Supplemental Information

I. Datasets

We analyzed published raw datasets specified below. Please refer to associated references for details regarding sample selection, preparation, and expression profiling. In some cases, we analyzed subsets of patient samples reported (e.g. those with adequate clinical followup). Actual datasets that were analyzed are provided (e.g. Dataset A). The accompanying Microsoft Excel spreadsheet contains pages with actual results. These pages are refered to as Web Spreadsheet (i.e. Microsoft Excel spreadsheet) followed by a letter (e.g. Web Spreadsheet A) throughout the following text.

a. Global Cancer Map – 64 primary adenocarcinomas and 12 metastatic adenocarcinomas (lung, breast, prostate, colon, ovary, and uterus) from unmatched patients prior to any treatment. Clinical stage of primary tumors and outcome unknown [1]. See Dataset A. This study was previously published in 2001 by Ramaswamy et al.

b. Lung – 62 stage I and II sporadic, primary lung adenocarcinomas with > 4 years clinical followup after surgical resection. Clinical endpoint: Overall survival [2]. See Dataset B. This study was previously published in 2001 by Bhattacharjee at al.

c. Breast – 78 stage I sporadic, primary breast adenocarcinomas with > 5 years clinical followup after lumpectomy. Clinical endpoint: Time to metastasis [3]. See Dataset C. This study was previously published in 2002 by Van't Veer et al.

d. Prostate – 21 stage I sporadic, primary prostate adenocarcinomas with > 4 years clinical followup after radical prostatectomy. Clinical endpoint: Time to PSA relapse after radical prostatectomy [4]. See Dataset D. This study was previously published in 2002 by Singh et al.

e. Medulloblastoma – 60 medulloblastomas with > 5 years clinical followup after multi-modality treatment. Clinical endpoint: Overall survival [5]. See Dataset E. This study was previously published in 2002 by Pomeroy et al.

f. Large B-cell Lymphoma – 58 large B-cell lymphomas with >5 years clinical followup after combination CHOP chemotherapy. Clinical endpoint: Overall survival [6]. See Dataset F. This study was previously published in 2002 by Shipp et al.

II. Primary versus Metastatic Adenocarcinoma Comparison

We first compared 64 primary adenocarcinomas, treated as one class, to all 12 metastatic adenocarcinomas treated as a separate class using the signal-to-noise (S_x) statistic (Dataset A). The primary tumors were from different sites as designated in the accompanying dataset. The 12 metastases arose from the same spectrum of sites, but were resected from a variety of end-organs:

Sample	Site of Origin	Resection Site
Met__Unknown_CUP_1	**Breast**	**Lung**
Met__Unknown_CUP_16	**Colon**	**Retroperitoneum**
Met__Unknown_CUP_22	**Lung**	**Adrenal**
Met__Unknown_CUP_12	**Lung**	**Kidney**
Met__Unknown_CUP_14	**Ovary**	**Omentum**
Met__Unknown_CUP_19	**Uterus**	**Omentum**
Met__Metastases_9912c062_Rb	**Colon**	**Ovary**
Met__Metastases_HCTN_19274	**Ovary**	**Colon**
Met__Metastases_MetCaP109	**Prostate**	**Bone**
Met__Metastases_MetCaP125	**Prostate**	**Bone**
Met__Metastases_MetCaP128	**Prostate**	**Bone**
Met__Metastases_MGH_4934	**Breast**	**Lung**

Re-scaling of data

The raw expression data are Affymetrix's GeneChip software (MAS 4) “average difference” units. Each dataset was re-scaled to account for different microarray intensities in a given set. Each column (sample) in the dataset was multiplied by *1/slope* of a least squares linear fit of the sample versus the reference (the first sample in the dataset). This linear fit is done using only genes that have 'Present' calls in both the sample being re-scaled and the reference. The sample chosen as reference is a typical

one (i.e. one with the number of "P" calls closer to the average over all samples in the dataset).

Gene mapping

The primary and metastatic adenocarcinoma comparison was done on the Affymetrix Hu6800 / Hu35KsubA oligonucleotide microarray platform. Our primary lung adenocarcinoma dataset was, however, analyzed using the Affymetrix U95A oligonucleotide microarray platform. We therefore determined the universe of genes that are common to both microarray types (Web Spreadsheet A). Mapping was performed by matching Affymetrix probe set names to GenBank accession numbers and then mapping the GenBank accession numbers to UniGene clusters (build #147). All genes falling into the same UniGene clusters from both platforms were considered “mapped” genes. These 9376 common “mapped” genes (between the Hu6800 / Hu35KsubA and U95A platforms) were then considered when performing downstream analysis including gene marker selection and supervised learning (see below) to identify the gene set best able to distinguish between primary and metastatic adenocarcinomas.

Pre-processing of data

Pre-processing of the data consisted of a thresholding step followed by a filtering step. For the upper threshold, a ceiling of 16,000 units was chosen for all experiments because we observe fluorescence saturation of the scanner above this level; thus, values above this cannot be reliably measured. We also used a minimum threshold of 20 units for low expression values below this level. After thresholding, gene expression values were subjected to a variation filter that excluded genes showing minimal variation

across the samples being analyzed. The variation filter tests for a fold-change and absolute variation over samples (comparing max/min and max-min with predefined values and excluding genes not obeying both conditions), and thus excludes genes that vary minimally across the dataset. We used a max / min < 3 and max – min < 100 (8176 of 9376 genes passed this variation filter).

Gene marker selection

Genes correlated with a particular class distinctions (e.g. class 0 and class 1) were identified by sorting all of the genes on the array according the signal-to-noise statistic [7] $S_x = (\mu_{class\ 0} - \mu_{class\ 1}) / (\sigma_{class\ 0} + \sigma_{class1})$ where μ and σ represent the mean and standard deviation of expression, respectively, for each class. The colorgram depicted in Figure 1 displays the top 128 genes differentially expressed by primary adenocarcinomas and metastatic adenocarcinomas, ordered by their signal-to-noise ranking. These genes are also listed in Web Spreadsheet B (also Web Spreadsheet C) with their signal-to-noise score. An unexpected finding was that the metastasis-associated gene expression pattern appeared to be present in some primary tumors (not the “striped” pattern in the colorgram). This observation suggested the hypothesis that the gene expression program of metastasis may already be present in some primary tumors at the time of diagnosis.

III. Confirming Metastasis Gene Expression In Primary Tumors

To test the hypothesis that a metastasis program is expressed in primary tumors, we analyzed genes that are differentially expressed by primary and metastatic

adenocarcinomas in an independent set of 62 stage I / II primary lung adenocarcinomas [2] (Dataset B).

Gene marker selection

We began by analyzing the expression levels in the dataset that contained the primary and metastatic adenocarcinomas (Dataset A). Expression levels in this dataset below 20 units were assigned a value of 20, and those exceeding 16,000 units were assigned a value of 16,000. Genes whose expression did not vary across the dataset were removed (i.e. if max / min < 3 or max – min < 100). Genes correlated with a particular class distinctions (e.g. class 0 and class 1) were identified by sorting all of the genes in the dataset according the signal-to-noise statistic [7] $S_x = (\mu_{class\ 0} - \mu_{class\ 1}) / (\sigma_{class\ 0} + \sigma_{class1})$ where μ and σ represent the mean and standard deviation of expression, respectively, for each class. Graded numbers of these marker genes were then used to build a weighted-voting classifier.

Supervised learning: Prediction

We next used supervised learning to determine the best number of genes capable of distinguishing primary and metastatic adenocarcinomas. Supervised learning was performed using a weighted voting algorithm and evaluated using leave-one-out cross-validation [7]. Supervised learning incorporates the knowledge of class label information to make distinctions of interest. A training data set is used to select those features that best make a distinction. These features are then applied to an independent test data set to validate the ability of selected features to make that distinction. We selected a subset of

expressed genes best able to distinguish primary from metastatic adenocarcinomas and built a computational model using a weighted-voting algorithm that uses these selected genes to differentiate between these two classes.

Weighted-voting algorithm

The weighted voting algorithm [7] makes a weighted linear combination of relevant "marker" or "informative" genes obtained in the training set to provide a classification scheme for new samples. The selection of features (marker genes) is accomplished by computing the signal-to-noise statistic S_x (described above). The class predictor is uniquely defined by the initial set of samples and marker genes. In addition to computing S_x, the algorithm also finds the decision boundaries (half way) between the class means: $b_x = (\mu_{class0} + \mu_{class1}) / 2$ for each gene. To predict the class of a test sample y, each gene x in the feature set casts a vote: $V_x = S_x (g_x^y - b_x)$ and the final vote for class 0 or 1 is $sign$ $(\Sigma_x V_x)$.

Leave-one-out cross-validation

We used leave-one-out cross-validation to determine the number of genes best capable of distinguishing primary from metastatic adenocarcinomas. Briefly, one withholds a sample, builds a predictor based only on the remaining samples, and predicts the class of the withheld sample. The process is repeated for each sample, and the cumulative error rate is calculated.

Supervised learning using this weighted-voting leave-one-out cross-validation approach was performed using graded numbers of top primary versus metastasis markers. The top 128 markers (64 overexpressed in primary adenocarcinomas and 64 overexpressed in metastatic adenocarcinomas) yielded the best primary versus metastasis prediction using cross-validation (Web Spreadsheet D), and these markers were used in subsequent analysis (see below). As seen, a number of primary tumors are mis-classified at metastases.

P-values for this prediction were assigned based upon the frequency with which models generated and tested on 1000 random permutations of the class labels performed better than models generated using the observed class labels (Web Spreadsheet E). After each permutation of the class distinction, weighted voting predictors with a range of feature numbers were used to predict the identity of the held out sample. Statistics, including the maximum, minimum, and mean accuracies, of the models generated from the permuted data are shown. By performing 1000 random permutations in this manner, the strength of the observed association was measured against chance. When predicting the primary tumor versus metastases distinction, the best weighted voting model made 15 errors out of 76 samples using 128 features (using leave-one-out cross-validation for testing). When limiting consideration to a 128 feature model for predicting the tumor versus metastases distinction in the permuted samples, only 12 of the 1000 permutations of the class labels made 15 or less errors ($P = 0.012$). A more conservative estimate of the p-value for the predictor can be found when the weighted voting predictor is allowed to use any number of features to predict the permuted class labels. In this case, 149 of the 1000 permutations of the class labels made 15 or less errors ($P = 0.149$).

We then used unsupervised learning, or clustering, to examine the expression of these top markers (derived from a comparison between primary and metastatic tumors) in an independent set of 62 stage I / II primary lung adenocarcinomas (Dataset B). Since these lung tumors were initially profiled on the Affymetrix U95A microarray, we had to match the 128 Hu6800 / Hu35KsubA probe sets to corresponding probe sets on the U95 microarray (Web Spreadsheet F). This corresponded to 169 probe sets of the U95A platform (due to redundant probe sets), which were then used in the cluster analysis described next.

Unsupervised learning: Clustering

Unsupervised learning, or clustering, involves the aggregation of a diverse collection of data into clusters based on different features in a data set. For example, one could divide a group of people into clusters based on any combination of eye color, waist size, or height. Similarly, one can gather data about the various expressed genes in a collection of tumor samples and then cluster the samples as best as possible into groups based on the similarity of their aggregate expression profiles. Alternatively, one could cluster genes across all samples, to identify genes that share similar patterns of expression in varying biologic contexts. This approach has the advantage of being unbiased and allows for the identification of structure in a complex data set without making any a priori assumptions. We used the Cluster and TreeView software [8] to perform average linkage clustering, which organizes all of the data elements into a single tree with the highest levels of the tree representing the discovered classes. For pre-processing, we median centered genes and arrays twice (median polished) and then normalized genes [8]. We used a weighted centered correlation for arrays and performed average linkage clustering.

Hierarchical clustering in the space of these 169 top metastasis probe sets (corresponding to the 128 genes discovered on the Hu6800/35KsubA microarray set) identified two major clusters of primary tumors (a “primary” type and a “metastatic” type) differentially expressing two major gene clusters (C0 and C1) (Web Spreadsheet G and Web Spreadsheet H). A Fisher test demonstrated the highly significant correlation between gene expression in either primaries or metastases and membership in the C0 or C1 gene cluster (and thus “primary” type or “metastatic” type primary tumors) (P = 0.002, Figure 2*a*) (Web Spreadsheet I). This finding confirmed the hypothesis that the metastasis gene expression program is indeed detectable in a subset of primary lung tumors.

We next examined the two major clusters of primary tumors by creating a Kaplan-Meier survival plot.

Kaplan-Meier survival analysis

Kaplan-Meier survival plots were computed using the S-Plus statistical software package (http://www.insightful.com/products/splus/) and S-Plus 2000, Guide to Statistics Volume 2, chapter 9. The p-values for the prediction of outcome groups were computed using a log-rank test (Mantel-Haenszel method, chapter 9 in the same reference).

If the presence of the metastasis program in a primary tumor is biologically significant, one might expect the clinical outcome of patients with primary tumors that express this program to be worse. Indeed, patients whose primary tumors bore the metastasis gene expression program had significantly worse survival compared with patients whose

tumors lacked it (P = 0.009, Figure 2*b* and Web Spreadsheet G), consistent with the fact that death from lung cancer is in most cases attributable to metastasis. The metastasis signature could be reduced to a subset of 17 genes that largely recapitulated the observed outcome distinction (P = 0.010, Figure 2*c*; see below and Web Spreadsheet J). Importantly, random selection of 17 genes failed to generate such distinctions (P = 0.004; see below and Web Spreadsheet K), indicating that this result could not have been achieved by chance alone. Similarly, clustering the primary lung cancers in the space of all varying genes on the microarray (9248 genes resulting from an absolute variation filter of maxVal – minVal >100) failed to yield an outcome distinction (P = 0.8, Figure 2*d* and Web Spreadsheet L). These results suggest that some primary tumors are pre-configured to metastasize, and this propensity is detectable at the time of initial diagnosis.

Identifying the reduced 17-gene metastasis expression signature

We reasoned that many of the top 128 metastasis genes might not be contributing significantly to the clustering of primary lung adenocarcinoma samples, and might thus represent “noise” in the analysis. We therefore considered the two groups (clusters) or primary lung cancers and plotted the signal-to-noise values for each gene based on expression in these 2 clusters (Web Spreadsheet M). We noted that 21 Affymetrix probe sets (corresponding to 17 unique genes) were significantly correlated with the primary lung tumor cluster distinction with a $S_x > 0.4$. When we clustered the primary lung adenocarcinoma samples using this reduced metastasis gene set, and analyzed the resulting sample clusters in a Kaplan-Meier plot, we were able to largely recapitulate the previously observed outcome distinction (Web Spreadsheet J).

Permutation testing to determine the statistical significance of the 17-gene metastasis expression signature

We next asked whether this 17-gene metastasis gene expression signature yielded outcome distinctions for primary lung adenocarcinomas that were better than what would be observed by chance alone. We therefore selected 1000 random sets of 17-genes from the pool of 11388 highly varying genes (Thresholded (Minimum = 2, Maximum = 16000) and filtered (Mimimum variation = 3-fold and Absolute difference = 50)) (this pre-processing was loosened compared to that described above to allow for all genes in the 17-gene signature to pass). These gene sets were then used to perform 1000 independent clusterings of the primary lung adenocarcinomas, and each clustering was subject to Kaplan-Meier survival analysis as described above. Notably, survival distinctions that exceeded our observed *P*-value of 0.010 were seen only 4 out of 1000 permutations ($P = 0.004$) (Web Spreadsheet N). This observation demonstrates that the 17-gene metastasis signature is relatively unique in describing subsets of primary lung adenocarcinoma with differential overall survival.

IV. Applying The 17-gene Metastasis Signature To Other Solid Tumor Types

To explore the generality of the metastasis signature, we applied it to other tumor types. We first studied gene the expression profiles of 78 stage I primary breast adenocarcinomas (Dataset C). These tumors were subjected to microarray gene expression profiling using the 24481-gene Rosetta inkject micorarray. We mapped the 17-gene lung metastasis signature to the Rosetta platform using UniGene build #147, which resulted in 16 successfully mapped genes (Web Spreadsheet O). We then

performed cluster / Kaplan-Meier survival analysis as described above using these genes (median polishing and average linkage clustering). Again, tumors bearing the metastasis signature at diagnosis were more likely to develop distant metastases than those lacking this signature (*P* = 0.024, Figure 3*a* and Web Spreadsheet P). A similar result was seen in 21 prostate adenocarcinomas (using the 17 gene signature, since both datasets were created with the U95A platform) (*P* = 0.022, Figure 3*b* and Web Spreadsheet Q), and was even seen in a series of 58 medullobastomas (using the 21 probe sets present on the Hu6800 platform only (Web Spreadsheet R)) (*P* = 0.029, Figure 3*c* and Web Spreadsheet S), even though these tumors are not adenocarcinomas. These results argue for the existence of generic metastasis gene expression programs rather than distinct mechanisms of metastasis in different tumor types. Interestingly, however, the metastasis signature defined herein did not predict outcome in diffuse large B-cell lymphoma (also on the Hu6800 platform; 21 probe sets) (Figure 3*d* and Web Spreadsheet T), consistent with the idea that hematopoietic tumors have specialized mechanisms for navigating the hematologic and lymphoid compartments. Of note, the use of clustering as the analytical method allowed us to define a metastasis signature on one microarray platform (Affymetrix HU-6800 and HU-35k subA) and apply it to samples analyzed using different microarray designs (Rosetta inkjet arrays and Affymetrix U95A arrays).

V. The 17-gene Metastasis Signature

The component genes of this signature are listed in Table 1 in the paper.

The metastasis signature consisted of 8 up-regulated and 9 down-regulated genes (Table 1 in the paper). Of note, none of the genes represent particularly striking individual markers of metastasis (see below); rather, as demonstrated above, the signature taken as a whole appears to contain biologically important information. Four of 8 upregulated genes are components of the protein translation apparatus (*SNRNPF*, *EIF4EL3*, *HNRNPAB*, *DHPS*), consistent with reports of amplification and overexpression of translation initiation factors in invasive cancers. The Securin gene (*PTTG1*), encoding an inhibitor of the enzyme Separase, is similarly overexpressed in metastases. Separase function is required for sister chromatid separation during cell division, and degradation of Securin at the metaphase-anaphase transition is essential for proper chromosome segregation. A role for Securin in cancer pathogenesis is also supported by the observation that increased Securin expression is seen in tumors with increased vascularity and local invasion. It is not yet clear whether these properties and the reported ability of Securin over-expression to transform NIH 3T3 cells are related to its role in mitosis or to some other, yet to be identified mechanism.

The metastasis signature is also notable for the significant proportion of the signature that appears to be derived from non-epithelial components of the tumor. Specifically, the Type I collagens *COL1A1* and *COL1A2* (whose expression is restricted to fibroblasts), Actin γ2 (*ACTG2*) and Calponin (*CNN1*) (markers of smooth muscle), MHC class II DP-β1 (*HLA-DPB1*) and *RUNX1* (unique to hematopoietic cells) are prominent components of the metastasis signature. The up-regulation of collagen genes in primary tumors with metastatic potential is consistent with the recent observations that epithelial-mesenchymal interactions are critical determinants of tumor cell behavior. This

observation is also in line with reports of increased levels of Type 1 collagen in metastatic lesions and in the serum of patients with metastatic diseases.

Similarly, the apparent down-regulation of MHC class II expression in tumors that metastasize likely reflects decreased numbers of infiltrating professional antigen presenting cells (e.g. dendritic cells, macrophages) that are critical for effective anti-tumor immune responses. Interestingly, *RUNX1* is also down-regulated in metastasis-prone tumors and is both a putative tumor suppressor and regulates MHC class II expression in hematopoietic cells (T.R.G. personal observation). These observations taken together suggest that the metastasis gene expression signature arises from both malignant and stromal elements in primary tumors. Of note, this major component of the metastasis signature would have been missed had the malignant epithelial cells been isolated (e.g. by laser capture microdissection) prior to expression profiling.

Statistical analysis of individual components of the 17-gene metastasis signature

We performed two-tailed T-tests (using S-plus) to determine the correlation between individual genes in the metastasis signature and clinical outcome in each solid tumor dataset to determine whether any single gene was solely capable of yielding clinical outcome differences. The p-values for these T-tests are presented (Web Spreadsheet T). As can be seen, few individual genes were associated with clinical outcome at a statistically significant level ($p < 0.05$) in any dataset. We also present values for our signal-to-noise feature selection statistic (see above) for each of these genes in each data set (Web Spreadsheet V). This shows the direction of correlation of the individual genes with the outcome signature in each of the datasets.

Overlap between our Metastases signature and Rosetta's list of 70 prognostic markers

We mapped Rosetta's list of 70 prognostic markers [3] for breast cancer metastasis to probe set accession numbers on Affymetrix's HU6800 / HU35KsubA chip set to evaluate whether there were any genes in common with our metastases signature. Thirty-seven of Rosetta's list of 70 prognostic markers could be properly mapped to 53 probe sets on Affymetrix's HU6800 / HU35KsubA chip set (Web Spreadsheet W). Of these 37 unique genes, none were in either our reduced 17 gene metastases signature or in the larger 128 gene signature. It is nevertheless possible that a large number of remaining genes that could not be successfully mapped are commonly present in the two analyses.

Outcome clustering summary

A summary of the outcome clustering in each dataset is shown in Web Spreadsheet X and Web Spreadsheet Y. These tables show how each sample clustered in the datasets, provides a confusion matrix summarizing each clustering result, and provides a summary of P-value statistics for these analyses. Also summarized in this table are the definitions for the clusters and observed outcome distinctions.

VI. References

1. Ramaswamy, S. et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 15149-15154 (2001).
2. Bhattacharjee, A. et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 13790-13795 (2001).
3. van 't Veer, L.J. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536 (2002).

4. Singh, D. et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203-209 (2002).

5. Pomeroy, S.L. et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436-442 (2002).
6. Shipp, M.A. et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **8**, 68-74 (2002).
7. Golub, T.R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).
8. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863-14868 (1998).