Gene expression profiling identifies clinically relevant subtypes of prostate cancer

Jacques Lapointea,b,c, Chunde Lid, John P. Higginsa, Matt van de Rijnb, Eric Bair*, Kelli Montgomerya, Michelle Ferraric, Lars Egevadd, Walter Rayfordf, Ulf Bergerheimg, Peter Ekmand, Angelo M. DeMarzoh, Robert Tibshiranih,i, David Botsteinj, Patrick O. Brownb,k, James D. Brooksc,l, and Jonathan R. Pollacka,m

Departments of aPathology, bBiochemistry, cUrology, dStatistics, Health Research and Policy, and eGenetics and fHoward Hughes Medical Institute, Stanford University, Stanford, CA 94305; gSection of Urology, Department of Surgical Science, Karolinska Institutet, SE-171 77 Stockholm, Sweden; Department of Urology, Louisiana State University Health Sciences Center, New Orleans, LA 70112; hSection of Urology, Department of Biomedicine and Surgery, Linköping University Hospital, 581 83 Linköping, Sweden; and iDepartment of Pathology, Johns Hopkins University, Baltimore, MD 21205

Prostate cancer, a leading cause of cancer death, displays a broad range of clinical behavior from relatively indolent to aggressive metastatic disease. To explore potential molecular variation underlying this clinical heterogeneity, we profiled gene expression in 62 primary prostate tumors, as well as 41 normal prostate specimens and nine lymph node metastases, using cDNA microarrays containing ~26,000 genes. Unsupervised hierarchical clustering readily distinguished tumors from normal samples, and further identified three subclasses of prostate tumors based on distinct patterns of gene expression. High-grade and advanced stage tumors, as well as tumors associated with recurrence, were disproportionately represented among two of the three subtypes, one of which also included most lymph node metastases. To further characterize the clinical relevance of tumor subtypes, we evaluated as surrogate markers two genes differentially expressed among tumor subgroups by using immunohistochemistry on tissue microarrays representing an independent set of 225 prostate tumors. Positive staining for MUC1, a gene highly expressed in the subgroups with “aggressive” clinicopathological features, was associated with an elevated risk of recurrence (P = 0.003), whereas strong staining for AZGP1, a gene highly expressed in the other subgroup, was associated with a decreased risk of recurrence (P = 0.0008). In multivariate analysis, MUC1 and AZGP1 staining were strong predictors of tumor recurrence independent of tumor grade, stage, and preoperative prostate-specific antigen levels. Our results suggest that prostate tumors can be usefully classified according to their gene expression patterns, and these tumor subtypes may provide a basis for improved prognostication and treatment stratification.

Worldwide, prostate cancer is the third most common cancer and the cause of 6% of cancer deaths in men (1). Its incidence and mortality vary in different parts of the world and are highest in Western countries (2). In the United States, it is the most frequently diagnosed and the second leading cause of cancer death in men (3). Despite these high death rates, prostate cancer is often an indolent disease, and patients can remain asymptomatic for years. The widespread use of serum prostate-specific antigen (PSA) screening has led to identification of an increasing number of asymptomatic low-stage tumors in younger men (4, 5). An important clinical question has become whether and how aggressively to treat such patients with localized prostate cancer.

Currently, prognostication and treatment stratification at the time of diagnosis are based on clinical stage, biopsy Gleason grade (a measure of tumor differentiation), and serum PSA levels. In cases treated by radical prostatectomy, prognosis can be refined by using pathological stage and grade. However, these prognostic indicators do not accurately predict clinical outcome for individual patients. Improved markers are needed to determine which patients might benefit from a more aggressive treatment, and which patients might be spared unnecessary and potentially harmful interventions.

The observed clinical heterogeneity of prostate cancer is likely to reflect underlying molecular heterogeneity among tumors, which, although largely invisible under the light microscope, might be captured by profiling gene expression using DNA microarrays. Indeed, microarray profiling studies have identified clinically relevant gene-expression subtypes in leukemia (6, 7), lymphoma (8), breast cancer (9, 10), and lung cancer (11–13). Although DNA microarray studies of prostate cancer have identified genes differentially expressed in tumor compared to nontumor samples (14–18) and genes whose expression correlates with tumor grade, metastasis, and disease recurrence (14, 17, 19, 20), to date, tumor subtypes based on gene expression have not been appreciated.

Here we report a cDNA microarray-based study in prostate cancer leading to the identification of biologically and clinically relevant gene-expression tumor subtypes. Furthermore, we demonstrate that the protein expression levels for two genes, serving as surrogate markers for tumor subtypes, are strong predictors of tumor recurrence, independent of known risk factors. Our results support the existence of distinct gene expression subtypes in prostate cancer, and their potential use in disease diagnosis and management.

Materials and Methods

Gene Expression Profiling. Freshly frozen prostate surgical specimens were obtained from Stanford University, Karolinska Institute, and Johns Hopkins University, with institutional review board approval from the involved centers (see Supporting Note 1, which is published as supporting information on the PNAS web site). In total, we selected for study 62 primary prostate tumors (61 adenocarcinomas and one adenoid cystic tumor), 41 matched normal prostate tissues (from the noncancerous region of the prostate), and nine unmatched (i.e., different patient) pelvic lymph node metastases. Detailed pathological and clinical data for specimens are provided in Table 2, which is published as supporting information on the PNAS web site. Gene expression profiling was performed essentially as described (9), by using cDNA microarrays containing 26,260 different human genes (UniGene clusters). More detailed information, including data selection and manipulation methods, is available in

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: PSA, prostate-specific antigen; FDR, false discovery rate.

*To whom correspondence may be addressed at: Department of Urology, Stanford University School of Medicine, 5287, 300 Pasteur Drive, Stanford, CA 94305-5118. E-mail: jbrooks1@stanford.edu.

**To whom correspondence may be addressed at: Department of Pathology, Stanford University School of Medicine, CSSR 3245A, 269 Campus Drive, Stanford, CA 94005-5176. E-mail: pollack@stanford.edu.

© 2004 by The National Academy of Sciences of the USA
Tissue Microarrays. A tissue arrayer (Beecher Instruments, Sun Prairie, WI) was used to construct a prostate cancer tissue microarray comprising an independent set of 225 formalin-fixed, paraffin-embedded primary prostate tumor cases selected from diagnostic radical prostatectomy specimens collected at Stanford University, with institutional review board approval. Duplicate 0.6-mm tumor cores represented each case, and the series was associated with a minimum clinical follow-up of 5 years and a median follow-up of 8 years. Primary antibodies directed against MUC1 (SC73335, Santa Cruz Biotechnology) and AZGP1 (SC-11242, Santa Cruz Biotechnology) were used for immunohistochemical staining. More detailed information is available in Supporting Note 3, which is published as supporting information on the PNAS web site.

Results

Identification of Prostate Tumor Subtypes. To survey the molecular variation among prostate tumors, we profiled gene expression in 112 prostate tissues, including 62 primary tumors, 41 matched normal prostate tissues, and nine unmatched lymph node metastases. To gain further insight into the biological and clinical relevance of new molecular features, we examined the distribution of clinicopathological parameters among primary tumors (Fig. 1b). High grade and advanced stage tumors were more highly represented among tumor subgroups II and III (P = 0.04, χ² test; Table 4, which is published as supporting information on the PNAS web site). Seven tumors associated with early recurrence in our dataset were also found to reside only within subgroups II and III. Consistent with these observations, a supervised analysis using the significance analysis of microarrays (SAM) method (ref. 26; see Supporting Note 2) identified sets of genes whose expression were associated with high-grade, advanced stage or early tumor recurrence (Figs. 1m and 2); these genes were predominantly located within gene expression features characterizing tumor subtypes II and/or III (Fig. 1e, g, j, and l; also see Supporting Note 6 and Table 5, which are published as supporting information on the PNAS web site).

In all, we identified 41 genes associated with high grade (Fig. 2a) and 11 genes associated with advanced stage (Fig. 2b), with false discovery rates (FDR) of 2% and 8%, respectively. Among the genes associated with high grade, COL1A2, SPARC, ABCA5 and BGN were reported by Singh et al. (17) to correlate with tumor grade; however, that same study had not identified genes correlated with tumor stage. We also identified four genes positively and 19 genes negatively associated with early recurrence (Fig. 2c), with a FDR of 16%; the relatively high FDR is likely attributable to the short period of clinical follow-up. Although there was no overlap between this gene set and five outcome predictor genes reported by Singh et al. (17), our set of 23 genes correlating with outcome accurately predicted recurrence for patients in their study (Fig. 6, which is published as supporting information on the PNAS web site).

Surrogate Markers for Tumor Subtypes Predict Recurrence. To further characterize the clinical relevance of tumor subtypes, we evaluated as surrogate markers two genes differentially expressed among tumor subtypes, by using immunohistochemistry on tissue microarrays comprising an independent set of 225 primary prostate tumors with a minimum clinical follow-up of 5 years and a median follow-up of 8 years. The two genes were selected based on their differential expression across subtypes and the availability of specific antibodies. MUC1, encoding the mucin 1 transmembrane protein, resided within a gene expression feature that characterized tumor subtypes II and III (Fig. 1j), both associated with more “aggressive” clinicopathological features. MUC1 protein expression, determined by immunohistochemistry on the tissue microarray, was variable across tumors (Fig. 3a); in Kaplan–Meier survival analysis, positive MUC1 staining was associated with significantly shorter time to recurrence (P = 0.003; Fig. 3c). In contrast, AZGP1, encoding zinc-α-2-glycoprotein, resided within a feature characterizing tumor subtype I (Fig. 1f), and strong immunostaining of AZGP1 was associated with significantly prolonged time to recurrence (P = 0.0008; Fig. 3d). The widest separation in recurrence-free survival curves was attained by combining immunostaining data with tumor grade and stage.
Hierarchical cluster analysis of prostate samples. (a) Thumbnail overview of the two-way hierarchical cluster of 112 prostate specimens (columns) and 5,153 variably expressed genes (rows). Mean-centered gene expression ratios are depicted by a log2 pseudocolor scale (ratio fold-change is indicated); gray denotes poorly measured data. The complete data set depicted here is available at http://microarray-pubs.stanford.edu/prostateCA. (b) Enlarged view of the sample dendrogram. Terminal branches for normal prostate samples are colored pink, and those for tumor samples are colored according to gene expression subgroups: III (purple), I (yellow), and II (dark blue). Two tumors clustering with normal samples (see text) are colored light blue. Clinicopathological features associated with individual tumor samples are indicated by black boxes below the dendrogram (asterisks indicate missing data). High grade indicates Gleason grade $\geq 4$; advanced stage indicates pathological stage $\geq T3$; tumor recurrence indicates PSA rise after surgery or clinical metastasis. (c–l) Selected gene expression “features” extracted from cluster (locations indicated by vertical colored bars). Because of space limitations, only selected genes are indicated. Genes are annotated as indicated if associated in supervised analysis with high-grade (blue circles), advanced stage (green squares), short time to recurrence (red triangles), or long time to recurrence (red inverted triangles). Genes positively and negatively associated with epithelial cell content are indicated by colored text (dark blue and light blue, respectively; see Supporting Note 1). Genes characterized by immunohistochemistry are indicated with arrow. m, moving average (41-gene window) plots for the t-test statistic (grade and stage) and Cox's proportional hazards partial likelihood score (recurrence-free survival) shown for the 5,153 genes in the cluster. Note that peaks (high grade, advanced stage, early recurrence) and valleys frequently correspond to gene expression features characterizing tumor subtypes.
free survival, independent of tumor grade and stage. AZGP1 has strong immunostaining was associated with longer recurrence-more highly differentiated tumors. We had selected AZGP1 normal prostate, suggesting that this subgroup may represent features defining subtype I (Fig. 1 aggressive subclass. Indeed, one of the two gene expression on an independent set of samples, suggests that these subtypes on distinct patterns of gene expression. Moreover, the distribu-

The main objective of our study was to survey the molecular variation of prostate cancer, to gain new insight into the underly-

on MUC1 and AZGP1 (Fig. 3e), suggesting an additive value for prog nostication. Despite the association we noted between tumor subtypes and tumor grade and stage, no significant association was identified between MUC1 or AZGP1 protein expression and tumor grade or stage (Table 6, which is published as supporting information on the PNAS web site). To determine whether MUC1 and AZGP1 expression added prognostic information over and above known prognostic factors, we performed multivariate proportional hazards analysis. MUC1 and AZGP1 staining were found to be strong predictors of tumor recurrence [odds ratios = 2.4 (1.3–4.2) and 0.38 (0.21–0.69), respectively; \( P < 0.001 \)], independent of tumor Gleason grade, stage and preoperative serum PSA (Table 1).

**Discussion**

The main objective of our study was to survey the molecular variation of prostate cancer, to gain new insight into the underlying biology of this clinically heterogeneous disease. We used unsupervised two-way hierarchical clustering to discover that primary prostate tumors stratify into three robust subtypes based on distinct patterns of gene expression. Moreover, the distribution of clinicopathological features, as well as the performance of surrogate immunohistochemical markers for these subtypes on an independent set of samples, suggests that these subtypes are associated with distinct biological and clinical behavior.

We have characterized subtype I tumors as the clinically least aggressive subclass. Indeed, one of the two gene expression features defining subtype I (Fig. 1f) includes genes expressed in normal prostate, suggesting that this subgroup may represent more highly differentiated tumors. We had selected AZGP1 expression as a surrogate marker for this tumor subtype, and strong immunostaining was associated with longer recurrence-free survival, independent of tumor grade and stage. AZGP1 has previously been reported to be expressed in primary prostate tumors, and to a lesser extent in metastases (27, 28). Hale et al. (28) found an inverse association with tumor stage and grade, whereas Gagnon et al. (27), as did we, found no such association. Given that subgroup I, although predominantly comprising low-grade tumors, also included higher-grade tumors, we speculate that expression profiling may identify a molecular signature of differentiation not apparent by histology.

We have determined subtype III, along with subtype II, to represent a clinically aggressive tumor subclass. Notably, primary tumors within subtype III shared features of gene expression with unmatched lymph node metastases (Fig. 1g). Distinct gene expression signatures between primary prostate tumors and metastases have been reported (14, 19). Interestingly, in both of these studies, a small proportion of primary tumors also clustered together with metastases. Recently, Ramaswamy et al. (20) reported a gene expression signature of metastasis present in a subset of primary solid tumors, including prostate cancer. Our findings are consistent with their conclusion that a metastatic phenotype may preexist within the bulk tumor population for a subset of primary tumors. Importantly, however, because both tumor subtypes II and III are associated with tumor recurrence, our data suggest that at least in prostate cancer, this metastatic signature may represent only one of at least two distinct signatures associated with poor outcome (also see Supporting Note 7, which is published as supporting information on the PNAS web site).

Subtype II tumors represent the second clinically aggressive tumor subclass, and the gene expression feature that characterizes this subgroup included several genes identified in supervised analysis to be associated with both high grade and advanced stage, such as HDAC9 and DIO2 (Fig. 1e). Other genes associated with high grade (e.g., NOX4) and advanced stage (e.g., F5) resided within two features of gene expression also shared...
with subtype III (Fig. 1 j and l). The role of these genes in the development or progression of prostate cancer remains to be determined. Nevertheless, it is worth noting that nearly half of the tumors within subgroup II are low grade and early stage, suggesting again that gene expression features may represent molecular signatures of biological processes relevant to tumor progression that are not appreciable by pathological analysis.

One expression feature of particular interest (Fig. 1 j), its expression shared by subgroups II and III, included several genes involved in cellular invasion and angiogenesis, such as F2R, MUC1, NRPI, LOX, ANGPT2, and TEM7 (29–35). We had selected MUC1 expression as a surrogate marker for this feature, characterizing the clinically aggressive subgroups II and III, and found positive immunostaining to be associated with shorter recurrence-free survival, independent of tumor grade and stage. Increased expression of MUC1 has previously been associated with poor prognosis in other types of carcinoma (36–39). It has been proposed that overexpression of MUC1 increases the metastatic potential of cancer cells by reducing E-cadherin and integrin-mediated cell adhesion (34, 35). However, a role of MUC1 in prostate cancer has not been established. One study reported MUC1 expression to correlate with prostate cancer grade and stage (40), whereas another found an association with intratumoral angiogenesis but not with grade (41). Interestingly, coexpression of MUC1 along with multiple angiogenic factors has been observed in non-small cell lung tumors (42). In our data set, MUC1 was coexpressed along with several genes involved in cellular invasion and angiogenesis, and this expression feature may represent the signature of these biological processes, which are important for prostate cancer progression.

In this study, we have further characterized the expression and prognostic value of two genes, functioning as surrogate markers for our newly identified tumor subtypes. In an independent set of 225 prostate tumors assessed by immunohistochemistry on tissue microarrays, MUC1 (a surrogate for subtypes II and III) and AZGP1 (subtype I) were found to be strong predictors of tumor recurrence. Importantly, these genes were found in multivariate analysis to add additional prognostic information over and above the known risk factors of tumor grade, stage, and preoperative PSA. Interestingly, these genes also provided prognostic value independent of one another, suggesting that using two genes improves the accuracy of tumor subtyping and prognostication. It remains to be determined whether adding yet additional genes might further improve prognostication, and, of course, it will be important to validate our findings prospectively and on preoperative tumor biopsy samples (the specimens most relevant for treatment stratification). Nonetheless, our results suggest that prostate tumors can be usefully classified according to their gene expression patterns, and that these tumor subtypes may provide a basis for improved prognostication and treatment stratification.

Table 1. Multivariate proportional hazards analysis

<table>
<thead>
<tr>
<th>Variable</th>
<th>Hazard ratio (95% confidence interval)</th>
<th>P value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gleason grade, ≥4 + 3 vs. ≤3 + 4</td>
<td>3.11 (1.70–5.70)</td>
<td>0.0002</td>
</tr>
<tr>
<td>Pathological stage, ≥T3 vs. ≤T2 stage</td>
<td>2.80 (1.46–5.35)</td>
<td>0.002</td>
</tr>
<tr>
<td>Preoperative serum PSA, per ng/ml†</td>
<td>1.03 (1.01–1.05)</td>
<td>0.0004</td>
</tr>
<tr>
<td>MUC1 staining, positive vs. negative</td>
<td>2.35 (1.30–4.24)</td>
<td>0.0005</td>
</tr>
<tr>
<td>AZGP1 staining, strong vs. weak/negative</td>
<td>0.38 (0.21–0.69)</td>
<td>0.002</td>
</tr>
</tbody>
</table>

*Wald test.
†Serum PSA was used as a continuous variable.
We thank laboratory members for many helpful comments and suggestions. We also thank Janet Mitchell and the Stanford Tissue Bank for collection of tissues, Mike Fero and the staff of the Stanford Functional Genomics Facility for providing high-quality cDNA microarrays, and Gavin Sherlock and the Stanford Microarray Database group for providing outstanding database support. We are grateful to Helen Fedor, Gertrun March, and Marcella Southerland for assisting with tissue procurement and frozen sectioning at Johns Hopkins University. This work was supported by National Institutes of Health Grant U01CA585129, the Howard Hughes Medical Institute, the Grove Foundation, the Swedish Cancer Society, the Cancer Society in Stockholm, the Konung Gustaf V:s Jubileumsfond, Wallström and Osterman Foundations in Karolinska Institutet, and the Swedish Medical Society. The Johns Hopkins tissue bank is funded in part by the National Cancer Institute/National Institutes of Health Specialized Programs of Research Excellence (prostate) P50CA58236. J.L. was supported in part by a fellowship from the Canadian Institute of Health Research. P.O.B. is an investigator of the Howard Hughes Medical Institute.