A molecular signature of metastasis in primary solid tumors

Sridhar Ramaswamy[1,2], Ken N. Ross[1], Eric S. Lander[1,3] & Todd R. Golub[1,2]

Published online 9 December 2002; doi:10.1038/ng1060

Metastasis is the principal event leading to death in individuals with cancer, yet its molecular basis is poorly understood[1]. To explore the molecular differences between human primary tumors and metastases, we compared the gene-expression profiles of adenocarcinoma metastases of multiple tumor types to unmatched primary adenocarcinomas. We found a gene-expression signature that distinguished primary from metastatic adenocarcinomas. More notably, we found that a subset of primary tumors resembled metastatic tumors with respect to this gene-expression signature. We confirmed this finding by applying the expression signature to data on 279 primary solid tumors of diverse types. We found that solid tumors carrying the gene-expression signature were most likely to be associated with metastasis and poor clinical outcome ($P < 0.03$). These results suggest that the metastatic potential of human tumors is encoded in the bulk of a primary tumor, thus challenging the notion that metastases arise from rare cells within a primary tumor that have the ability to metastasize[2].

The prevailing model of metastasis holds that most primary tumor cells have low metastatic potential, but rare cells (estimated at less than one in ten million) within large primary tumors acquire metastatic capacity through somatic mutation[2]. The metastatic phenotype includes the ability to migrate from the primary tumor, survive in blood or lymphatic circulation, invade distant tissues and establish distant metastatic nodules. This model is primarily supported by animal models in which poorly metastatic cell lines can spawn highly metastatic variants if the process is facilitated by the isolation of rare metastatic nodules, expansion of the cells *in vitro* and injection of these selected cells into secondary recipient mice[3,4]. No direct evidence of this genetic selection model has, however, been documented in human tumors.

To study the molecular nature of metastasis, we analyzed the gene-expression profiles of 12 metastatic adenocarcinoma nodules of diverse origin (lung, breast, prostate, colorectal, uterus, ovary) and compared them with the expression profiles of 64 primary adenocarcinomas representing the same spectrum of tumor types obtained from different individuals. This comparison identified an expression pattern of 128 genes that best distinguished primary and metastatic adenocarcinomas (Fig. 1). Notably, this gene-expression pattern associated with metastases also seemed to be present in some primary tumors, resulting in these tumors being misclassified as metastases (see Web Note A and Web Table A). This observation suggested the hypothesis that a gene-expression program of metastasis may already be present in the bulk of some primary tumors at the time of diagnosis.

To test this hypothesis, we analyzed the metastases-derived gene-expression program in several large gene-expression data sets containing molecular profiles of primary solid tumors. First, we analyzed 62 stage I/II primary lung adenocarcinomas[5] for expression of the metastases-associated genes defined above. Hierarchical clustering in the space of these 128 genes identified two main clusters of primary tumors with gene-expression profiles that were highly correlated with the original primary-tumor versus metastases distinction ($P = 0.002$; Fig. 2*a* and see Web Note A and Web Table A online). This finding confirmed that the gene-expression program associated with metastasis was detectable in a subset of primary lung tumors.

If the presence of the metastasis program in a primary lung tumor is biologically significant, one might expect the clinical outcome of individuals with that gene-expression profile to be worse, as death from lung adenocarcinoma is in most cases attributable to metastasis[6]. Indeed, individuals whose primary tumors bore the metastases-associated gene-expression program had significantly shorter survival times compared with individuals whose tumors lacked it ($P = 0.009$; Fig. 2*b*).

We next sought to refine the metastases-associated gene-expression signature to a smaller set of genes that reflected the structure of the 128-gene set, reasoning that many of these 128 genes might not contribute greatly to the observed distinction. A reduced set of 21 probes representing 17 unique genes nearest the centroids of the two lung cancer clusters largely recapitulated the observed outcome distinction ($P = 0.010$; Fig. 2*c*). Notably, we found that random sets of 17 genes did not generate such distinctions ($P = 0.004$; see Web Note A and Web Table A online), indicating that the distinction was probably not achieved by chance alone. Similarly, clustering the primary lung cancers in the space of all genes on the microarray did not yield an outcome distinction ($P = 0.8$; Fig. 2*d*). These results support the idea that some primary tumors are pre-configured to metastasize, and that this propensity is detectable at the time of initial diagnosis.

To explore the generality of the refined gene-expression signature associated with metastasis, we applied it to other tumor types. In 78 small stage I primary breast adenocarcinomas[7], tumors bearing the

[1]*Whitehead Institute/MIT Center for Genome Research, One Kendall Square, Building 300, Cambridge, Massachusetts 02139, USA.* [2]*Dana-Farber Cancer Institute/Harvard Medical School, 44 Binney Street, Boston, Massachusetts 02115, USA.* [3]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence should be addressed to S.R. (e-mail: sridhar@genome.wi.mit.edu) or T.R.G. (golub@genome.wi.mit.edu).*

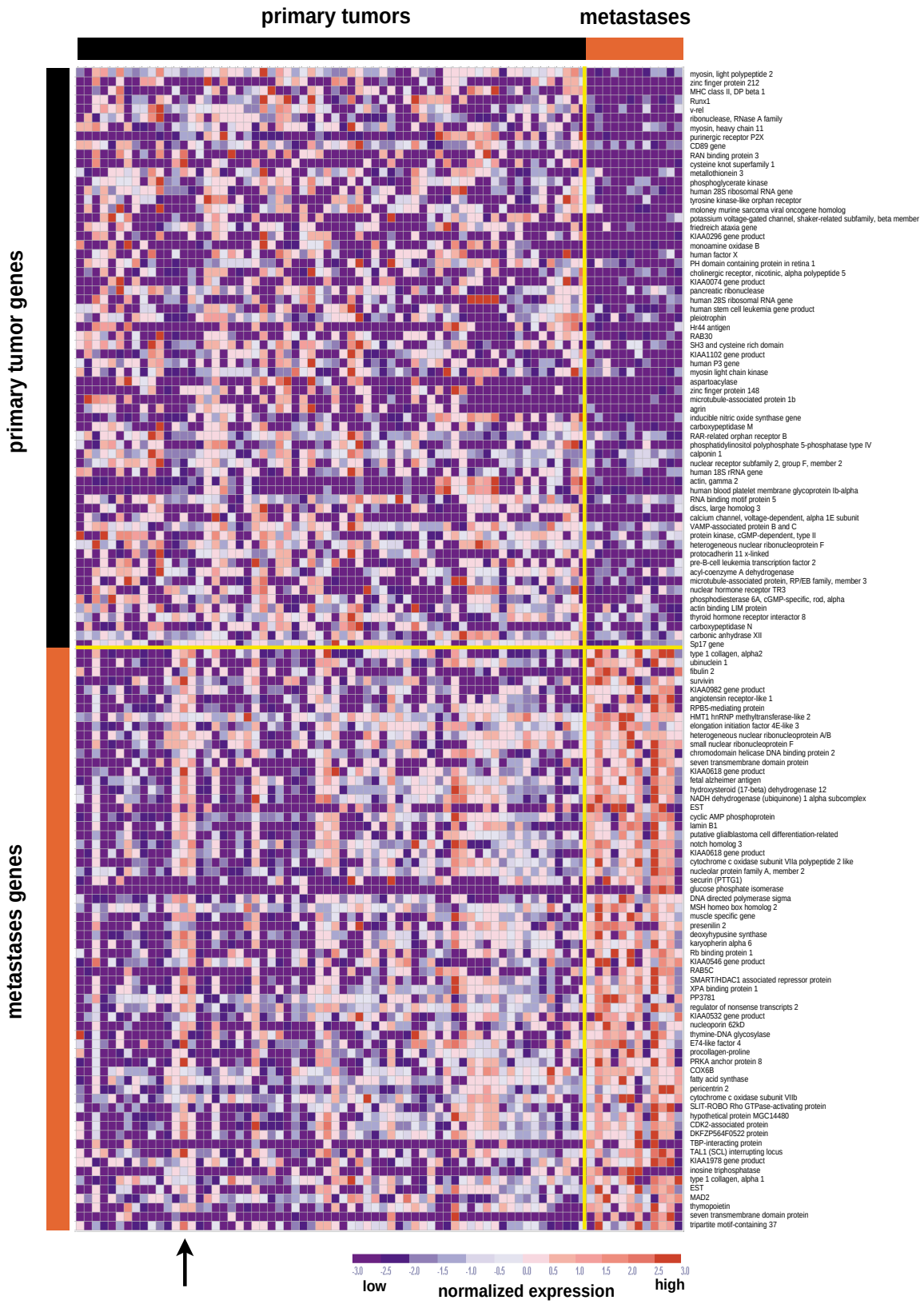

Fig. 1 Genes associated with metastases. Columns represent human tumor samples (64 primary and 12 metastatic adenocarcinomas); rows represent the 128 genes (64 overexpressed and 64 underexpressed in metastases) that best distinguished primary from metastatic tumors using a weighted-voting algorithm in leave-one-out cross-validation[28] (cross-validation accuracy = 80%, $P = 0.012$ by permutation testing; see Web Note A and Web Table A online). Colorgram depicts high (red) and low (blue) relative levels of gene expression. A 'striped' pattern was observed in some primary tumors (arrow), indicating the presence of a gene-expression program associated with metastases.

gene-expression signature at diagnosis were more likely to develop distant metastases than those lacking this signature ($P = 0.024$; Fig. 3*a*). We observed a similar result in 21 prostate adenocarcinomas[8] ($P = 0.022$; Fig. 3*b*). This result was also observed in a series of 60 medulloblastomas[9] ($P = 0.029$; Fig. 3*c*), despite that fact that these tumors are not adenocarcinomas. Collectively, these results argue for generic gene-expression programs related to metastasis rather than distinct mechanisms of metastasis in different tumor types. Notably, the refined metastasis signature did not predict outcome in diffuse large B-cell lymphoma[10] ($P = 0.497$; Fig. 3*d*), consistent with the idea that hematopoietic tumors have specialized mechanisms for navigating the hematologic and lymphoid compartments. It is important to note that 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).

The refined gene-expression signature associated with metastasis contained eight upregulated and nine downregulated genes (Table 1). None of these genes represent individual markers of metastasis; rather, it was the signature taken as a whole that seemed to contain predictive information (see Web Note A and Web Table A online). Four of eight upregulated genes are components of the protein translation apparatus (*SNRPF*, *EIF4EL3*, *HNRPAB*, *DHPS*), consistent with reports of amplification and overexpression of translation factors in tumor growth and invasion[11]. The gene encoding securin (*PTTG1*), 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[12,13]. A role for securin in cancer pathogenesis is also supported by the observation that increased expression of securin has been observed in tumors with increased vascularity and local invasion[14]. It is not yet clear whether these properties and the reported ability of securin overexpression to transform NIH3T3 cells are related to its role in mitosis or to some other mechanism[15].

A considerable proportion of the refined gene-expression signature associated with metastasis seems to be derived from non-epithelial components of the tumor. Specifically, these include the genes encoding the type I collagens (*COL1A1* and *COL1A2*), whose expression is restricted to fibroblasts, actin γ2, myosin heavy chain, myosin light chain kinase and calponin (markers of smooth muscle), MHC class II DP-β1 and the transcription factor RUNX1 (unique to hematopoietic cells). The upregulation of collagen genes in primary tumors with metastatic potential is consistent with recent observations that epithelial–mesenchymal interactions are critical determinants of tumor cell behavior[16,17]. High levels of type 1 collagen in metastatic lesions and in the serum of individuals with metastatic disease have also been reported[18,19]. Similarly, the downregulation of expression of MHC class II in primary tumors that metastasize probably reflects lower numbers of infiltrating professional antigen-presenting cells (for example, dendritic cells and macrophages) that are critical for effective anti-tumor immune responses[20]. *RUNX1* is also downregulated in metastasis-prone tumors and is both a putative tumor suppressor[21] and regulator of MHC class II expression in hematopoietic cells (T.R.G., unpublished data). The gene-expression signature associated with metastasis may thus arise from both malignant and stromal elements in primary tumors. Notably, the large stromal component of the signature would have been missed had only malignant epithelial cells been isolated (for example, by laser capture microdissection) before expression profiling.

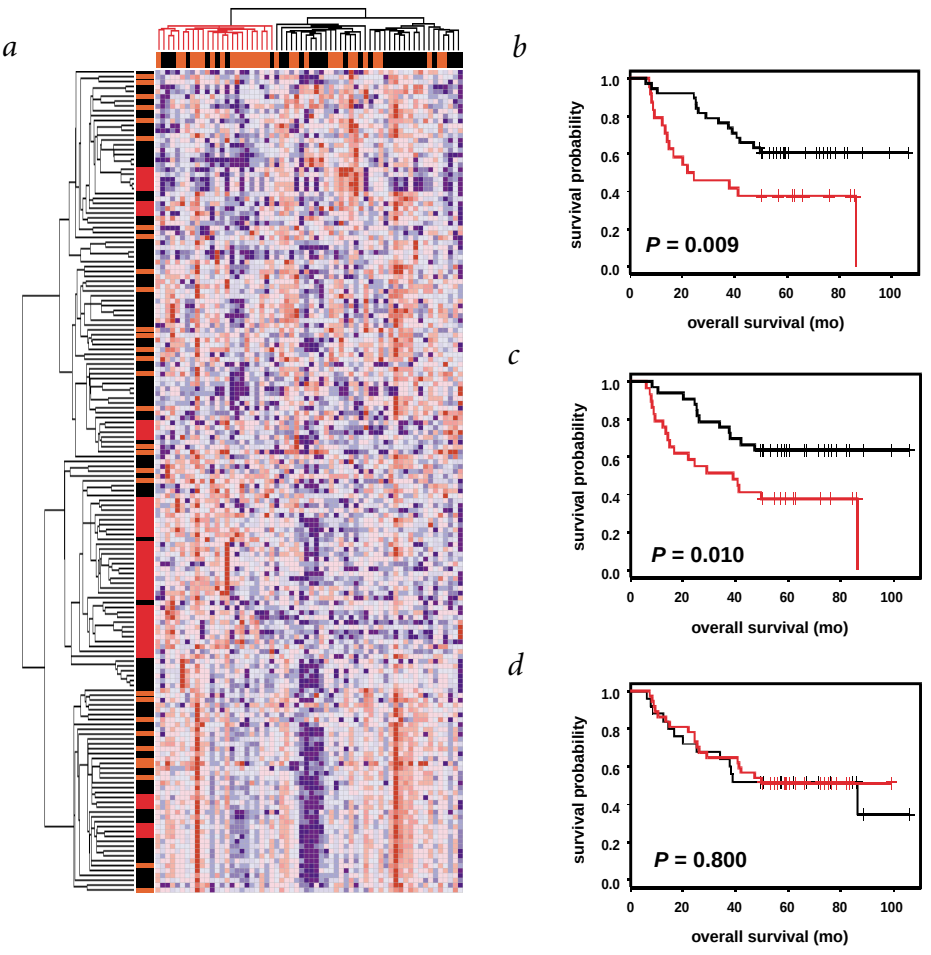

Fig. 2 Hierarchical clustering and Kaplan–Meier survival analysis in lung adenocarcinoma. ***a***, Hierarchical clustering of 62 primary lung adenocarcinomas using 128 metastases-derived genes defined two predominant primary-tumor groups in the resulting dendrogram. Colorgram depicts high (red) and low (blue) relative levels of gene expression. Vertical bar (left) indicates genes that were originally expressed in primary tumors (black) or metastases (red). Horizontal bar (top) indicates samples from individuals whose cancer was observed to be non-recurrent (black) or recurrent (red). ***b***, Kaplan–Meier survival analysis of clusters of individuals defined by 128 genes. ***c***, Kaplan–Meier survival analysis of clusters of individuals defined by the refined 17-gene signature. ***d***, Kaplan–Meier survival analysis of clusters of individuals defined by 9,248 highly varying genes in the data set.

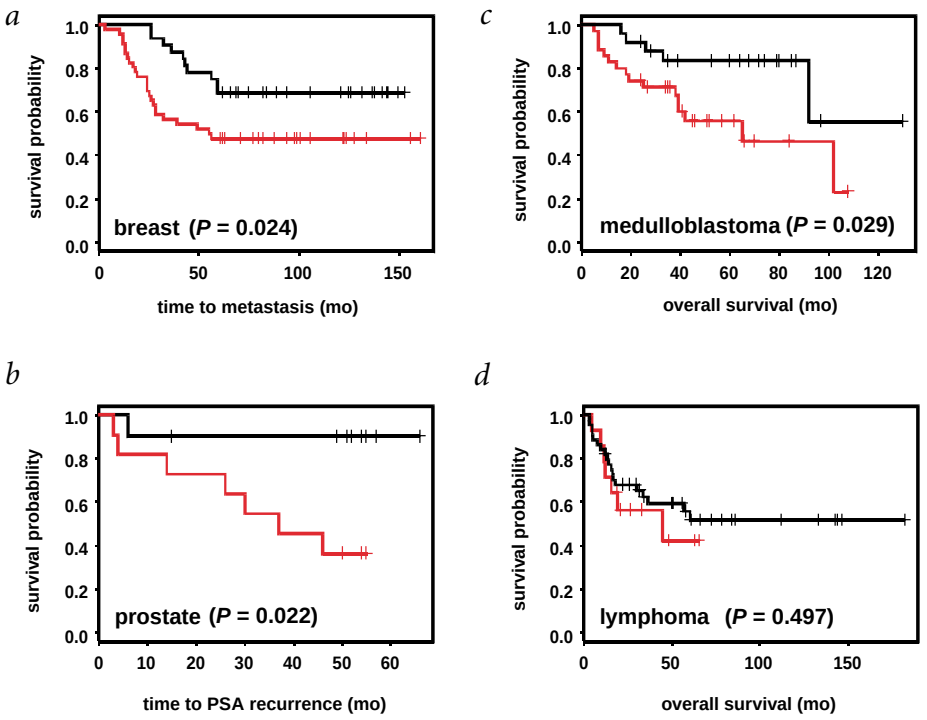

Fig. 3 Broad diagnostic utility of the signature associated with metastasis in solid tumors. Kaplan–Meier analyses of cluster-defined primary-tumor subsets using the 17-gene signature associated with metastasis in ***a***, breast adenocarcinoma (78 individuals); ***b***, prostate adenocarcinoma (21 individuals); ***c***, medulloblastoma (60 individuals); and ***d***, large B-cell lymphoma (58 individuals).

The data presented here support a model in which the propensity to metastasize reflects the predominant genetic state of a primary tumor rather than the emergence of rare cells with the metastatic phenotype. Such a model has recently been suggested on theoretical grounds, but firm genetic data supporting this view has been lacking[22]. The prevailing model predicts that the incidence of metastasis is related to the number of cells susceptible to metastasis-promoting mutations, and hence to tumor size. Micro-metastases have, however, been observed in many individuals with small, low-stage tumors[23]. Metastases have also been found in the absence of clinically detectable primary tumors[24]. It is thus possible that the ability to form distant metastases is a consequence of particular mechanisms of transformation as opposed to a selection process favoring the metastatic phenotype *per se*. Further study will be required to determine whether any signature genes are mechanistically important in the cascade of events driving metastasis. It should be noted that, although it yielded statistically significant results, our outcome predictor was imperfect, suggesting that additional correlates of tumor behavior have yet to be identified or that larger data sets will be required to generate more robust classifications.

The gene-expression signature associated with metastasis described here was probably identifiable only because the primary and metastatic tumors in this analysis were unmatched. The inclusion of tumors of different anatomical origin probably also enabled the definition of a tissue-independent signature associated with metastasis. Although this signature has the potential to be developed as a clinical diagnostic test, larger numbers of samples will be required to refine it and to determine if it is sufficiently robust for clinical implementation.

Our findings should be distinguished from recent reports that primary tumors and metastases from the same individual are more similar to each other than either is to tumors from other individuals[25]. These observations probably reflect the fact that metastases harbor many random genetic changes that first arose in the primary tumor from which they were derived. By contrast, our findings here concern comparisons across different tumor types and an expression signature that classifies a subset of primary solid tumors with a metastatic phenotype.

These findings support the emerging notion that the clinical outcome of individuals with cancer can be predicted using the gene-expression profiles of primary tumors at diagnosis[7,26]. Most previous studies have predicted response to therapy, however, and it is not clear whether such gene expression–based predictors reflect sensitivity to treatment or more fundamental

Table 1 • The 17-gene signature associated with metastasis

Gene	Gene name	GenBank ID
Upregulated in metastases		
SNRPF	Small nuclear ribonucleoprotein F	AI032612
EIF4EL3	Elongation initiation factor 4E-like 3	AF038957
HNRPAB	Heterogeneous nuclear ribonucleoprotein A/B	M65028
DHPS	Deoxyhypusine synthase	U79262
PTTG1	Securin	AA203476
COL1A1	Type 1 collagen, α1	Y15915
COL1A2	Type 1 collagen, α2	J03464
LMNB1	Lamin B1	L37747
Downregulated in metastases		
ACTG2	Actin, γ2	D00654
MYLK	Myosin light chain kinase	U48959
MYH11	Myosin, heavy chain 11	AF001548
CNN1	Calponin 1	D17408
HLA-DPB1	MHC Class II, DPβ1	M83664
RUNX1	Runt-related transcription factor 1	D43969
MT3	Metallothionein 3	S72043
NR4A1	Nuclear hormone receptor TR3	L13740
RBM5	RNA binding motif 5	AF091263

aspects of tumor cell biology. The fact that the gene-expression signature described herein is predictive of metastasis argues that the clinical behavior of solid tumors is governed at least in part by the intrinsic biological behavior of tumor cells rather than simply by differential chemo- or radiosensitivity. In addition, no previous study has provided evidence for a molecular signature that is biologically informative in multiple tumor types. The gene-expression signature described here may represent a composite of multiple, tissue-specific metastasis programs. The findings are also consistent with the existence of a molecular program of metastasis that is shared by multiple solid-tumor types, suggesting the possible existence of therapeutic targets common to different cancers.

Methods

Cross-platform gene mapping. Analyzing the prognostic value of genes associated with metastasis required defining common genes present on multiple distinct microarray platforms. We carried out the initial primary-tumor versus metastases comparison on Affymetrix Hu6800 and Hu35KsubA oligonucleotide microarrays (16,063 genes; ref. 27). Primary-tumor gene-expression outcome data sets were initially created using Affymetrix U95A (12,600 genes; lung[5] and prostate[8] adenocarcinoma), Affymetrix Hu6800 (6,817 genes; medulloblastoma[9] and large B-cell lymphoma[10]) and Rosetta inkjet (24,479 genes; breast adenocarcinoma[7]) oligonucleotide microarrays. We carried out cross-platform mapping of genes using UniGene build #147. We considered all genes that fell into the same UniGene clusters from both platforms to be 'mapped' genes (for example, there were 9,376 common mapped genes between the Hu6800/Hu35KsubA and U95A platforms; see Web Note A and Web Table A online).

Pre-processing of data. We re-scaled each data set to account for different microarray intensities in a given set. We multiplied each column (sample) in the data set by 1/slope of a least-squares linear fit of the sample versus a reference (the first sample in the data set). We did this linear fit using only genes that had 'Present' calls in both the sample being re-scaled and the reference. We chose a typical sample (that is, one with the closest number of 'Present' calls to the average over all samples in the data set) as reference. Pre-processing of the data consisted of a thresholding step and then a filtering step. We did thresholding using a ceiling of 16,000 units and a floor of 20 units. We then subjected gene-expression values to a variation filter that excluded genes with minimal variation across the samples being analyzed by testing 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. We used a max/min < 3 and max – min < 100 (for example, 8,176 of 9,376 genes passed this variation filter for the initial lung adenocarcinoma analysis).

Supervised prediction. We first compared 64 primary adenocarcinomas (breast, prostate, lung, colon, uterus and ovary) to 12 metastatic adenocarcinomas (from the same spectrum of sites but resected from a variety of end-organs) in the mapped and filtered gene space using the signal-to-noise (S_x) statistic (see Web Note A and Web Table A online). We defined primary tumors and metastases as classes 0 and 1, respectively. We identified the genes that best distinguished metastases from primary tumors using a signal-to-noise metric: $S_x = (\mu_{class0} - \mu_{class1} / \sigma_{class0} + \sigma_{class1})$ where, for each gene, μ_{class0} represents the mean value and σ_{class0} represents the standard deviation for that gene in all samples of class 0. We then applied a weighted-voting classification algorithm as previously described and tested it by 'leave-one-out' cross-validation[28]. Briefly, the weighted-voting algorithm makes a weighted linear combination of relevant 'marker' or 'informative' genes obtained in the training set to provide a classification scheme for new samples after marker-gene selection using the signal-to-noise statistic (S_x). In addition to computing S_x, the algorithm also finds the decision boundaries (halfway) 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 $sign (\Sigma_x V_x)$. We calculated the total number of prediction errors in cross-validation (for the primary-tumor versus metastases distinction) using graded numbers of genes, and found that a final 128-gene model yielded the minimal cross-validation error rate (see Web Note A and Web Table A online). We used these 128 genes for analysis.

Hierarchical clustering. We used the Cluster and TreeView software to carry out 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[29]. For pre-processing, we median-centered the genes and arrays twice (median-polished) and then normalized the genes. We used a weighted centered correlation for arrays to carry out clustering.

Selection of the 17-gene signature associated with metastasis. The 128 genes identified by supervised learning for the metastases versus primary-tumor distinction (on the Affymetrix Hu6800/Hu35KsubA oligonucleotide microarray set) had 169 analogs on the Affymetrix U95A oligonucleotide microarray (owing to probe-set redundancy). We used the signal-to-noise metric to determine the individual correlation for each of these 169 probe sets with the two primary lung tumor clusters defined through hierarchical clustering. We selected the top 21 probe sets with $S_x > 0.4$, which corresponded to 17 unique genes (see Web Note A and Web Table A online).

Permutation testing of the 17-gene signature associated with metastasis. We selected 1,000 random sets of 17 genes from the pool of 11,388 highly varying genes. We then used these gene sets to carry out 1,000 independent clusterings of the primary lung adenocarcinomas, and subjected each clustering to Kaplan–Meier survival analysis.

Statistical analysis. We created Kaplan–Meier survival curves using S-Plus. We used the Mantel–Haenszel log-rank test to calculate the statistical significance (P value) of differences between survival curves. We carried out two-tailed t-tests using S-Plus to determine the correlation between individual genes in the signature associated with metastasis and clinical outcome in each solid-tumor data set to determine whether any single gene was solely capable of yielding statistically significant clinical outcome differences.

URLs. Further details on data sets and analysis are available at http://www-genome.wi.mit.edu/cancer/solid_tumor_metastasis. Information about S-Plus is available at http://www.insightful.com.

Note: Supplementary information is available on the Nature Genetics website.

Acknowledgments

We thank R. A. Weinberg, P. Tamayo and M. A. Gillette for helpful comments. This work was supported in part by a grant from the US National Institutes of Health to T.R.G.

Competing interests statement

The authors declare that they have no competing financial interests.

Received 29 July; accepted 13 November 2002.

1. Hellman, S., DeVita, V.T. & Rosenberg, S.A. Cancer: principles & practice of oncology. (Lippincott-Raven, Philadelphia, 2001).
2. Poste, G. & Fidler, I.J. The pathogenesis of cancer metastasis. *Nature* **283**, 139–146 (1980).
3. Fidler, I.J. & Kripke, M.L. Metastasis results from pre-existing variant cells within a malignant tumor. *Science* **197**, 893–895 (1977).
4. Clark, E.A., Golub, T.R., Lander, E.S. & Hynes, R.O. Genomic analysis of metastasis reveals an essential role for RhoC. *Nature* **406**, 532–535 (2000).
5. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* **98**, 13790–13795 (2001).
6. Schiller, J.H. *et al.* Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. *N. Engl. J. Med.* **346**, 92–98 (2002).
7. van't Veer, L.J. *et al.* Gene-expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
8. Singh, D. *et al.* Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203–209 (2002).
9. Pomeroy, S.L. *et al.* Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436–442 (2002).
10. 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).
11. Anand, N. *et al.* Protein elongation factor *EEF1A2* is a putative oncogene in ovarian cancer. *Nat. Genet.* **31**, 301–305 (2002).

12. Zou, H., McGarry, T.J., Bernal, T. & Kirschner, M.W. Identification of a vertebrate sister-chromatid separation inhibitor involved in transformation and tumorigenesis. *Science* **285**, 418–422 (1999).
13. Jallepalli, P.V. *et al*. Securin is required for chromosomal stability in human cells. *Cell* **105**, 445–457 (2001).
14. Heaney, A.P. *et al*. Expression of pituitary-tumour transforming gene in colorectal tumours. *Lancet* **355**, 716–719 (2000).
15. Bernal, J.A. *et al*. Human securin interacts with p53 and modulates p53-mediated transcriptional activity and apoptosis. *Nat. Genet.* **32**, 306–311 (2002).
16. Skobe, M. & Fusenig, N.E. Tumorigenic conversion of immortal human keratinocytes through stromal cell activation. *Proc. Natl. Acad. Sci. USA* **95**, 1050–1055 (1998).
17. Olumi, A.F. *et al*. Carcinoma-associated fibroblasts direct tumor progression of initiated human prostatic epithelium. *Cancer Res.* **59**, 5002–5011 (1999).
18. Brown, L.F. *et al*. Vascular stroma formation in carcinoma *in situ*, invasive carcinoma, and metastatic carcinoma of the breast. *Clin. Cancer Res.* **5**, 1041–1056 (1999).
19. Jensen, B.V., Johansen, J.S., Skovsgaard, T., Brandt, J. & Teisner, B. Extracellular matrix building marked by the N-terminal propeptide of procollagen type I reflect aggressiveness of recurrent breast cancer. *Int. J. Cancer* **98**, 582–589 (2002).
20. Pardoll, D.M. Spinning molecular immunology into successful immunotherapy. *Nat. Rev. Immunol.* **2**, 227–238 (2002).
21. Song, W.J. *et al*. Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. *Nat. Genet.* **23**, 166–175 (1999).
22. Bernards, R. & Weinberg, R.A. Metastasis genes: a progression puzzle. *Nature* **418**, 823 (2002).
23. Braun, S. *et al*. Cytokeratin-positive cells in the bone marrow and survival of patients with stage I, II, or III breast cancer. *N. Engl. J. Med.* **342**, 525–533 (2000).
24. Hainsworth, J.D. & Greco, F.A. Treatment of patients with cancer of an unknown primary site. *N. Engl. J. Med.* **329**, 257–263 (1993).
25. Perou, C.M. *et al*. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
26. Alizadeh, A.A. *et al*. Distinct types of diffuse large B-cell lymphoma identified by gene-expression profiling. *Nature* **403**, 503–511 (2000).
27. Ramaswamy, S. *et al*. Multiclass cancer diagnosis using tumor gene-expression signatures. *Proc. Natl. Acad. Sci. USA* **98**, 15149–15154 (2001).
28. Golub, T.R. *et al*. Molecular classification of cancer: class discovery and class prediction by gene-expression monitoring. *Science* **286**, 531–537 (1999).
29. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).