Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence

Qi Long2,3,†, Jianpeng Xu1,†, Adeboye O. Osunkoya3,4,5,6, Soma Sannigrahi4, Brent A. Johnson2, Wei Zhou3,4,7,8, Theresa Gillespie3,9, Jong Y. Park10, Robert K. Nam11, Linda Sugar11, Aleksandra Stanimirovic11, Arun K. Seth11, John A. Petros3,4,5,6,7, and Carlos S. Moreno1,3,4,*

1Department of Biomedical Informatics, Emory University, Georgia
2Department of Biostatistics and Bioinformatics, Emory University, Georgia
3Winship Cancer Institute, Emory University, Georgia
4Department of Pathology and Laboratory Medicine, Emory University, Georgia
5Urology Department, Emory University, Georgia
6Atlanta VA Medical Center, Decatur, Georgia
7Department of Hematology and Medical Oncology, Emory University, Tampa, FL
8Department of Human Genetics, Emory University, Tampa, FL
9Department of Surgery, Emory University, Tampa, FL
10Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL
11Department of Laboratory Medicine and Pathobiology, University of Toronto, and Department of Anatomic Pathology, Sunnybrook Health Sciences Centre, Toronto, CA

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 (FFPE) 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 cancer with five specifically associated with prostate cancer (BTG2, IGFBP3, SIRT1, MX1 and FDPS). 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 (BCR). Our work also identified differences in gene expression between

*To whom correspondence should be addressed: Carlos S. Moreno, Ph.D., Associate Professor, Department of Pathology & Laboratory Medicine, Winship Cancer Institute, Emory University, Whitehead Research Building, Rm. 105J, 615 Michael St., Atlanta GA, 30322, (404) 712-2809 (Ph), (404) 727-8538 (Fax), cmoreno@emory.edu.
†Equal contribution.

Conflict of Interest: The authors have no conflicts of interest to disclose.
Gleason Pattern 4+3 and 3+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.

**Keywords**

Prostate Cancer; Recurrence; Biomarkers

**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 prostate cancer patients 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 in order to aid clinicians in deciding on the appropriate course of treatment and in accurate prognostication for patients.

We recently published a panel of ten protein-coding genes and two miRNA genes that could be used to separate patients with and without biochemical recurrence after prostatectomy (7). Postoperative biochemical recurrence (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 1146 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 VA Medical Center, Sunnybrook Health Sciences Centre at the University of Toronto, and Moffitt Cancer Center), and prepared libraries for RNA sequencing. We performed RNA-sequencing on 106 prostatectomy samples from 100 patients (61 from AVAMC, 29 from Toronto, and 10 from Moffitt) using the Illumina HiSeq2000 instrument, and identified a new set of biomarkers of biochemical recurrence composed of a 24-gene panel. This panel showed significant improvement of prediction of BCR over clinical
parameters alone and over previously described biomarker panels based on cellular proliferation (8).

**Materials and Methods**

**Sample Selection**

We identified 150 cases at the Atlanta Veterans Administration Medical Center (AVAMC) hospital between 1990–2000 that could potentially be used for this project. We were able to locate slides and formalin-fixed paraffin embedded (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 Moffitt Cancer Center (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 IRB consent waiver. MCC subjects were prostate cancer patients who underwent radical prostatectomy at the Moffitt Cancer Center between 1987–2003. Prostate cancer cases were men 21 years and older who had surgery for their disease at the Moffitt Cancer Center and had pathologically confirmed primary prostate cancer. AVAMC cases were prostate cancer patients who underwent radical prostatectomy between 1990–2000. University of Toronto (UT) cases were prostate cancer patients who underwent radical prostatectomy at the Sunnybrook Health Science Center between 1998–2006. Exclusion criterium 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 neo-adjuvant 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 biochemical recurrence (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 ten samples from MCC. MCC supplied total FFPE RNA that was prepared from five micron unstained sections. H&E stained slides were reviewed by a board certified Anatomic Pathologist with expertise in...
genitourinary pathology (AOO for AVAMC, LS 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, Carlsbad, CA). FFPE RNA was quantitated using a Nanodrop spectrophotometer (Wilmington, DE), and tested for RNA integrity and quality by Taqman analysis of the RPL13a ribosomal protein on a HT7900 real-time PCR instrument (Applied Biosystems, Foster City, CA). Samples with sufficient yield (>500 ng), $A_{260}/A_{280}$ ratio > 1.8 and RPL13a $C_T$ values less than 31 cycles were used for preparation of RNAseq libraries.

RNA Sequencing

We prepared RNA sequencing libraries using 106 prostatectomy RNA samples from 100 patients and performed 50bp paired-end sequencing using the Illumina HiSeq platform. RNA sequencing libraries were prepared from 2 µg of Total RNA using the TruSeq kit (Illumina, Inc) with the following modification. Instead of purifying poly-A RNA using poly-dT primer beads, we removed ribosomal RNA using the Ribominus kit (Ambion). All other steps were performed according to the manufacturer’s protocols. RNAseq libraries were analyzed for QC and average size of inserts were approximately 200–300 bp. Samples were multiplexed into three samples per lane on the Illumina Version 3 Flowcells on the HiSeq2000 platform. Samples were sequenced at the Emory GRA Genomics Core Facility, Hudson Alpha Institute, and at the Southern California Genotyping Consortium (SCGC). The complete dataset can be accessed in the GEO database with accession number GSE54460.

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) (11), and Cufflinks software (12) was used to generate fragment per kilobase 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 5265 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 Knowledgebase (13). Transcriptome coverage was computed based on a transcriptome size of 81Mb from the 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,
since substantial variability was observed for both 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., biochemical recurrence), we adopted the following strategy, similar to our earlier work (7). In the training step, the prediction models were built for time to biochemical recurrence 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 biomarkers and the resulting prediction score, where the tuning parameter for lasso was selected using a leave-one-out cross-validation technique. Based on 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 prostate cancer patients (17). To compare performance of different prediction models we employed three metrics for 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 Atlanta Veterans Administration Medical Center (AVAMC) hospital between 1990–2000, and also obtained prostatectomy FFPE cores from the Sunnybrook Health Science Center (Toronto, ON) and prostatectomy FFPE RNA samples from the Moffitt Cancer Center (MCC) in Tampa, FL. We prepared RNA from two FFPE 1mm 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 biochemical recurrence (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 Supplemental Table 1. Detailed clinical data are provided in Supplemental Table 2.
Since RNA from FFPE samples is typically fragmented, conventional methods for sequencing library preparation that utilize 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 were 55.4M reads/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 Supplemental Table 3.

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 Figure 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 (TCGA) (Supplementary Figure 1). 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 Figure 2). 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) (Figure 2).

We also prepared RNA from separate cores from the same six patients, and prepared separate sequencing libraries on different days. Analysis of the fragment per kilobase of transcript per million mapped reads (FPKM) values from these replicate sequence analyses indicated very strong correlation ($r^2 = 0.70–0.96$) for the 5,265 genes that were robustly detected in at least 80% of samples and used in our biomarker analyses (Supplementary Figure 3). The pair of samples with the lowest correlation (UTPC034) had the greatest difference in number of mapped reads (18M vs. 112M), while 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) (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 Figure 4). 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 samples or 80% of non-BCR samples included a set of 5265 protein-coding or non-coding 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 pre-selection 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 4). 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 (Figure 3A and 3B), and with both clinical parameters and RNaseq data (Figure 3C and 3D). Kaplan-Meier analysis (Figure 3A and 3C) demonstrated that these markers could significantly discriminate patients at higher and lower risk of recurrence by the log-rank test (p = 1.45e-21) in our training data, more significant than using clinical variables alone (p = 5.39e-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 prostate cancer patients (17). Using the data from Taylor et al. (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 (Figure 3B), showing significant discriminative performance in the validation data as well (p = 2.85e-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 = 7.87e-5, Figure 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 = 4.94e-8, Figure 3E) than our biomarker panel at a level similar to clinical parameters alone. In the validation set from Taylor et al. (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 = 1.4e-4, Figure 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 CoxPH q-value < 0.10 (Supplementary Table 5) and subjected it to pathway enrichment and upstream analyses. The most highly enriched biological function pathways in this gene set (Supplementary Table 6) included RNA processing (p = 5.1e-13), RNA expression (p = 9.4e-13), apoptosis (p = 1.3e-8), cellular proliferation (p = 5.0e-8), and cell autophagy (p = 1.4e-6). The most significantly enriched canonical pathways from this analysis (Supplementary Table 7) included mTOR signaling (p = 6e-6), p70S6K signaling (p = 5.3e-5), PI3K/AKT signaling (p = 6.0e-5), Actin regulation by ARP-WASP (p=1.0e-4), ILK signaling (p = 1.1e-4), embryonic stem cell pluripotency (p = 6.5e-4), and Integrin Signaling (p = 6.6e-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 8) including sirolimus, topotecan, HIF1A, dihydrotestosterone, and TP53. Of the top twelve upstream regulators, ten included the androgen receptor in the mechanistic network, and all twelve 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 = 7.59e-05) with fifteen genes annotated for this disease (BTG2, COL15A1, COL3A1, FDPS, HIST1H1C, IFT57, IGFBP3, ITPR1, MARCH5, MED4, MEMO1, PTN, SIRT1, SYNM), as well as for pelvic cancer (p = 2.98E-06) with a subset of 10 genes having this annotation.

Gleason score analysis

The grading system for prostate cancer is unique in that the final pathological grade is a Gleason sum obtained by assigning a single Gleason grade to the most prevalent pattern, known as the primary patterns 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 Gleason seven patients 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) to 22 samples with Gleason 4+3 (primary pattern 4). We identified 304 genes differentially expressed between these two patient groups (Supplementary Table 12). There were several genes differentially expressed relevant to prostate cancer, including up-regulated (miR10A, Twist, HOXC6, AR) and down-regulated (ERG, EGF, SOX9, WIF1, WNT5A, SHH) genes in Gleason 4+3 compared to 3+4 cases. IPA pathway analysis also determined that the top biological functions 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 biochemical recurrence in prostate cancer patients. Biomarker studies in prostate cancer have faced a challenge in that the number of available frozen specimens with very long term follow up (> 7 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, ITPR1, LBH, MED4, MEMO1, MXI1, PTN, SACM1L, SIRT1, SRSF3, and SYN) have been previously associated with some type of cancer, and five (BTG2 (24), IGFBP3 (25–27), MXI1 (28), FDPS (29), and SIRT1 (30, 31)) have been previously associated with prostate cancer. There are several biological 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 that includes SYN (32), COL15A1, COL3A1, PTN (33), IFT57 (34), and MXI1 (35). Next, there is a set of cell cycle-related genes that includes CDC37L1, BTG2 (24), and SACM1L. In addition, there are several genes associated with apoptosis, PI3K, Insulin, metabolism, and steroid hormone signaling that includes IFT57 (34), ITPR1 (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 that 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 a 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-amplicon or capture sequencing in a CLIA 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 is 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 physicians an idea as to 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 to 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 biological 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Selvi Ramalingam for technical assistance.

Grants Support: This research was supported by DOD grant W81XWH-010-1-0090, NIH R01CA106826, NIH U01 CA168449, NIH R03CA173770, NIH R03CA183006, R01CA128813, Canadian Cancer Society Research Institute Grant 019038, as well as NCI Cancer Center Support Grant P30CA138292 for the Winship Cancer Genomics Shared Resource and the Emory Integrated Genomics Core.

References

6. Herman CM, Kattan MW, Ohori M, Scardino PT, Wheeler TM. Primary Gleason pattern as a predictor of disease progression in Gleason score 7 prostate cancer: a multivariate analysis of 823


Figure 1.
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.
Figure 2.
FPKM expression values for the ERG (blue), ETV1 (red), ETV4 (green), ETV5 (magenta), and SPINK1 (cyan) genes that have been previously shown to exhibit translocations in prostate cancers. Outlier samples are clearly evident, and increased expression is mutually exclusive as expected. Proportions of numbers of cases with increased expression of each gene also are in accordance with previous reports, with high expression of ERG in 45%, ETV1 in 6%, ETV4 in 4%, and SPINK1 (10%) of samples.
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 et al (17) (B, D, and F). K–M curves using clinical parameters alone are shown in panels A and B. K–M curves using clinical parameters combined with the 24 RNA biomarker genes as described in Table 2 are shown in panels C and D. K–M curves using 31 biomarkers from Myriad Genetics are shown in panels E and F.
Table 1
RNA biomarkers of biochemical recurrence following prostatectomy identified by RNAseq analysis. Selected genes and estimated coefficients using all expressed genomic markers.

<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&amp;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>0.403303</td>
</tr>
<tr>
<td>FDPS</td>
<td>farnesyl diphosphate synthase</td>
<td>Sterol Biosynthesis</td>
<td>Breast, Ovarian, Head&amp;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>na</td>
<td>0.022612</td>
</tr>
<tr>
<td>MARCH5</td>
<td>membrane-associated ring finger (C3HC4) 5</td>
<td>E3 ubiquitin ligase of the mitochondria</td>
<td>na</td>
<td>−0.25456</td>
</tr>
<tr>
<td>MED4</td>
<td>mediator complex subunit 4</td>
<td>Transcription, Chromatin</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>MXI1</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>SACML</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</td>
<td>Prostate, Breast, many others</td>
<td>−0.08227</td>
</tr>
<tr>
<td>SNORA20</td>
<td>small nucleolar 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>
Table 2

The prediction model using RNA biomarkers and clinical variables is shown, where the predictive score is calculated from Table 1.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA predictive score</td>
<td>2.35147</td>
</tr>
<tr>
<td>Gleason</td>
<td>0.123248</td>
</tr>
<tr>
<td>( \log(\text{prePSA}+1) )</td>
<td>0.492755</td>
</tr>
<tr>
<td>Age</td>
<td>0.012269</td>
</tr>
<tr>
<td>SMS</td>
<td>0.651851</td>
</tr>
<tr>
<td>( \text{pStage} &gt; 2 )</td>
<td>0.679813</td>
</tr>
</tbody>
</table>
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 integrated discrimination improvement (IDI), net reclassification improvement (NRI) and median improvement in risk score for censored survival outcomes (18). A positive value in all three metrics including IDI and NRI indicates an improvement over the second model. Significant p-values (bold font) indicate a statistically significant improvement of the full model over other models in prediction of BCR.

<table>
<thead>
<tr>
<th></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>Training Set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<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>
<tr>
<td>Validation Set</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<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>-</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>0.439</td>
<td>0.051</td>
<td>0.652</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>