Prognostic Biomarker Discovery using RNA-Seq in Two Cohorts of Breast Cancer Patients
Genomic Health, Inc. Redwood City, CA

1. Introduction

- Gene expression biomarkers are the basis for clinical diagnostic tests such as the 21 gene Oncotype DX® breast cancer assay
- Transcript profiling by next generation sequencing (RNA-Seq) has the potential to greatly expand the number of transcripts that can be surveyed in biomarker discovery studies compared to earlier methods such as RT-PCR and microarrays
- We report here that RNA-Seq of RNA from 212 formalin-fixed paraffin-embedded (FFPE) breast tumor specimens up to 23 years old has sufficient sensitivity and precision to rediscover biomarkers discovered previously as well as thousands of new transcripts arising from protein-coding, intronic, and intergenic regions of the genome

2. Study Design

- Two cohorts of breast tumors (136 and 76 patients) with associated clinical data were analyzed by RNA-Seq
- Standardized hazard ratios for individual coding, intronic, and intergenic 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 Oncotype DX® biomarkers
- Identification of RefSeq transcripts associated with breast cancer recurrence (FDR < 10%) was employed to investigate co-expressed gene networks

3. Methods

Sequencing

- Total RNA was extracted from three 10 µm sections of FFPE tumor specimens 3 – 23 years after fixation using the MasterPure™ kit from Epicentre™ Biotechnologies
- 100 ng of total RNA was depleted of ribosomal and mitochondrial RNA as previously described (AGBT 2011 and MS submitted)
- RNA-Seq libraries were prepared using ScriptSeq™ kits (Epicentre Biotechnologies) with 6 base index sequences for sample multiplexing
- Two libraries were multiplexed for cluster generation in each lane of Illumina flow cells using TrueSeq™ 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

- Primary analysis of sequence data was performed using CASAVA 1.7 from Illumina
  - 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
  - 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, RefSeq RNA counts and adding the cohort mean 3rd quartile similar to the method of Bullard et al. (2010)
  - Hazard ratios for breast cancer recurrence were calculated using univariate Cox proportional hazards regression analysis (Cox 1972) with the robust standard error estimate of Lin and Wei (1989) and false discovery rates (FDRs) were assessed using the method of Storey (2002)

Cohort A: RNA-Seq recovers Oncotype DX genes and hundreds of others that associate with disease recurrence

- Intergenic transcripts
  - Approach 1: Counts that mapped to 2,500 well-documented lincRNAs (Khalil et al., 2009) 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

4. Results

Clinical characteristics of the two cohorts are very different with respect to distribution of nodal status and tumor size

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>No. patients/cohorts (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor size (cm)</td>
<td>55/78 (71%)</td>
</tr>
<tr>
<td>No. lymph nodes at primary diagnosis</td>
<td>23/78 (29%)</td>
</tr>
<tr>
<td>Histological tumor grade</td>
<td>23/78 (29%)</td>
</tr>
<tr>
<td>Histopathological grade</td>
<td>23/78 (29%)</td>
</tr>
<tr>
<td>Clinical tumor stage, date due to recurrence or death</td>
<td>23/78 (29%)</td>
</tr>
</tbody>
</table>

Cohort A: 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 chr17p23-24
- D. genes mapping to chr8p21-24
- E. genes mapping to chr9p22
- F. 134 transcripts heavily enriched in pre-miRNA and splicing receptor genes

Cohort B: RNA-Seq recovers Oncotype DX genes and hundreds of others that associate with disease recurrence

- The effect size is plotted as the standardized hazard ratio and statistical significance as the p-value
- The effect size is plotted as the standardized hazard ratio and statistical significance as the p-value
- The effect size is plotted as the standardized hazard ratio and statistical significance as the p-value
- The effect size is plotted as the standardized hazard ratio and statistical significance as the p-value
- The effect size is plotted as the standardized hazard ratio and statistical significance as the p-value

Concordance of Hazard Ratios for cohort A patients as determined by RNA-Seq and RT-PCR across 184 mRNAs

- Relationships 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
- The effect size is plotted as the standardized hazard ratio and statistical significance as the p-value
- The effect size is plotted as the standardized hazard ratio and statistical significance as the p-value
- The effect size is plotted as the standardized hazard ratio and statistical significance as the p-value
- The effect size is plotted as the standardized hazard ratio and statistical significance as the p-value

5. Conclusions

- Whole transcriptome RNA-Seq profiling is feasible for biomarker discovery studies using small amounts of fixed paraffin embedded tissue
- The methods we developed have sufficient sensitivity and precision to re-discover Oncotype DX genes in archival FFPE breast tumors
- RNA-Seq identified hundreds of additional protein-coding and non-protein-coding transcripts that strongly associated with breast cancer recurrence in these patient cohorts
- Intronic RNAs were a particularly rich source of “hits” and may carry novel recurrence risk information
- The set of RNA-Seq transcripts associated with breast cancer recurrence in cohort A overlapped significantly with published microarray data from an independent group of 337 breast cancer patients and with the set of RefSeq transcripts identified in cohort B
- 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.