Supporting Information

Markert et al. 10.1073/pnas.1117029108

SI Materials and Methods

Signature Curation. Signatures were collected from the literature to represent currently known pathway aberrations in prostate and other cancers. An overview of sources can be found in Table S1*A*, and Table S1*B* contains a complete list of genes contained in each signature. For several features we found multiple signatures in the literature. In these cases we selected several signatures and assigned them to signature groups [i.e., signatures "ESC1" and "ESC2" in the signature group "ESC", characterizing embryonic stem cells (ESC)] (Table S1*A*).

Gene Set Enrichment Analysis. We downloaded the gene expression data from Gene Expression Omnibus (GEO) (series GSE16560) and normalized it by mean centering across samples (data were preprocessed by log transform and mean centering along samples). Whole-transcript gene expression data for GSE21034 were downloaded and normalized by log transform and mean centering along samples followed by mean centering across samples. Gene expression values were averaged over probes per gene. To optimize comparability of results, only genes represented on the Illumina custom chip used for GSE16560 were selected among the genes represented on the Affymetrix platform used for GSE21034. We then performed gene set enrichment analysis (GSEA) with each signature in the library on each of the samples, obtaining a matrix of signature scores. GSEA was done by ordering each sample by normalized expression values and calculating the positive and negative scores according to the distribution of signature genes within the ordered sample. The scores are given as the maximum and minimum of the score function (partial sums),

$$\mathrm{SCORE^{+}}_{\mathrm{SIG}}(\mathrm{sample}) = \max_{1 \le k \le N}\Bigl(\sum_{1 \le j \le k} d_{\mathrm{SIG}}(j)\Bigr),$$

$$\mathrm{SCORE^{-}}_{\mathrm{SIG}}(\mathrm{sample}) = \min_{1 \le k \le N}\Bigl(\sum_{1 \le j \le k} d_{\mathrm{SIG}}(j)\Bigr),$$

where $d_{\mathrm{SIG}}(j) = +\sqrt{(|\mathrm{SIG}|/(N - |\mathrm{SIG}|))}$ if gene j is in signature SIG, and $d_{\mathrm{SIG}}(j) = -\sqrt{((N - |\mathrm{SIG}|)/|\mathrm{SIG}|)}$ otherwise. Here N is the total number of genes and $|\mathrm{SIG}|$ is the number of genes in the signature SIG. This definition is the same as in Mizuno et al. (1). The P values for significance of score assignments are determined by a random test using $n = 100{,}000$ random gene sets of the same length as the signature gene set in question and calculating the significance of the difference of the positive and the negative score compared with the distribution of the random score differences. Significance was stored in log-scale and used for representing signature scores in heatmaps (Figs. 1 *A* and *B*, 2*A*, and 4*A*; average values were used for signature groups with multiple signatures). For clustering analysis, both positive and negative signature scores were collected for each signature and each sample, producing a matrix of signature scores of the size (no. samples) × (2 × no. signatures).

Bayesian Clustering Optimizer. The *Bayesian clustering optimizer* is based on a Bayesian formulation of a Gaussian mixture model (ref. 2, section IV). It is composed of a set of self-consistent equations—[clusters + data → model parameters] and [model parameters + data → clusters]—that are applied until a convergence criterion is reached (here, until the relative change of the free energy score, defined below, is $<10^{-6}$). These self-consistent equations are analog to those used by the *K*-means clustering method or the expectation maximization (EM) algorithm (3), except for additional Bayesian corrections. Furthermore, as in those methods, an initial clustering of the samples in groups is required as input to start iterating the self-consistent equations. These initial clusterings can be generated assuming a given number of clusters *R* and using a given library of clustering methods. The consistency of each clustering is further quantified by a score function, also referred to as free energy in ref. 3 because of its resemblance to the concept of free energy in physics. The method resulting in the best free-energy score is then selected, and the associated optimized clustering reported. Note that the number of returned clusters may be smaller than what was given as input, because the Bayesian clustering may judge that the initial clustering was overfitting the data. Furthermore, samples may end in different groups after iteration of the self-consistent equations.

The input of the Bayesian clustering optimizer is thus the data and an initial clustering of the samples into a predefined number of groups *R*, and the outcome is the optimal (effective) number of sample groups, the assignment of each sample to a group, the best preclustering method, and the free-energy score associated with the final clustering. It was implemented in Matlab, making use of its built-in clustering methods toolbox.

Group Signature Scores. In general, there is more than one reported gene signature for each molecular feature [e.g., ESC, induced pluripotent stem cell (iPSC), etc.]. Applying the Bayesian optimizer to the subset of signatures associated with a molecular feature, we obtained a clustering of the samples according to that molecular feature. Specifically, we applied the Bayesian clustering optimizer using the library of clustering methods in Table S2*B* and *R* values from one to five clusters. The input resulting in the lowest free-energy score was selected and the associated optimized number of groups and clustering reported. The resulting clusters were ordered in increased order of the average over all individual scores in the cluster. We assigned evenly spaced discrete values to the clusters starting from 0 for the cluster with the lowest average scores to 1 for the cluster with the highest. For example, if the optimal number of groups would be four, the group values would be 0, 0.25, 0.5, 0.75, and 1. This approach produces a group signature score for each molecular feature and sample, the lowest group signature scores indicating a low consensus representation and the highest group signature scores indicating a high consensus representation.

Unsupervised Clustering Methodology. The unsupervised clustering of the prostate tumor samples was performed using the Bayesian clustering optimizer with the signature scores as input data, different numbers of predefined groups *R*, and a wide library of preclustering methods (Fig. S1). A subset of the preclustering methods used as input the signature scores and standard clustering methods suitable for continuous data (Table S2*A*). Another subset used instead the group signature scores and standard clustering methods suitable for continuous and discrete data (Table S2*B*). Finally, for each preclustering method, *R* values between 2 and 15 were considered.

Table S3 demonstrates the stabilization of the optimized output as *R* increases. We listed normalized mutual information (total correlation) for each optimized output clustering with the previous one to indicate the degree of agreement [$\mathrm{MI}(X, Y) = I(X, Y)/\min(H(X), H(Y))$, where *H* is entropy and *I* is mutual information, satisfying $\mathrm{MI}(X, Y) = 1$ for identical clusterings].

Table S3 contains the values for clustering samples using all signatures (producing the results in Fig. 2) and for clustering using only the stemness signatures (producing the results in Fig. 1*B*). In the latter case stabilization also occurs at five clusters; however, for lack of sensible biological interpretation of these five clusters from the stemness signature patterns alone we chose the three clusters obtained from $R = 3$ that represent the obvious patterns (stem-like, differentiated, and neither of them).

Statistical Analysis. To analyze the overall association of a cluster with a signature or signature group, we calculated the average gene expression on each cluster from the original gene expression and used these average expression vectors as input for gene set enrichment analysis with the signatures. This method produced *P* values for the significance of the assignment. The *P* values were used for illustration of the cluster profiles in Figs. 2*B* and 4*B*. Average values were used where signature groups contained multiple signatures.

We calculated the means of clinical values on each cluster and obtained *P* values for the significance of these through Fisher's exact test for discrete variables and Student's *t* test for continuous variables (both two-tailed). We used the Matlab scripts kmplot and logrank to perform Kaplan–Meier analysis of survival. We used overall survival times (follow-up time) with censoring for the Kaplan–Meier plots in Figs. 1*C* and 3 *A*–*C*. Cox proportional hazard ratios were calculated using the Matlab routine coxphfit.

We counted and listed genes in the overlap between the most strongly correlated signature groups (Pearson's correlation, not included) in Tables S4*A* and S4*B*. Overlap sizes were generally small, typically between 0 and 5%, showing that the correlations found in data are functional and not systematic.

1. Mizuno H, Spike BT, Wahl GM, Levine AJ (2010) Inactivation of p53 in breast cancers correlates with stem cell transcriptional signatures. *Proc Natl Acad Sci USA* 107: 22745–22750.
2. Vazquez A (2008) Bayesian approach to clustering real value, categorical and network data: Solution via variational methods. arXiv:0805.2689v3.
3. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B Methodol* 39(1):1–38.

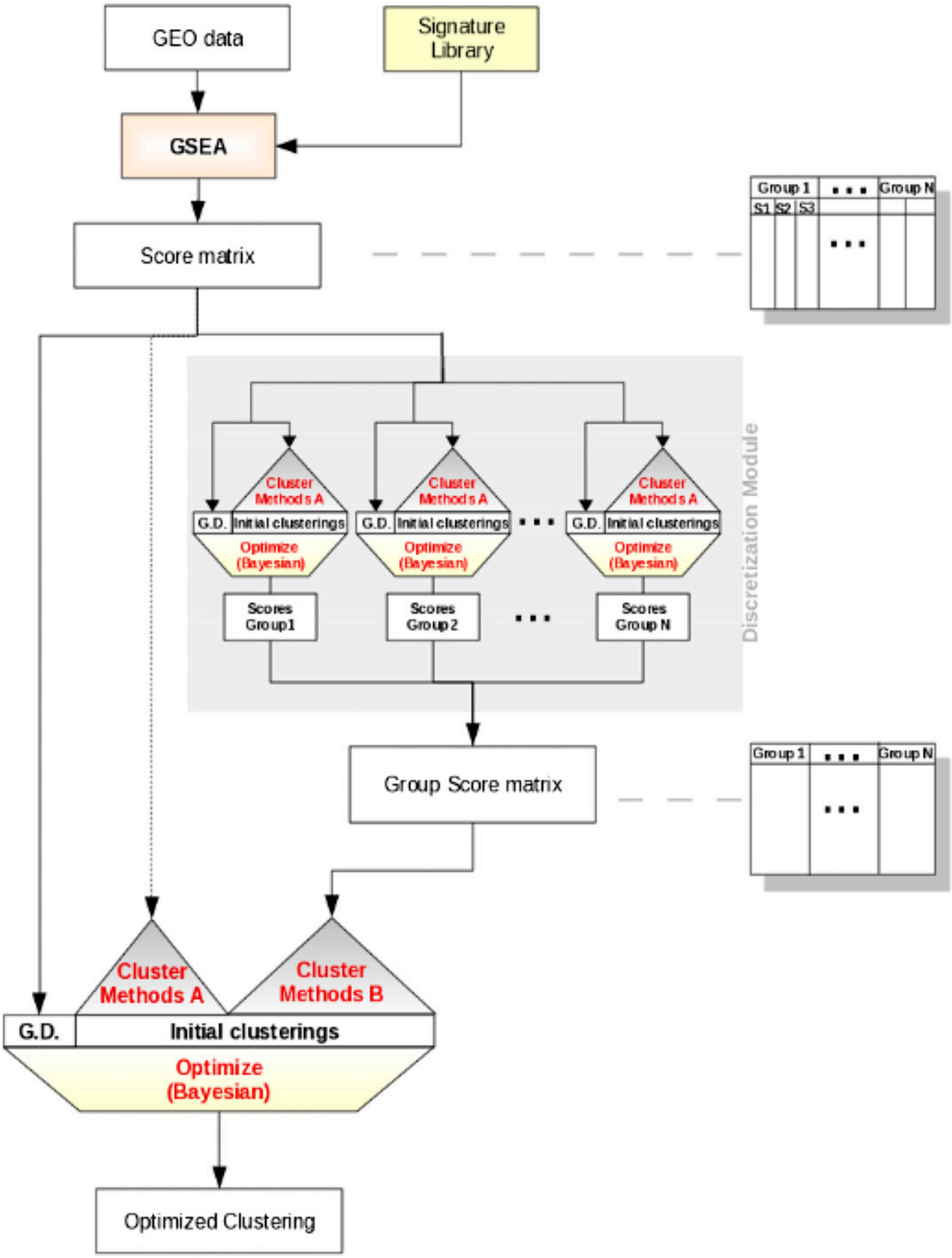

Fig. S1. Flowchart of the clustering algorithm. Clusters are created using a Bayesian optimization step that takes as input a Gaussian data matrix (G.D., the basis for the optimization procedure) and initial clusterings (starting points for the optimization). Gaussian data are always taken from the original score matrix, indicated by input arrows (G.D.). We create initial clusterings from the original score matrix (dashed line) and from a discrete group score matrix (right-hand side). Group score creation (discretization module) compiles the core information of several signatures within one feature group, reduces noise, and returns a matrix of easily interpretable group scores (0, feature absent, . . . , 1, feature present). This step is done by breaking the score matrix into the groups and applying Bayesian optimized clustering into the maximal five clusters for each signature group. Clusters are then used to produce the discrete group scores (binning into up to five bins). The discrete group score matrix is used to produce additional initial clusterings for the final clustering of samples.

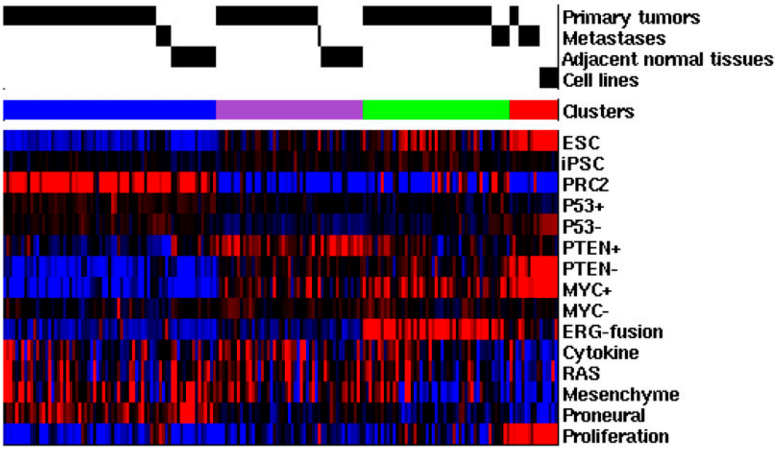

Fig. S2. Heatmap of signature scores for clustering of all 185 samples in the Taylor et al. (1) dataset. This heatmap is the equivalent to Fig. 4*A*, showing the results for all samples including (unmatched) normal tissue samples and cell lines. Normal tissue samples clustered with the PRC2 | differentiated group and the cytokine | transitional group, whereas cell line samples all clustered with the ESC | (P53$^-$) | PTEN$^-$ group.

[1] Taylor BS, et al. (2010) Integrative genomic profiling of human prostate cancer. *Cancer Cell* 18:11–22.

Table S1*A*. Overview of signature library used for this analysis, with sources

Table S1*A* (DOC)

Table S1*B*. Collection of all gene signatures used in this analysis

Table S1*B* (DOC)

Table S2*A*. Methods for creating initial clusterings and signature group scores, as indicated in Fig. S1: methods suitable for continuous data, used for both preclustering and signature score creation

Table S2*A* (DOC)

Table S2*B*. Methods for creating initial clusterings and signature group scores, as indicated in Fig. S1: methods used for preclustering signature scores

Table S2*B* (DOC)

Table S3. Stabilization of optimal clusterings

Table S3 (DOC)

MI-normal refers to normalized mutual information (total correlation) between the optimized output clusterings in the given range and those in the previous one. Effective number of clusters, optimal free energy values (FE), winning initial clustering method, and the *P* values for the split in Kaplan–Meier survival curves between the most lethal resulting cluster and the rest are listed.

Table S4*A*. Sizes of overlaps between most correlated signature groups

Table S4*A* (DOC)

Overlaps are relatively small compared with signature sizes.

Table S4*B*. Genes in the overlaps of signatures

Table S4*B* (DOC)

The Ras-pathway signature shares two oncogenes (FOS and JUNB) and the immune regulator MYD88 with the cytokine signatures. The PTEN^{-} signature and the ESC signatures share genes involved in cell cycle and cell division.