Supporting Information

Boutros *et al.* 10.1073/pnas.0809444106

SI Text

Prognostic Signature Identification by Modified Steepest Descent. To identify a signature comprising genes that are not ranked by some univariate criterion, we developed a discrete, greedy gradient-descent algorithm, which we termed modified steepest descent (mSD). mSD begins by considering all possible classifiers (signatures) of 1 dimension (gene), and selecting the best gene. Once this optimal single-gene classifier is identified, the algorithm proceeds to add additional dimensions (genes) sequentially, testing all possible subsets of 2 genes that contain the optimal single-gene classifier. This corresponds to testing all supersets of the single-gene classifier and taking the largest discrete step to improve classifier performance. This procedure iterates through higher dimensions, evaluating successive supersets of the best n-gene classifier identified thus far. The algorithm terminates when an n gene classifier is discovered whose performance is not exceeded by any $n + 1$ gene superset of itself. At each stage of the feature selection, classifier performance is evaluated by using k-medians clustering with $k = 2$ to separate patients into 2 groups. Note that clustering is used here as an exploratory technique, not as a significance-testing method (1, 2). Next, survival differences between these 2 groups are assessed by using the log-rank test. Gene selection was made on the basis of the χ^2 statistic from the log-rank test, and thus the termination criterion corresponds to finding an n gene classifier whose χ^2 score cannot be exceeded by adding any single additional gene. The final output of the algorithm is a subset of prognostic genes, along with a separation of patients into a group with good survival (the "good prognosis group") and a group with poor survival (the "poor prognosis group"). A Cox proportional hazards model including stage was then fit to these group assignments. Hazard ratios for the classification were extracted, along with p values based on the Wald test. Feature selection was implemented in Perl (v5.8.7) and was run on AIX (v5.2.0.0) on an IBM p690. Clustering used the Algorithm::Cluster (v1.31) C library (3) via its Perl bindings. Survival analysis used the survival package (v2.20) in R (v2.0.1).

Comparison of Patient Classification By Using 3- and 6-Gene Signatures. Two genes (*STX1A* and *HIF1A*) from this signature overlap with our previously reported linear risk-score analysis (4). Because we used the same training dataset for both algorithms we are able to investigate the effect this overlap has on patient classifications. We compared the patient-by-patient predictions of our earlier risk-score-derived 3-gene signature and our current 6-gene signature (Table S2). The 3-gene signature did not classify 10 patients from the initial cohort of 147, leaving 137 patients classified by both methods. Of these, 108 (79%) were classified identically by both methods. Most of the 29 mismatches (24/29 = 83%) were classified as poor prognosis by the 3-gene signature and good prognosis by the 6-gene signature. Similar proportions of adenocarcinomas and squamous cell carcinomas were divergently classified (22.6% vs. 20.2%, $P = 0.904$). The 2 classifiers showed somewhat greater divergence for stage I than stage II or III patients, although this was not statistically significant (25.6% vs. 13.7%, $P = 0.154$). The few divergences observed reflect the use of median dichotomization in the risk-score analysis. Median dichotomization is a common statistical procedure used when the training groups that cannot be defined as a priori, and forces the good and poor prognosis groups to be equally sized in the training dataset. By contrast the semisupervised approach used by the mSD algorithm finds groups that reflect the strongest trend within the training dataset, regardless of group sizes. This is done by using unsupervised pattern-recognition (clustering). As a result mSD identifies groups of unequal size (92 good and 55 poor prognosis patients) whereas the risk-score analysis identified groups of equal size (68 good and 69 poor prognosis patients). Despite this underlying algorithmic difference, these data show that the 2 classifiers concur on the classifications for the majority of patients and that the few divergent classifications are not strongly biased according to any clinical covariates.

Prognostic Signature Cross-Validation. By using the normalized dataset, each of the 147 patients was sequentially removed from the sample. The mSD algorithm was then trained on the remaining 146 patient samples to select a prognostic subset of genes, as outlined above. The Euclidean distance between the expression profile of the omitted patient and the median expression profiles of the good and poor prognosis groups of patients were then calculated. The patient was classified into the nearer of these 2 groups, and the entire procedure was repeated 147 times so that each patient was omitted once. A survival curve of the resulting classifications was then plotted, and a stage-adjusted Cox proportional hazards model fitted as above. Cross validation was performed in *R* (v2.4.1) by using the survival package (v2.31).

Independent Validation Datasets. Four independent, publicly available datasets were used to validate the 6-gene classifier identified by modified steepest-descent (5–8). These datasets were not used to select the 158 genes in our study and thus each constitutes an independent validation dataset. Two validation datasets were generated by using Affymetrix microarrays (7, 8) and two using custom cDNA arrays (5, 6). Two are comprised primarily of adenocarcinomas (5, 7) and two exclusively of squamous cell carcinomas (6, 8). In each case, the normalized data were downloaded from the Gene Expression Omnibus (GEO) repository. ProbeSets or spots representing the genes involved in the signature were identified by using NetAffx annotation for Affymetrix arrays (7, 8) and BLAST analysis against UniGene build Hs.199 (5, 6) for cDNA arrays. When multiple ProbeSets for a single gene were present, the Pearson's correlation between their vectors was calculated. If they were strongly correlated ($R > 0.75$) they were collapsed by averaging; otherwise bl2seq analysis against the RefSeq mRNA for the gene in question was used to identify the best match. Median scaling was performed as described in ref. 9. House-keeping gene normalization was used for the 2 Affymetrix array platforms, as described above for the PCR analysis. Because only 2 of the 4 house-keeping genes used were available on the custom cDNA platforms, this normalization step was omitted.

For each validation dataset, the distance between the expression profile for each patient and the cluster centers (medians) identified from the training dataset were calculated. A patient was classified into the nearer cluster if the ratio of the distances between the profile and the 2 clusters was at least 0.9. This quality criterion was not used for the 2 studies with small sample sizes where 1 signature gene was not present on the array platform (5, 6). The resulting classifications were then tested to determine whether our prognostic signature resulted in significant survival differences by using Cox proportional hazards model with adjustment for stage in R (v2.4.1) by using the survival library (v2.33) as previously described.

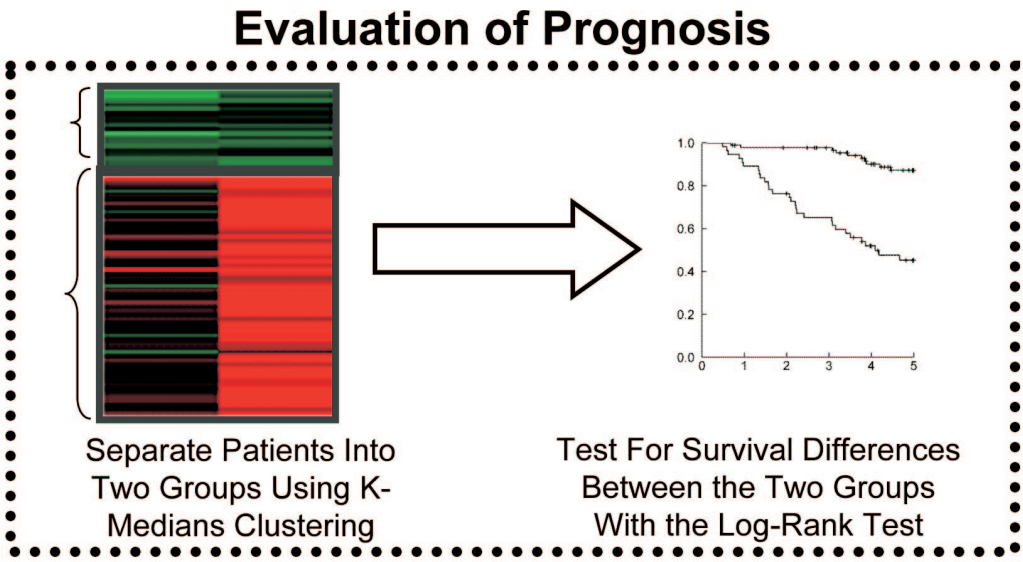
Pooled Validation. Several smaller expression studies of nonsmall cell lung cancer were also available but, because of their limited number of patients, were not useful as validation datasets. To leverage these resources, we combined all patients from the 4 studies described above, along with datasets from the Mayo Clinic and Washington University (10), and 2 additional studies of mRNA expression in NSCLC (11, 12). In each of these cases, the raw data (CEL files) were downloaded and preprocessed by using the RMA algorithm (13) as implemented in v1.6.7 of the BioConductor affy package (14) for the *R* statistical environment (v2.1.1). One dataset (11) included highly-correlated technical replicates for some samples, which were collapsed through ProbeSet-wise averaging. The resulting dataset of 589 patients was then subject to the same nearest-centre classification described above. Survival between the 2 groups was tested by using Cox proportional hazards model with adjustment for stage. The normalized data and clinical annotations for all patients used in this article are presented in Table S2.

Permutation Analysis. To determine the number of 6-gene classifiers (signatures) that could be generated from our 158-gene expression dataset, we performed a large permutation analysis. We tested the prognostic capability of 10 million combinations of 6 genes. For each signature we used the methodology described above: *k*-means clustering to divide patients into 2 groups and log-rank analysis to estimate the separation between the 2 groups. Study of all combinations is not possible for larger subset sizes because of the combinatorial explosion. This analysis was performed in the *R* statistical environment (v2.6.1) by using the survival package (v2.34).

To test each signature we used the clusters defined in our training cohort to classify patients from 4 additional datasets (7, 8, 11, 12), again by using Euclidean distances and log-rank analysis. The normalized data for each of these datasets was extracted for the genes in each signature. Euclidean distances were calculated between each patient and the centre of the 2 training clusters, and the patient was classified into the nearest cluster. Survival differences between good and poor prognosis clusters were then assessed by using log-rank analysis.

Finally, to consider the generalizability of each prognostic signature across all 4 testing datasets, we used percentile analysis. First, for visualization purposes we calculated and plotted the Gaussian kernel density of prognostic signatures in each validation dataset. Next, we calculated the percentile rank of each signature in each of the 4 validation datasets. The product of these ranks provides an estimate of the overall validation of a classifier across all 4 datasets, and we plotted the Gaussian kernel densities of these ranks. The performance of the 6-gene mSD-signature was then treated in the same manner and its location marked on plots with an arrow to indicate its performance relative to the distribution of all potential prognostic markers.

1. Boutros PC, Okey AB (2005) Unsupervised pattern recognition: An introduction to the whys and wherefores of clustering microarray data. *Brief Bioinform* 6:331–343.
2. Simon R, Radmacher MD, Dobbin K, McShane LM (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 95:14–18.
3. de Hoon MJ, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. *Bioinformatics* 20:1453–1454.
4. Lau SK, *et al.* (2007) Three-gene prognostic classifier for early-stage non small-cell lung cancer. *J Clin Oncol* 25:5562–5569.
5. Larsen JE, *et al.* (2007) Gene expression signature predicts recurrence in lung adenocarcinoma. *Clin* Cancer Res 13:2946–2954.
6. Larsen JE, *et al.* (2007) Expression profiling defines a recurrence signature in lung squamous cell carcinoma. *Carcinogenesis* 28:760–766.
7. Bild AH, *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439:353–357.
8. Raponi M, *et al.* (2006) Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res* 66:7466–7472.
9. Barsyte-Lovejoy D, *et al.* (2006) The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer Res* 66:5330–5337.
10. Lu Y, *et al.* (2006) A gene expression signature predicts survival of patients with stage I nonsmall cell lung cancer. *PLoS Med* 3:e467.
11. Bhattacharjee A, *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 98:13790–13795.
12. Beer DG, *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 8:816–824.
13. Irizarry RA, *et al.* (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31:e15.
14. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20:307–315.



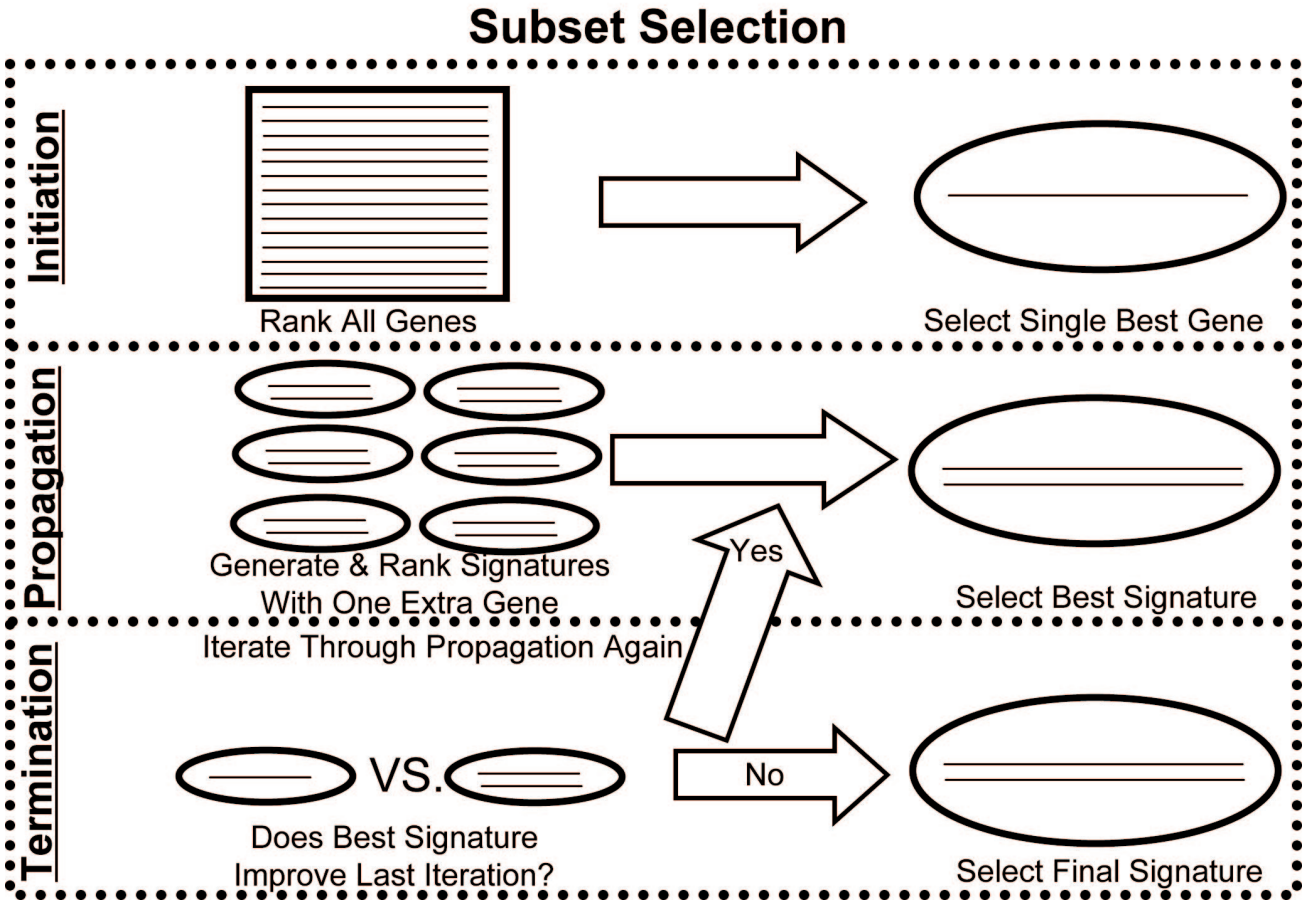


Fig. S1. Outline of the mSD procedure. The modified steepest-descent algorithm has 2 components: a prognosis-prediction component and a feature-selection component. First, given a set of one or more features, mSD estimates prognosis in a semisupervised way. Patients are clustered using *k*-medians clustering into 2 groups and the survival difference between these 2 groups is measured with the χ^2 output of a log-rank test. Features are ranked according to this χ^2 statistic. Second, features are selected by using a gradient-descent approach. The initial feature is chosen based on the univariate ranking of all features. After this initiation phase, features are added one-by-one by greedy descent. Once a local minimum has been reached, the algorithm terminates.

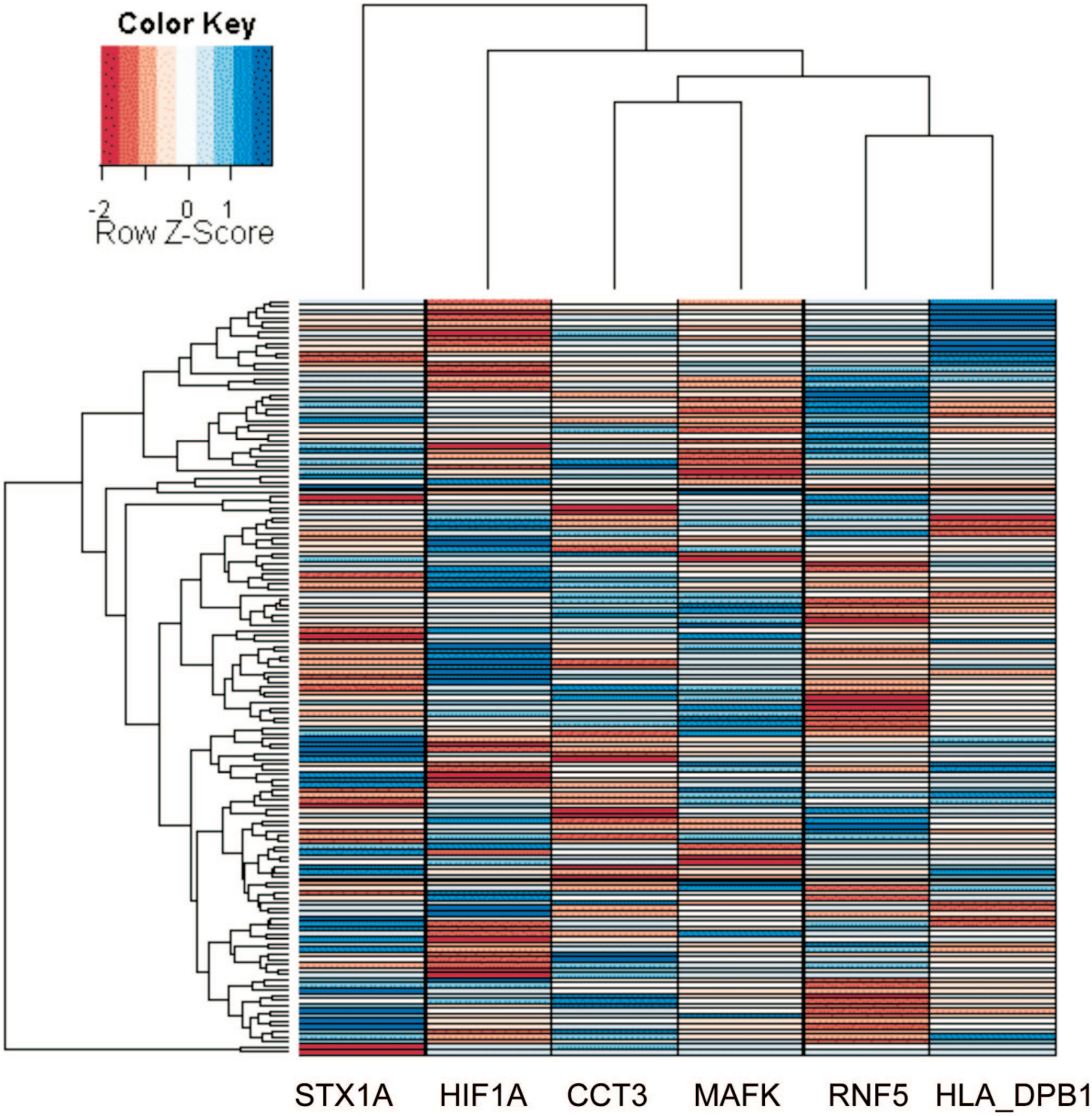
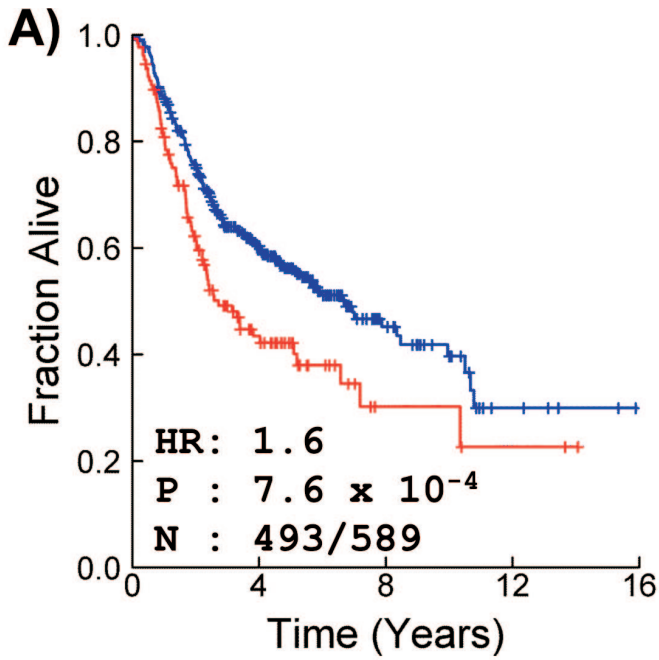
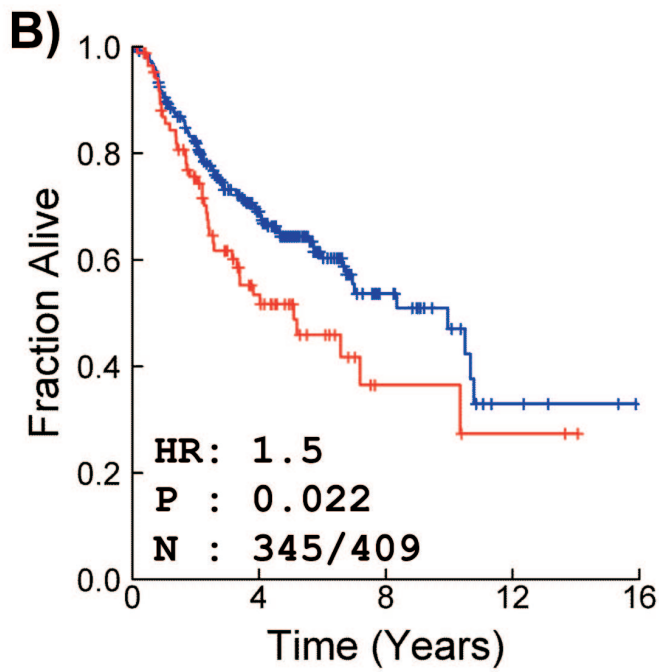


Fig. S2. Clustering of the training dataset. The expression profiles of the 6-genes from the mSD-signature for the 147 patients of the training dataset were subjected to unsupervised pattern-recognition. Agglomerative hierarchical clustering by using complete linkage was performed. The columns represent genes and the rows represent individual patients. The 6 genes all show unique expression patterns, as indicated by the long terminal arms of the column dendrogram. Patients do not fall into 1 or 2 large clusters, but rather into a diversity of small, nonlinear ones, as indicated by the row dendrogram.



Good	365	159	30	5	0
Poor	128	35	4	2	0



Good	259	127	22	4	0
Poor	86	30	4	2	0

Fig. S3. Classifier validation in a pooled dataset. Data from 8 studies was pooled into a dataset of 589 patients. The 6-gene classifier separated all (*A*) and stage I patients (*B*) into groups with significantly different survival. The number of patients at risk in each molecularly-defined group is indicated at each time-point. The stage-adjusted hazard ratio (HR) and *P* value (Wald test), and the number of patients successfully classified (*N*) are also shown.

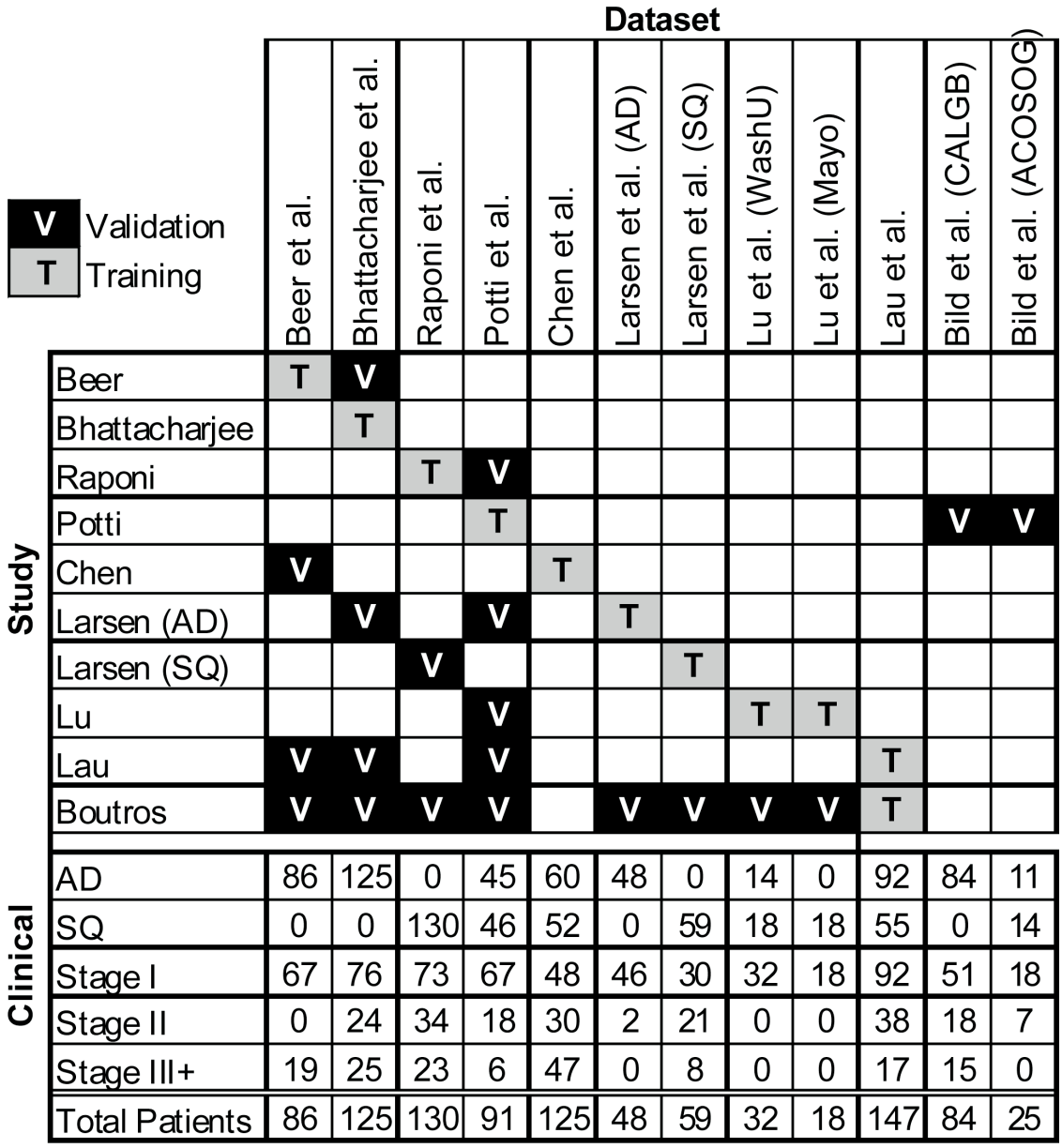


Fig. S4. Summary of validation datasets. A number of public NSCLC datasets exist. These datasets are listed along the top of the chart, while various papers are listed along the side, identified by the first author. Each dataset is annotated according to which studies used it. Training datasets are marked with gray, whereas validation datasets are marked with solid black. The current study is highly validated, assessing 8 distinct datasets. Some key clinical characteristics of each dataset are listed. AD = adenocarcinoma. SQ = squamous cell carcinoma.

Other Supporting Information Files

Table S1 (PDF)
Table S2 (PDF)
Table S3 (PDF)
Table S4 (PDF)
Table S5 (PDF)