Automated multidimensional phenotypic profiling using large public microarray repositories

Min Xu1, Wenyuan Li1, Gareth M. Jamesb, Michael R. Mehan1, and Xianghong Jasmine Zhou1,2

1Molecular and Computational Biology, Department of Biological Sciences, and 2Marshall School of Business, University of Southern California, Los Angeles, CA 90089

Edited by Michael S. Waterman, University of Southern California, Los Angeles, CA, and approved June 1, 2009 (received for review January 26, 2009)

Phenotypes are complex, and difficult to quantify in a high-throughput fashion. The lack of comprehensive phenotype data can prevent or distort genotype–phenotype mapping. Here, we describe “PhenoProfiler,” a computational method that enables in silico phenotype profiling. Drawing on the principle that similar gene expression patterns are likely to be associated with similar phenotype patterns, PhenoProfiler supplements the missing quantitative phenotype information for a given microarray dataset based on other well-characterized microarray datasets. We applied our method to 587 human microarray datasets covering >14,000 samples, and confirmed that the predicted phenotype profiles are highly consistent with true phenotype descriptions. PhenoProfiler offers several unique capabilities: (i) automated, multidimensional phenotype profiling, facilitating the analysis and treatment design of complex diseases; (ii) the extrapolation of phenotype profiles beyond provided classes; and (iii) the detection of confounding phenotype factors that could otherwise bias biological inferences. Finally, because no direct comparisons are made between gene expression values from different datasets, the method can use the entire body of cross-platform microarray data. This work has produced a compendium of phenotype profiles for the National Center for Biotechnology Information GEO datasets, which can facilitate an unbiased understanding of the transcriptome-phenotype mapping. The continued accumulation of microarray data will further increase the power of PhenoProfiler, by increasing the variety and the quality of phenotypes to be profiled.

The fundamental aim of modern genetics is linking genotype to phenotype. With the rapid accumulation of genomics data, the lack of phenotype data has become the bottleneck of this process (1). Phenotyping, especially for human subjects, is a laborious process (2). Moreover, researchers often gloss over the complexity of human phenotypes by reporting only those traits specifically relevant to their studies. For example, a given dataset may provide survival information but not the patients’ ages. Inferences derived from such data could be biased or even invalidated by undocumented or poorly documented phenotypic traits. Furthermore, most available phenotype characterizations are qualitative (categorical) rather than quantitative (continuous). This practice is problematic for 2 reasons: The boundaries between categories are often vague or arbitrary (3), and any phenotypic information distinguishing data within a category is lost.

In this article, we address the above issues by developing “PhenoProfiler,” a computational framework for predicting the quantitative phenotype information missing from a genomic dataset. In particular, this method associates each sample of a given dataset with the relative intensity of a specific phenotype trait. The quantitative measures of samples across the whole dataset is referred to as a “phenotype profile” (PP). Examples include the body weights of individuals, degrees of malignancy in tumor samples, and the quantitative responses of patients to drug treatments.

The principle of PhenoProfiler is that similar genomic patterns are likely to be associated with similar phenotypic patterns (4). Thus, we can supplement the (incomplete) phenotypic information in a given genomics dataset with traits recorded in other well-characterized datasets. In particular, we focus on the vast accumulation microarray data. The National Center for Biotechnology Information Gene Expression Omnibus (GEO) (5), for example, currently contains >2000 human microarray datasets that systematically document the transcriptome basis of phenotypes as diverse as heart diseases, mental illness, infectious diseases, and a variety of cancers.

The intuition behind our method is as follows. Given a training dataset with known sample description of a phenotype P, for each gene we can derive an association between its expression profile and this phenotype P. We denote as “signature genes,” those genes whose expressions are strongly associated with the phenotype in the training dataset. Given a new microarray dataset that is known to be related to the phenotype P, but the phenotype description of its individual samples are unknown, we aim to estimate the PP by constructing a sample profile as a real-valued vector that is most similar to the expression profiles of those “signature genes” in the new dataset. Fig. 1 illustrates this approach.

Because the information we borrow from the training dataset is only the association between the gene expressions and sample phenotypes, we do not directly compare the expression values between the training and prediction datasets, thus bypassing the data incompatibility problem between cross-platform and cross-laboratory microarray datasets. PhenoProfiler can therefore use as many microarrays as available in the public repositories in the training stage, and construct a new dataset’s profiles for a wide range of phenotypes.

We applied our method to 587 human microarray datasets, covering >14,000 microarray samples. The predicted phenotype profiles were highly consistent with known phenotype descriptions. We showed that PhenoProfiler can robustly provide multidimensional characterization of the phenotypes missing from a dataset, and can facilitate the discovery of confounding factors for the transcriptome-phenotype mapping. The comprehensive phenotypic data generated by this approach will vastly increase the value of published and forthcoming genomics data.

Results

Overview of Method. As illustrated in Fig. 1, PhenoProfiler consists of 2 steps: (i) Given a microarray dataset D1 whose samples have known descriptions of the phenotype P, for each gene i we calculate a coefficient wi that indicates the degree of


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

1M.X. and W.L. contributed equally to this work.

2To whom correspondence should be addressed. E-mail: xzhou@usc.edu.
association between its expression profile and the phenotype \( P \). The appropriate way to calculate \( w_i \) depends on the structure of the phenotype descriptions. If the phenotype description is binary, i.e., the dataset compares 2 phenotype groups, we can simply use the 2-sample \( t \) statistic as \( w_i \). If the phenotype description is continuous, we can use the correlation between the phenotype values and the gene expression values as \( w_i \). Genes with a large magnitude of \( w_i \) are termed the *signature genes* of the phenotype \( P \). (ii) To predict the phenotype profile (PP) of a new microarray dataset \( D_2 \), we use the following constrained optimization approach. For a dataset with \( m \) individual samples, the PP is defined as a normalized, real-valued vector of \( m \) values \((p_1, p_2, \ldots, p_m)^T\) (denoted \( \mathbf{p} \)). We also define the normalized expression vector of the gene \( i \) as \((e_{i1}, e_{i2}, \ldots, e_{im})^T\), denoted \( e_i \); and we denote the expression matrix containing all such expression vectors as \( \mathbf{E} \). Given a set of coefficients \( w_i \) determined from the training data, the objective is to find a profile \( \mathbf{p} \) that minimizes the weighted least-squares difference \( \sum_w |w_i| \sum_j (\text{sgn}(w_i) e_{ij} - p_j)^2 \). (The function \( \text{sgn} \) is +1 or −1 depending on the sign of its argument; we want the phenotype profile \( \mathbf{p} \) to be close to \( e_i \) when \( w_i \) is positive, and close to \(-e_i \) when \( w_i \) is negative.) A series of matrix computations (see *Methods*) yields the optimal solution \( \mathbf{p} \), which is essentially the normalized weighted (by \( w_i \)) sum of gene expression values across samples.

To assess whether the predicted profile \( \mathbf{p} \) captures the expression trend of those signature genes, we calculate an association score \( \hat{c} \), defined as the Pearson’s correlation between the coefficients \( w_i \) and \( \mathbf{E} \mathbf{p} \) (detailed explanation in *Methods*). To assess the statistical significance of \( \mathbf{p} \), we compare the \( \hat{c} \) to those calculated using the same expression matrix \( \mathbf{E} \) and 1,000 random permutations of coefficients \( w_i \).

**Fig. 1.** The principles of PhenoProfiler. Each row of an expression matrix corresponds to a gene, and each column corresponds to a sample. The magnitude of gene expression is indicated by a color scale running from green (low) to red (high). From the training dataset (at left), we obtain for each gene \( i \) a coefficient \( w_i \) describing the degree of association between its expression level and the phenotype. Here, genes 1 and 2 have high positive coefficients. Gene 3 shows no clear association with the phenotype, so its coefficient is close to zero. Gene 4 has a high negative coefficient. Therefore, genes 1, 2, and 4 are signature genes of the phenotype. Given a new dataset, for each sample we aim to estimate the relative intensity of the association between this sample and the phenotype such that the derived intensity values of all samples are most similar to the expression values of signature genes. We term such a sample profile a phenotype profile. The phenotype profile vector is depicted by the gray cells, where brighter colors indicate estimated tighter association with the phenotype. After sorting the samples based on the phenotype profile, it can be seen that the expression profiles of genes 1 and 2 are strongly correlated with phenotype profile, whereas that of gene 4 is anti-correlated.

**An Illustrative Case: Reconstructing the Temporal Order of the Yeast Log-to-Stationary Growth Transition.** As an illustrative example, consider the 2 microarray datasets GDS18 and GDS283. Both study the log-to-stationary growth transition of yeast, but with different microarray platforms. Both datasets measure gene expression starting at the logarithmic phase and extending through the stationary phase. Here, our goal is to predict changes in yeast phenotype from log to stationary transition, responding to the depletion of nutrients. Naturally, the temporal order of the samples serves as a good means of validating the prediction.

Using one dataset for training, we computed the Spearman’s rank correlation between individual gene expression profiles and the temporal order of the samples. These statistics are used as the training coefficients. In the other dataset, we then predicted the phenotype response profile based on gene expressions, hoping to recover the correct temporal order of samples. In both cases, the predicted PP was highly consistent with the actual sequence of samples (Spearman’s rank correlation was 0.83 and 0.79, depending on which dataset was used for training). Fig. 2 shows the predicted and the original sample order of dataset GDS18. Two subgroups are visible in the predicted profile.
accurately reflecting the logarithmic and stationary phases. The sole exception is at the transition between the 2 phases.

Intriguingly, experiment GDS283 stopped taking measurements only 17 h after the yeast entered the stationary phase. Experiment GDS18, however, continued measurements for another 12 days. So it is remarkable that a phenotype signature derived from GDS283 can accurately sort the phenotype progression of GDS18. This result demonstrates that the essential physiological changes occurring within and between the logarithmic and stationary growth phases can be extrapolated and interpolated.

Large-Scale Prediction of Phenotype Profiles. To test the general applicability of our method, we performed a large-scale analysis of 587 human microarray datasets (see Methods for details on data collection and processing). Datasets containing at least 2 disjoint sample groups, representing a phenotype and its baseline (P and P'), each with at least 10 samples, were selected as training datasets (D2). If a dataset contains n sample groups, we can generate (2^n) distinct training datasets. Because the phenotype values in these datasets are categorical, we use the 2-sample t statistic as coefficients w. By setting the threshold P value for the predicted PP to 0.001 and the association score ĉ ≥ 0.25, a total of 37,852 PPs were associated to the 587 datasets.

To validate the method, for each training dataset D2, we also need a testing dataset D1 that contains the sample descriptions on exactly the same phenotypes P and P'. Among all 587 datasets, we only identified 4 training-testing dataset pairs meeting this criterion, in which each of the testing datasets also contains 2 sample groups of P and P'. To assess whether the predicted phenotype profile is consistent with the known distribution of phenotypes P and P' in the testing dataset, we used the Wilcoxon rank sum test. Specifically, in the testing dataset, the 2 sample groups' (P and P') predicted phenotype values are compared using the Wilcoxon rank sum test. A small Wilcoxon P value indicates that there is a significant difference between the distributions of predicted phenotype values for the 2 groups, therefore the predicted profile is consistent with known phenotype information. Among the 4 training-testing pairs, all predicted PPs were highly consistent with the known phenotype groups (Wilcoxon P < 10^-4).

To obtain a general assessment using more validation data, we relaxed the requirement that the description of the testing dataset exactly matches the training data phenotypes. In fact, if the phenotypes of a given dataset were even moderately similar to the training phenotype, the predicted profiles were found to agree well with known phenotype groups in the testing dataset. This implies a strong interdependence among related phenotypes. We quantify the similarity between the training and testing phenotype with 2 measures: (i) the percentage γ of Unified Medical Language System (UMLS) concepts of the merged sample group descriptions of D2 shared with the dataset description of D1; and (ii) the similarity between the descriptions of corresponding sample groups in D1 and D2, denoted as s. The latter is defined as the cosine of the angle between 2 term frequency-inverse document frequency (tf-idf) vectors of mapped UMLS terms (see Methods for details). Using these measurements, we identified 32 training-testing dataset pairs with similarity thresholds s > 0.4 and γ > 0.6. Among these, 81% of predicted phenotype profiles were consistent with prior phenotype descriptions (Wilcoxon test P < 0.05). This result highlights the effectiveness of our method in exploiting the interdependence of similar phenotypes.

We further studied the robustness of our method against the perturbation of the training dataset. We randomly selected a training dataset and a testing dataset, and then calculated the correlation between the PP constructed with the original training dataset and that with a certain amount of training samples randomly removed. Repeating this test 10,000 times with 10% (and 20%) sample removal produced an average correlation of 0.98 (and 0.95) between the resulting PPs and those without any samples removed. Even for those training datasets with a small size of 10 samples in each of the 2 phenotype groups, the obtained PP correlations were still >0.9 for both 10% and 20% sample removal, demonstrating the robustness of our method.

Multidimensional Profiling of Complex Phenotypes. As previously mentioned, a total of 37,852 PPs were derived and assigned to the 587 datasets. On average, each dataset is assigned 65 PPs. In some cases, related training datasets generated highly correlated PPs, further enhancing our confidence in the prediction. Two examples are described below.

Dataset GDS2855 studies various forms of muscular dystrophy. Three training sets (GDS609, 610, and 612) generated highly correlated PPs (average correlation 0.88) for GDS2855. All 3 training datasets describe the difference between Duchenne muscular dystrophy and normal muscle tissues, although they were measured with different platform technologies. Furthermore, all 3 predicted PPs were highly consistent with the original sample description of GDS2855 (Wilcoxon test P < 10^-4).

Dataset GDS1962 studies gliomas of different grades, and was assigned 4 highly correlated PPs (average correlation 0.9) by datasets GDS1975, GDS1976, GDS1815, and GDS1816. All 4 training datasets focused on comparing grade III and grade IV glioma samples. Remarkably, the predicted PPs not only did a good job of separating grade III from grade IV samples in the testing dataset, but also separated grade II from grade III samples. In addition, the separations followed the order of tumor grades. This example shows that our method captures the essential difference between high- and low-grade tumors, and thus can be extrapolated to tumors of grades beyond those represented in the training data. This ability to extrapolate from the training dataset represents a significant advantage over traditional classification methods.

A testing dataset is often (78% of cases) assigned multiple uncorrelated PPs (correlation <0.1) describing different properties of a complex phenotype. For example, dataset GDS843 contains 49 samples comparing patients with abnormal karyo-
types to patients with normal karyotypes to study adult acute myeloid leukemia (AML). The samples were collected from peripheral blood or bone marrow. Its predicted phenotypes include 3 uncorrelated profiles (see Fig. 3), which are detailed below.

1. Training dataset GDS842 also studied abnormal versus normal karyotypes in adult AML patients. The derived phenotype profile is consistent with the known sample description of this phenotype in the testing dataset (Wilcoxon $P = 0.04$), thus validating our method.

2. Training dataset GDS2118 compared individuals with refractory anemia to normal individuals. The PP trained by this comparison is highly correlated (correlation $>0.9$) with two other PPs that also come from training datasets that studied refractory anemia. In fact, the recently proposed WHO classification of hematologic malignancies merged the disease “refractory anemia with excess blasts in transformation” (RAEB-T) into the category AML. However, this new disease classification is controversial. Although RAEB-T and AML share similar clinical parameters, a study pointed out that their biological bases are different (e.g., RAEB-T is distinguished from AML by a significant increase in apoptosis), and it suggested that RAEB-T should be regarded as a distinct disease entity (6). Therefore, the derived PP may uncover hidden patient information and possibly help to differentiate RAEB-T from AML, which could further lead to improved treatment design.

3. Training set GDS1221 studies the patient response to the drug Imatinib. Imatinib was designed to treat chronic myeloid leukemia by reducing the tyrosine kinase activity of the well-known bcr–abl fusion gene. Our phenotype profile could therefore be used to identify patients that would be more likely to respond to Imatinib treatment.

In summary, although the first PP serves as an internal validation of the method, the other two PPs provide insights into the pathologic and therapeutic properties of sample phenotypes in the dataset GDS843. The specific phenotype properties represented by the above PPs can be further confirmed by examining genes whose expressions are significantly correlated ($P < 0.001$) with these profiles. For example, the AML PP (number 1 above), has 10 significantly correlated genes that are known to be associated with the UMLS concept “Leukemia, Myelocytic, Acute.” A particularly interesting gene is FLT3. A study suggested that in patients with karyotype alterations, a reciprocal translocation was not sufficient to cause acute promyelocytic leukemia, and that an additional mutation in FTL3 may be required (7). For the refractory anemia activation profile, there are 12 significantly correlated genes known to be involved in “Anemia.” These include 3 Fanconi Anemia genes (FANCA, FANCD2, FANCG) and TGFB1, which may affect the progression of refractory anemia specifically (8). Among the genes correlated with the Imatinib response profile, three are tyrosine kinases, which is consistent with the target of Imatinib (9).

Discovery of Hidden Confounding Factors in Microarray Studies. Due to the scarcity of phenotype information in many microarray datasets, confounding phenotype variables may not be well documented. Thus, caution should be exercised in deriving inferences from microarray datasets. The following cases provide representative scenarios.

Dataset GDS1673 examines normal lung tissue from 23 donors, including smoking and nonsmoking individuals. Interestingly, we found that a predicted PP trained on male vs. female skeletal muscle samples (dataset GDS914) was able to separate the smoking and nonsmoking samples of GDS1673 (Wilcoxon $P = 0.0002$). After obtaining additional phenotype information on the GDS1673 subjects, it turns out that among nonsmokers, which made up almost 2/3 of the sample, females outnumbered males by $>2:1$, whereas among smokers the numbers of the 2 genders were approximately equal. Thus, simply comparing the expression profiles of the smoking versus nonsmoking groups would not derive the signature of smoking, but rather the mixed signatures of smoking and gender.

As another example, the goal of the GDS1887 study was to build a prognosis model for rectal cancer cells responding to
radio therapy. According to the GEO annotation, its 46 samples had been separated into training and test groups for model construction and validation. Surprisingly, we found that the original training and test samples from this study could be well separated (average Wilcoxon $P = 0.001$) by 4 highly correlated PPs (average correlation 0.95). All of those 4 PPs come from training datasets that compare the cancer to normal tissue or that compare cancers of different malignancies. This strongly suggests that there were systematic differences in cancer malignancy between the training and testing samples, even though they were supposed to be generated by random partition. Any such sampling bias would negatively impact the accuracy of the prognosis model.

Of course, sampling bias can often be traced to the very limited availability of phenotype data in the first place. Our compendium of predicted phenotype profiles (http://zhoulab.usc.edu/PhenoProfiler) provides a comprehensive description of a large proportion of the datasets in the GEO database. This knowledge can facilitate an unbiased understanding of the transcriptome-phenome mapping. It can also serve as the starting point for the identification of molecular mechanisms shared by different diseases and phenotypes.

**Discussion**

Phenotypes are complex, and difficult to quantify in a high-throughput fashion. The lack of comprehensive phenotype data can prevent or distort genotype–phenotype mapping. This article describes a unique approach to perform in silico phenotype profiling. Our method provides numerous advantages, which we outline here. (i) For most datasets we were able to predict multiple phenotype profiles, which could help researchers to reveal different aspects of complex diseases and facilitate treatment design. (ii) We can provide a quantitative phenotype description of the sample characteristics. Although “categorical” phenotype description is prevalent, in reality phenotypes constitute a continuous spectrum. (iii) Our method can extrapolate the profiling to classes beyond those represented in the training data, as illustrated in the glioma case study. This is an advantage over traditional classification methods. (iv) PhenoProfiler avoids direct comparison of gene expression values from different datasets, and thus can use almost all available microarray data regardless of platform or laboratory. In contrast, traditional regression methods cannot be directly applied to microarray datasets from different platforms.

The continued accumulation of microarray data will further increase the power of PhenoProfiler in 2 aspects: the variety of phenotypes to be profiled, and the confidence of its predictions. The latter benefit derives from having several mutually correlated PPs from similar datasets. The principles of our method can be easily applied to other types of genomics data (e.g., proteomics or metabolomics) as they become increasingly available. The present work focuses on linear gene-phenotype associations, but more complex relationships can be devised depending on the data characteristics.

Our univariate method for constructing the gene coefficients from the training samples is only one of many possible approaches. For example, one could consider constructing coefficients using a multivariate procedure that takes into account correlations among the gene expression levels, such as Fisher’s linear discriminant procedure (i.e., discriminant function analysis for 2 groups). However, such an approach requires estimating a covariance matrix for the gene expressions which is not practical given that there are thousands of genes and a limited number of samples (typically on the order of 10) per dataset. Fan and Fan (10) prove that, when the dimension of the feature space is high, a univariate 2-sample $t$ test procedure, similar to our approach, is often superior to a multivariate method. Alternatively, when the important phenotype information can be characterized using a small number of linear combinations of the genes, dimension reduction techniques like Nonnegative Matrix Factorization (11, 12) may also produce meaningful phenotype predictions.

**Methods**

**Predicting Phenotype Profiles by Constrained Optimization.** From a training microarray dataset, we derive a vector $w = (w_1, w_2, \ldots, w_n)^T$ that contains the gene-phenotype association coefficients of $n$ genes. Given a new dataset with $n$ genes and $m$ samples, and the normalized gene expression matrix $E = (e_{ij})_{n \times m}$, we aim to obtain the optimal phenotype profile (PP) of the $m$ samples, where PP is a normalized, real-valued vector $p = (p_1, p_2, \ldots, p_m)^T$ that show high similarity to the expression profiles of those genes that have high magnitude of gene-phenotype association coefficients $w$ (signature genes)

$$Q_1 : \min_{p \in \mathbb{R}^m} \quad d(E, w, p) = \sum_{i} w_i \sum_{j} (\text{sgn}(w_i) e_{ij} - p_j)^2$$

subject to

$$\sum_{j} p_j = 0$$
$$\frac{1}{m-1} \sum_{j} p_j^2 = 1$$

Let $b = (b_1, b_2, \ldots, b_m)^T$ be the weighted sum of gene expression values for each sample, $b_j = \sum w_i e_{ij}$. The following theorem provides the solution to the minimization problem.

**Theorem.** The solution $\hat{p}$ to problem $Q_1$ is a vector that is the normalized form of $b$. That is,

$$\hat{p} = \frac{b - \bar{b}}{\sigma(b)}$$

where $b$ and $\sigma(b)$ are the mean and standard deviation of $b$ respectively.

Proof: By expanding the function $d$, we have $d(E, w, p) = \sum w_i (e_{ij} - \bar{b})^2 + \sum w_i \sum_{j} (\text{sgn}(w_i) e_{ij} - p_j)^2 - 2 \sum w_i e_{ij} p_j$. Because $p$ is normalized and $E$ and $w$ are fixed, the first 2 terms are fixed. So the minimization problem $Q_1$ can be simplified to an equivalent maximization problem:

$$Q_2 : \max_{p \in \mathbb{R}^m} \quad c(E, w, p) = w^T E p$$

subject to

$$p^T 1 = 0$$
$$p^T p = m - 1$$

where $1$ is the vector whose elements are all $1$.

Let $b - E 1w$. Let the Lagrangian function for $Q_2$ be.

$$L(p, \lambda_1, \lambda_2) = b^T p + \lambda_1 (p^T 1 + \lambda_2 (p^T p - (m - 1))$$

where $\lambda_1$ and $\lambda_2$ are Lagrangian multipliers. According to the Karush–Kuhn–Tucker conditions (13) (as the functions $B^T p$, $p^T 1$, and $p^T p$ are all convex), the solution to Eq. 1 contains the global optimum of $Q_2$.

$$\nabla L(p, \lambda_1, \lambda_2) = 0$$

Eq. 1 results in 2 solutions: $p = \pm (b - \bar{b})/\sigma(b)$. Because $Q_2$ is a maximization problem, it is easy to show that the solution of $Q_2$ is $p = (b - \bar{b})/\sigma(b)$, and so is the solution of $Q_1$.

$\hat{p}$ is regarded as the PP of the new dataset because among all vectors in $\mathbb{R}^m$, $\hat{p}$ is the one that most resembles the normalized expression profiles of the signature genes that were defined by the training data. We calculate an association score $\epsilon$ as the Pearson correlation between $w$ and $\hat{p}$. The score is derived from the maximization problem $Q_2$. $\hat{p}$ provides the association between expression profiles and the predicted phenotype profile in the testing dataset. Thus, higher $\epsilon$ indicates higher consistency of gene-phenotype associations derived from the training and testing datasets.
Data Collection and Processing. We collected 587 human microarray datasets, each containing at least 5 samples, from the National Center for Biotechnology Information Gene Expression Omnibus (GEO) (5). For data generated with the Affymetrix platforms, we increased any values <10 to 10 and performed a log transform of the gene expression values. For genes with multiple probesets present, the expression values of those probesets were averaged. We then normalized each dataset by converting the expression values of each gene to Z scores (zero mean and unit variance). The 587 datasets yielded 537 training datasets. A training dataset contains 2 disjoint sample groups, each of which contains at least 10 samples, representing a phenotype and its baseline. When performing PP prediction, we discarded those training-testing dataset pairs that share <100 genes in common.

Automatic Processing of Phenotype Annotations. In GEO, a dataset is usually annotated by a short description paragraph; a sample group is annotated by a word or a short phrase; and a sample is usually annotated by 1 sentence. To systematically categorize the phenotype information associated with each microarray dataset, we used the Unified Medical Language System (UMLS) (14, 15). We mapped the dataset description, sample group descriptions, and sample descriptions onto UMLS concepts via the MetaMap Transfer program (16). To reduce noise we focused on disease-related concepts, including the MeSH vocabulary and the semantic types “Pathologic Function,” “Injury or Poisoning,” “Anatomical Abnormality,” “Body Part, Organ, or Organ Component,” “Tissue,” and “Cell.” In general, the higher a concept is on the UMLS hierarchy, the broader is the concept. Disease concepts at the fine granularity level may be associated with more clinical significance. To infer higher-order links between datasets, all of the ancestor concepts of mapped concepts were included.

Measuring Phenotype Annotation Similarity. We measure phenotype similarity between sample groups (the 2 groups of a training dataset and the 2 groups of a testing dataset) by the following procedure. (i) For each group, we map its title and member sample descriptions onto UMLS concepts. (ii) We then construct a term frequency-inverse document frequency (tf-idf) vector (17) for each sample group. (iii) Suppose that \( U_{11} \) and \( U_{12} \) are 2 tf-idf vectors corresponding to the 2 sample groups in the training dataset, and that \( U_{21} \) and \( U_{22} \) are tf-idf vectors corresponding to the 2 sample groups in the testing dataset. The similarity between the sample groups is then calculated as \( \text{max}(U_{11}, U_{12}) + (U_{22}, U_{22}) \), where \((a, b)\) denotes the cosine similarity (18) of 2 vectors \(a\) and \(b\), calculated as a normalized dot product \( (a, b) = \frac{a^T b}{||a|| ||b||} \). Essentially, this measure identifies the best match between the sample groups in the training dataset and testing dataset while taking into account the possibility that the groups could be matched in reverse order.

ACKNOWLEDGMENTS. We thank Qiang Song for creating the database of the phenotype profile compendium and Chao Cheng, Caleb Finch, Huanying Ge, Haifeng Li, Chun-Chi Liu, Todd Morgan, Rebecca Nugent, Juan Nunez-Iglesias, Xiting Yan, and anonymous reviewers for constructive comments and suggestions. This work was supported by National Institutes of Health Grants R01GM071463, P50HG002790, and U54CA112952 and NSF Grants 0515936, 0747475 and DMS-0705312. X.J.Z. is an Alfred Sloan Fellow.