AWARD NUMBER: W81XWH-14-1-0080

TITLE: Total RNA Sequencing Analysis of DCIS Progressing to Invasive Breast Cancer

PRINCIPAL INVESTIGATOR: Christopher B. Umbricht, MD, PhD

CONTRACTING ORGANIZATION: Johns Hopkins University
Baltimore, MD 21205-1832

REPORT DATE: NOVEMBER 2018

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOVEMBER 2018	Final	01Sept2014 - 31Aug2018

4. TITLE AND SUBTITLE
Total RNA Sequencing Analysis of DCIS Progressing to Invasive Breast Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-14-1-0080

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Christopher B. Umbricht, MD, PhD

E-Mail: cumbrich@jhmi.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Johns Hopkins University

Baltimore, MD 21205-1832

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project is designed to complement a multi-institutional, NIH-funded study of genetic and epigenetic alterations of pre-invasive DCIS that did or did not progress to invasive breast cancer, with an in-depth analysis of gene expression data, including the characterization of multiple isoforms and splice-variants. During the current reporting period, we have completed, in collaboration with Dr. C. Perou at the U. of North Carolina, the RNA-sequencing of our DCIS sample collection from 5 collaborating institutions using the Illumina TruSeq RNA Access Library Preparation kit and the Illumina NextSeq500 sequencer. We received the final batch of sequence data from 32 study samples last week, for a final total of analyzable samples passing all Q/C steps, including mapped read counts, of 133 DCIS samples (67 cases and 66 controls). We are proceeding with our bioinformatic analysis, which will be complemented with our genome-wide methylation and copy number variation data obtained from the same samples obtained through our NIH-funded sister project.

15. SUBJECT TERMS
NONE LISTED

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
Unclassified	Unclassified	Unclassified	Unclassified	22	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Page

1. Introduction……. 1

2. Keywords……. 2

3. Accomplishments……..……... 2

4. Impact……...…… 16

5. Changes/Problems...…..…… 16

6. Products, Inventions, Patent Applications, and/or Licenses..……. 16

7. Participants & Other Collaborating Organizations…… 17

8. Special Reporting Requirements……17

9. Appendices……17

1. Introduction

A previously reported, we have completed the DNA and RNA extractions of 98 cases (DCIS progressing to invasive breast cancer) and 98 controls (no further breast cancer or DCIS with a 10 year negative follow-up) passing our Q/C testing. We have completed a genome-wide analysis of the DCIS methylomes using Illumina 450K methylome arrays, and report the results below. In light of the often limiting amounts of nucleic acids we can obtain from our archival tissue samples, particularly given the often small DCIS lesion sizes, we have also proceeded with our development and investigation of a computational approach we have called EPICOPY to obtain reliable copy number variation (CNV) data from the methylome array data, thereby decreasing the DNA requirements in half. A manuscript describing EPICOPY has been accepted for publication by the Journal of Computational Biology last week.
Our characterization of the DCIS transcriptome is still ongoing. Initial attempts using the Illumina DASL array platform had to be abandoned because the vendor withdrew the platform without warning and before the complete cohort could be processed. A subsequent attempt using the Affymetrix HGA2.0 expression arrays failed in spite of promising pilot experiments when we detected a pervasive artifactual background signal that was indirectly proportional to the quality control metrics of our RNA samples, and our bioinformaticians informed us this could not be mitigated computationally. Standard RNA-sequencing performed at our core facility yielded insufficient mappable reads. We then initiated a collaborative effort with UNC's Lineberger Comprehensive Cancer Center, were we successfully piloted a new exon-targeted RNA sequencing method developed by Illumina (TruSeq **RNA Access** Library Prep Kit), and have now completed 133 samples (67 cases and 66 controls) for which acceptable quality libraries were obtained. Preliminary results are summarized below, including a novel RNA splicing-variant analysis made possible by the available exon-level data provided by the Access NGS methodology, which we are using to complement the differential gene expression analysis.

The SPECIFIC AIMS unchanged from the updated aims in the 2016 report:

Aim 1: **Develop novel methods of assessing quality of samples and performing normalization across FFPE samples of variable quality**.
Aim 1a: Apply more stringent quality control parameters for enrichment of samples with high quality data.
Aim 1b: Optimize thresholds of qRT-PCR-based QC analysis of FFPE samples for identification of samples that will yield reproducible data.
Aim 1c: Integrate transcriptomic, methylome, and copy number data to identify biomarkers of progression in DCIS samples.

Aim 2: **Perform multi-omic analysis of transcriptome, methylome, and copy number data of DCIS**.
Aim 2a: Develop novel approaches, including non-parametric methods, to analyzing FFPE data with variable quality.
Aim 2b: Identify subtypes across DCIS samples and learn molecular alterations unique to those subtypes across all three molecular platforms through exploratory data analysis.
Aim 2c: Integrate transcriptomic, methylome, and copy number data to identify biomarkers of progression in DCIS samples.

Aim 3: **Perform RNA Access on a subset of DCIS samples, which allows for both comparative assessment of RNA species across methodologies and technical validation of genes of interest.**

Aim 3a: Perform sample-to-sample assessment of HTA2 and RNA Access data to identify commonalities, as well as differences across platforms.

Aim 4: **Validate genes of interest and biomarkers.**
Aim 4a: Develop bench-based assays and perform technical validation on a phenotypically-stratified subset of DCIS samples.
Aim 4b: External validation of biomarkers in DCIS validation cohort.

2. Keywords

Preinvasive breast cancer (DCIS); Invasive breast cancer (IBC); Transcriptome; Prognostic markers; splice variant analysis; formalin-fixed paraffin-embedded (FFPE) tissue; Receiver Operator Characteristic (ROC), Area under the Curve (AUC); Estrogen Receptor (ER), EPICOPY.

3. Accomplishments

Methods

Sample processing, quantification, and quality control
Upon the completion of sample accrual from 5 institutions, including Johns Hopkins Hospital (JHH), University of Iowa at Iowa City (UIowa), University of Southern California (USC), University of Alabama Birmingham (UAB) and University of Hawaii Honolulu (UHawaii), (See summary in previous report) we have processed the FFPE material. FFPE slides were macrodissected to enrich for >70% tumor and DNA and RNA were extracted using the Allprep DNA/RNA FFPE Kit (Qiagen, Valencia, CA).
DNA was quantified using the Qubit dsDNA BR assay (Life Technologies, Frederick, MD) and assessed for methylation array suitability using the Illumina FFPE QC kit (Illumina, San Diego, CA). Samples that pass quality control (QC) are processed for Illumina Human Methylation 450K microarray and samples that failed to reach QC detection limits are saved for validation.
RNA is quantified using the Qubit RNA BR assay (Life Technologies) and further assessed for integrity using the Experion StdSens Analysis kit (Bio-Rad, Hercules, CA).
A total of 98 progressive DCIS cases, 98 non-progressive DCIS controls, 12 DCIS adjacent normal tissue, and 5 reduction mammoplasty samples were processed for gene expression and methylation analysis.

DNA 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.

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.

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).

Epicopy

Raw data

The Epicopy method is developed using The Cancer Genome Atlas (TCGA) thyroid, breast, and lung small cell carcinoma datasets. Raw Level-1 450K microarray data are downloaded from Broad Institute's "Firehose Genome Data Analysis Center" (http://gdac.broadinstitute.org/) and analyzed using the R statistical environment [1]. 450K array data is read into R and preprocessed using the "minfi" package [2].

Comparison with SNP array

Affymetrix SNP 6.0 processed Level-3 data is downloaded and used as a "gold standard" for comparison. GISTIC 2.0 [3] was used to summarize Epicopy generated data into gene-level copy number variation (CNV) and frequently occurring CNV within each tumor type.

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.

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.

Transcriptome analysis

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.

Affymetrix HTA2 microarray
FFPE-derived RNA was processed per manufacturer recommended protocols using the WT Pico kit for global amplification of the RNA and hybridization on the HTA2 microarray. Based on our results from the titration experiment, 10ng total RNA were used as input.
HTA2 data processing
Per manufacturer recommendation for FFPE-derived RNA, data was processed using the Affymetrix Expression Console using the SST transformation, GCCN correction, and RMA normalization. Batch effects across processing plate were adjusted using COMBAT. Manufacturer recommended QC was performed and the positive vs negative AUC measure of 0.7 was used as a threshold to filter against samples of poor performance and principal component analysis (PCA) was used to identify outliers. A single sample was removed from further analysis, with low positive vs negative AUC and behaving as outlier on PCA analysis.
The Affymetrix HTA-2 Probeset Annotation (Release 36) was used to map probe sets to known genomic features.
Differential expression analysis
Differential expression analysis was performed using linear models for microarray analysis (limma) by constructing a model comparing progressive versus non-progressive DCIS.
Next Generation Sequencing using the TruSeq RNA Access *pipeline*
RNA fragment distribution was analyzed by the Tapestation and found to be highly degraded, as is expected for FFPE samples, eliminating the need for fragmentation before library preparation.
Library preparation and sequencing
FFPE-derived RNA was processed per manufacturer recommended protocols using the Illumina TruSeq RNA Access Library Preparation kit for global amplification of the RNA. Since the kit captures coding regions, no rRNA subtraction or poly(A)capture steps are required. The maximum recommended amount of total RNA (200ng) was used because of the typically low DV200 values observed in the DCIS RNA

samples. Sequencing was performed using Illumina NexSeq500 on a pooled library of 4 samples to produce approximately 150 million paired-ended sequencing reads of 48 base pairs per sample.

Transcriptome Analysis

Transcriptomic data collected by RNA Access library sequencing from 67 'Case' and 66 'Control' samples were analyzed to determine the genes that are present in each sample and condition, their expression levels, and the differences between expression levels among different experiment conditions. In a preliminary analysis, reads were mapped to the human genome version hg38 with the alignment tool Tophat2 v.2.1.0 [4], which allows for large 'gaps' in the alignment, representing introns. The aligned reads were assembled with the software CLASS2 v.2.1.7 [5] to create partial gene and transcript models (transfrags). Transfrags from all samples were further merged with Cuffmerge (ref-cufflinks) and mapped to the GENCODE v. 22 gene models, to create a unified set of gene annotations for differential analyses.

Differentially expressed genes (transcripts) were determined with the tool DESeq [6], while differentially spliced introns were inferred with LeafCutter [7].

Results

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 1a).

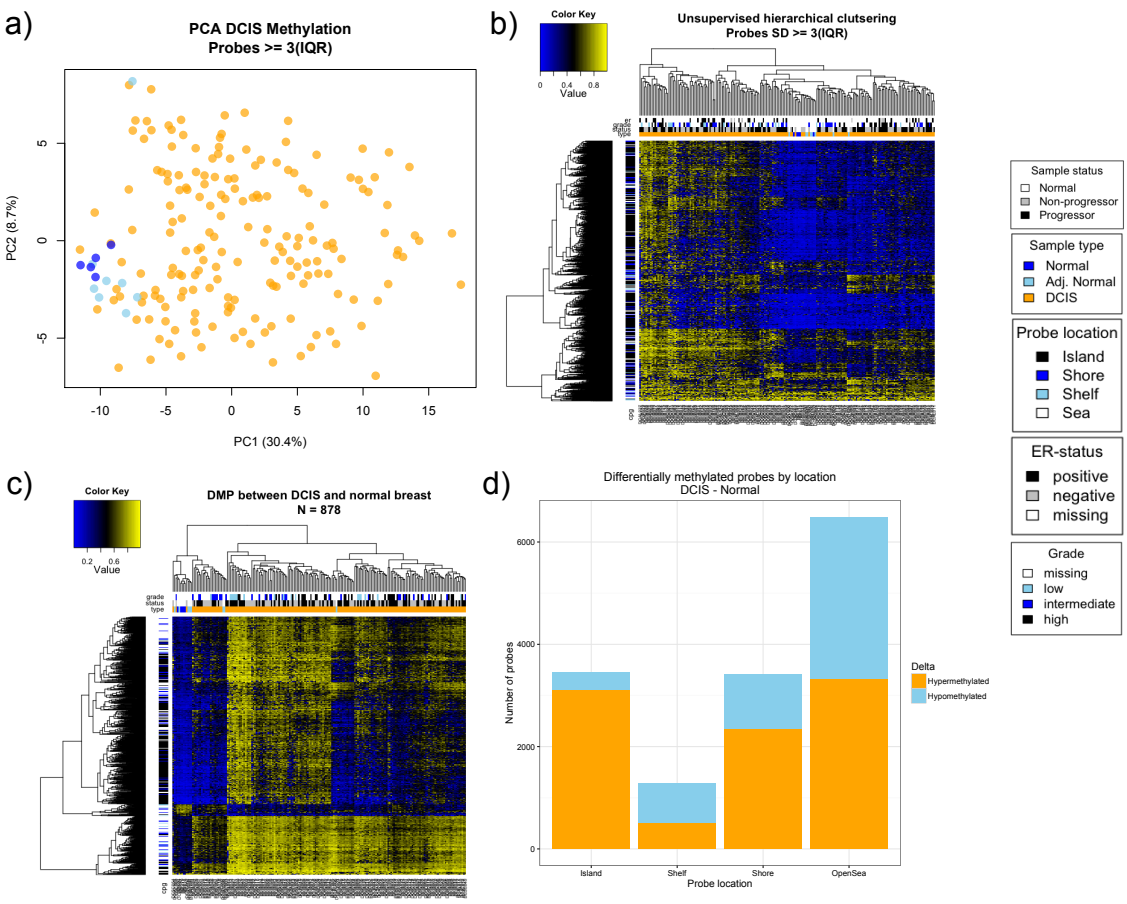


Figure 1: Distinct methylation profiles between normal and DCIS tissues.
a) PCA analysis on probes with SD > 3 IQR showed clustering of normal samples on both PC1 and PC2. b) Unsupervised hierarchical clustering using complete linkage revealed that normal tissues are unmethylated in most of these CpG 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).

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 1b). DCIS samples did not cluster by progression status, which may suggest that progression status is 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 1c) 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 1d, 9-fold), and to a smaller extent, shores (2.2-fold).
DMPs identified from this analysis include probes located in promoter regions or gene bodies of genes implicated in breast cancer development, including RASSF1A, TP73, CDKN2A (p16),

GSTP1, MGMT, APC, and HOX family genes. We also observed DMPs in genes related to estrogen receptor (ER) signaling; ESR1, RUNX3, and NCOR2.
We 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 2). Taken together, this suggests that global oncogenic methylation changes occur in DCIS, before the development of IDC.

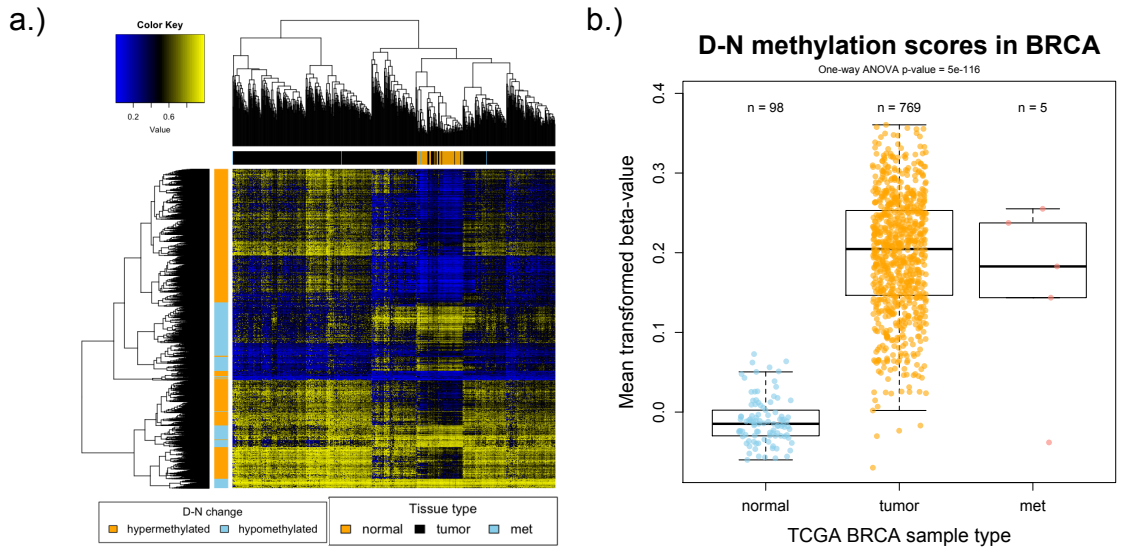


Figure 2: DMPs in D-N show consistent change in invasive breast cancer.
a). 14,652 DMPs between DCIS and normal tissue in the TCGA BRCA dataset. Methylation profiles in the tumors are concordant with the methylation profiles in DCIS when compared to normal breast tissue. b) D-N methylation scores calculated 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) [8-10]. 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 is 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 [11, 12]. Consistent with this theory, we observed concerted differentially methylated neighboring probes in many genes and performed DMR analysis using *DMRcate* [11]. We identified a total of 678 DMRs comparing DCIS and normal tissues, including in regions where aberrant methylation has been observed in breast cancer [8, 13]. Promoter hypermethylation of the HOX family genes of master regulators was commonly observed. HOX genes have been implicated in the oncogenesis and aggressive phenotypes in breast cancer [14-17]. Recently, work in our lab has shown that HOX genes regulate cell fate transition [18], invasiveness [19], and endocrine therapy resistance [20]. Furthermore, tumor suppressor genes involved in DNA repair, such as TP73, APC, CCND2, and MGMT, were also differentially methylated. DMRs were also present in the estrogen responsive genes, ESR1, RUNX3, and FOXA2, suggesting that at least a subset of the DCIS have dysregulated ER signaling.

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 1). Interestingly, within a subset of D-N DMPs, methylation profiles intermediate to that between DCIS and normal tissue were present in the series of tumor-adjacent normal tissue. Indeed, when calculating 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 3a), and this result was statistically significant (Bonferroni adjusted pairwise Wilcox test, $p < 0.05$ across all comparisons).

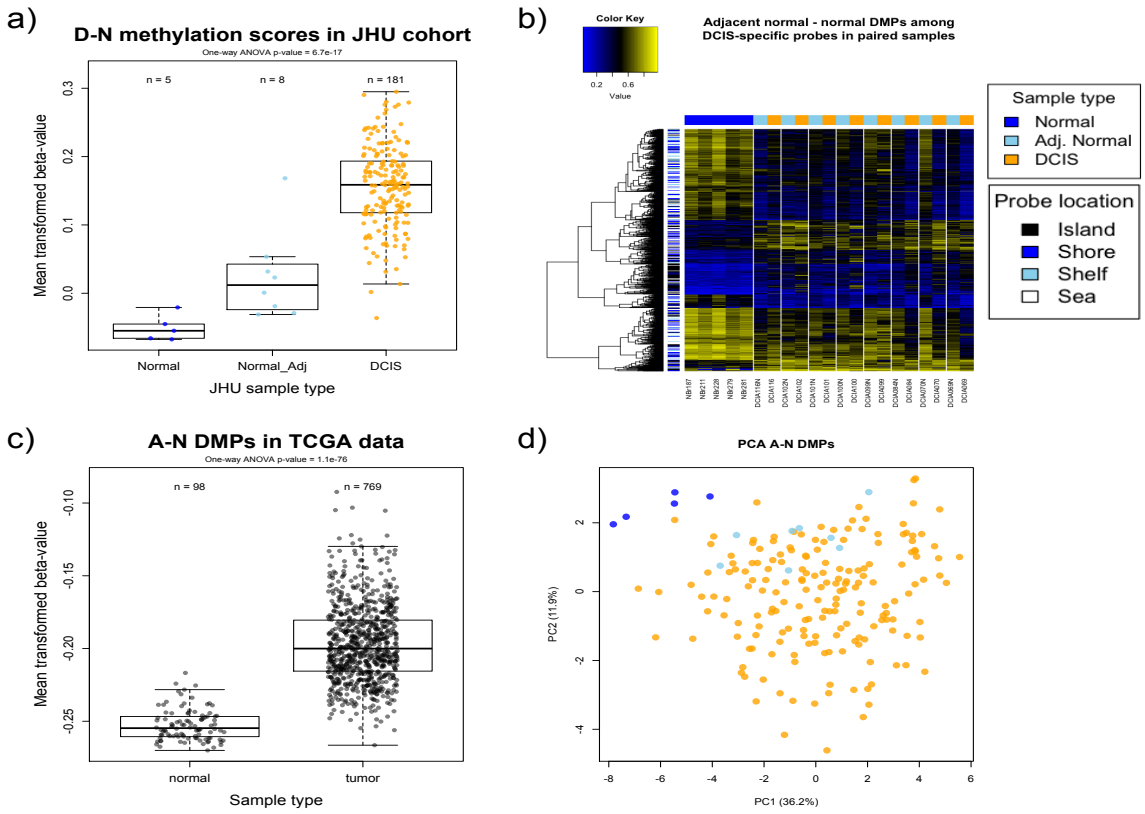


Figure 3: DCIS and oncogenic methylation observed in DCIS adjacent normal

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.

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 3b). To identify if there are oncogenic differences between probes differentially methylated between DCIS-adjacent normal and normal breast tissue (A-N), we 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 3c). This signature also distinguishes normal tissue from DCIS (Figure 3d).
This observation is 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 3c). 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 [21]. Furthermore, other studies have shown field cancerization [22] in breast cancer, where cancer-associated genetic [23], epigenetic [24, 25], transcriptomic [26, 27], and telomeric [28] 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.

Unsupervised clustering identifies four DCIS methylation clusters or Epitypes
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 and 5). Hierarchical clustering with complete linkage was used to identify clusters of probes, and this revealed 3 major probe 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 1, Figure 4). 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 [29]. 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 [30].

Table 1: 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)

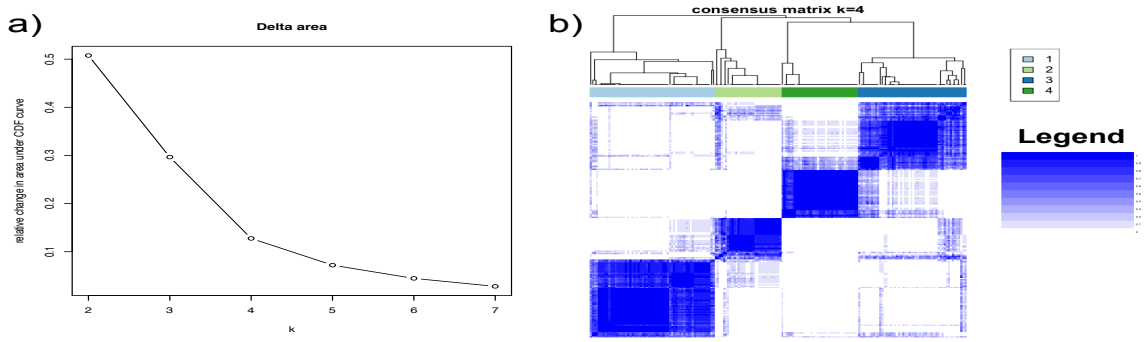


Figure 4: Consensus cluster metrics for selection of optimal K
a) Change in area under the CDF curve identifies k = 4 as the inflection point. b) Consensus matrix for k = 4 shows good consensus and stability across four clusters.

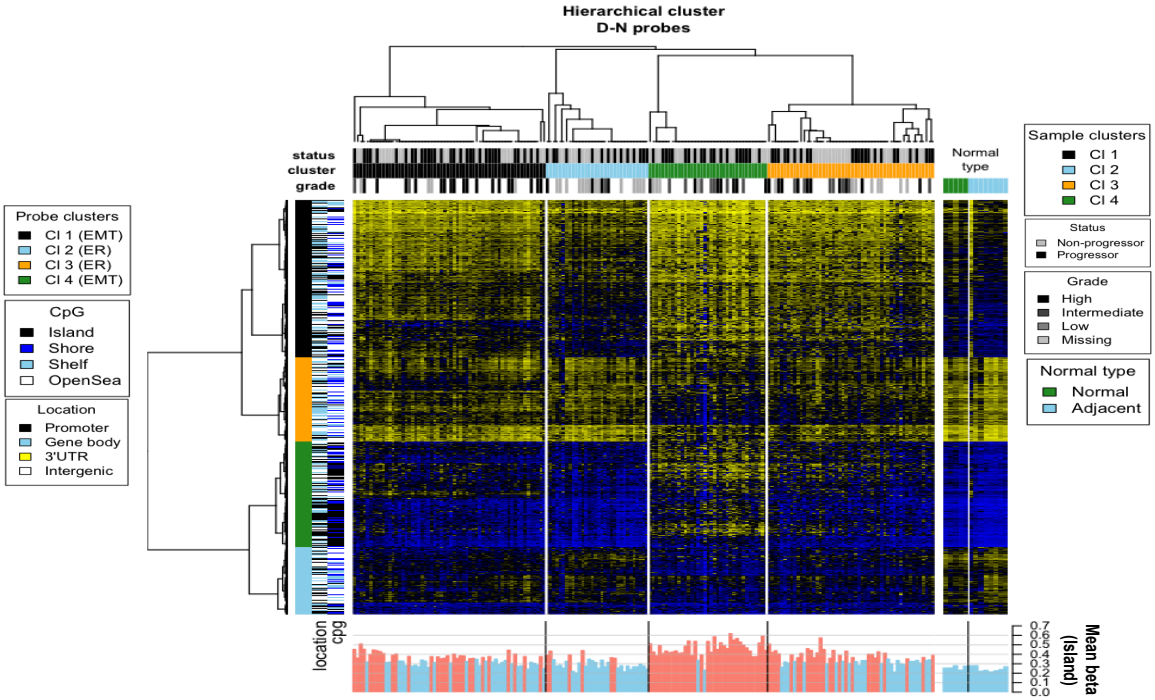


Figure 5: Clustering of DCIS samples
Clustering of DCIS samples of 14,529 probes differentially expressed between DCIS and normal samples. Samples were clustered using consensus clustering and four stable clusters were identified. Probes were clustered using hierarchical clustering with the Ward clustering 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.

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 2). 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 3, Figure 6).

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. These data will be reanalyzed using a multiomic approach when the ongoing gene expression and splice variation analyses are complete.

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 2. DMPs comparing case and control in DCIS-specific probes

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

Table 3. Probes associated with progression status as identified by supervised PCA

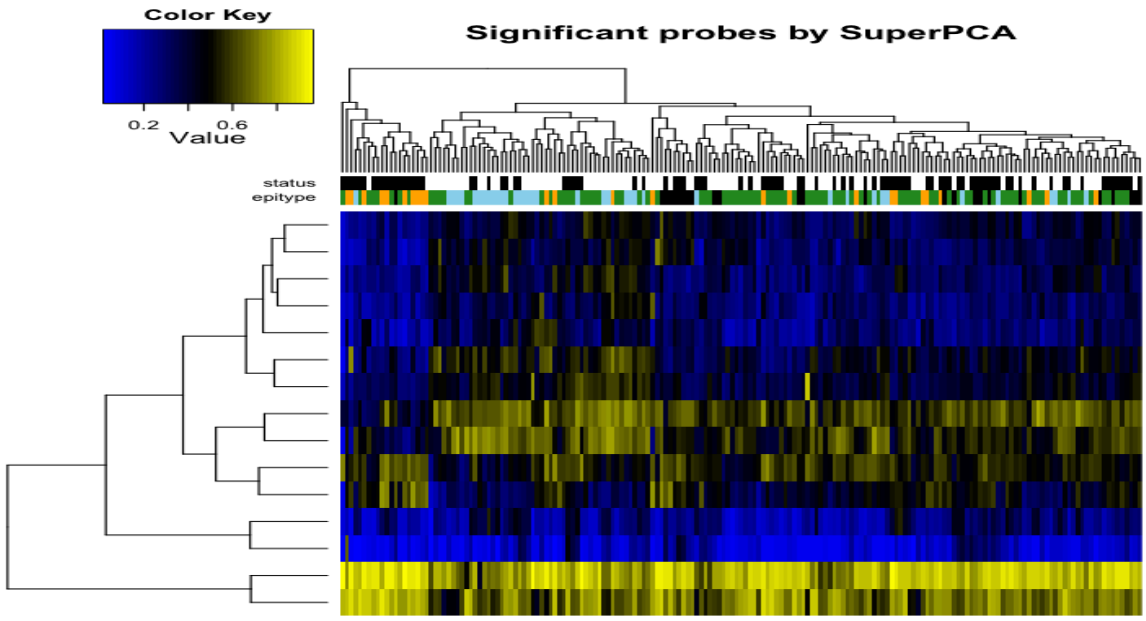


Figure 6: Supervised principal component analysis
Super PCA identified 15 probes associated with progression status across 3 principal components. Interestingly, these probes had greater association with epitype than with progression status.

CNV data recapitulate previously identified recurrent CNVs in DCIS
Please refer to our previous progress reports for a description of the Epicopy methodology. The CNV profiles of all JHU 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 on previously published DCIS CNV studies [31]. 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 7).

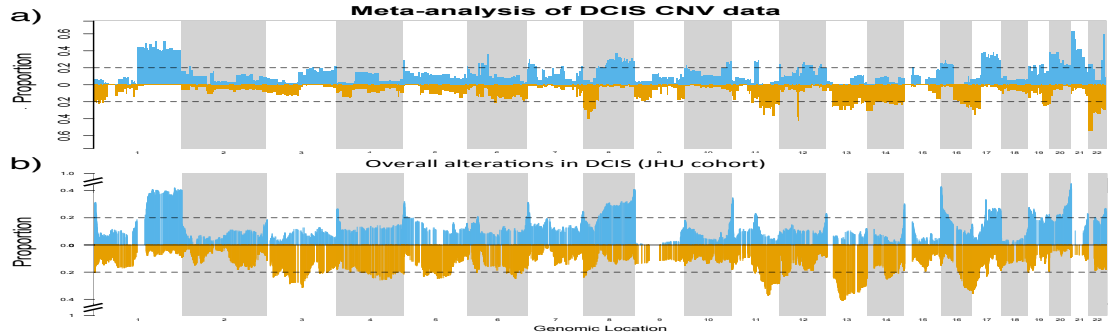


Figure 7: CNV events by incidence in previously published studies and JHU cohort
a) Proportion of previously published DCIS samples from a meta-analysis [31] which showed CNV alterations at specific regions of the genome. b) The same information for the proportion of JHU DCIS samples. Many of the events 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.

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 8). 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.

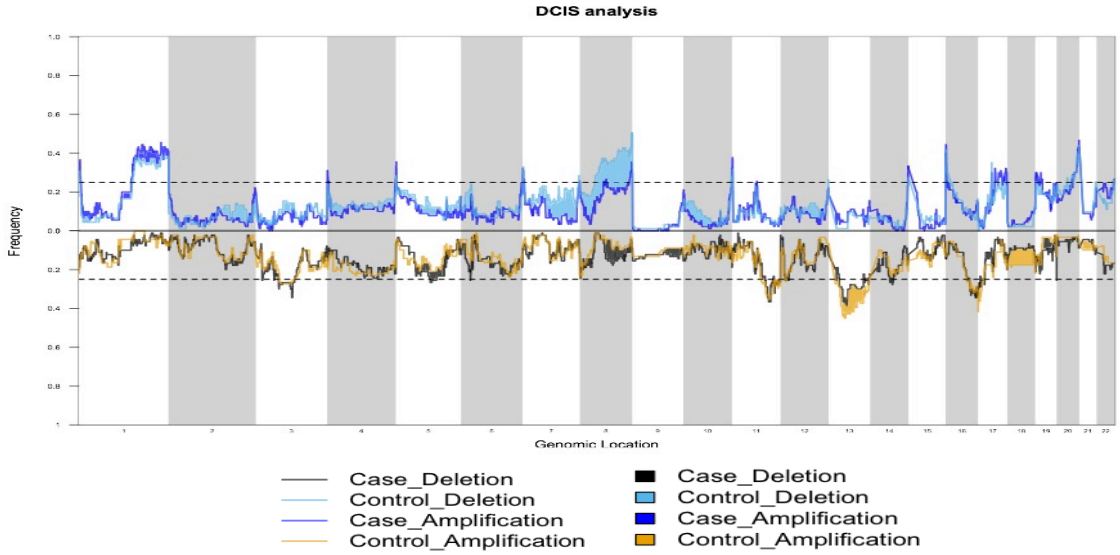


Figure 8: Comparison of CNV incidences across case and controls in JHU DCIS cohort

Determining the DCIS transcriptome using Illumina RNA-Access pipeline at UNC

We described the unsatisfactory results of our Affymetrix HTA2 array-based transcriptome analysis in our last report. As reported, this led us to pursue a collaboration with Dr. Perou's group at UNC, a national center for next-generation sequencing. We described on our initial pilot results in the last report, and are presenting the 2 key figures from that report again for continuity. The bar graphs below (Figure 9) plot the number of reads (Millions) per sample. It was apparent that large sample-to-sample variation in read counts were present, which were due to competition between high and low quality RNA samples during the library generation process, which occurs in batches of 4 as per Illumina's instructions. As reported, we then experimented with binning of samples of similar RNA quality to minimize this effect, achieving insufficient improvement, and ultimately proceeded to single sample library preparation to maximize the library quality for these unique and irreplaceable samples. This changes was instrumental in achieving a 75% success rate with these very challenging samples.

Figure 10 illustrates the gene-level expression data were received from our UNC collaborators. The heatmaps shows expression levels of the 500 most variable genes across 12 DCIS samples that have completed the standard data analysis pipeline.

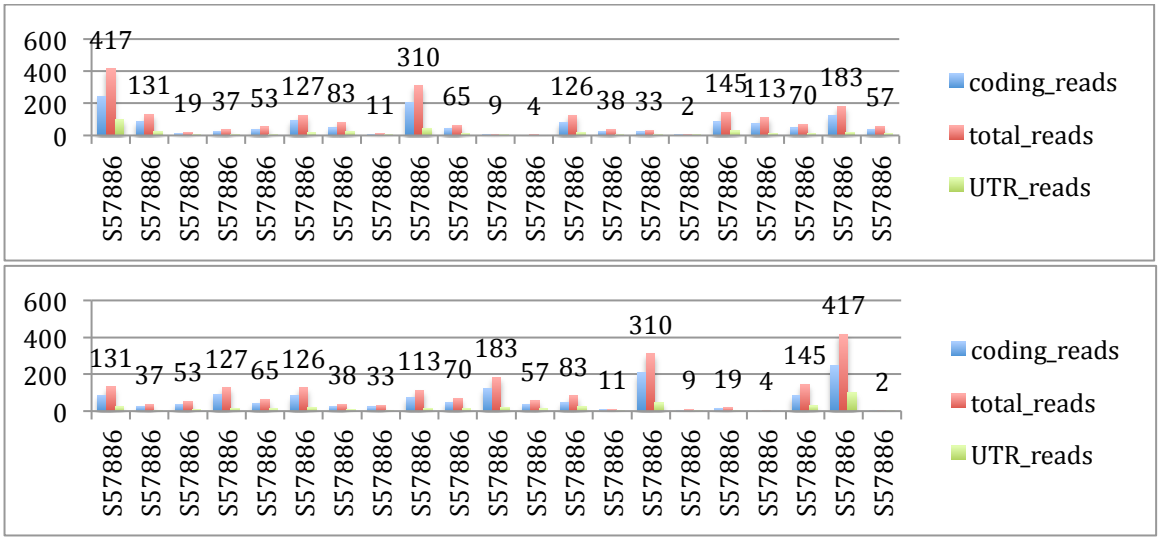


Figure 9. Total read count yields from first batch of study DCIS samples.

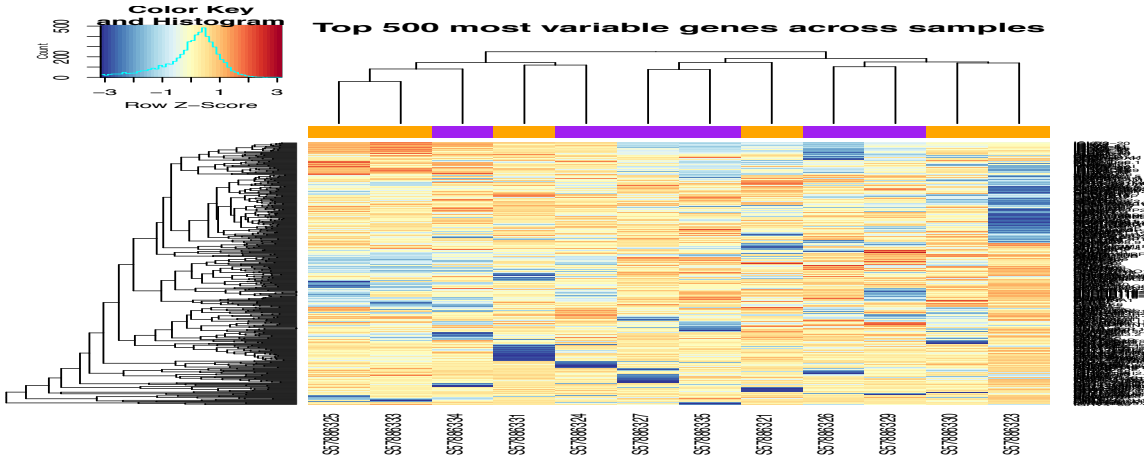
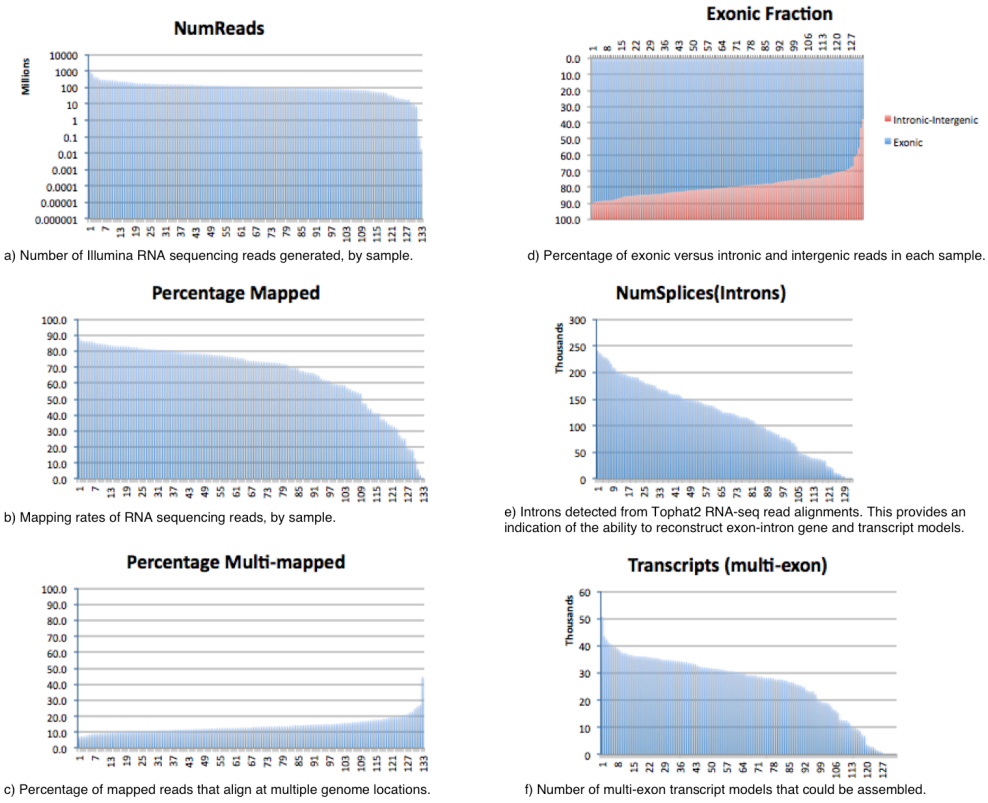


Figure 10. Heatmap showing expression of top 500 most variable genes from first batch of study DCIS samples.

We can now report that we have successfully completed the RNA-Access-based sequencing of our entire available DCIS cohort, of which 75% yielded mappable coding read counts. Our bioinformatic analysis of the DCIS transcriptomes, including isoforms and splice variants, will be based on 59 DCIS cases that progressed to invasive breast cancer, and 58 DCIS controls with at least 10 years of documented negative follow up. Our bioinformaticians, Drs. Leslie Cope and Liliana Florea, have assessed the emerging data and report that it will allow us to determine not just overall expression, but splice-variant analysis as well.

Initial results of transcriptomic and splice variant analysis
Mapping rates varied between 60-88.7% for 98 samples (73.7%), between 40-60% for 19 samples (14.3%) and were below 40% for 17 samples (12.8%), the latter being excluded from analyses. Except for four samples, between 61.2-89.9% of the aligned reads were exonic, and therefore could be effectively used to reconstruct genes and transcripts. Alignments provided evidence for 164,000-241,311 splices (introns) per sample; 11 samples had <20,000 splices, and were excluded from the merged set of gene annotations, but were used in the differential splicing and

expression analyses. DESeq found 2,934 potentially differentially expressed genes for a p-value cutoff of 0.05, and that number was reduced to 103 when restricting the numbers of false positives (p_adj <= 0.1). Lastly, LeafCutter reported 2,948 genes with potentially differentially spliced introns (p-val<=0.05), which was reduced to 226 when adjusting for multiple testing to control the rate of false positives (p_adj<=0.1).



a) Number of Illumina RNA sequencing reads generated, by sample.

d) Percentage of exonic versus intronic and intergenic reads in each sample.

b) Mapping rates of RNA sequencing reads, by sample.

e) Introns detected from Tophat2 RNA-seq read alignments. This provides an indication of the ability to reconstruct exon-intron gene and transcript models.

c) Percentage of mapped reads that align at multiple genome locations.

f) Number of multi-exon transcript models that could be assembled.

Figure 11. RNA-Access Mappable read counts

Ongoing Experiments

The following changes in the bioinformatics analysis protocol will be applied to ensure a comprehensive, accurate and robust analysis of the data for final reporting and publication, which due to time constraints could not be applied. ***First***, reads will be quality checked and trimmed of primer and/or adaptor sequence, and for low-quality bases. This pre-processing step is expected to significantly increase the mapping rate across the samples. ***Second*****,** multiple tools will be used in the alignment, transcript assembly, and differential splicing and gene expression steps, to ensure accuracy and robustness, and that the most suitable tools are being used that address characteristics of the data. These include the mappers HiSAT2 and STAR; assemblers StringTie and PsiCLASS, the latter a novel multi-sample transcript assembly tools that produces more consistent annotations across all samples and a more precise set of meta-annotations to serve as the merged annotation database; and differential analysis tools Cuffdiff, for gene expression, and rMATS and JULiP for differential splicing, the latter of which has been developed in the Florea lab to address the analysis needs of very large collections of samples. ***Third***, we will explore different alignment filtering techniques, for instance excluding putative PCR duplicate reads

(identified with the tool Picard). ***Lastly***, we will more carefully consider the range of samples to be included in each step of the analysis, for instance identifying gene expression profile outliers, and establishing further criteria for sample and data quality.

Integrative analysis

One of the strengths of the DCIS study is the simultaneous profiling of genetic, epigenetic, and transcriptomic changes in our discovery cohort. As such, integrated analysis of the different molecular phenotypes will allow us to measure concurrent changes within a single pathway or identify potential drivers with higher confidence when more than one molecular phenotype contains the change. For example, correlations between increased copy number and increased gene expression can be used to identify the most relevant gene within an amplified region of DNA.

In addition, integrative cluster analysis using iCluster[11] will be performed across all three datasets to identify subtypes across the arrayed DCIS samples. A comprehensive clustering approach will allow us to combine consistent genomic, epigenetic, and transcriptomic alterations to discover novel subtypes, and revisit our preliminary results reported above.

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, which could be the result of confounding influences from the molecularly complex phenotypes in breast cancers and DCIS. We will revisit this possibility using integrative data analysis once the matching transcriptomic analysis is complete.

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 is prevalent in progressors while copy number gain is enriched in the non-progressors.

4. Impact

To be determined. A molecular signature reflecting the risk of progression of DCIS to invasive breast cancer will improve personalized management of DCIS, limiting the morbidity of treatment in low risk cases and improving outcome in high risk cases.

5. Changes/Problems

See discussion of our results in section 1.

6. Products

Manuscript:
Soonweng Cho, Hyun-seok Kim, Martha A. Zeiger, Christopher B. Umbricht, Leslie M. Cope. Measuring DNA copy number variation using high-density methylation microarrays. In press at: Journal of Computational Biology.

7. Participants & Other Collaborating Organizations

Charles M. Perou, Ph.D, The May Goldman Shaw Distinguished
Professor of Molecular Oncology Departments of Genetics, and
Pathology & Laboratory Medicine
Lineberger Comprehensive Cancer Center
125 Mason Farm Road
The University of North Carolina at Chapel Hill Chapel Hill, NC 27599

8. Special Reporting Requirements

N/A

9. Appendices

References cited:

1. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al*: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**(10):R80.
2. 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**(10):1363-1369.
3. 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**(4):R41.
4. 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**(4):R36.
5. Song L, Sabunciyan S, Florea L: **CLASS2: accurate and efficient splice variant annotation from RNA-seq reads**. *Nucleic Acids Res* 2016, **44**(10):e98.
6. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome Biol* 2010, **11**(10):R106.
7. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, Pritchard JK: **Annotation-free quantification of RNA splicing using LeafCutter**. *Nat Genet* 2018, **50**(1):151-158.
8. Jones PA, Baylin SB: **The fundamental role of epigenetic events in cancer**. *Nat Rev Genet* 2002, **3**(6):415-428.
9. 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 medicine* 2014, **6**(8):61.

10. 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**(8):768-775.
11. 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.
12. 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**. *International journal of epidemiology* 2012, **41**(1):200-209.
13. Jovanovic J, Ronneberg JA, Tost J, Kristensen V: **The epigenetics of breast cancer**. *Molecular oncology* 2010, **4**(3):242-254.
14. Shah N, Sukumar S: **The Hox genes and their roles in oncogenesis**. *Nature reviews Cancer* 2010, **10**(5):361-371.
15. Jin K, Sukumar S: **HOX genes: Major actors in resistance to selective endocrine response modifiers**. *Biochimica et biophysica acta* 2016, **1865**(2):105-110.
16. Jin K, Sukumar S: **BRCA1: linking HOX to breast cancer suppression**. *Breast cancer research : BCR* 2010, **12**(4):306.
17. Chen H, Sukumar S: **HOX genes: emerging stars in cancer**. *Cancer Biol Ther* 2003, **2**(5):524-525.
18. 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.
19. 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**(17):5449-5458.
20. 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**(9):944-959.
21. Ma XJ, Dahiya S, Richardson E, Erlander M, Sgroi DC: **Gene expression profiling of the tumor microenvironment during breast cancer progression**. *Breast cancer research : BCR* 2009, **11**(1):R7.
22. Dotto GP: **Multifocal epithelial tumors and field cancerization: stroma as a primary determinant**. *The Journal of clinical investigation* 2014, **124**(4):1446-1453.
23. 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**(2):131-139.
24. 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**(26):3348-3353.
25. 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**. *Nature communications* 2016, **7**:10478.
26. 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**(7):1596-1601.
27. 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. *International journal of cancer* 2011, **129**(6):1310-1321.
28. 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**. *International journal of cancer* 2006, **119**(1):108-116.
29. 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 : news and reviews in molecular, cellular and developmental biology* 2014, **36**(2):134-140.
30. Esteller M: **Epigenetics in cancer**. *N Engl J Med* 2008, **358**(11):1148-1159.
31. Rane SU, Mirza H, Grigoriadis A, Pinder SE: **Selection and evolution in the genomic landscape of copy number alterations in ductal carcinoma in situ (DCIS) and its progression to invasive carcinoma of ductal/no special type: a meta-analysis**. *Breast Cancer Res Treat* 2015, **153**(1):101-121.