SUPPLEMENTAL FIGURE LEGENDS

Supplemental Fig. 1 Tumor versus normal marker genes

The top 50 genes with high expression in tumor as well as the top 50 genes with high expression in normal are listed ranked according to the variation of the signal-to-noise statistic (tumor vs. normal). The expression of each gene (rows) in each sample (columns) is represented by the number of standard deviations above (red) or below (blue) the mean for that gene across all 102 samples.

Upon quantification, the epithelial content of tumor samples (79±14%) was significantly higher than in normal samples (27±21%, $p<0.0001$). As a result, some genes, whose expression correlated with epithelial content, may have the appearance of differential expression based solely on epithelial content. To identify these genes, we calculated the correlation of each gene with percent epithelium in both the normal samples and the tumor samples. We then used permutation testing to determine what level of correlation was significantly better than expected by chance alone with a p value of 0.05. Genes in blue type had a significant ($p\leq0.05$) correlation with epithelial content in both tumor and normal samples.

Supplemental Fig. 2 Tumor vs. normal prediction

A) The prediction accuracy of *k*-NN models using 1 to 256 genes. For each gene number tested, models were built and tested using leave-one out cross validation. The x-axis indicates number of genes used in model building, and the y-axis indicates the frequency of success. The success rate (correct predictions divided by total predictions) in the observed data is shown (red solid line). The mean success rate +/- the standard deviation (bottom dashed line) and maximum success rate (top dashed line) obtained using 1000 permutations is shown.

B) Frequency of gene use in the 16-gene models built during leave-one out cross validation. For each sample left out, the 16 genes best distinguishing between tumor and normal in the remaining 101 samples were identified and used to predict the identity of the left out sample. A list of all genes used in the 102 16-gene models built during leave-one out cross validation is shown. The x-axis represents the frequency of gene use in all models.

Supplemental Fig. 3 Tumor-Normal Class Prediction Validation:

A variation of signal-to-noise metric was used to develop 4 and 16-gene prediction models based on the initial 102 samples. These models were applied to an independent validation set of 35 samples (8 normals, 27 tumors) and the identity of each sample was predicted based on the expression of 4 and 16 genes. Because of large differences in array intensity between the initial and validation sets, both raw data and normalized data were used. For normalized data, the average difference for each gene was normalized across all samples used in the analysis with the mean expression set at 0 and all subsequent expression values expressed as standard deviations away from the mean. Two-by-two table depicts the predicted and actual class membership. Fisher's exact test was used to determine if the class prediction algorithm perform better than expected by chance alone.

Supplemental Fig. 4 Genes Correlating with Clinical and Pathological Features of

Prostate Cancer

A) Patient's age and serum PSA level together with the tumor sample's Gleason score were treated as continuous variables and the Pearson coefficient was calculated between each of these three variables and the expression of each gene across tumor samples. After the Pearson coefficient was calculated, the samples labels were randomly permuted 10000 times and the Pearson coefficient for the correlation between each variable and gene expression that would be expected by chance alone at a frequency of 0.001was determined. The number of genes having a Pearson coefficient greater than expected by chance alone (at 0.001 for either positive or negative correlations) was determined for age, serum PSA, and Gleason score. The expected number of genes to reach a Pearson coefficient at the set significance of 0.001 by chance alone was determined by taking the total number of genes included in the analysis (n = 5265) and multiplying it by the frequency for both positive and negative correlation (0.001 X 2 tails = 0.002).

B) For dichotomous variables, signal-to-noise metrics were used to identify genes with expression best discriminating between tumors with or without the pathological features of capsular penetration, positive surgical margins, or perineural invasion similar to the analysis performed for Suppl. Figure 1. Permutation testing, by randomizing the sample labels, was used to compare the observed data to that expected by chance alone. The number of genes matching each of the presented class distinctions better than expected by chance alone with a frequency of 0.001 is presented. The expected number of genes was calculated by multiplying the frequency by the total number of genes included in the analysis.

Supplemental Fig. 5 Independent Validation of Co-Expressed Genes Correlated with Gleason Score

The 29 genes whose expression correlated with Gleason score in the initial data set ($p < 0.001$) were used to cluster samples from the initial dataset and a validation data set. Expression data for the 29 genes from both data sets were thresholded, imported into Gene Cluster (Eisen's software), and log-transformed. Expression Data was mean centered and normalized across genes and arrays. Hierarchical clustering (Kendall's Tau metric) was performed on the datasets independently. The Two classes of genes in the training set were identified (Type 1 or T1 (pink) vs. Type 2 or T2 (black)) and compared to the two classes of genes from the Validation set (Validation 1 or V1 and Validation 2 or V2). Pink and black designations in the dendrogram for the Validation set indicate the original class assignment (T1 or T2). The genes comprising these two classes were almost identical in both data sets ($p < 0.0001$ by Fischer's Exact Test) suggesting that there is consistent co-expression of these genes in two independent data sets. In both sets, tumors of high Gleason score (> 8 identified by arrows) tend to be associated with expression of the Type 1 genes.