 min, recovering the DNA. Keep on ice (it is single stranded).
   h. Prepare a 1:5 aliquot: Mix 2 µl eluted DNA with 8 µl water, and quantitate the 1:5 dilution by using a nanodrop instrument, using a factor of 40 (like for RNA; dilute in water). Use this value to adjust the undiluted stock DNA to 75 ng/µl for methylation microarray.
   i. Use 2µl of the 1:5 dilution for MSP (20 ng/reaction is ideal) to verify that the DNA is amplifiable, indicating that bisulfite conversion was successful.
   j. Freeze bisulfite converted DNA at -70° C or less. Submit 10 µl of DNA at ~30 ng/µl for RESTORATION. It is possible to use a lower concentration of DNA, but array results may be suboptimal.

MMLV reverse transcription

Dnase I Treatment

Prepare reaction mix:

	1x
RNA	Amount for input of interest (1ug generally, 50ng for GAPDH QC)
10x Buffer	1 ul
DNase I	1 ul
DEPC Water	8 – RNA ul
Total	10 ul

1. Incubate at room temp < 15 mins
2. Add 1ul 25mM EDTA
3. Heat inactivate at 65°C for 10 mins

Reverse Transcription

1. Add primers
    a. 4ul random primers (final 0.5ug/1ug RNA) – dilute 1:8 1ug/ul stock
    NOTE: Must use random primers for FFPE-derived RNA
    b. 4ul oligoDT (final 0.5ug/1ug RNA) – dilute 1:20 2.5ug/ul stock
2. Incubate at 70°C for 5 minutes
3. Quick chill on ice for 1 minute and spin briefly
4. Make M-MLV mix (per):
    a. 5 ul M-MLV 5x Rxn Buffer
    b. 0.5 ul 25mM dNTP
    c. 1 ul M-MLV
    Note: For control, do not add M-MLV. Add all other components of the mix and water for M-MLV.
    d. 3.5 ul DEPC water
5. Mix by flicking and spin briefly;
6. Incubate for 1 hr.
    a. 37°C for random primers
    b. 42°C for oligoDT
7. Heat inactivate RTase @ 70°C for 15 mins.

Illumina FFPE QC Kit

Sample preparation

1. Measure the concentration of DNA using fluorescence dye assays such as Picogreen or Qubit.
2. Dilute samples to 1ng/µL

Standard preparation

1. Thaw QCP and QCT to room temperature.
    a. Make six 10µL aliquots of QCT in a 1.7mL Eppendorf tube and store at -20°C
2. Take a fresh 10µL aliquot of QCT and add 990µL DiH_2O to create a 100-fold dilution.
3. Vortex and quick spin.

Assay

1. Prepare the QPCR mix as follows:

	10µL reaction volume	20µL reaction volume
2x qPCR Master Mix	5 µL	10 µL
QCP	1 µL	2 µL
DiH_2O	2 µL	4 µL
Total volume per well	8 µL	16 µL

2. For 10µL reactions, add 8µL of reaction mix into each well that will be used. For 20µL reactions add 16µL.
3. For 10µL reactions, pipette 2µL of sample (QCT, sample, or water as NTC) in triplicate into the wells. For 20uL reactions, pipette 4µL of sample in triplicate.
4. Seal and quick spin plate.
5. Run the following QPCR program, using the appropriate reference dye
    a. Activation : 50°C for 2 min
    b. Denaturation : 95°C for 10 min
    Cycle 40 times:
    c. Denaturation : 95°C for 30 sec
    d. Priming : 57°C for 30 sec
    e. Extension : 72°C for 30 sec

Data analysis

1. Flag and remove any replicates which C_t diverge more than 0.5 cycles.
2. Calculate average C_t and calculate ΔC_t against the QCT controls.
3. Check NTC for negative amplification. Data is <u>acceptable</u> if NTC samples are >10 cycles after the QCT samples.
4. Samples with <8 ΔC_t (manufacturer recommends <5) can be used for methylation microarray. Prioritize samples with lower ΔC_t as it fits the study design.

Bioinformatics pipelines

The source code for analysis pipelines are available at a public Github repository: https://github.com/sean-cho/scu_pipelines.

II: Abbreviations

AR	Androgen receptor
BCS	Breast conserving surgery
BRCA	Breast cancer
CN	Copy number
CNV	Copy number variation
DCIS	Ductal carcinoma in situ
ER/ESR1	Estrogen receptor
FFPE	Formalin-fixed paraffin embedded
FTC	Follicular thyroid cancer
FVPTC	Follicular variant papillary thyroid cancer
GSA	Gene set analysis
GSEA	Gene set enrichment analysis
HM450K	Illumina Human Methylation 450K microarray
IDC	Invasive ductal carcinoma
IQR	Interquartile range
LRR	Log R ratio
LUSC	Lung squamous cell carcinoma
MAF	Minor allelle frequency
NGS	Next-generation sequencing
PK	Proteinase K
PR/PGR	Progesterone receptor
PTC	Papillary thyroid cancer
RCT	Randomized clinical trial
SNP	Single nucleotide polymoprhisms
SNP6	Affymetrix SNP6 microarray
Tam	Tamoxifen
TCGA	The Cancer Genome Atlas
THCA	Thyroid cancer
TNBC	Triple negative breast cancer
VPNI	Van Nuys prognostic index

Soonweng (Sean) Cho
810 Saint Paul St, Baltimore MD 21202 soonwengcho@gmail.com 909-618-7354

Educational History

Ph.D.	2016	Program in Cellular and Molecular Medicine	Johns Hopkins School of Medicine
B.S.	2008	Biotechnology Minor in Chemistry	California State Polytechnic University Pomona

Other Professional Experience

Summer Internship	2016	Pfizer Oncology, Computational Biology Group
Research Associate	2008 - 2010	Lab of Dr. Michael Jensen, City of Hope
Undergraduate Research	2006 - 2008	Lab of Dr. Wei-Jen Lin, Cal Poly Pomona
Summer Internship	2006	Bioprocessing Group, Sime Darby Tech. Ctr, Malaysia
Undergraduate Research	2005	Chemistry Lab, Technology Park Malaysia Academy

Academic Honors

2006 - 2008 Dean's List & President's Honors Rolls Cal Poly Pomona

Publications

Kim HS, Umbricht CB, Illei PB, Cimino-Matthews A, **Cho S,** Chowdhury N., et al. (2016) "Optimizing the use of gene expression profiling in early stage breast cancer." JCO.

Teo WW, Merino VF, **Cho S**, Korangath P, Liang X, Wu RC, Neumann NM, Ewald AJ, Sukumar S. (2016) "HOXA5 determines cell fate transition and impedes tumor initiation and progression in breast cancer through regulation of E-cadherin and CD24." Oncogene.

Aragon Han P, Kim HS, **Cho S**, et al. (2016) "Association of BRAF(V600E) Mutation and MicroRNA Expression with Central Lymph Node Metastases in Papillary Thyroid Cancer: A Prospective Study from Four Endocrine Surgery Centers." Thyroid.

Merino VF, Nguyen N, Jin K, Sadik H, **Cho S**, Korangath P, et al. (2016) "Combined Treatment with Epigenetic, Differentiating, and Chemotherapeutic Agents Cooperatively Targets Tumor-Initiating Cells in Triple-Negative Breast Cancer." Cancer Research.

Jin K, Park S, Teo WW, Korangath P, **Cho SS**, Yoshida T, et al. (2015) "HOXB7 Is an ERα Cofactor in the Activation of HER2 and Multiple ER Target Genes Leading to Endocrine Resistance." Cancer Discovery.

Barrio-Real L, Benedetti LG, Engel N, Tu Y, **Cho S**, Sukumar S, Kazanietz MG. (2014) "Subtype-specific overexpression of the Rac-GEF P-REX1 in breast cancer is associated with promoter hypomethylation." Breast Cancer Res.

Han L, Diehl A, Nguyen NK, Korangath P, Teo W, **Cho S**, et al. (2014) "The Notch pathway inhibits TGFβ signaling in breast cancer through HEYL-mediated crosstalk." Cancer Research.

Fackler MJ, Lopez Bujanda Z, Umbricht C, Teo WW, **Cho S**, Zhang Z, et al. (2014) "Novel methylated biomarkers and a robust assay to detect circulating tumor DNA in metastatic breast cancer." Cancer Research

Avraham A, **Cho SS**, Uhlmann R, Polak ML, Sandbank J, Karni T, et al. (2014) "Tissue specific DNA methylation in normal human breast epithelium and in breast cancer." PLoS One.

Shah N, Jin K, Cruz LA, Park S, Sadik H, **Cho S**, Goswami CP, et al. (2013) "HOXB13 mediates tamoxifen resistance and invasiveness in human breast cancer by suppressing ERα and inducing IL-6 expression." Cancer Research.

Publications in submission or preparation

Merino VF, **Cho S,** Liang X, Park S, Jin K, Chen Q, Pan D, Zahnow C, Rein A, Sukumar S. (In submission) "Inhibitors of STAT3, B-catenin, and IGF-1R sensitize mouse PIK3CA mutant breast cancer to PI3K inhibitors."

Han L, Korangath P, Diehl A, **Cho S**, Teo W, Okamura H, O'Hagan R, Myal Y, Gesseler M, Romer L, Sukumar S. (In submission) "Identifying HEYL as a key angiogenesis regulator through integrative meta-analysis of tumor angiogenesis signatures."

Sengupta S, Nagalingam A, Bonner M, Kuppusamy P, Goguen D, Shriver M, Muniraj N, **Cho S**, Begum A, et al. (In submission) "Activation of tumor suppressor LKB1 by honokiol abrogates cancer stem-like phenotype in breast cancer via inhibition of oncogenic Stat3."

Cho S, Kim HS, Umbricht CB, Cope L. (In preparation) "Comprehensive assessment in the derivation of copy number variation information from methylation microarrays."

Cho S, Kim HS, Umbricht CB, Cope L. (In preparation) "Epicopy: copy number variation from Illumina methylation microarrays."

Cho S, Florea L, Cope L, Wang Y, Zeiger M, Bishop J, Revilla M, Hayes N, Umbricht CB (In preparation) "Molecular Characterization of Metastatic Follicular Thyroid Cancer by RNA-Sequencing."

Merino VF*, **Cho S***, Nguyen N, Sadik H, Talbot C, Cope L, Zhang Z, Gyorffy B, Sukumar S, (In preparation) "Induction of cell cycle arrest and inflammation by combined treatment with epigenetic, differentiating, and chemotherapeutic agents in triple negative breast cancer."

Kim H, Wilsbach K, Marti A, **Cho S**, Najafian A, Meeker AK, et al. (In preparation) "DNA copy number variation and driver mutation patterns of follicular thyroid tumors"

Cho S, Fackler M, Wilsbach K, Chowdhury N, et al. (In preparation) "Identification of chemotherapy independent prognostic biomarkers and functionally relevant molecular landscapes in ER-negative breast cancer."

Invited Talks

Cho, S. (2014) "Epicopy: Measuring DNA copy number using Illumina 450K methylation microarrays". Computational Genomics Symposium, Johns Hopkins Hospital.

Posters

Cho, S., et al. (2016) "Molecular characterization of metastatic follicular thyroid cancer by RNA-seq". American Thyroid Association Annual Meeting 2016, Denver Colorado.
Finalist for Trainee Poster Contest

Cho, S., et al. (2016) "Characterization of metastatic follicular thyroid cancer by RNA-seq". Surgical Fellows Research Symposium 2016, Johns Hopkins Hospital.

Cho, S., et al. (2015) "Epicopy: Measuring DNA copy number variation using Illumina high density methylation microarrays." AACR Annual Meeting 2015.

Cho, S., et al. (2015) Epicopy: Measuring DNA copy number variation using Illumina high density methylation arrays. Surgical Fellows Research Symposium 2015, Johns Hopkins Hospital.

Competitive Grants

"Identification of Early Epigenetic Changes on African-American Progenitor Cells and Their Role on Breast Cancer Initiation". **DOD-Breakthrough Award** - Funding Level 2 (BC141315, 10/14-10/17).

"Identification of early epigenetic changes in the African American progenitor cells and their role in breast cancer aggressiveness". **Safeway Pilot Grant** (07/14-06/15).

Service and leadership

2012 – 2013	Pollard's Scholar, Molecular Biology and Genomics
2010 – 2012	Incentive Mentoring Program
2007 – 2008	President, Malaysian Students' Association
2006 – 2007	Vice President, Malaysian Students' Association
2006 – 2008	ASI Tutor: Biochemistry, Molecular Biology, Genetics