Breast Cancer Recurrence Risk Probed by Whole Transcriptome Next Generation Sequencing in 136 Patients

Joffre Baker,[1] Mei-Lan Liu,[1] Michael Crager,[1] James Stephans,[1] Mylan Pho,[1] Jennie Jeong,[1] Aaron Scott,[1] Ranjana Ambannavar,[1] John Morlan,[1] Robert Pelham,[1] Kunbin Qu,[1] Raul Mena,[2] Jose Esteban,[2] Francois Collin,[1] and Dominick Sinicropi[1]

[1]Genomic Health, Inc., Redwood City, CA; [2]Providence Saint Joseph Medical Center, Burbank, CA

San Antonio Breast Cancer Symposium- December 6-10, 2011

Background

- Gene expression biomarkers are the basis for clinical diagnostic tests such as the 21 gene Onc*otype* DX® breast cancer assay.
- Transcript profiling by next generation sequencing (RNA-Seq) has potential to greatly expand the number of transcripts that can be surveyed in biomarker discovery studies, relative to earlier methods such as RT-PCR and microarrays.
- Whether RNA-Seq using fixed paraffin embedded (FPET) tissue has sufficient sensitivity and precision for clinical biomarker discovery is an open question.
- We report here:
  - RNA-Seq using small quantities (0.1 ug) of RNA from FFPE breast tumor specimens up to 12 years old
  - that RNA-Seq provides the precision and sensitivity required to re-discover Onc*otype* DX® biomarkers with similar statistical significance and effect sizes as originally discovered using RT-PCR technology
  - the identification of hundreds of new candidate expression biomarkers arising from protein-coding, intronic, and intergenic regions of the genome

Study Design Summary

- A cohort of 136 breast tumors with associated clinical data was analyzed by RNA-Seq.
- Standardized hazard ratios for individual transcripts were calculated using univariate Cox proportional hazards regression analysis.
- Hazard ratios and p-values determined using RNA-Seq were compared with those determined using RT-PCR for Onc*otype* DX biomarkers.
- Hierarchical clustering of RefSeq transcripts associated with breast cancer recurrence was employed to investigate co-expressed gene networks.

Methods

Laboratory

- Total RNA was extracted from three 10 um sections of FFPE tumor specimens 5 – 12 years after fixation using the MasterPure™ Kit from Epicentre® Biotechnologies.
- 100 nanograms of total RNA were depleted of ribosomal RNA by hybridization to complementary DNA oligonucleotides and RNaseH treatment (Morlan et al., manuscript in preparation).
- RNA-Seq libraries were prepared using ScriptSeq™ Kits (Epicentre Biotechnologies) with six-base index sequences for sample multiplexing.
- Two libraries were multiplexed for cluster generation in each lane of Illumina flow cells using TruSeq™ SR Cluster Kits v2 (Illumina, Inc.).
- Sequencing was performed on an Illumina HiSeq® 2000 instrument for 50 cycles in a single direction followed by 7 cycles in the opposite direction for the index sequences.

Bioinformatics and Statistics

- Primary analysis of sequence data was performed using CASAVA 1.7 from Illumina
  - de-multiplexing of samples was done with a mismatch tolerance of 1
  - FASTQ sequences were trimmed on both the 5' and 3' ends to address adapter contamination and other artifacts
  - trimmed sequences were mapped to version hg19 of the human genome using ELAND2 from Illumina with two mismatches allowed in a seed of 32 base pairs and gap penalties allowed in alignment extension
  - sequences that mapped to multiple locations in the human genome were excluded from analysis
- Data were analyzed in three categories: consolidated exons and introns of RefSeq genes, and intergenic transcripts
  - Exons and Introns
    - The number of sequences that mapped uniquely within the coordinates of exons and introns of RefSeq genes were counted as a measure of transcript expression ("counts" or "reads").
    - RNAs for which none of the 136 specimens yielded 5 or more counts were excluded from analysis.
    - Raw counts were normalized by subtracting the 3rd quartile of the $\log_2$ RefSeq RNA counts and adding the cohort mean 3rd quartile similar to the method of Bullard et al.[1]

Methods (cont.)

- Hazard ratios for breast cancer recurrence were calculated using univariate Cox proportional hazards regression analysis[2] with the robust standard error estimate of Lin and Wei[3] and false discovery rates (FDRs) were assessed using the method of Storey.[4]
  - Intergenic transcripts
    - Approach 1: Counts that mapped to 2,500 well-documented lincRNAs[5] were analyzed by the method described above for exons and introns.
    - Approach 2: An algorithm was developed to identify clusters of reads that map to intergenic regions in one or more tumor specimens and the number of reads mapped to these clusters was used as a measure of the expression of these putative intergenic transcripts, which were analyzed as above for exonic and intronic RNAs.

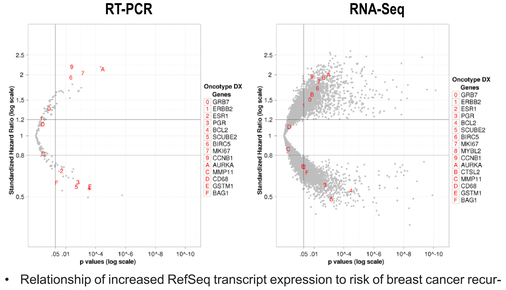
Results

Table 1: Case characteristics and outcomes

Characteristic		No. patients/no. analyzed
Tumor size (cm)	0-2	81/136 (60%)
	2-5	49/136 (36%)
	> 5	6/136 (4%)
No. lymph nodes at primary diagnosis	0	110/136 (81%)
	1-9	11/136 (8%)
	10-15	9/136 (7%)
	16-20	6/136 (4%)
Adjuvant tamoxifen	Yes	54/136 (40%)
	No	77/136 (57%)
	Unknown	5/136 (4%)
Adjuvant chemotherapy	Yes	51/136 (38%)
	No	79/136 (58%)
	Unknown	6/136 (4%)
ER Status*	ER positive	111/136 (82%)
	ER negative	25/136 (18%)
Vital Status	Distant recurrence, death due to breast cancer, or death due to unknown cause: Total	26/136 (19%)
	ER positive	16/136 (12%)
	ER negative	10/136 (7%)
	Alive without distant recurrence or death due to other cause: Total	110/136 (81%)
	ER positive	95/136 (70%)
	ER negative	15/136 (11%)

*ER status determined by RNA-Seq analysis

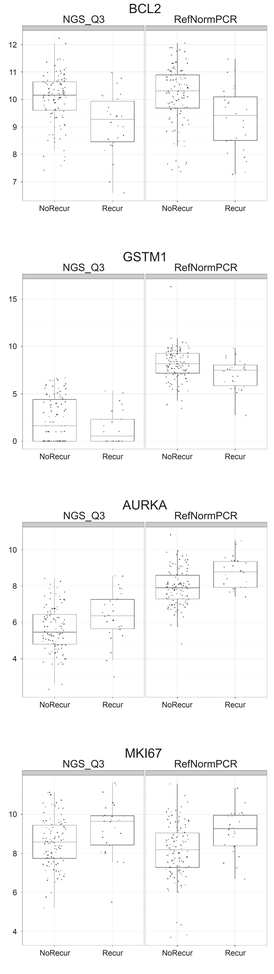
Figure 1: RNA-Seq rediscovers Onc*otype* DX genes and hundreds of additional transcripts



- Relationship of increased RefSeq transcript expression to risk of breast cancer recurrence in 136 patients
- The effect size is plotted as the standardized hazard ratio and statistical significance as the p-value.
- Left panel: Analysis of 185 genes using RT-PCR
- Right panel: Analysis of 20,462 RefSeq transcripts using RNA-Seq
- Onc*otype* DX genes are labeled with red alphanumeric symbols in both plots.

Results (cont.)

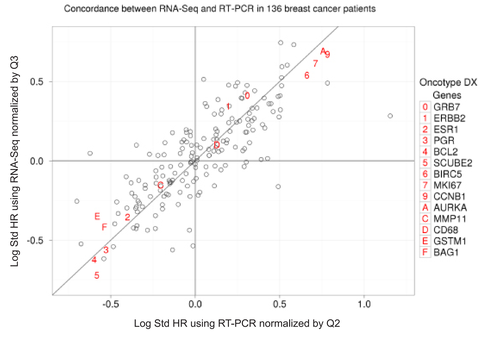
Figure 2: Relative distribution of tumor mRNA levels between patients who did and did not have disease recurrence: Similarity between RNA-Seq and RT-PCR results



- Examples of box plots of normalized expression values, stratified by recurrence status, as determined by RNA-Seq and RT-PCR for Onc*otype* DX genes

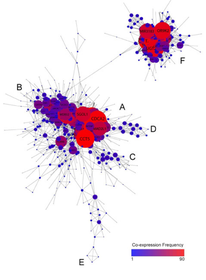
Results (cont.)

Figure 3: Concordance between RNA-Seq and RT-PCR



- Scatterplot of recurrence risk hazard ratios of 184 transcripts determined by RT-PCR (x-axis) and RNA-Seq (y-axis)
- Pearson R = 0.813, Lin concordance correlation 0.810

Figure 4: Co-expressed gene networks associated with breast cancer recurrence



Cytoscape was used to evaluate the subset of 597 transcripts that co-expressed with at least one other RNA at R ≥ 0.6

Prominent networks:

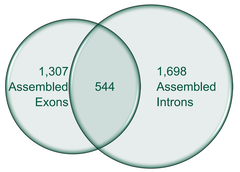
A. associated with cell proliferation
B. co-expressing with estrogen receptor
C. genes mapping to chr17q23-24
D. genes mapping to chr8q21-24
E. genes mapping to chr9q22
F. 134 transcripts heavily enriched in pre-miRNAs and olfactory receptor genes

A significant fraction of the identified RefSeq transcripts are also associated with disease recurrence in published DNA microarray datasets

- The set of RefSeq transcripts associated with breast cancer recurrence was further evaluated by comparison with published microarray data from an independent group of 337 breast cancer patients (the "NKI dataset") merged from two publications.[6,7]
- Metastasis-free survival data were available for 319 patients.
- About half of the RefSeq genes identified by RNA-Seq were detected as present in the microarrays (11,659 genes).
- For RNA-Seq transcripts present at > 100 counts, 44% were associated with metastasis-free survival in the NKI dataset at FDR < 10%.
- Likelihood of no correlation $p < 6 \times 10^{-16}$

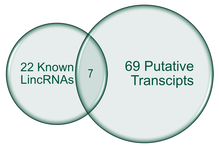
Results (cont.)

Figure 5: Many of the RNAs associated with disease recurrence risk are intronic



- 1307 assembled exons and 1698 assembled introns were associated with breast cancer recurrence (FDR < 10%).
- Approximately 2/3 of the discovered intronic transcripts belong to genes for which their cognate exons are not also discovered.
- Overall, introns and their cognate exons tend to co-express (R = 0.67) but the strength of co-expression varies over a very wide range.

Figure 6: Intergenic transcripts associated with breast cancer recurrence



- We used two approaches to search for intergenic RNAs associated with breast cancer recurrence:
  - Reads mapping to 2,500 well-documented lincRNAs were interrogated.[5] Twenty-two of these associated with recurrence risk at FDR < 10%.
  - A computational algorithm was developed to identify clusters of reads that map to intergenic regions of the genome. A total of 2,101 putative intergenic transcripts were identified of which 775 overlapped with lincRNAs. Expression of 194 of these intergenic transcripts correlated with disease recurrence at FDR < 10% and the list was further condensed by merging clusters that were separated by < 1000 bp and were strongly co-expressed. This produced a list of 69 putative intergenic transcripts of which 7 overlapped with the 22 lincRNA "hits."

Summary and Conclusions

- A method for whole transcriptome RNA-Seq profiling has been developed; results of this 136 patient study indicate that the method:
  - is feasible for biomarker discovery studies using small amounts of fixed paraffin embedded tissue.
  - has sufficient sensitivity and precision to identify Onc*otype* DX genes in archival FFPE breast tumors.
- Effect sizes and statistical significance of transcripts associated with breast cancer recurrence determined by RNA-Seq were similar to RT-PCR.
- RNA-Seq identifed hundreds of additional protein-coding and non-protein-coding transcripts that strongly associated with breast cancer recurrence in this patient cohort.
- Co-expression analysis of RefSeq RNAs revealed networks of genes associated with breast cancer recurrence in earlier published datasets as well as previously unrecognized RNA networks.
- Intronic RNAs were a particularly rich source of "hits" and may carry qualitatively novel prognostic information.
- Recognizing the challenges for development of robust gene signatures associated with clinical variables and the fact that many genes provide redundant recurrence risk information, the transcripts identified here should be explored in additional, independent cohorts of breast cancer patients.

References

1. Bullard JH, Purdom E, Hansen KD, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*. 2010;11:94.
2. Cox DR. Regression models and life tables. *Journal of the Royal Statistical Society, Series B* (Methodological) 1972;34:479.
3. Lin DY and Wei LJ. The robust inference for the Cox Proportional Hazards Model. *Journal of the Royal Statistical Society, Series B*. 1989; 84:1074.
4. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*. 2002;64(3):479-98.
5. Khalil AM, Guttman M, Huarte M, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci USA*. 2009;106(28):11667-72.
6. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*. 2002;347(25):1999-2009.
7. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415(6871):530-6.

This presentation is the intellectual property of the authors. Contact jbaker@genomichealth.com for permission to reprint and/or distribute.

Genomic Health and Oncotype DX are registered trademarks of Genomic Health, Inc.