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

REPORT DATE: September 2017

TYPE OF REPORT: Annual

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

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE September 2017	**2. REPORT TYPE** Annual	**3. DATES COVERED** 1Sep2016 - 31Aug2017

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

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 expression data on the entire range of informative RNA categories. During the current reporting period, we have further analyzed an Affymetrix HTA 2.0 array-based comprehensive transcriptome assay of samples from 5 collaborating institutions. Despite promising results from a smaller pilot experiment, and initially promising data reported in the prior report, further analysis lead us to conclude that much of the signal from these arrays have to be attributed to technical variation, and we are limited in the depth of biological information we can obtain from these arrays. In collaboration with Affymetrix scientists, we were able to determine Q/C measures that are predictive of subsequent array data quality, but at the cost of losing a large number of irreplaceable samples in our cohort. We have therefore continued our collaboration with Dr. C. Perou at UNC to maximize the possibility of a successful RNA Sequencing effort, and have confirmed the encouraging pilot results using the new Illumina TruSeq RNA Access Library Preparation kit followed by RNA sequencing performed using the Illumina NextSeq500. We have now completed the initial batch of 48 study samples, of which 75% yielded good read counts. We describe additional optimization steps we are implementing to complete the analysis of this unique sample cohort.

15. SUBJECT TERMS

Nothing listed

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

Table of Contents

1. Introduction.

Our overall goal remains to develop predictive markers that will be useful in identifying the minority of cases of preinvasive breast cancer (DCIS), that do in fact progress to invasive disease (IBC), and complements our multi-institutional, NIH-funded study of genetic and epigenetic alterations of pre-invasive DCIS that either progressed to invasive breast cancer IBC (cases) or had no further breast cancer events (controls).

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; non-coding RNA; formalin-fixed paraffin-embedded (FFPE) tissue; Receiver Operator Characteristic (ROC), Area under the Curve (AUC); Estrogen Receptor (ER).

3. Accomplishments

In previous reporting periods, we reported the completion of the accrual, initial characterization and processing of samples from 5 collaborating institutions. We also reported on the successful DNA methylome assessment using Illumina's 450K microarray, and an initial assessment of DNA copy number variation (CNV) based on a computational method we developed called Epicopy. We have since refined this method to adapt to the specific challenges posed by FFPE-derived DNA.

As previously reported, we initially changed our transcriptome analysis strategy from full RNA sequencing, which yielded insufficiently robust data, to the HTA2.0 microarray from Affymetrix, after obtaining good results in a titration pilot study using a subset of our DCIS samples. Unfortunately, despite promising results from a smaller pilot experiment, further analysis of the HTA2.0 microarray data lead us to conclude that much of the signal from these arrays have to be attributed to technical variation, and we are limited in the depth of biological information we can obtain from these arrays. While the result reported last year are likely valid, they represent only part of the biological information we are seeking, because only the strongest signals were detectable in our data. In collaboration with Affymetrix scientists, we were able to determine Q/C measures that are predictive of subsequent array data quality, but at the cost of losing a majority of samples in our cohort. We have therefore continued our collaboration with Dr. C. Perou at UNC to maximize the possibility of a successful RNA Sequencing effort, and have confirmed the encouraging pilot results using the new Illumina TruSeq RNA Access Library Preparation kit followed by RNA sequencing performed using the Illumina HiSeq 2500 we briefly reported last year. We have now completed the initial batch of 48 study samples, of which 75% yielded good read counts. We are further implementing additional optimization steps to address a remaining flaw in our results, which is due to the fact that following the manufacturer's protocol for the library preparation, sets of 4 samples are combined. When there are significant differences in RNA quality in the samples, good quality RNA outcompetes lesser quality samples, decreasing the available reads from the latter in the subsequent sequencing process. We are therefore attempting to bin samples of similar RNA quality, and if unsatisfactory, we will run each sample in individual library preparations. This will marginally increase the cost of the experiment from approximately $800 to $1000 per sample, since most of the cost is incurred at the sequencing step, not the library preparation. This will allow us to complete the analysis of this unique sample cohort with expression data of much improved quality, and enable a meaningful multi-omic analysis combining transcriptome data with the already available methylome and DNA copy number variation data.

Methods

Patient identification and sample collection

Table 1: Sample distribution

Source Institution:	Johns Hopkins	UA Birmingham	U Hawaii	U Iowa	USC	**Totals**:
Discovery Phase:						
Case	6	7	1	64	20	98
Control	6	8	1	54	29	98

Normal breast				8		8
Total	12	15	2	126	49	
Validation Phase:						
Case	22	6	23	64	12	127
Control	22	6	23	62	12	125
Normal breast						
Total	44	12	46	126	24	

For discovery experiments, 181 DCIS and 8 normal tissue samples passed QC and were used for the analysis.

DNA/RNA co-extraction

After evaluation of DCIS area by resident pathologist, DCIS epithelial cells were enriched via macrodissection with a clean scalpel. DNA and RNA were extracted using Qiagen (Hilden, Germany) Allprep RNA/DNA FFPE kit with a modified deparaffinization protocol, where samples were deparaffinized in xylene for 3x 10 minute washes instead of manufacturer recommended 10 minute wash.

DNA and RNA yield were quantified using Qubit fluorometer (Qiagen), with the broad range RNA and DNA reagents.

DNA and RNA quality assessment

DNA quality for methylation profiling was performed using the Illumina (San Diego CA) FFPE DNA QC kit and samples with a delta CT ≤ 6 compared to the provided control. RNA quality was assessed using Experion (Biorad, Hercules CA) on a randomly sampled subset of samples with varying yield and age of FFPE block. The RNA assessment showed no significant difference in RIN score or distribution of the RNA fragments across these samples.

Data analysis

Analyses were performed using the R statistical software [1] with base, Bioconductor [2] and custom functions and packages where necessary.

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 cases.

Gene Set Enrichment Analysis

A rank-based GSEA-like [4] approach was used to perform gene set analysis. Briefly, moderated t-statistics from the DCIS progressive vs. non-progressive limma analysis restricted on RefSeq genes were used to rank the genes. These scores were used to calculate enrichment against the hallmark geneset curated by the Molecular Signatures Database (MSigDB) [5, 6] to identify biologically relevant gene set differences between progressive and non-progressive DCIS.

Estrogen receptor (ER)-classification of DCIS samples

A k top-scoring pairs (KTSP) approach implemented by the switchbox [7] package was used to build an ER classifier for both methylome and transcriptome datasets. Briefly, an ER classifier was built using invasive breast cancer data obtained from TCGA with unambiguous ER-status using a 10-fold cross validation scheme for parameterization. Two parameters were optimized using cross-validation approaches: 1) the number of features (genes or probes) in the search space (termed feature number, F) and 2) k pairs to use in the classifier (k).

Feature number or search space optimization

Feature number was optimized using a 10-fold cross-validation approach where the ER-positive and ER-negative samples were split proportionally into 10 sets, where 9 sets were used as training sets and the remaining set was used as a validation set. The feature number was optimized by altering the search space to obtain a KTSP score for each of the validation samples and assessing prediction accuracy using an ROC analysis to maximize AUC. The number of pairs, k, was allowed to vary between 3 (minimum requirement) and the rounded up square root of F.

k optimization

Following feature number optimization, the optimal number of k TSPs were identified using a similar schema, where $k \in 3 \ldots F$ was used to maximize the AUC of an ROC analysis in the validation dataset.

Voting scheme

Since previous measurements of prediction potential of the KTSP classifier was performed using ROC analyses, no thresholds were required for making a prediction call. In the application of this classifier in an unknown dataset, a threshold for classification is necessary. The classifier implements a majority vote in its decision process.

Classifier validation & classification

The ER classifier is then evaluated for predictive accuracy by using it to classify a subset of DCIS with known ER-status. An empirical threshold for AUC was set at 0.8 for the ability to predict ER-status in these samples to constitute success, before using the same classifier for the rest of the DCIS samples. Following validation, the ER-status for all the DCIS samples was predicted.

Illumina Human Methylation 450k microarray (Illumina 450K array)
FFPE-derived DNA were restored using the Illumina FFPE DNA restoration kit per manufacturer's recommendation. Restored DNA samples were then hybridized and scanned according to manufacturer provided protocol.

Illumina 450K data processing
Raw Idat files of the Illumina 450K array were provided by the SKCCC microarray core and were read using the minfi package. Sample-wise call rate was calculated using a detection p-value cutoff of 1e-05 and density plots were used to evaluate the distribution of beta-values. Samples with <80% call rate or have an aberrant beta value distribution were excluded from downstream analyses.
Pre-processing was performed using functional normalization. Probe-wise call rate was calculated using a detection p-value cut-off of 1e-05 for all probes, and probes with call rates of < 99% (failed in 2 or more samples) were dropped from the study. Probes within 3 base pairs of a known SNP with 5% minor allele frequency (MAF) were removed from the study.

TCGA data
Processed RNA-seq and Illumina 450K methylation data [3] were obtained from the Firehose GDAC hosted by the Broad Institute, with the data downloaded in August 2015.

Copy number analysis in DCIS
Epicopy (Cho S. et al, submitted for publication) was used to obtain copy number information from Illumina 450K data. To adjust for FFPE-derived DNA, a more stringent threshold for minimum probe number per segment and fold change was implemented to obtain high quality segment calls. GISTIC 2.0 was used to identify and quantify recurrent copy number variation (CNV) across all DCIS samples. The meta-analysis results from Rane et al. (2015) [8] were obtained for use in a comparative Manhattan plot as the known CNVs in DCIS. A comparative analysis between progressive and non-progressive DCIS was performed by taking the difference of the frequencies of CNV observed across both groups.

Transcriptome analysis using total RNA-Access in DCIS
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 is expected for FFPE samples, eliminating the need for fragmentation before library preparation.

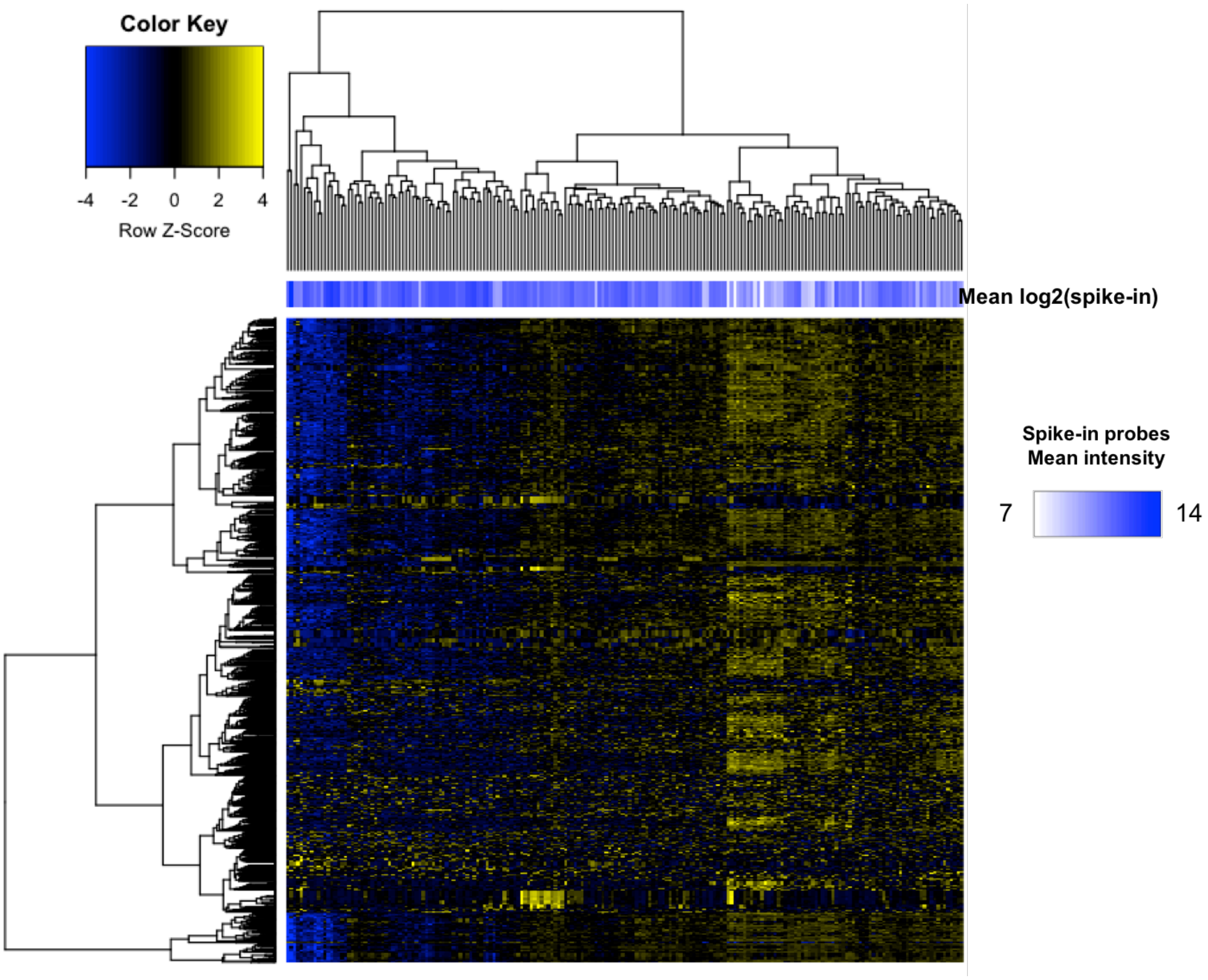
Library preparation and sequencing

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.

Results

Unsupervised analysis of HTA2 array analysis

Unsupervised clustering and principal component analysis (data not shown) revealed that signal intensities were affected by technical variation of unknown origin.



T

Figure 1. Heatmap of HTA2 array showing top 2000 most variable features in DCIS.

he heatmap shows the top 2000 most variable features (x-axis) from the microarray across all the DCIS samples (y-axis). The blue/white color bar on the top represents the combined signal from a mix of control probes ranging from low to high amounts, spiked into the sample as a quality control measure. We observed correlation of expression in large blocks of features with this quality control measure. This behavior expands to most of the features obtained from this array and up to the second principal component (data not shown). Thus, we conclude that a majority of the signal from these arrays are attributed to technical variation, and we are unlikely to obtain useful biological information from these arrays.

Using the Affymetrix Quality Assessment Kit to predict data quality in HTA2 arrays

We established a collaboration with Affymetrix scientists to investigate our findings. We repeated a subset of array experiments comparing results with read-outs from the Affymetrix RT-PRC-based RNA Q/C kit, since the standard RNA Q/C assessment with Bioanalyzer-derived RIN-values is not useful for FFPE-derived RNA samples. The following data is representative of the type of predictive value one can expect when using the FFPE RNA Quality Assessment Kit prior to running Affymetrix® microarrays.

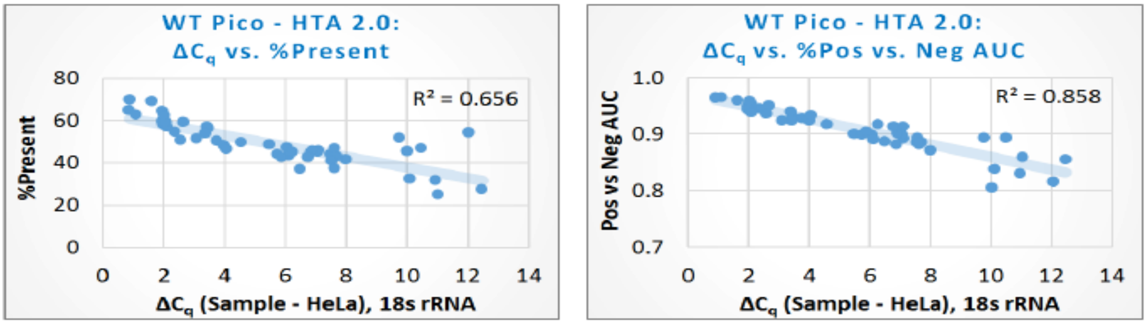


Figure 2. Correlation plots of ΔCq values (x-axis) vs.% Present Calls (A, y-axis), ΔCq values (x-axis) vs. Pos_vs._Neg AUC (B, y-axis).

	Samples	
ΔCq	38	*%*
<1	18	*47%*
1-2	3	*8%*
2-3	4	*11%*
3-4	3	*8%*
4-5	2	*5%*
5-6	4	*11%*
6-7	1	*3%*
7-8	2	*5%*
>8	1	*3%*
sum #	*38*	*100%*

The Cq value is inversely proportional to the number of amplifiable templates. 74% of all FFPE samples tested generated ΔCq values less than 4 and should be considered as of acceptable quality (with predicted % Present calls around 50 based on Affymetrix in-house data).

Our analysis indicates that ΔCq values above 3-4 in this assay showed the strongest correlation between signal strength of transcript features and the spike-in controls.

Three-way comparison of Illumina DASL vs Affymetrix HTA2 vs Affymetrix ClariomD gene expression arrays

While assessment of the transcriptome using FFPE-derived RNA has always been challenging, our lab has had excellent results with the now unfortunately discontinued DASL arrays by Illumina. We therefore decided to compare one of our existing, DASL-derived datasets obtained from Triple-negative invasive ductal carcinomas, with the new Affymetrix platforms, both the HTA2 arrays, which we had used for our DCIS project as described above, as well as Affymetrix' newest iteration of gene expression arrays, the ClariomD. RNA stored at -80°C from the same preparations used for the DASL array was used in these experiments.

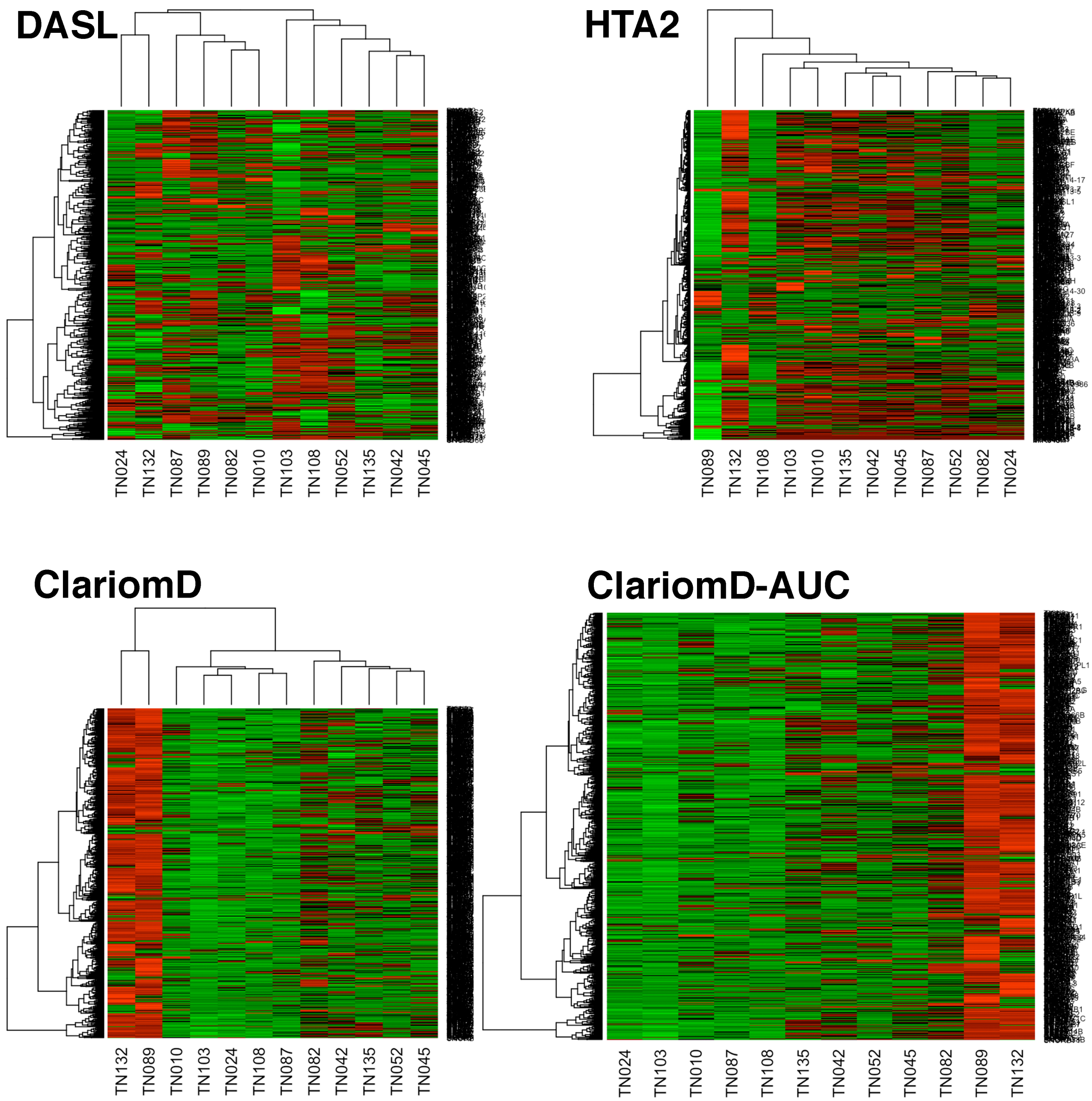


Figure 3. Comparison of DASL, HTA2, and ClariomD expression arrays in TriNeg Breast Cancer.

Overall, the signal intensities are about 2-fold higher on the DASL array than on the 2 Affymetrix arrays, which may be an artifact of normalization, or scanner settings of the different vendors etc., and are adjusted for in these heatmaps.

The heatmaps show the 1000 most differentially expressed genes/features in the samples. An unsupervised cluster analysis of the DASL array data suggests that samples are weakly divided into two groups, consistent with the study hypothesis that recurrent vs non-recurrent TriNeg Breast Cancer have distinguishing features. This finding is obscured on the HTA2 and ClariomD arrays.

Intensities on the HT2.0 array are much lower, and the dynamic range seems quite small. In general, the 25th percentile of probe intensities should reflect background, while the 75% percentile should be unambiguously expressed, but those values are very close to one another on this array. These values would be much lower than expected on previous generations of Affymetrix chips, such as the hgu133.

The ClariomD platform also has very low expression intensities, thought the dynamic range is slightly higher. The unsupervised cluster analysis using the most variable genes shows distinctive, vertical red and green stripes, a phenomenon that was evident in the DCIS HT2.0 dataset.

The vertical stripes visible in the ClariomD heatmap are strongly associated with RNA quality. The heatmap presented in the bottom right is rearranged in order of RNA quality as measured by positive_vs_negative_AUC (Area Under the curve).

The quantile normalization in RMA has successfully coerced the data into a common distribution. Despite this, the most variable genes capture a systematic bias in expression levels that is associated with RNA. The association with RNA quality is not evident if the most variable genes from the DASL array are selected instead.

Integrative correlations provide a way of measuring gene-level agreement, across studies/platforms, even when there are no samples in common. Briefly, a gene has a high integrative correlation if it exhibits the same co-expression patterns with other genes in each dataset. In this instance, we have the same samples on all 3 platforms and can directly calculate gene-specific correlation coefficients between platforms, so we use integrative correlation as a general measure of agreement between platforms rather than to guide selection of informative genes.

Integrative correlations between datasets were generally low (data not shown), probably indicating both heterogeneity among samples and lack of agreement between platforms.

One characteristic distinguishing the DASL platform and the new Affymetrix arrays (but not the older hgu133 arrays) is that the DASL and hgu133 arrays only captured 3’-end features of transcripts, whereas HTA2 and ClariomD are designed to detect multiple features along the entire transcript. It is not how if this is related to the artifacts were are observing with lower quality RNA preparations, but we are in the process of running this sample set on one additional new Affymetrix platform, the ClariomS, which is limited to 3’-end transcript features.

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

In light of these results, we decided to pursue the successful collaboration with Dr. Perou's group at UNC we initiated last year using the Illumina RNA-Access pipeline. The initial pilot results were reported in last years progress report. We now have the initial results of the first batch of 48 samples of the DCIS discovery cohort, of which 75% yielded mappable coding read counts. Our bioinformaticians, Drs. Leslie Cope and Liliana Florea, have assessed the emerging data and report that it should allow us to determine not just overall expression, but splice-variant analysis as well. The bar graphs below plot the number of reads (Millions) per sample.

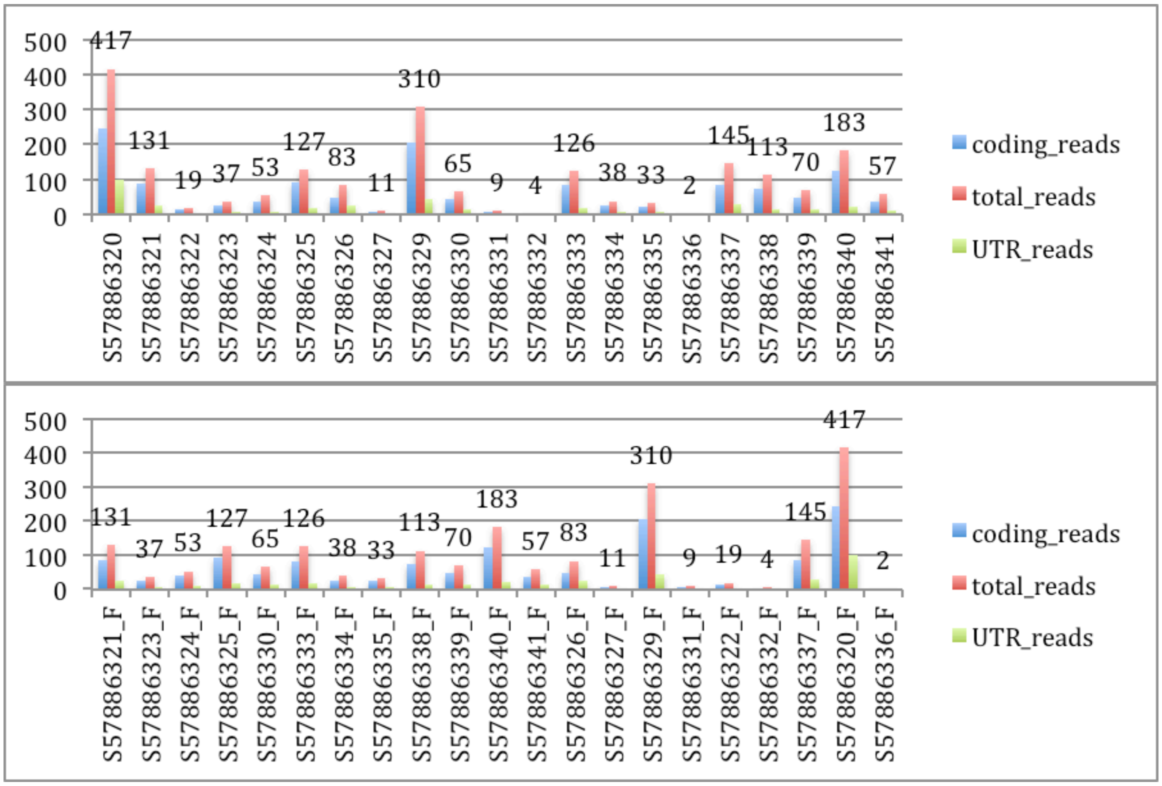


Figure 4. Total read count yields from first batch of study DCIS samples.

It is immediately apparent that large sample-to-sample variation in read counts are present, which are 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. We are currently experimenting with binning of samples of similar RNA quality to minimize this effect, as well as repeating the library preparations of samples with low read counts. If necessary, we will proceed to single sample library preparation to maximize the library quality for these unique and irreplaceable samples. Figure 5 illustrates the data were are receiving 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 – significantly, no banding is evident for these samples.

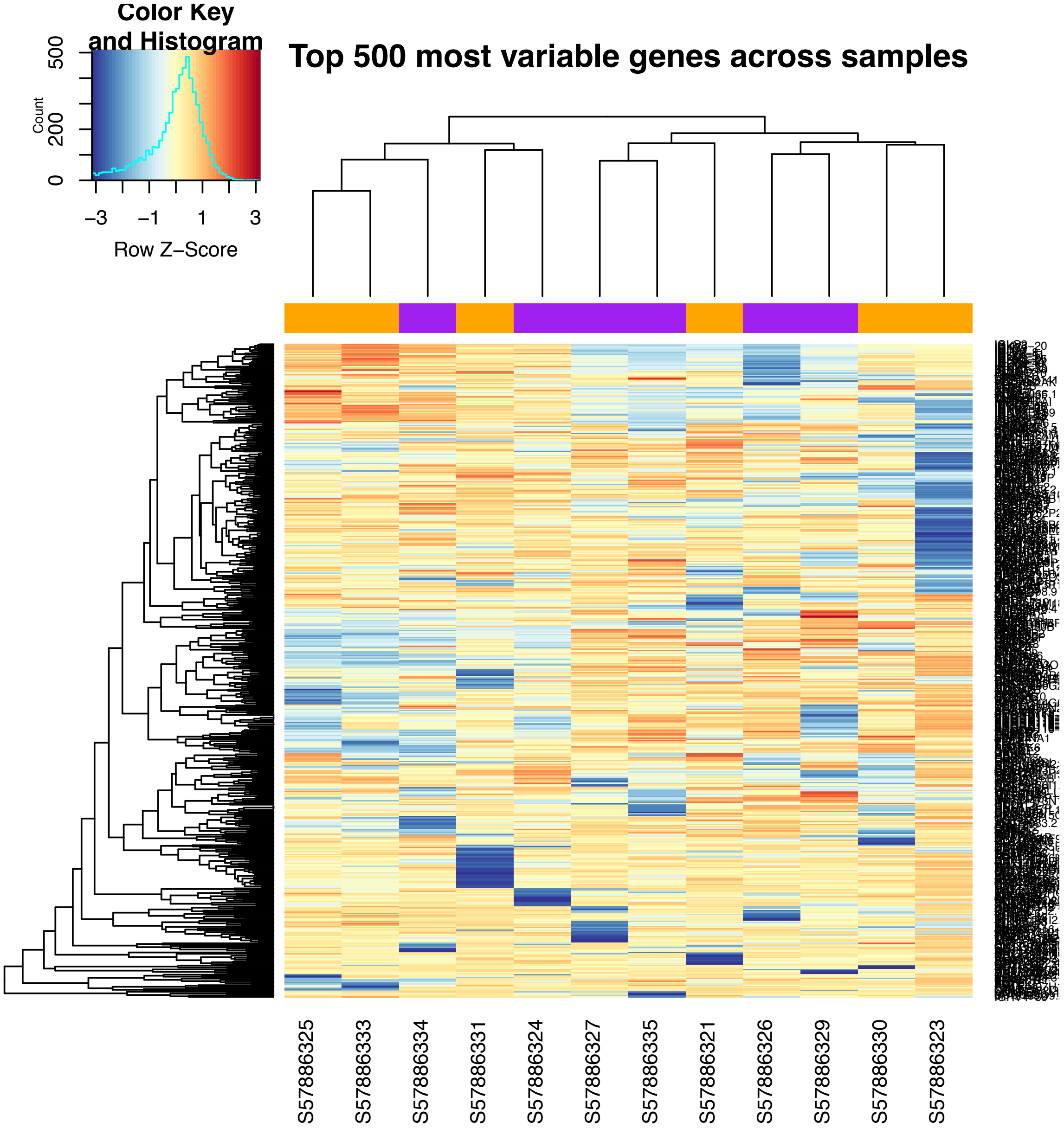


Figure 5. Heatmap showing expression of top 500 most variable genes from first batch of study DCIS samples.

Conclusion

Taken together, we conclude that despite technical challenges that still persist in the data, we were able to detect biologically relevant signals from these DCIS samples, which we can augment with high quality methylation and copy number data. Furthermore, our continued assessment and improvement of FFPE-derived RNA analysis technologies yielded valuable technical insights into new array platforms, and confirmed promising results using the RNA-Access experimental pipeline, which we are now applying to the discovery set of DCIS samples.

4. Impact

N/A

5. Changes/Problems

See discussion of our results in section 3.

6. Products

N/A

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
1. R-Core-Team: R: A language and environment for statistical computing.: R Foundation for Statistical Computing; 2016.
2. 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:R80.
3. Cancer Genome Atlas N: Comprehensive molecular portraits of human breast tumours. Nature 2012, 490:61-70.
4. 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.
5. 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.
6. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP: Molecular signatures database (MSigDB) 3.0. Bioinformatics 2011, 27:1739-1740.
7. Marchionni L, Afsari B, Geman D, Leek JT: A simple and reproducible breast cancer prognostic test. BMC Genomics 2013, 14:336.

8. 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, 153:101-121.