Metagene projection for cross-platform, cross-species characterization of global transcriptional states

Pablo Tamayo*, Daniel Scanfeld*, Benjamin L. Ebert*, Michael A. Gillette*†, Charles W. M. Roberts‡, and Jill P. Mesirov*§

*Eli and Edythe L. Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02141; †Pulmonary and Critical Care Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114; and ‡Department of Pediatric Oncology, Dana–Farber Cancer Institute, Boston, MA 02115

Communicated by Edward M. Scolnick, The Broad Institute, Cambridge, MA, February 6, 2007 (received for review December 7, 2006)

The high dimensionality of global transcription profiles, the expression level of 20,000 genes in a much small number of samples, presents challenges that affect the sensitivity and general applicability of analysis results. In principle, it would be better to describe the data in terms of a small number of metagenes, positive linear combinations of genes, which could reduce noise while still capturing the invariant biological features of the data. Here, we describe how to accomplish such a reduction in dimension by a metagene projection methodology, which can greatly reduce the number of features used to characterize microarray data. We show, in applications to the analysis of leukemia and lung cancer data sets, how this approach can help assess and interpret similarities and differences between independent data sets, enable cross-platform and cross-species analysis, improve clustering and class prediction, and provide a computational means to detect and remove sample contamination.

cancer | dimension reduction | expression analysis | noise reduction | sample contamination

A major challenge in the analysis of global transcription profiles is the high level of noise and the lack of reproducibility across data sets, which results from fitting models to small numbers of samples in a high-dimensional space (i.e., thousands of genes). Ideally we would prefer to reduce the data to a small number of metagenes that better capture the essential behavior of the samples.

There are many advantages to such a metagene approach. By capturing the major, invariant biological features and reducing noise, metagenes provide descriptions of data sets that allow them to be more easily combined and compared. This is especially important when we are considering cross-platform or cross-species data. Ultimately, this can result in more sensitive clustering and classification. In addition, interpretation of the metagenes, which characterize a subtype or subset of samples, can give us insight into underlying mechanisms and processes of a disease.

Here, we describe a general methodology, metagene projection, that creates a low-dimensional representation of a training (model) data set using nonnegative metagene factors into which an independently obtained new (test) set of samples or data can be projected and analyzed. The metagene factors are a small number of gene combinations that distinguish expression patterns of subclasses in a data set. We obtain the factors by the application of nonnegative matrix factorization (NMF) (1, 2) used to extract facial features from images. We showed (3) how NMF can extract metagenes that provide stable, robust clustering of expression data. Moreover, by using gene set enrichment analysis (GSEA) to annotate the metagene factors themselves, we can gain insight into the underlying biology of both the training and test data sets.

Importantly, we illustrate the utility of metagene projection by its application to leukemia and lung cancer data sets. We show how the projection of new data sets into the space of metagene factors reduces noise and emphasizes relevant biological correlations and thus (*i*) enables cross-platform analysis by removing technological noise from data, (*ii*) enables cross-species analysis and the assessment of disease models, (*iii*) improves the accuracy of classification and prediction methods in the mapping of diseases types, and (*iv*) detects contamination in tumor samples.

Results

Overview of Method. We consider a gene expression data set consisting of a collection of N_M model samples, which we use to characterize a domain of biological (transcriptional) states of interest. The model data are represented as an $n_M \times N_M$ matrix, M, whose rows contain the expression levels of the n_M genes in the N_M samples.

Using NMF, we find a small number, k, of metagenes, positive linear combinations of the N_M genes, which can be used to distinguish the transcription profiles of the subtypes contained in the model data set. Mathematically, this corresponds to finding an approximate factoring, $M \approx W_M \times H_M$, where both factors have only positive entries. W_M is an $n_M \times k$ matrix that defines the metagene decomposition model and whose columns specify how much each of the n_M genes contributes to each of the k metagenes. H_M is a $k \times N_M$ matrix whose entries represent the expression levels of the k metagenes for each of the N_M samples. This model selection is done in an unsupervised fashion by using either a knowledge-based or data-driven model selection approach. One can set k equal to the number of known phenotypes in the model set. Alternatively, optimal values of k can be determined based on projection stability by using consensus clustering techniques as described (3).

From the factoring of M, we are able to construct a mapping that allows us to project a data set into the space of the metagenes derived above. Mathematically, this can be accomplished by using the Moore–Penrose generalized pseudoinverse (4) of W_M, so that, $\hat{H}_M = (W_M)^{-1} \times M$, where $\hat{H}_M \approx H_M$. For simplicity in notation we refer to the projected matrix as H_M. After elimination of outlier samples and model refinement, we can apply the final resulting pseudoinverse to a new individual sample or entire data set and analyze that data in the context of the metagenes, which characterized the model data.

We summarize the three major steps in the metagene projection method below (Fig. 1). More detail can be found in *Methods*. The software is freely available from The Broad

Author contributions: P.T. and J.P.M. designed research; P.T., D.S., B.L.E., C.W.M.R., and J.P.M. performed research; M.A.G. contributed data; P.T., D.S., and J.P.M. analyzed data; and P.T., B.L.E., and J.P.M. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Abbreviations: GSEA, gene set enrichment analysis; NMF, nonnegative matrix factorization; SVM, support vector machine.

§To whom correspondence should be addressed. E-mail: mesirov@broad.mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0701068104/DC1.

GENETICS

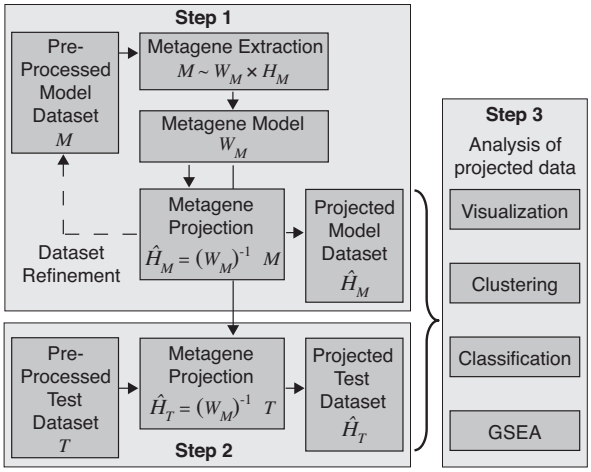

Fig. 1. Schematic of the metagene projection methodology.

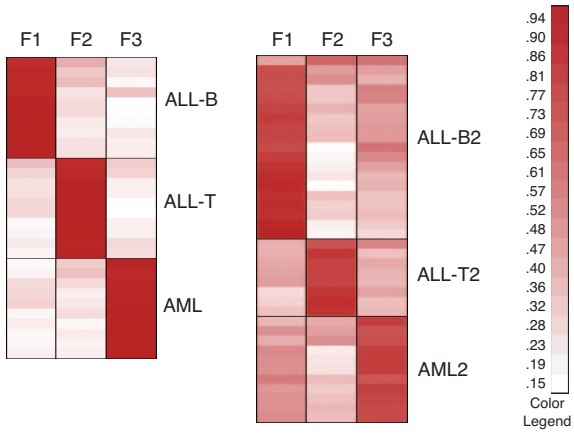

Fig. 2. Heat maps of metagene projection of leukemia samples. These heat maps of the H_M and H_T matrices show the metagene expression levels for each sample. Each factor clearly corresponds to same leukemia subtype in both model (*Left*) and test (*Right*) sets.

Institute web site as both R-code and a module in the GenePattern software package.

Step 1. Metagene Factor Extraction and Refinement of the Model Data Set. We start with standard data preprocessing: thresholding and eliminating genes that do not vary sufficiently across the model set and rank normalizing to minimize platform idiosyncrasies. We apply NMF to factor the resulting expression matrix and derive the Moore–Penrose pseudoinverse of W_{M}. Next, we project the model data set into metagene space and, by using a support vector machine (SVM) (5) classification step, trim outliers from the model set (model data set refinement). Finally, we refactor the expression matrix M of the refined model set, $M \approx W_{\mathrm{M}} \times H_{\mathrm{M}}$, and define a refined pseudoinverse or projection map. We use this refinement of the projection map in the analysis of new test data sets.

Step 2. Metagene Factor Projection of the Test Data Set. We threshold the expression values as in step 1 and then match the genes in each test set to the corresponding genes in the model set. We then rank normalize the test samples to yield the corresponding columns in the test data expression matrix, T. Finally, we apply the pseudoinverse $(W_{\mathrm{M}})^{-1}$ to both M and T to obtain H_{M} and H_T, their projections into metagene space.

Step 3. Analysis of Model and Test Data Set Projection Results. In our experience, the use of metagenes, instead of genes, as features for analysis, increases the signal-to-noise ratio and yields more robust, accurate results. Now that both the model and test data are represented in the lower dimensional metagene space, there are a variety of analyses we can apply. These include the following:

Visualization. Model and test samples can be characterized and compared by using heat maps of the H matrices.

Clustering model and test projections. The projection can provide a sample's class assignment by identifying the metagene with maximum expression. Alternatively, we can cluster the columns of H_{M} and H_T.

Classification of test samples. We can use the projected data to build a multiclass predictor and assess any data set of test samples. Below, we use a one-versus-all SVM classifier (6, 7) to predict phenotypes by using the k metagenes as the input features. This method provides a predicted class and a predictive confidence by using a modified Brier score (see *Methods* for details).

GSEA-based metagene interpretation. To gain biological insight into the different metagene factors, we use a variation of our GSEA methodology (8). Using the expression profile of a metagene, i.e., the corresponding row of the H_{M} matrix, as a template, we sort the genes according to the correlation of their expression profile from the M matrix with the metagene template. We can then evaluate the "enrichment" of gene sets representing a pathway or other biological process at the top of that ranked list by using GSEA. For each metagene, one obtains a list of "enriched" gene sets and their statistical significance [see supporting information (SI) *Text*].

Examples. Here, we describe three applications of the metagene projection method to highlight its utility in three cross-platform analyses, to validate disease models, to improve classification of cross-platform data sets, to assess the similarities and differences of subtypes across data sets, and to detect contamination. We start with a simple example. We then describe two more innovative results.

Cross-Platform Clustering of Leukemia Data. We analyzed two leukemia data sets from different microarray platforms to test the method and demonstrate its power to enable cross-platform classification and to improve sensitivity in clustering. Often clustering of cross-platform data reveals the platform or originating lab as the strongest differentiating signal in the data. Importantly, we establish that the method was able to cluster the cross-platform data correctly and that these results are because of the metagene representation rather than the rank normalization step.

We considered two data sets containing samples representing three leukemia subclasses: B and T cell acute lymphoblastic leukemia (ALL-B, ALL-T) and acute myeloid leukemia (AML). The model data set consisted of 30 samples (10 ALL-B, 10 ALL-T, 10 AML) (from refs. 9 and 10). The test data set contained the 38 samples (19 ALL-B, 8 ALL-T, 11 AML) from ref. 11. The two data sets came from different laboratories and were acquired on different microarray technologies, Affymetrix U-133 for the model set and Affymetrix HU6800 for the test set (Affymetrix, Santa Clara, CA).

We applied the metagene projection methodology as described above. In particular, we noted that the model data set is very consistent, and no model refinement was necessary. Because the number of subtypes was known, we used $k = 3$ metagene factors. Fig. 2 shows the resulting heat maps for the projected model and test sets. Clearly, the metagenes are associated with the biological phenotypes (F1 $\approx$ ALL-B, F2 $\approx$ ALL-T, F3 $\approx$ AML) in both.

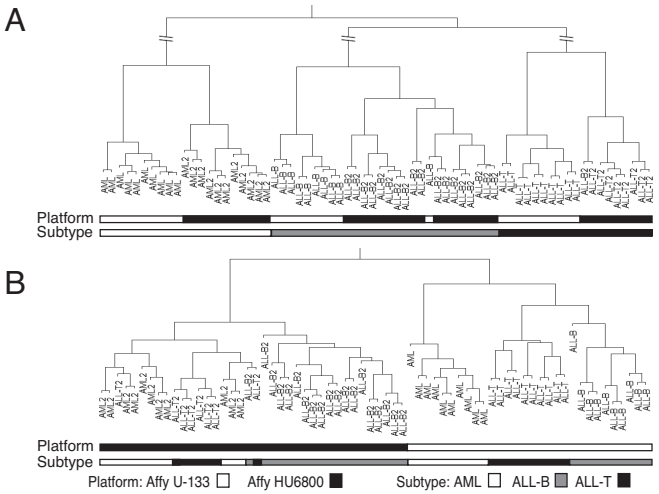

Fig. 3. Hierarchical clustering of the leukemia model and test samples. (*A*) Clustering of the merged test and model data sets after metagene projection, i.e., columns of the merged H_M and H_T matrices. (*B*) Clustering of merged model and test sets normalized but without projection. For clarity, some dendrogram vertical lines have been truncated in *A*; for full dendrograms see SI Fig. 7.

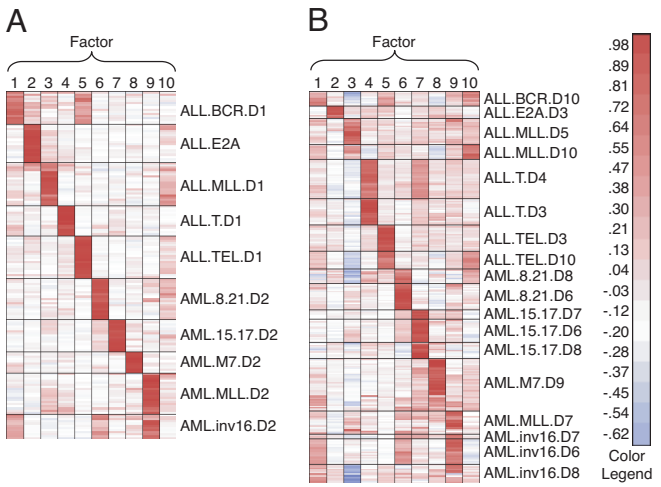

Fig. 4. Leukemia subclasses metagene projection. Heat maps of the model (*A*) and test (*B*) sets after metagene projection show consistent representation of subtype structure across technology platform and laboratory group. *SI Text* contains a detailed description of the different leukemia subtypes shown here.

Postprojection clustering of the model samples demonstrates reduction of noise and greater emphasis of the biologically invariant signal in the data. The clusters corresponding to each phenotype have higher intracluster correlation and greater intercluster distance than obtained with the original data (SI Fig. 6). More importantly, clustering of the merged set of projected model and test samples produces very clear results with the major three clusters consisting of each leukemia subtype independent of the data set of origin (Fig. 3*A* and SI Fig. 7*A*).

We next sought to confirm that this consistency of subtype clusters across the data sets was due to the metagene projection and not just the result of preprocessing and rank normalization. To this end, we performed two additional clusterings: one merging the model and test samples after rank normalization and clustering in the space of all filtered genes without using metagene projection (Fig. 3*B* and SI Fig. 7*B*) and another clustering the merged and rank normalized data in the space of the top-500 marker genes of each of the three subtypes in the model set, 1,500 genes in total (SI Fig. 8). This last procedure is often used for cross-platform analysis. In both alternative clusterings, not using metagene projection, the samples first split according to their data set of origin before the biological subclassification appears.

Leukemias: Improving Cross-Platform Classification and Interpretation of Subtypes. We sought to ascertain whether metagene projection would be an effective procedure for unsupervised feature extraction (12) and dimension reduction to enable more robust and accurate classifiers. To this end, we considered 10 subclasses of leukemia (5 subtypes of ALL and 5 subtypes of AML) as represented in a model set of 170 samples from refs. 9 and 10. The test set consisted of 297 samples (13–20), obtained from eight independent published data sets. The model set samples were all acquired on the same platform in the same laboratory, whereas the test set came from multiple labs and three different microarray platforms (see SI Table 1).

We set the number of metagene factors to the number of known phenotypes in the model set, $k = 10$. Metagene projection, followed by model refinement, resulted in elimination of eight outlier samples from the model set [2 of 21 AML t (8, 21); 4 of 23 AML MLL; 2 of 14 AML inv (16)] (for more detail see *Methods*). Fig. 4 shows the metagene expression matrices for both the model and test data sets after projection. Strikingly, we found that each leukemia subtype was characterized by essentially one metagene.

Next, we sought to determine whether we could build a classifier using the metagene projections that would accurately predict the subtype of the cross-platform samples in the test set. We noted that the data-driven model selection technique described in our previous work (3) indicated that $k = 13$ was the best choice (SI Fig 9). Thus, we evaluated SVM classifiers using both the 10- and 13-metagene models and compared them with SVM and K-nearest neighbor (K-NN) classifiers using all genes in common between the model and test data sets. SI Fig. 10 shows the comparative performance of the 10- and 13-metagene SVMs with the all-gene classifiers.

Our metagene-based classifier outperformed the classifiers based on all-genes or markers selected in all-gene space. The 13-metagene classifier attained the "best" performance, with a correct call accuracy of 88% and fewer errors than the 10-metagene model. The 10-metagene, all-gene SVM, and K-NN classifiers' correct call accuracies were 86%, 82%, and 72% respectively. We note that the SVM classifier using all common genes made fewer "confident" calls but made correspondingly fewer errors. We used 0.3 as the confidence threshold for all of the SVM multiclass predictors. Increasing this threshold will reduce both the number of correct calls and the number of errors. (SI Tables 2 and 3 contain details).

Closer examination of the confusion matrices for the 10- and 13-metagene classifiers revealed that two thirds of the errors resulted from placing ALL-BCR, AML-t (8, 21), AML-M7, and AML-MLL samples into the AML-inv16 class. We believe this results from shared metagene signals, which can be seen in the heat map in Fig. 4*B*. A GSEA analysis of the metagene factors, described below, uncovered a biological interpretation for some of the errors. This also led us to explore the extent to which cross talk between the AML and ALL data in the model might be affecting our ability to predict the classes in the test set. Interestingly, we found that building 10-metagene, five-class classifiers for just the ALL [respectively AML] subtypes improved accuracy substantially to 97% (130 samples) with 1.5% no calls (2 samples) and 1.5% errors (2 samples) [92% (150 samples) with 3% no calls (4 samples) and 5% errors (9 samples)]. The all-gene SVM and 9-NN predictors also improved

accuracy, but the metagene-based classifier continues to make more correct calls and fewer no-calls (SI Fig. 10).

These are remarkably good multiclass, cross-platform classification results. It was difficult to make direct comparisons with other approaches in the literature, because the specific data sets or data preparation were not always available. However, the metagene-based approach appears to outperform other leukemia cross-platform classification approaches: 93–96% accuracy on ALL subtypes and 68–78% on AML subtypes (21); ≈40% accuracy on AML subtypes (22).

Finally, we applied GSEA analysis to help interpret the metagenes characterizing the leukemia subtypes. Interestingly, many of the results agreed with the current understanding of these subclasses, and others posed new hypotheses. We present them as an illustration of the power of the metagene projection method to provide biological insights. The top two gene sets enriched in F4 (i.e., high in ALL T Cell) are (*i*) a set of E2F1 targets known to be activated in T Cell ALL (23) and (*ii*) a set of genes down-regulated by ET-743 treatment, which is known to induce apoptosis in acute T cell leukemia Jurkat cells (24). Metagene F9, high in AML-MLL, shows enrichment for chromosome band 11q13, which is known to be frequently coamplified with MLL in AML patients (25).

F6 is highly expressed in t (8, 21) and also up-regulated in inv (16) subclasses of AML. The mechanism of leukemogenesis in AML in both these subtypes is disruption of the core binding factor (CBF) transcriptional complex, comprised of the RUNX1 and CBFB proteins. In t (8, 21), *RUNX1* is fused to the *CBFA2T1* gene, and inv (16) causes a *CBFB-MYH11* fusion gene. Both fusion genes disrupt the CBF complex, which is required for normal hematopoietic differentiation. Patients with t (8, 21) and inv (16) also have similar clinical features: both subclasses are associated with a relatively good prognosis and particular benefit from consolidation chemotherapy with high-dose cytarabine. F6 therefore identifies patients harboring distinct cytogenetic abnormalities with a common molecular mechanism and clinical phenotype. Intriguingly, F9 also shows strong correlation with both AML.MLL and AML-inv (16). This leads us to speculate some common program of these two AML subtypes.

In this example, we have shown that metagene projection is an effective approach to building multiclass classification models across different platforms and sources of data that are accurate, robust, and interpretable.

Lung Cancer: Cross-Platform Comparison, Contamination Detection, and Interpretation of Cell Line Models. We next investigated whether metagene projection would enable us to evaluate consistency in a collection of cross-platform data sets, validate cell lines as good models for different tumor types, and, importantly, provide a method to computationally extract some of the expression signal of normal tissue contamination from tumor samples.

For our model set, we used a subset of data set A from ref. 26, BOS, consisting of 30 lung adenocarcinomas, 20 squamous tumors, and 17 normal lung samples. Our test set derived from seven independent data sets (refs. 27–32 and one unpublished set, see SI Table 4). Note that these data sets were acquired on four different microarray platforms by six different laboratories.

We first built a four-metagene model from the BOS model set as described above. Although the model set included three major subtypes, the data-driven NMF model selection procedure indicated that four factors was the smallest optimal solution greater than the number of known phenotypes (SI Fig. 11). After SVM model refinement, one outlier adenocarcinoma sample was removed from the model set, and the metagene factors recalculated. Fig. 5*A* contains the H_M matrix of metagene expression levels. From the H_M matrix, we can see that metagenes F2, F3, and F4 characterize the adenocarcinoma, squamous, and normal samples respectively, whereas the F1 metagene picks up an additional signature in a subset of the adenocarcinoma and squamous samples. Next, we projected all of the test data sets into metagene space (H_T in Fig. 5*A*) and found an unexpected result. The normal test samples NL-STA continued to be characterized by F4. However, although the adenocarcinoma and squamous samples still showed F2 and F3 metagene signatures, respectively, they also showed significant expression in the F4 "normal" metagene. This led us to speculate that these samples might have varying degrees of contamination by stroma or normal tissue, which we might be able to extract computationally.

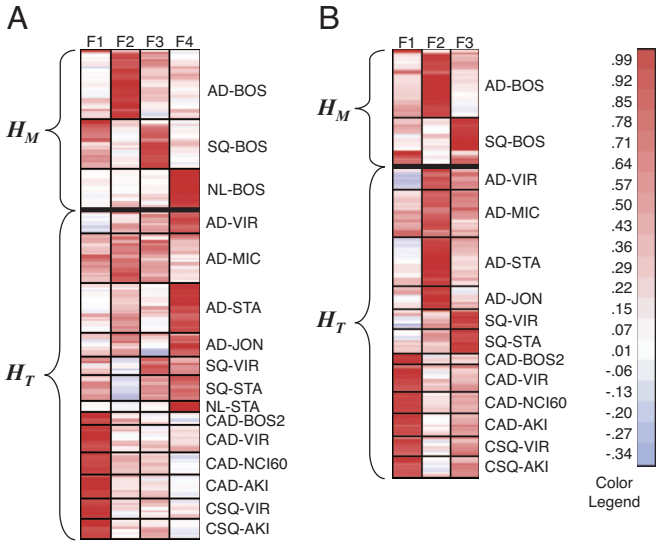

Fig. 5. Metagene projection of the lung cancer data set. Heat maps showing projection of model and test data sets into four-metagene space F1-F4 (*A*) and three-metagene space F1-F3 after numerical removal of normal component F4 and reconstruction of model (*B*). AD, adenocarcinoma; SQ, squamous; C, cell lines; NL, normal lung.

To remove the normal signature, we set the F4 metagene factor coefficient in the H_M matrix to zero and multiplied it by the original W_M to yield a matrix $\tilde{M}$ that reproduces the original data but without the contribution of F4. We then excluded the normal tissue samples from the model data set because they only had residual values, factored the resulting data matrix to extract the three remaining metagene factors, and projected all of the samples as was done before. The resulting expression profiles of the metagenes in the model and test sets are seen in Fig. 5*B*. Eliminating the contribution of the F4 metagene, we find the dominant signatures in the adenocarcinoma and squamous samples are F2 and F3, respectively, as in the model set, and F1 retains its role as the signature of the cell lines. Thus, we were able to numerically "modulate" a specific metagene to computationally reduce contamination in the tumor samples.

The most striking feature of the metagene projection of the test samples is that the adenocarcinoma and squamous cell lines do not project with the corresponding tumor classes. This has been reported in the literature (27). Using the GSEA approach we described above, we can gain some biological insight into the metagene, F1, which characterizes the cell lines. SI Table 5 shows the top-20 gene sets enriched in F1.

Metagene F1 is enriched in gene sets associated with rapamycin response (mTOR activation), protein production (genes down-regulated by amino acid starvation), lack of differentiation, the mitochondria, oxidative phosphorylation, and BRCA1 signaling. We have observed some of these gene sets before as part of a group of gene sets enriched in poor-outcome lung adenocarcinoma patients in three different data sets (8). This

leads us to speculate that F1 represents transcriptional programs associated with hyperactivation of AKT/mTOR, an associated mTOR-mediated increase of protein production and high proliferation, and a lack of differentiation.

In this example, we have shown the power of the metagene projection to define a common space of transcriptional variation in which we can analyze and assess multiple data sets across different technology platforms and laboratories. Despite the diversity of platforms, sample sources, and different experimental conditions, most test samples project with their biological counterparts. Moreover, we have shown that metagene projection provides a method for computationally reducing sample contamination, which enables more coherent projection of tumor samples. Finally, the combination of metagene projection and GSEA analysis allows us to gain insights into more robust, invariant biological features of different phenotypes and tumor subtypes.

Discussion

Traditional approaches to microarray analysis focus on identifying marker genes, which are correlated with a phenotype of interest, and on using them to build classifiers for samples whose phenotype may be unknown or to gain some insight into the underlying biology of a cellular state. These strategies often fail when classifiers are applied to data from other laboratories or derived on different technology platforms or when used to try to assess the validity of a disease model.

Lower-dimensional projections and decompositions of DNA microarray data, such as principal component analysis, singular value decomposition, and NMF, have been used to analyze transcriptional states (3, 33–37). Primarily, these approaches were applied in the context of a single data set for clustering or visualization.

We introduced a metagene projection method to assess the validity of a Snf5 knockout mouse as a murine model for Snf5-deficient human rhabdoid tumors (38), and found that the murine Snf5 model samples were closely related to the human rhabdoid samples (from both model and test sets) and distinct from the controls. The model and test sets were obtained on different microarray platforms in addition to being cross-species. This approach combined our previous work, using NMF to identify a small number of gene combinations (metagenes) whose profiles best represent the most distinguishing features of the expression patterns of the subclasses in a data set, with our previously published gene expression data set derived from a collection of human pediatric brain tumors (rhabdoid, medulloblastoma, glioma, and normal cerebellum) (33). A corresponding projection map, the Moore-Penrose generalized pseudoinverse of one of the factor matrices, allowed us to analyze new data in the context of the space of metagenes arising from the original data set.

This article presents a refinement of that method, which is more sensitive, robust, and broadly applicable to cross-platform and cross-species analysis and classification (see *SI Text*). In addition, we have shown how the projection can be used to highlight the biologically invariant aspects and commonalities of the subclasses, assess the similarities and differences between suitable chosen sets of model and test samples, and, surprisingly, to computationally remove contaminating signals from tumor data.

The method, as presented here, has a number of advantages over other approaches. Metagene projection, together with NMF, reduces dimensionality and summarizes the salient features of a data set with coherent patterns shared by multiple genes and samples. In contrast to approaches using principal component analysis or singular value decomposition, it yields a sparser representation of the original model data set optimized for the number of factors specified. NMF factors are nonnegative and more localized and therefore easier to interpret and analyze. We note here that Alter and Golub (39) applied the pseudoinverse to genomic data by using the singular value decomposition.

There is complementary work of Huang (40) and Bild (41), which is conceptually similar to ours in the sense of combining dimensionality reduction and classification models, but has distinct objectives. Their main goal is to provide an exquisitely specific predictor of pathway activation, which has been experimentally characterized by the overexpression of a single gene. In contrast, our goal is to model global transcriptional states, rather than specific pathways, and to use them to describe an entire range of biological behavior, e.g., different morphologies, lineages, etc. Thus, the specific methodologies and techniques we use are also quite different.

Classifiers built in metagene, rather than all-gene, space are more robust, reproducible, and generalizable across platforms and laboratories because the projection can reduce noise and technology-based variation more than simple normalization. In particular, we found this approach to be very sensitive in the complex, cross-platform, multiclass setting of the leukemia data sets. Others have studied cross-platform classification in lung cancer (42, 43). However, they use the test data explicitly to choose similarly correlated genes as features, rather than relying solely on the model set.

Most importantly, metagene models built on previously acquired or published data sets enable the use of prior knowledge to help characterize and analyze new data. This is seen in our work validating a mouse model for human rhabdoid tumors (38). We also used this approach to analyze samples from malaria-infected patients using signatures derived from publicly available yeast data (P.T., D.S., J.P.M., unpublished work). Thus, we see that this metagene projection method not only decreases noise by reducing the dimensionality of microarray data, but can also provide a powerful knowledge-based approach to the cross-platform, cross-species analysis of microarray data.

Methods

Data Set Preprocessing and Normalization. For Affy Hu6800 and U133 microarrays, we threshold at 20 and 100,000 units. Gene filtering excludes genes with <5-fold and 500 units of maximum difference for the first leukemia example, 8-fold/800 for the second leukemia example, and 3-fold/300 for the lung. We rank the genes according to their expression levels and replace the value by $10{,}000 \times (\text{rank(gene)} - 1)/(\text{number of genes} - 1)$.

Metagene Factor Extraction. We use NMF with 2,000 iterations and stopping criterion as described (3).

Metagene Model Selection. We select k based either on the known number of phenotypes or by using the values determined by projection stability described (3). Optimal solutions are peaks in the cophenetic coefficient as a function of k.

Data Set Refinement. We train a SVM on H_M to predict each class, and we remove samples that are errors (known phenotypes) or no calls (discovered classes). In our experience, the number of outliers is quite small compared with the size of the classes if the number of metagenes is chosen as described above.

Calculating the Pseudoinverse of W_M. We use "ginv" from R's MASS package.

Metagene Projection of Model and Test Set Samples. To project the model set, we use the pseudoinverse of W_M. For each data set in the test set, we match the genes to the corresponding rows of W_M (i.e., genes in the model set). We calculate the pseudoinverse for that set of rows and apply it to obtain the corresponding columns of H_T for that specific data set. This procedure adapts the projection to the particular test data set and, by tolerating unmatched genes between model and test set, supports the projection of data sets from

different platforms. If too many unmatched genes result in weak amplitudes in H_T, we rescale the columns of H_T so the sum of the squares of their row-entries is equal to one. This postnormalization is optional.

Clustering. We use "hclust" (complete linkage) from R's STATS package.

Classification and Prediction Confidence. We use the "svm" function from R's e1071 package (one vs. all, radial function kernel). The predicted class is the one with the highest probability, and a predictive confidence $1 \geq C_p \geq 0$ is computed by using a modification of the Brier skill score (44):

$$C_P = 1 - \frac{\left((1 - P_1)^2 + \sum_{i=2}^{k} P_i^2\right)}{(1 - 1/k)^2 + (k - 1)(1/k)^2}, \qquad [1]$$

where $P_1 > P_2 > \ldots > P_k$ is the sorted list of k output probabilities for a given sample. $C_p < 0.3$ is a no call. The K-NN classifier in the leukemia example used 50 marker genes and nine nearest neighbors. For the SVM using all genes we use a "linear" kernel.

We thank J. P. Brunet, T. Golub, E. Lander, and M. Meyerson for helpful conversations and for reviewing this manuscript.

1. Lee DD, Seung HS (1999) *Nature* 401:788–791.
2. Lee DD, Seung HS (2001) *Adv Neural Info Proc Syst* 13:556–562.
3. Brunet JP, Tamayo P, Golub TR, Mesirov JP (2004) *Proc Natl Acad Sci USA* 101:4164–4169.
4. Ben-Israel A, Greville TNE (2003) *Generalized Inverses: Theory and Applications* (Springer, New York).
5. Cristianini N, Shawe-Taylor J (2000) *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* (Cambridge Univ Press, New York).
6. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, *et al.* (2001) *Proc Natl Acad Sci USA* 98:15149–15154.
7. Rifkin R, Mukherjee S, Tamayo P, Ramaswamy S, Yeang CH, Angelo M, Reich M, Poggio T, Lander ES, Golub TR, Mesirov J (2003) *SIAM Rev* 45:706–723.
8. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) *Proc Natl Acad Sci USA* 102:15545–15550.
9. Ross ME, Zhou X, Song G, Shurtleff SA, Girtman K, Williams WK, Liu HC, Mahfouz R, Raimondi SC, Lenny N, *et al.* (2003) *Blood* 102:2951–2959.
10. Ross ME, Mahfouz R, Onciu M, Liu HC, Zhou X, Song G, Shurtleff SA, Pounds S, Cheng C, Ma J, *et al.* (2004) *Blood* 104:3679–3687.
11. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, *et al.* (1999) *Science* 286:531–537.
12. Guyon IM, Gunn SR, Nikravesh M, Zadeh L (2006) *Feature Extraction, Foundations and Applications* (Physica, Springer, Heidelberg).
13. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, *et al.* (2002) *Cancer Cell* 1:133–143.
14. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ (2002) *Nat Genet* 30:41–47.
15. Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J, Foa R (2004) *Blood* 103:2771–2778.
16. Bullinger L, Dohner K, Bair E, Frohling S, Schlenk RF, Tibshirani R, Dohner H, Pollack JR (2004) *N Engl J Med* 350:1605–1616.
17. Valk PJ, Verhaak RG, Beijen MA, Erpelinck CA, Barjesteh van Waalwijk van Doorn-Khosrovani S, Boer JM, Beverloo HB, Moorhouse MJ, van der Spek PJ, Lowenberg B, Delwel R (2004) *N Engl J Med* 350:1617–1628.
18. Gutierrez NC, Lopez-Perez R, Hernandez JM, Isidro I, Gonzalez B, Delgado M, Ferminan E, Garcia JL, Vazquez L, Gonzalez M, San Miguel JF (2005) *Leukemia* 19:402–409.
19. Bourquin JP, Subramanian A, Langebrake C, Reinhardt D, Bernard O, Ballerini P, Baruchel A, Cave H, Dastugue N, Hasle H, *et al.* (2006) *Proc Natl Acad Sci USA* 103:3339–3344.
20. Fine BM, Stanulla M, Schrappe M, Ho M, Viehmann S, Harbott J, Boxer LM (2004) *Blood* 103, 1043–9.
21. Nilsson B, Andersson A, Johansson M, Fioretos T (2006) *Haematologica* 91, 821–4.
22. Warnat P, Eils R, Brors B (2005) *BMC Bioinformatics* 6:265.
23. Lemasson I, Thebault S, Sardet C, Devaux C, Mesnard JM (1998) *J Biol Chem* 273, 23598–604.
24. Gajate C, An F, Mollinedo F (2002) *J Biol Chem* 277, 41580–9.
25. Zatkova A, Ullmann R, Rouillard JM, Lamb BJ, Kuick R, Hanash SM, Schnittger S, Schoch C, Fonatsch C, Wimmer K (2004) *Genes Chromosomes Cancer* 39, 263–76.
26. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, *et al.* (2001) *Proc Natl Acad Sci USA* 98, 13790–5.
27. Virtanen C, Ishikawa Y, Honjoh D, Kimura M, Shimane M, Miyoshi T, Nomura H, Jones MH (2002) *Proc Natl Acad Sci USA* 99, 12357–62.
28. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN, *et al.* (2001) *Proc Natl Acad Sci USA* 98, 10787–92.
29. Gemma A, Li C, Sugiyama Y, Matsuda K, Seike Y, Kosaihira S, Minegishi Y, Noro R, Nara M, Seike M, *et al.* (2006) *BMC Cancer* 6:174.
30. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, *et al.* (2002) *Nat Med* 8:816–824.
31. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, *et al.* (2001) *Proc Natl Acad Sci USA* 98:13784–13789.
32. Jones MH, Virtanen C, Honjoh D, Miyoshi T, Satoh Y, Okumura S, Nakagawa K, Nomura H, Ishikawa Y (2004) *Lancet* 363:775–781.
33. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, *et al.* (2002) *Nature* 415:436–442.
34. Kim PM, Tidor B (2003) *Genome Res* 13:1706–1718.
35. Alter O, Brown PO, Botstein D (2000) *Proc Natl Acad Sci USA* 97:10101–10106.
36. Moloshok TD, Klevecz RR, Grant JD, Manion FJ, Speier WFT, Ochs MF (2002) *Bioinformatics* 18:566–575.
37. Dueck D, Morris QD, Frey BJ (2005) *Bioinformatics* 21 (Suppl 1):i144–i151.
38. Isakoff MS, Sansam CG, Tamayo P, Subramanian A, Evans JA, Fillmore CM, Wang X, Biegel JA, Pomeroy SL, Mesirov JP, Roberts CW (2005) *Proc Natl Acad Sci USA* 102:17745–17750.
39. Alter O, Golub GH (2006) *Proc Natl Acad Sci USA* 103:11828–11833.
40. Huang E, Ishida S, Pittman J, Dressman H, Bild A, Kloos M, D'Amico M, Pestell RG, West M, Nevins JR (2003) *Nat Genet* 34:226–230.
41. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, *et al.* (2006) *Nature* 439:353–357.
42. Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E (2004) *Clin Cancer Res* 10:2922–2927.
43. Hayes DN, Monti S, Parmigiani G, Gilks CB, Naoki K, Bhattacharjee A, Socinski MA, Perou C, Meyerson M (2006) *J Clin Oncol* 24:5079–5090.
44. Brier GW (1950) *Monthly Weather Rev* 78:1–3.