Global Transcriptome Analysis of Formalin-Fixed Prostate Cancer Specimens Identifies Biomarkers of Disease Recurrence

Qi Long1,2, Jianpeng Xu1, Adeboye O. Osunkoya3,4,8,9, Soma Sannigrahi3, Brent A. Johnson2, Wei Zhou3,5,6,8, Theresa Gillespie7,8, Jong Y. Park10, Robert K. Nam11,12, Linda Sugar11,12, Aleksandra Stanimirovic11,12, Arun K. Seth11,12, John A. Petros4,5,8,9, and Carlos S. Moreno1,3,8

Abstract

Prostate cancer remains the second leading cause of cancer death in American men and there is an unmet need for biomarkers to identify patients with aggressive disease. In an effort to identify biomarkers of recurrence, we performed global RNA sequencing on 106 formalin-fixed, paraffin-embedded prostatectomy samples from 100 patients at three independent sites, defining a 24-gene signature panel. The 24 genes in this panel function in cell-cycle progression, angiogenesis, hypoxia, apoptosis, PI3K signaling, steroid metabolism, translation, chromatin modification, and transcription. Sixteen genes have been associated with prostate cancer, with five specifically associated with prostate cancer (BTG2, IGFBP3, SIRT1, MXI1, and FDP5). Validation was performed on an independent publicly available dataset of 140 patients, where the new signature panel outperformed markers published previously in terms of predicting biochemical recurrence. Our work also identified differences in gene expression between Gleason pattern 4 and 3 and 3 and 4 tumors, including several genes involved in the epithelial-to-mesenchymal transition and developmental pathways. Overall, this study defines a novel biomarker panel that has the potential to improve the clinical management of prostate cancer. Cancer Res; 74(12); 3228–37. ©2014 AACR.

Introduction

Prostate cancer remains the most common cancer diagnosed for U.S. males, and ranks second among tumor site-specific mortality with over 28,000 deaths per year (1). Currently, both clinicians and patients are faced with difficult treatment decisions when a diagnosis is made of prostate cancer because it is very hard to predict whether a patient is likely to progress to aggressive, metastatic disease. The most common clinical parameters that are used to make prognoses for patients with prostate cancer include age, stage, prostate-specific antigen (PSA) level, positive margins, perineural invasion, and Gleason score. Gleason score is generated by summing the grades of the two most prominent patterns, and somewhat takes into account the heterogeneity of prostate cancers (2, 3). Patients with a Gleason score of eight or higher generally have a poor prognosis, and those with Gleason scores of six or lower generally have favorable prognoses, but it is much more difficult to predict outcomes for the very large number of patients with a Gleason score of seven. Several studies have supported the concept that the Gleason 4 + 3 pattern tumors have worse outcomes than Gleason 3 + 4 tumors (4–6), but clinical parameters alone are not sufficient for accurate prediction of recurrence. Thus, biomarkers that can predict the likely clinical outcome and recurrence for patients after surgical therapy are urgently needed to aid clinicians in deciding on the appropriate course of treatment and in accurate prognostication for patients.

We recently published a panel of 10 protein-coding genes and two miRNA genes that could be used to separate patients with and without biochemical recurrence (BCR) after prostatectomy (7). Postoperative BCR is defined as two detectable PSA readings (>0.2 ng/mL), and generally precedes clinical recurrence with metastases, although BCR does not always predict clinical recurrence, and patients with BCR do not always have clinical metastases. Importantly, this panel of 12 biomarkers could significantly predict clinical recurrence for patients with a Gleason score of seven, but this analysis was limited to a selected set of approximately 522 prostate cancer–relevant genes and 1,146 human miRNAs. To take a more global approach to identify biomarkers of recurrence, we collected formalin-fixed paraffin-embedded (FFPE) prostatectomy samples from three independent sites (Atlanta Veterans Administration Medical
Materials and Methods

Sample selection

We identified 150 cases at the AVAMC hospital between 1990 and 2000 that could potentially be used for this project. We were able to locate slides and FFPE blocks for 100 of those cases, and identified regions of cancer and benign tissue in slides for each of them. These samples were then submitted for processing to obtain 1-mm tissue cores. We prepared RNA from 99 AVAMC samples, of which 79 passed our RNA quality control (QC) analysis for genomic profiling. We also obtained 81 samples from the Sunnybrook Health Science Center (Toronto, ON) and 99 RNA samples from the MCC in Tampa, FL. Approximately 80% of samples passed QC analysis.

Those samples that were included had sufficient RNA yield, met specific inclusion criteria, had available tissue specimens, documented long-term follow-up, and consented to participate or were included by Institutional Review Board consent waiver. MCC subjects were patients with prostate cancer who underwent radical prostatectomy at the MCC between 1987 and 2003. Prostate cancer cases were men 21 years and older who had surgery for their disease at the MCC and had pathologically confirmed primary prostate cancer. AVAMC cases were patients with prostate cancer who underwent radical prostatectomy between 1990 and 2000. University of Toronto (UT) cases were patients with prostate cancer who underwent radical prostatectomy at the Sunnybrook Health Science Center between 1998 and 2006. Exclusion criterion included recurrent patients who underwent hormonal or radiation treatment before radical prostatectomy. The cases were assigned prostate ID numbers to protect their identities. These patients did not receive neoadjuvant or concomitant hormonal therapy. Their demographic, treatment, and long-term clinical outcome data have been collected and recorded in an electronic database. Clinical data recorded include PSA measurements, radiological studies and findings, clinical findings, tissue biopsies, and additional therapies that the subjects may have received. In total, 106 samples from 100 patients were sequenced by next-generation sequencing: 61 from the AVAMC, 35 from Toronto, and 10 from MCC. Of these cases, 49 had BCR and 51 had no BCR. Of those 100 patients, 97 had complete PSA data and were included in construction of biomarker prediction models. We used only 10 of the 99 MCC cases for several reasons including limited RNA yield, the fact that only three cases had BCR, and funding limitations to conduct sequencing analyses.

RNA preparation

Tissue cores (1 mm) were used for RNA preparation (two cores per case) rather than sections because of the heterogeneity of samples and the opportunity for obtaining cores with very high percentage tumor content, except for the 10 samples from MCC. MCC supplied total FFPE RNA that was prepared from 5-μm unstained sections. Hematoxylin and eosin-stained slides were reviewed by a board-certified anatomic pathologist with expertise in genitourinary pathology (A.O. Osunkoya for AVAMC and L. Sugar for UT) to identify regions of cancer to select corresponding areas for cutting of cores from paraffin blocks as previously described (9). Total RNA was prepared at the Winship Cancer Genomics Shared Resource from FFPE cores as previously described (10), using the Omega Biotek FFPE RNA methodology in 96-well format on a MagMax 96 Liquid Handler Robot (Life Technologies). FFPE RNA was quantitated using a Nanodrop spectrophotometer, and tested for RNA integrity and quality by Taqman analysis of the RPL13a ribosomal protein on a HT7900 real-time PCR instrument (Applied Biosystems). Samples with sufficient yield (>500 ng), A260/A280 ratio > 1.8, and RPL13a C_{T} values less than 31 cycles were used for preparation of RNAseq libraries.

Bioinformatics analysis

FASTQ files generated from the Illumina HiSeq2000 were analyzed for quality using FASTQC, mapped to the human genome using TopHat software (version 2.0.8) and Bowtie (version 2.1.0; ref. 11), and Cufflinks software (12) was used to generate fragment per kilobase of transcript per million reads (FPKM) values. Genes were filtered to determine if they were detected (defined as FPKM > 1) in each sample. Genes that were detected in 80% of BCR samples or 80% of non-BCR samples were retained, leaving a set of 5,265 genes for further analysis. Duplicates from six patient samples were removed, and three patients had incomplete clinical data on PSA and pathologic stage, leaving 97 samples for biomarker analysis. Pathway and upstream analyses were performed using the Ingenuity Pathway Analysis (IPA) Knowledgebase (13). Transcriptome coverage was computed based on a transcriptome


Downloaded from cancerres.aacrjournals.org on April 12, 2017. © 2014 American Association for Cancer Research.
size of 81 Mb from the University of California at Santa Cruz (UCSC) known genes set (build hg19). We performed differential gene expression analysis between Gleason pattern 4 + 3 and Gleason pattern 3 + 4 tumors using DESeq (14) in R Bioconductor.

**TaqMan validation**

Individual TaqMan assays were obtained for TMSB10 (Hs1005565), PTN (Hs383235), and SYNM (Hs2326749) and FFPE RNA from 22 samples were tested for correlation with FPKM values. Direct Ct values were used for correlation analysis without a reference gene, because substantial variability was observed for both beta actin (ACTB) and 18S with constant RNA inputs.

**Statistical analysis**

PSA data were skewed with extreme outliers and were log-transformed, log(PSA + 1), in all analyses. To identify important RNA biomarkers and build and evaluate prediction models for prostate cancer recurrence (i.e., BRC), we adopted the following strategy, similar to our earlier work (7). In the training step, the prediction models were built for time to BRC using our RNAseq data. Specifically, we first fit a univariate Cox proportional hazard (PH) model for each individual RNA biomarker, and a set of important RNA biomarkers were then preselected based on a false discovery rate (FDR) threshold of 0.10. Next, we fit a lasso Cox PH model (15, 16) using the preselected RNA biomarkers to identify the optimal panel of RNA biomakers and the resulting prediction score, where the tuning parameter for lasso was selected using a leave-one-out cross-validation technique. On the basis of the RNA biomarker panel, a final prediction model for recurrence was built to also include relevant clinical biomarkers, including pathologic stage, PSA, Gleason score, surgical margin status, and age, through fitting a Cox PH model. For comparison, we also built a prediction model using only clinical information, including pathologic stage, PSA, Gleason score, and age, through fitting a Cox PH model. In the validation step, we first used our samples to evaluate each prediction model; specifically, we calculated the predictive scores for subjects based on each prediction model, divided subjects into high (poor score) and low (good score) risk groups based on the median predictive score, and performed the log-rank test to compare BCR between the two risk groups. Subsequently, we repeated the validation procedure using an independent gene expression microarray study with data from 140 patients with prostate cancer (17). To compare performance of different prediction models, we used three metrics for

![Figure 1](cancerres.aacrjournals.org) Representative tracks of coverage of RNAseq reads from the IGFBP3 gene from the Integrated Genomics Viewer (IGV). Coverage is primarily focused in exons, but some intronic coverage is evident. Little, if any, coverage is observed in intergenic regions, indicating that intronic reads are likely from partially processed RNAs due to the lack of a poly-A selection step.
evaluating predictive accuracy, integrated discrimination improvement (IDI), net reclassification improvement (NRI), and median improvement in risk score for censored survival outcomes as described (18).

Results

Genome-wide RNA sequencing analysis

We identified prostatectomy cases and FFPE blocks at the AVAMC hospital between 1990 and 2000, and also obtained prostatectomy FFPE cores from the Sunnybrook Health Science Center (Toronto, ON) and prostatectomy FFPE RNA samples from the MCC. We prepared RNA from two FFPE 1-mm cores per block, and following QC analysis, prepared a total of 106 RNAseq libraries from 100 patients: 61 from the AVAMC, 35 libraries from 29 patients from Toronto, and 10 from MCC. Of these cases, 49 had BCR and 51 had no BCR. Of these 100 cases, 97 had complete associated clinical data and were used for predictive model building. A summary of the clinical characteristics of these samples is provided in Supplementary Table S1. Detailed clinical data are provided in Supplementary Table S2.

Because RNA from FFPE samples is typically fragmented, conventional methods for sequencing library preparation that use poly-dT hybridization to capture fully processed mRNA are not optimal. We tested alternative library preparation methods and developed a protocol for robust RNAseq analysis of FFPE samples using the Ribominus Kit (Ambion) to remove ribosomal RNA, followed by library preparation using the Illumina TruSeq Kit. We determined that multiplexing three samples per lane gave us adequate coverage for RNAseq analysis. In total, we generated approximately 490 billion base pairs (Gbp) of sequence, of which 294 Gbp mapped uniquely to the human genome (build 19, hg19). In total, we obtained 5.874 billion mapped reads. The average number of mapped reads was 55.4M reads per sample, providing an average coverage of 34.2× for the human transcriptome (UCSC known genes hg19). A summary of the sequencing output is provided in Supplementary Table S3.

Possibly due to the fact that we did not perform a poly-A selection step, we observed a significant number of reads that mapped to gene introns, likely from partially processed mRNAs. We did not observe large numbers of reads mapping to intergenic regions, indicating that there was no DNA contamination in our samples. A representative screenshot for the mapped reads is given in Fig. 1. Moreover, the level of intronic reads was similar to that observed in RNAseq data derived from fresh-frozen samples analyzed by The Cancer Genome Atlas project (Supplementary Fig. S1). To validate and verify the accuracy of our RNAseq data, we performed TaqMan analyses on a few select genes and observed high correlation of TaqMan and RNAseq data for these genes ($r^2 = 0.80–0.97$; Supplementary Fig. S2). In addition, we analyzed expression of genes that are typically involved in chromosomal translocations such as ERG, ETV1, ETV4, ETV5, and SPINK1. We observed mutually exclusive high levels of expression of ERG (45% of samples), ETV1 (6%), ETV4 (4%), and SPINK1 (10%) in proportions similar to previous studies (19–22; Fig. 2).
0.96) for the 5,265 genes that were robustly detected in at least 80% of samples and used in our biomarker analyses (Supplementary Fig. S3). The pair of samples with the lowest correlation (UTPC034) had the greatest difference in number of mapped reads (18M vs. 112M), whereas the paired samples with the highest correlation (UTPC004) both had very deep coverage (94M and 101M mapped reads each). Differential gene expression analysis using DESeq indicated very few differentially expressed genes between replicate sequencing libraries (14 genes on average). For our biomarker analysis, in each case, we used the library with the higher number of mapped reads.

Biomarker analysis

FASTQ files generated from the Illumina HiSeq2000 were mapped to the human genome using TopHat software (version 2.0.8) and Bowtie (version 2.1.0; ref. 11), and Cufflinks software (12) was used to generate FPKM values. Genes were filtered to determine if they were detected (defined as FPKM > 1) in each sample. First, we analyzed the distribution of detected genes, and observed a peak of genes that were detected in 80% or more of the samples (Supplementary Fig. S4). To avoid excluding those genes that were not expressed in one of the two groups, we elected to retain genes detected in either 80% of BCR or 80% of non-BCR samples. Genes that were detected in 80% of BCR

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Gene name</th>
<th>Function</th>
<th>Cancers</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTG2</td>
<td>BTG family, member 2</td>
<td>Cell cycle, apoptosis</td>
<td>Prostate, breast, lung</td>
<td>-0.08831</td>
</tr>
<tr>
<td>CDC37L1</td>
<td>Cell division cycle 37 homolog (S. cerevisiae)-like 1</td>
<td>Cell cycle</td>
<td>Head and neck</td>
<td>-0.13272</td>
</tr>
<tr>
<td>COL15A1</td>
<td>Collagen, type XV, alpha 1</td>
<td>Angiogenesis, extracellular matrix</td>
<td>Kidney</td>
<td>0.007957</td>
</tr>
<tr>
<td>COL3A1</td>
<td>Collagen, type III, alpha 1</td>
<td>Extracellular matrix</td>
<td>Breast, ovarian, lung, brain, gastric</td>
<td>0.27692</td>
</tr>
<tr>
<td>EIF2D</td>
<td>Eukaryotic translation initiation factor 2D</td>
<td>Translation</td>
<td>na</td>
<td>0.40303</td>
</tr>
<tr>
<td>FDPS</td>
<td>Farnesyl diphosphate synthase</td>
<td>Sterol biosynthesis</td>
<td>Breast, ovarian, head and neck</td>
<td>0.019765</td>
</tr>
<tr>
<td>HIST1H1C</td>
<td>Histone cluster 1, H1c</td>
<td>Chromatin</td>
<td>na</td>
<td>0.088821</td>
</tr>
<tr>
<td>HIST1H2BG</td>
<td>Histone cluster 1, H2bg</td>
<td>Chromatin</td>
<td>na</td>
<td>0.067189</td>
</tr>
<tr>
<td>IFT57</td>
<td>Intraflagellar transport 57 homolog (Chlamydomonas)</td>
<td>Apoptosis</td>
<td>GBM</td>
<td>0.274924</td>
</tr>
<tr>
<td>IGFBP3</td>
<td>Insulin-like growth factor binding protein 3</td>
<td>Insulin signaling PI3K</td>
<td>Prostate, breast, colon, many</td>
<td>0.076922</td>
</tr>
<tr>
<td>ITPR1</td>
<td>Inositol 1,4,5-triphosphate receptor, type 1</td>
<td>PIP3, apoptosis</td>
<td>Prostate, breast, leukemia</td>
<td>-0.27849</td>
</tr>
<tr>
<td>LBH</td>
<td>Limb bud and heart development homolog (mouse)</td>
<td>Transcription</td>
<td>Breast</td>
<td>0.098016</td>
</tr>
<tr>
<td>LOC284801</td>
<td>LOC284801</td>
<td>Unknown</td>
<td>E3 ubiquitin ligase of the mitochondria</td>
<td>na</td>
</tr>
<tr>
<td>MARCH5</td>
<td>Membrane-associated ring finger (C3HC4) S</td>
<td>Unknown</td>
<td>na</td>
<td>-0.25456</td>
</tr>
<tr>
<td>MED4</td>
<td>Mediator complex subunit 4</td>
<td>Transcription, chromatin signaling</td>
<td>Breast, cervical</td>
<td>-0.14</td>
</tr>
<tr>
<td>MEMO1</td>
<td>Mediator of cell motility 1; similar to mediator of cell motility 1</td>
<td>Insulin signaling, EMT, estrogen signaling</td>
<td>Breast, ALL</td>
<td>0.116525</td>
</tr>
<tr>
<td>MFI1</td>
<td>MAX interactor 1</td>
<td>Transcription, MYC activity</td>
<td>Prostate, melanoma, kidney, bladder</td>
<td>-0.03462</td>
</tr>
<tr>
<td>PTN</td>
<td>Pleiotrophin</td>
<td>Angiogenesis</td>
<td>Prostate, colorectal, bladder, breast</td>
<td>-0.16271</td>
</tr>
<tr>
<td>RPL23AP53</td>
<td>Ribosomal protein L23a pseudogene 53</td>
<td>Ribosomes</td>
<td>na</td>
<td>-0.12988</td>
</tr>
<tr>
<td>SACM1L</td>
<td>SAC1 suppressor of actin mutations 1-like (yeast)</td>
<td>Mitotic spindles, golgi</td>
<td>ALL</td>
<td>-0.02007</td>
</tr>
<tr>
<td>SIRT1</td>
<td>Sirtuin (silent mating type information regulation 2 homolog) 1 (S. cerevisiae)</td>
<td>Transcription, chromatin signaling</td>
<td>Prostate, breast, many others</td>
<td>-0.08227</td>
</tr>
<tr>
<td>SNORA20</td>
<td>Small nuclear RNA, H/ACA box 20</td>
<td>Ribosomes</td>
<td>na</td>
<td>-0.10131</td>
</tr>
<tr>
<td>SRSF3</td>
<td>Serine/arginine-rich splicing factor 3</td>
<td>Splicing, mRNA transport</td>
<td>Colon, ovarian</td>
<td>-0.02622</td>
</tr>
<tr>
<td>SYNM</td>
<td>Synemin, intermediate filament protein</td>
<td>Intermediate filament</td>
<td>Breast, GBM, liver</td>
<td>-0.0574</td>
</tr>
</tbody>
</table>

Abbreviations: na, not applicable; GBM, glioblastoma.
samples or 80% of non-BCR samples included a set of 5,265 protein-coding or noncoding genes for further analysis. Duplicates from six patient samples were removed, and three patients had incomplete clinical data on PSA and pathologic stage, leaving 97 samples for biomarker analysis.

Using the set of 5,265 genes from 97 samples, we built a 24-gene prediction model using a preselection step and a lasso Cox PH model (Table 1), and the final prediction model was built to include the predictive score based on this panel of 24 markers as well as the relevant clinical biomarkers, including pathologic stage, PSA, Gleason score, surgical margin status, and age (Table 2). For comparison, we also built a prediction model using only clinical information, namely, pathologic stage, PSA, Gleason score, surgical margin status, and age, through fitting a Cox PH model (Supplementary Table S4). We then performed log-rank tests to compare BCR between the low-risk (good score) and high-risk (poor score) groups with clinical variables only (Fig. 3A and B), and with both clinical parameters and RNAseq data (Fig. 3C and D). The Kaplan–Meier analysis (Fig. 3A and C) demonstrated that these markers could significantly discriminate patients at higher and lower risk of recurrence by the log-rank test \( P = 1.45 \times 10^{-8} \) in our training data, more significant than using clinical variables alone \( (P = 5.39 \times 10^{-8}) \). The improvements of the full model in Table 2 over the model using only clinical parameters in prediction as measured by IDI, NRI, and median improvement in risk score for censored survival outcomes (18) were all statistically significant (Table 3).

To validate this panel of biomarkers, we identified an independent gene expression microarray study with data from 140 patients with prostate cancer (17). Using the data from Taylor and colleagues (17), we evaluated the final prediction models obtained from the training phase. Each prediction model from the training phase was used to generate a predictive score for each subject in the testing data set, and subjects were subsequently divided into the high- and low-risk groups using the median predictive score. Log-rank tests were performed to compare time to BCR between the high (poor score) and low (good score) risk groups. We first tested our prediction model based on clinical variables alone (Fig. 3B), showing significant discriminative performance in the validation data as well \( (P = 2.85 \times 10^{-3}) \). In addition, we observed that the full panel including RNA biomarkers and clinical variables was very significantly prognostic in this independent validation set \( (P = 7.87 \times 10^{-5}) \), Fig. 3D).

**Comparison with existing biomarkers**

In addition, we performed a direct comparison of our 24 biomarker genes with a set of 31 cell-cycle progression genes developed by Myriad Genetics (8). The Myriad panel performed less well in our dataset \( (P = 4.94 \times 10^{-8}, \text{Fig. 3E}) \) than our biomarker panel at a level similar to clinical parameters alone. In the validation set from Taylor and colleagues (17), our panel of biomarker genes outperformed the 31 cell-cycle progression genes, which had less significant \( P \) value for the log-rank test \( (P = 1.4 \times 10^{-4}, \text{Fig. 3F}) \). Furthermore, analysis for IDI, NRI, and median improvement in risk score demonstrated that our panel was statistically significantly better in prediction of recurrence than the Myriad panel (Table 3).

**Pathway analysis**

To evaluate the pathways that might be associated with BCR, we selected a list of 894 genes with a univariate Cox PH \( q \) value < 0.10 (Supplementary Table S5) and subjected it to pathway enrichment and upstream analyses. The most highly enriched biologic function pathways in this gene set (Supplementary Table S6) included RNA processing \( (P = 5.1 \times 10^{-13}) \), RNA expression \( (P = 9.4 \times 10^{-13}) \), apoptosis \( (P = 1.2 \times 10^{-8}) \), cellular proliferation \( (P = 5.0 \times 10^{-8}) \), and cell autophagy \( (P = 1.4 \times 10^{-6}) \). The most significantly enriched canonical pathways from this analysis (Supplementary Table S7) included mTOR signaling \( (P = 6 \times 10^{-6}) \), p70S6K signaling \( (P = 5.3 \times 10^{-5}) \), PI3K/AKT signaling \( (P = 6.0 \times 10^{-5}) \), actin regulation by actin-related protein-Wiskott-Aldrich Syndrome Protein (ARP-WASP) \( (P = 1.0 \times 10^{-4}) \), integrin-linked kinase (ILK) signaling \( (P = 1.1 \times 10^{-4}) \), embryonic stem cell pluripotency \( (P = 6.5 \times 10^{-4}) \), and integrin signaling \( (P = 6.6 \times 10^{-4}) \). In addition, an upstream analysis of the compounds and transcription factors that affect this set of genes identified a number of compounds and pathways that are associated with prostate cancer progression (Supplementary Table S8), including sirolimus, topotecan, HIF1A, dihydrotestosterone, and TP53. Of the top 12 upstream regulators, 10 included the androgen receptor in the mechanistic network, and all 12 included TP53. Finally, we performed IPA analysis on the final set of 24 biomarker genes, and found that it was significantly associated with malignant neoplasm of the abdomen \( (P = 7.59 \times 10^{-5}) \) with 15 genes annotated for this disease \( (\text{BTG2, COL15A1, COL3A1, FDS, HIST1H1C, IFT57, IGBP3, ITPR1, MARCH5, MED4, MEMO1, MXII, P73, SIRT1, and SYNM}) \), as well as for pelvic cancer \( (P = 2.98 \times 10^{-6}) \) with a subset of 10 genes having this annotation.

**Gleason score analysis**

The grading system for prostate cancer is unique in that the final pathologic grade is a Gleason sum obtained by assigning a single Gleason grade to the most prevalent pattern, known as the primary pattern and then adding this to another single
Gleason grade assigned to the next most prevalent pattern, known as the secondary pattern, to obtain a sum known as the Gleason score. It has been suggested that primary Gleason 4 pattern and Gleason 3 pattern tumors represent different disease states (23), and several studies have supported the concept that the primary Gleason pattern of patients with a Gleason score of seven is predictive of outcome (4–6).

To investigate the differences in gene expression, we performed DEseq differential gene expression analysis comparing 43 samples with Gleason 3 + 4 (primary pattern 3) with 22 samples with Gleason 4 + 3 (primary pattern 4). We identified 304 genes differentially expressed between these two patient groups (Supplementary Table S12). There were several genes differentially expressed relevant to prostate cancer, including upregulated (miR16A, Twist, HOXC6, and AR) and downregulated (ERG, EGF, SOX9, WIF1, WNT5A, and SHH) genes in Gleason 4 + 3 compared with 3 + 4 cases. IPA pathway analysis also determined that the top biologic functions

Figure 3. Kaplan–Meier survival curves for 97 cases in the training set analyzed by RNAseq (A, C, and E) and for 140 cases in the validation set from Taylor and colleagues (17; B, D, and F). Kaplan–Meier curves using clinical parameters alone are shown in A and B. Kaplan–Meier curves using clinical parameters combined with the 24 RNA biomarker genes as described in Table 2 are shown in C and D. Kaplan–Meier curves using 31 biomarkers from Myriad Genetics are shown in E and F.
associated with the 304 differentially expressed genes were abnormal bone morphology \((P = 7.72e-8)\), cell differentiation \((P = 4.05e-7)\), genital tumors \((P = 1.66e-6)\), and prostatic bud formation \((P = 7.48e-6)\).

**Discussion**

There is a great need for robust biomarkers to predict which tumors are more likely to result in different clinical outcomes to optimize treatment decisions. Through RNAseq analysis of 100 FFPE prostatectomy samples, we have identified a new set of 24 biomarker genes that are highly predictive of BCR in patients with prostate cancer. Biomarker studies in prostate cancer have faced a challenge in which the number of available frozen specimens with very long-term follow-up \((> 7 \text{ years})\) is fairly limited. In this study, we have demonstrated that RNAseq analysis of FFPE prostate tissues is feasible and can provide important insight into prostate cancer progression.

Of the 24 biomarker genes, at least 16 \((BTG2, CDC27L1, COL15A1, COL3A1, FDPS, IFT57, IGFBP3, INTP1, LBH, MED4, MEMO1, MXI1, PTN, SACM1L, SIRT1, SRSF3, and SYNM)\) have been previously associated with some type of cancer, and five \((BTG2, \text{ref.} 24; IGFBP3, \text{refs.} 25–27; MXI1, \text{ref.} 28; FDPS, \text{ref.} 29; \text{and} SIRT1, \text{refs.} 30, 31)\) have been previously associated with prostate cancer. There are several biologic functions that can be used to functionally categorize these biomarker genes. First, there is a set of genes associated with the cytoskeleton, extracellular matrix, angiogenesis, and hypoxia, including SYNM \((32)\), COL15A1, COL3A1, PTN \((33)\), IFT57 \((34)\), and MXI1 \((35)\).

Next, there is a set of cell-cycle-related genes, including CDC27L1, BTG2 \((24)\), and SACM1L. In addition, there are several genes associated with apoptosis, PI3K, insulin, metabolism, and steroid hormone signaling, which includes IFT57 \((34)\), INTP1 \((36)\), IGFBP3 \((37)\), MEMO1 \((38)\), FDPS \((29)\), SRSF3 \((39)\), MARCH5 \((40)\), and MED4 \((41–43)\). Another functional category contains genes associated with ribosomes and translation, including RPL23AP53, SNORA20, and EIF2D. Finally, there is a set of genes relevant to transcription and chromatin modification, which includes HIST1H1C, HIST1H2BG, MED4 \((41–43)\), LBH \((44)\), SIRT1 \((30, 31)\), and MXI1 \((35)\).

FFPE specimens with long-term follow-up clinical data are much more abundant than frozen tissues, and successful sequence analysis of this resource opens up this large resource for further analysis. To date, very few studies have reported RNAseq analysis of FFPE tissues with the notable exception of breast cancer \((45)\). Development of biomarkers from FFPE tissues also increases the likelihood that they can be translated into clinically useful biomarkers that can be used in traditional clinical practice that may not have access to frozen specimens. Nevertheless, challenges remain in translating an RNAseq biomarker test to the clinic without additional validation using alternative platforms. Thus, further validation using TaqMan, NanoString, Digital Droplet PCR, BioMark PCR, or gene-focused RNAseq methods such as PCR-amplion or capture sequencing in a Clinical Laboratory Improvement Amendments laboratory environment would be important before developing this biomarker panel into a clinical test. This would be important both to simplify the assay and to provide independent validation that the data from RNAseq of FFPE samples are sufficiently robust across technologies. Future experiments are currently planned to determine the optimal approach for developing this biomarker panel.

RNAseq may be too complex and expensive for clinical translation, and thus, development of alternative assays such as TaqMan, digital PCR, or Nanostring type assays may be more realistic. In addition, the fact that approximately 20% of FFPE RNA samples failed to meet our QC criteria is another challenge. It remains to be seen whether alternative assays can accommodate more challenging specimens with more degraded RNA. Furthermore, additional validation of this panel in another large, independent set of patient samples will be important to increase the robustness of this panel. Validation studies on separate groups of patients will speed translation of these biomarkers into a clinical lab test that could be translated to widespread clinical application that would give

---

**Table 3.** Comparison of prediction performance of the full model, including the RNA biomarkers and clinical parameters with other models that include clinical variables alone, RNA biomarkers alone, or the MYRIAD model in terms of IDI, NRI, and median improvement in risk score for censored survival outcomes \((18)\)

<table>
<thead>
<tr>
<th>Training set</th>
<th>IDI</th>
<th>P value</th>
<th>NRI</th>
<th>P value</th>
<th>Median improvement in risk score</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model vs. clinical variables only</td>
<td>0.469</td>
<td>0.024</td>
<td>0.875</td>
<td>0.022</td>
<td>0.394</td>
<td>0.004</td>
</tr>
<tr>
<td>Full model vs. RNA biomarkers only</td>
<td>0.218</td>
<td>0.036</td>
<td>0.718</td>
<td>0.043</td>
<td>0.2</td>
<td>0.005</td>
</tr>
<tr>
<td>Full model vs. MYRIAD model</td>
<td>0.676</td>
<td>0.03</td>
<td>0.875</td>
<td>0.045</td>
<td>0.634</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Validation set</th>
<th>IDI</th>
<th>P value</th>
<th>NRI</th>
<th>P value</th>
<th>Median improvement in risk score</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model vs. clinical variables only</td>
<td>0.699</td>
<td>0.019</td>
<td>0.669</td>
<td>0.042</td>
<td>0.678</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Full model vs. RNA biomarkers only</td>
<td>0.005</td>
<td>0.643</td>
<td>–0.439</td>
<td>0.246</td>
<td>–0.002</td>
<td>0.986</td>
</tr>
<tr>
<td>Full model vs. MYRIAD model</td>
<td>0.601</td>
<td>0.027</td>
<td>0.625</td>
<td>0.051</td>
<td>0.652</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

**NOTE:** A positive value in all three metrics, including IDI and NRI, indicates an improvement over the second model. Significant P values (bold) indicate a statistically significant improvement of the full model over other models in prediction of BCR.
physicians an idea about what is the best course of treatment for patients with prostate cancer and help avoid unnecessary treatments. Future studies that apply these biomarkers to biopsy or biofluid samples from patients who undergo radiation or active surveillance could determine whether they are also useful for discriminating aggressive from indolent disease. In the long run, this will result in better patient outcomes and reduced healthcare costs and treatment side effects.

The comparison of Gleason 3 + 4 with 4 + 3 cases identified several interesting genes that suggest that increasing miR10A, Twist, HOXC6, and AR expression and decreasing SOX9, WIF1, and WNT5A are associated with increasing tumor grade. Moreover, the fact that the most significant biologic annotation was abnormal bone morphology suggests that higher Gleason primary pattern tumors intrinsically express genes that may facilitate metastasis to the bone. Future studies will be needed to determine the functional role of these genes associated with bone morphology in prostate cancer progression and metastasis.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors’ Contributions

References


Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): A.O. Osunkoya, W. Zhou, T. Gillespie, J.V. Park, R.K. Nam, A. Stanimirovic, J.A. Petros, C.S. Moreno

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): Q. Long, J. Xu, A.O. Osunkoya, B.A. Johnson, R.K. Nam, C.S. Moreno


Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): Q. Long, A.O. Osunkoya, S. Sannigrahi, T. Gillespie, I. Sugar, A.K. Seth, C.S. Moreno

Study supervision: A.O. Osunkoya, J.A. Petros, C.S. Moreno

Acknowledgments
The authors thank Sebri Ramalingam for technical assistance.

Grants Support
This research was supported by Department of Defense (DOD) grant W81XWH-09-1-0990, NIH R01CA100826, NIH U01 CA166449, NIH R01CA173779, NIH R01CA188006, R01CA128813. Canadian Cancer Society Research Institute grant 019038, as well as a National Cancer Institute Cancer Center Support grant P30CA138292 for the Winship Cancer Genomics Shared Resource and the Emory Integrated Genomics Core.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received September 18, 2013; revised March 19, 2014; accepted March 31, 2014; published OnlineFirst April 8, 2014.


44. Rieger ME, Simas AH, Coats ER, Clarke RB, Bridel KJ. The embryonic transcription cofactor LBH is a direct target of the Wnt signaling pathway in epithelial development and in aggressive basal subtype breast cancers. Mol Cell Biol 2010;30:4267–79.
