Supplemental Methods

RNA Isolation, Target preparation and Hybridization:

The OCT from 65 tumor/normal pairs was dissected off so as to include a minimal amount of OCT in subsequent steps of preparation. The samples were homogenized using a tissue shredder (Polytron® PT-MR 2100, Switzerland) in 3 mL of Trizol® (Life Technologies, Grand Island, NY). Total RNA was isolated using the manufacturer's Trizol protocol modified to include the use of glycogen and 1.5 mL gel phase seperation tubes (Phase Lock Gel™ Heavy, Eppendorf, Westbury, NY), and resuspended in DEPC water. RNA samples were analyzed by optical density measurements and gel electrophoresis. A total of 55 tumors and 53 corresponding normals yielded sufficient RNA to proceed with the target preparation. First strand cDNA synthesis was carried out from total RNA by reverse transcription using an oligo-dT primer that also contains a T7 polymerase recognition sequence. Double stranded cDNA (dscDNA) was synthesized by a modification of knick-initiated transcription, precipitated and used an *in vitro* transcrption (IVT) reactions with the T7 bacterial RNA polymerase and biotinylated nucleotides (bio-UTP and bio-CTP) to generate biotinylated cRNA. The cRNA was then fragmented with a high salt buffer and heat to create oligomers of approximately 50 nucleotides in length. The biotin-incorporated cRNA was hybridized to HU95Av2 microarrays for 16 hours at 40°C with constant rotation at 60 RPM. After washing the microarrays were stained using a three-step process including an initial streptavidin-phycoerythrin staining, a biotin-labeled, anti-biotin antibody, and a repeated streptavidin-phycoerythrin staining.

Average Difference Calculation:

Average differences were calculated using GeneChip Software (Affymetrix). Data quality measurements included the average pixel mean value for each probe set on the array, the average pixel standard deviation for each probe set on the microarray, the fraction of genes represented on the array called "present", the mean average difference for all genes called present on the arrays, and the standard deviation of the average difference for all genes called present on the arrays. Using these criteria, 3 additional microarray files were excluded from subsequent analysis as they consistently had values 2 standard deviations outside of all arrays. The final number of arrays available for subsequent analysis was 102 (50 normal samples and 52 tumor samples)

Scaling, Thresholding and Filtering

All expression files in a given experiment were scaled to a reference file (generally the file found to have the median value of expression) based upon the mean average difference for all genes present on the microarrays. The scaled files used in each experiment will be available at www-genome.wi.mit.edu/MPR/Prostate. All genes with average differences below the minimum threshold of 10 were set at the minimum threshold. The maximum threshold was set at 16,000. After thresholding, the relative variation of expression for each gene was determined by dividing the maximum expression for the gene among all samples (Max) by the minimum expression (Min) (Max/Min). The absolute variation in expression was determined by subtracting the Min from the Max (Max-Min). Filtering parameters of 5-fold change (Max/Min) and absolute difference of 50 (Max-Min) were used for all subsequent analysis.

Gene Ranking using the Signal-to-noise Statistic

Gene expression differences associated with a particular class distinction (i.e. Class 0 vs. Class 1) were measured, as described previously, using a variation of the signal-to-noise statistic $(\mu_{Class0}-\mu_{Class1})/(\sigma_{Class0}+\sigma_{Class1})$ where μ and σ are the mean and standard deviation of the expression for each gene (Golub et al., 1999). Using this statistic, genes were ranked for the dichotomous distinctions tumor vs. normal, recurrent vs. non-recurrent, positive vs. negative capsular penetration, positive vs. negative surgical margins, and present or absent perineural invasion. All calculations were performed using the GeneCluster software.

Permutation Testing

Standard tools for estimating statistical significance do not sufficiently account for the multiple hypothesis testing that occurs in situations where the number of variables (genes) greatly exceeds the number of samples (tumors or normal tissues). Permutation testing allows an empiric determination of the extent to which observed data demonstrating an association between a given class distinction and gene expression could be obtained by chance. For a detailed discussion please see supplementary materials of Pomeroy et. al. (http://www-genome.wi.mit.edu/mpr/publications/projects/CNS/Pomeroy_et_al_0G04850_11142001_suppl_info.doc). Briefly, for each class distinction the class labels for the actual data set were re-iteratively randomly reassigned. This process is automated in the GeneCluster software package (available at http://www-genome.wi.mit.edu/MPR/Software.html). After each reiteration, signal-to-noise metrics are re-calculated measuring the association of the expression pattern of each gene with the newly designated class labels. After 1000 permutations, summary statistics of the signal-to-noise measurements for the association of genes with the permuted class distinction are generated and compared to those obtained in the experimental data.

Similarly, boundaries for the statistical significance of Pearson correlation coefficients were determined using similar methods. In this case, the sample label designations were randomly reassigned, maintaining original distribution of these labels. For example, Gleason Scores were randomly reassigned maintaining a constant proportion of scores that were 6, 7, 8 etc. The Pearson correlation coefficient between gene expression and the permuted data was then determined. The measured frequency of Pearson coefficients obtained in each of 10000 permutations was determined and the value of the Pearson coefficient obtained in less than or equal to 1 in every 1000 permutations was compared to the observed data.

The performance of prediction models in leave-one-out cross validation was also compared to the performance obtained by models derived from permuted data. Here, after each permutation of the class distinction, two or three nearest neighbors were used to predict the identity of the held out sample. For each class distinction, nearest neighbor predictions were made in each of 1000 permuted data sets. The maximum and mean accuracies of the 1000 models generated from the permuted data is shown. By performing 1000 random permutations in this manner, the strength of the observed association was measured against chance alone with expected frequencies as low as 1 in 1000. These results are presented as experimentally determined “p” values for sake of simplicity throughout the manuscript unless otherwise noted.