Biomarker Discovery from Whole Transcriptome Analysis of Formalin-Fixed, Paraffin-Embedded Tumor and Normal Breast Specimens

Dominick Sinicropi, Kunbin Qu, John Morlan, Carl Millward, Francois Collin, James Stephans, Mei-Lan Liu, Mylan Pho, Jennie Jeong, Ranjana Ambannavar, Aaron Scott, and Joffre Baker

Genomic Health, Inc., Redwood City, CA

1. Introduction

- Our goal was to identify expression biomarkers that are prognostic for breast cancer recurrence
- As a first step we used RNA-Seq to identify differentially-expressed transcripts in tumor and normal breast tissues (**ΔTvN transcripts)**
- The prognostic significance of the protein-coding (RefSeq) ΔTvN transcripts was tested using published microarray data from ~3,000 patients
- The prognostic significance of selected protein coding and non-coding ΔTvN transcripts was investigated by RT-PCR in an independent cohort of 136 patients

2. Study Design

- RNA was isolated from 12 tumor (T1-T12) and 12 normal (N1-N12) formalin-fixed (FFPE) breast specimens that had been archived 10 – 13 years
- Two specimens were multiplexed in each lane of Illumina GA IIx flowcells
- Four specimens were analyzed in triplicate to assess technical variability

Lane	Flow Cell #1	Flow Cell #2
1	T1a + N1a	T1c + N1c
2	T1b + N1b	T2b + N2b
3	T2a + N2a	T2c + N2c
4	T3 + N3	T8 + N8
5	T4 + N4	T9 + N9
6	T5 + N5	T10 + N10
7	T6 + N6	T11 + N11
8	T7 + N7	T12 + N12

3. Methods

Preparation of Sequencing Libraries

1 μg total RNA extracted from Formalin-Fixed Paraffin-Embedded (FFPE) breast tumors
↓
Selective removal of 18S and 28S rRNA
↓
Illumina RNA-Seq Kit
↓
Two-sided SPRI selection
↓
18 PCR cycles

Data Analysis Pipeline

Illumina FASTQ Sequences
↓
Sequence Trimming
↓
Mapping to reference sequences by multiple algorithms: CASAVA & Bowtie
↓
Counting of uniquely mapped sequences using CASAVA and Genomic Health Differential Expression Algorithm
↓
Read counts for exons, junctions, introns, genes, ribosomal RNA, mitochondrial transcripts

4. Consistent Number of Mappable Reads Across All Samples

Normal

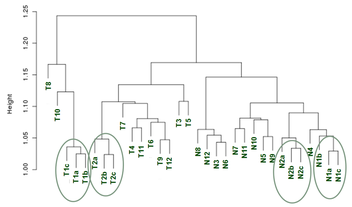
Sample	Number of Reads (Millions)	Percent Mapped*	Percent rRNA
N1 (X3)	59.5	70.4	5.3
N2 (X3)	55.7	64.2	4.9
N3	20.0	63.7	12.3
N4	18.4	69.8	2.8
N5	19.7	59.2	23.8
N6	17.5	62.6	21.1
N7	17.6	54.2	7.7
N8	19.6	63.0	22.6
N9	16.5	65.2	14.2
N10	17.4	58.7	13.2
N11	19.6	63.7	8.7
N12	15.9	60.5	23.7
Average	18.7	63.0	13.4

Tumor

Sample	Number of Reads (Millions)	Percent Mapped*	Percent rRNA
T1 (X3)	54.6	67.1	6.0
T2 (X3)	56.0	68.9	5.5
T3	17.2	69.5	3.9
T4	19.1	66.8	3.6
T5	19.2	65.5	6.5
T6	19.9	62.3	12.8
T7	20.3	70.8	5.5
T8	17.1	67.0	10.4
T9	23.2	71.5	3.6
T10	21.3	71.9	2.4
T11	19.8	64.3	5.6
T12	21.1	67.7	3.3
Average	19.1	67.8	5.7

* Uniquely mapped reads only

5. Technical Replicates Cluster as Expected Based on RefSeq Transcripts



Technical replicates are circled

6. Technical Precision of RNA-Seq

R^2 values

	N1a	N1b	N1c
N1a	1	0.9366	0.9383
N1b		1	0.9054
N1c			1

	T1a	T1b	T1c
T1a	1	0.9444	0.9424
T1b		1	0.9233
T1c			1

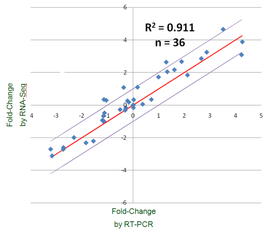
	N2b	N2c	N2a
N2b	1	0.9269	0.9018
N2c		1	0.9162
N2a			1

	T2b	T2c	T2a
T2b	1	0.9462	0.9092
T2c		1	0.9166
T2a			1

Average R^2 within flowcell = 0.9385

Average R^2 between flowcells = 0.9192

Concordance of RNA-Seq with RT-PCR



7. Hundreds of Differentially-Expressed ΔTvN Transcripts were Detected in Both Coding and Non-coding Regions of the Genome

	Differentially Expressed Transcripts*	
Transcript Source	Number of ΔTvN Transcripts	Percent of Total ΔTvN Transcripts
Exonic	2,873	46
Intronic	2,152	35
Intergenic	1,166	19
Total	6,191	100

- We have developed an algorithm to discover both coding and non-coding ΔTvN transcripts by:
  - identifying clusters of mapped transcripts across study samples
  - Defining regions by a maximal likelihood approach

*Parameters for Detection:

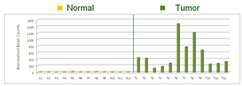
Gap between regions	>30 bp
Fold-Change between N and T	≥2
Number of Reads	≥ 10
Width of Region	≥ 50
Number of Reads/Width	≥ 0.03

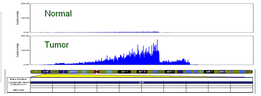
8. Examples of ΔTvN Transcripts

RefSeq Transcripts

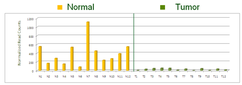
An intergenic non-coding transcript

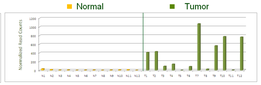
MMP11



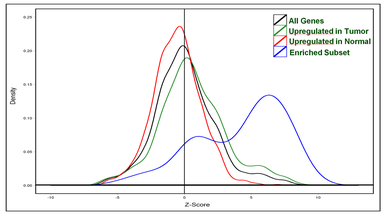


STAC2





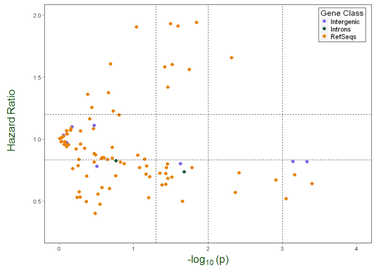
9. Association of RefSeq ΔTvN Transcripts With Clinical Outcome



- The prognostic significance of transcripts that mapped uniquely to RefSeq genes was surveyed using a collection of published microarray data[a] from ~3000 patients with known clinical outcome
- RefSeq ΔTvN Transcripts Were Only Modestly Enriched for Association With Poor Clinical Outcome
- Further Evaluation of ΔTvN (Up-in-Tumor) Transcripts by Gene Set Analysis[b] Produced a Gene Set with a Very High Percentage of Prognostic Genes
- Gene Set Analysis identified a set of 73 RefSeq genes of which 50 had Z-scores ≤ -3 or ≥ 3 for association with recurrence risk
- Functional metadata for the top 50 genes revealed a strong enrichment for genes associated with cell proliferation, chromatin remodeling, and regulation of transcription

[a] Wirapati et al., Breast Cancer Res. 10 (4): R65, 2008 [b] Irizarry et al., Stat. Meth. Med. Res. 18:565-575, 2009

10. Several Protein Coding and Non-Coding ΔTvN Transcripts Associated With Breast Cancer Recurrence as Determined by RT-PCR



- RT-PCR assays were designed for a total of 96 differentially-expressed RefSeq and non-coding transcripts
- A independent cohort of 136 FFPE breast tumors with known recurrence status was assayed using the panel of 96 RT-PCR assays

11. Summary and Conclusions

- We used RNA-Seq to identify in FFPE specimens hundreds of coding and non-coding transcripts that are differentially expressed in tumor versus normal breast tissue (ΔTvN transcripts). Results were concordant with RT-PCR spot check assays.
- Focusing on the ΔTvN **coding** transcripts, we tested the hypothesis that their expression frequently associates with breast cancer outcome, by consulting public DNA microarray datasets that link gene expression to disease outcome (in ~3,000 patients).
  - Selection of transcripts over-expressed in tumors compared with normal tissues produced only a modest enrichment for prognostic RefSeq transcripts
  - However, further evaluation of these ΔTvN transcripts by gene set analysis produced a set of transcripts highly enriched for association with recurrence of breast cancer
- In parallel, RT-PCR assays were designed for several ΔTvN **non-coding** transcripts and were used to screen a 136 patient cohort of breast cancer specimens, of which some proved to associate with breast cancer recurrence risk.
- **Conclusion**: RNA-Seq on FFPE tissue can be used productively to test hypotheses for biomarker discovery.

genomic health