doi:10.1093/nar/gks615

Co-clustering phenome–genome for phenotype classification and disease gene discovery

TaeHyun Hwang[1], Gowtham Atluri[2], MaoQiang Xie[3], Sanjoy Dey[2], Changjin Hong[4], Vipin Kumar[2] and Rui Kuang[2,*]

[1]Bioinformatics core at Masonic Cancer Center, [2]Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA, [3]College of Software, Nankai University, Tianjin, 300071, China and [4]Computational Biomedicine Division, Department of Medicine, Boston University, MA 02118, USA

Received March 23, 2012; Revised May 29, 2012; Accepted May 31, 2012

ABSTRACT

Understanding the categorization of human diseases is critical for reliably identifying disease causal genes. Recently, genome-wide studies of abnormal chromosomal locations related to diseases have mapped >2000 phenotype–gene relations, which provide valuable information for classifying diseases and identifying candidate genes as drug targets. In this article, a regularized non-negative matrix tri-factorization (R-NMTF) algorithm is introduced to co-cluster phenotypes and genes, and simultaneously detect associations between the detected phenotype clusters and gene clusters. The R-NMTF algorithm factorizes the phenotype–gene association matrix under the prior knowledge from phenotype similarity network and protein–protein interaction network, supervised by the label information from known disease classes and biological pathways. In the experiments on disease phenotype–gene associations in OMIM and KEGG disease pathways, R-NMTF significantly improved the classification of disease phenotypes and disease pathway genes compared with support vector machines and Label Propagation in cross-validation on the annotated phenotypes and genes. The newly predicted phenotypes in each disease class are highly consistent with human phenotype ontology annotations. The roles of the new member genes in the disease pathways are examined and validated in the protein–protein interaction subnetworks. Extensive literature review also confirmed many new members of the disease classes and pathways as well as the predicted associations between disease phenotype classes and pathways.

INTRODUCTION

Phenotypes, the observable characteristics (traits) of an organism, are believed to be determined by genetic materials (DNAs) under environmental influences (1,2). The key to achieving desired phenotypes such as favorable disease treatment outcomes lies in the understanding of the relation between phenotypes and the biological roles of genes (3–5). In the past two decades, promising bio-technologies such as microarray-based profiling (6–9) and second generation sequencing (10,11) were developed to hunt for potential phenotype–gene associations. Currently, in the most comprehensive disease, phenotype–gene relation database, Online Mendelian Inheritance in Man (OMIM) (2), nearly 2000 confirmed relations between around 6000 phenotypes and over 12 000 genes are documented. This knowledge base provides a new phenome (the collection of all phenotypes) perspective to study human diseases and their molecular mechanisms. Although most previous studies focused on predicting new disease phenotype–gene relations with OMIM data (12–19), we propose to cluster phenotypes and find gene modules associated with the phenotype clusters by integrating OMIM phenotype–gene relations with disease phenotype similarity network and the human gene interaction network as well as exiting disease categorization and molecular pathways. To effectively use all the sources of information, we design regularized non–negative matrix tri-factorization (R-NMTF) algorithms to tri-factorize the binary matrix of phenotype–gene relations into phenotype clusters, gene clusters and an association matrix representing the associations between phenotype clusters and the gene clusters (Figure 1). Since the matrix of known phenotype–gene relations is very sparse, constraints constructed from the prior knowledge and the phenotype/gene labels are introduced to regularize the NMTF models.

*To whom correspondence should be addressed. Tel: +1 612 6247820; Fax: +1 612 6250572; Email: kuang@cs.umn.edu

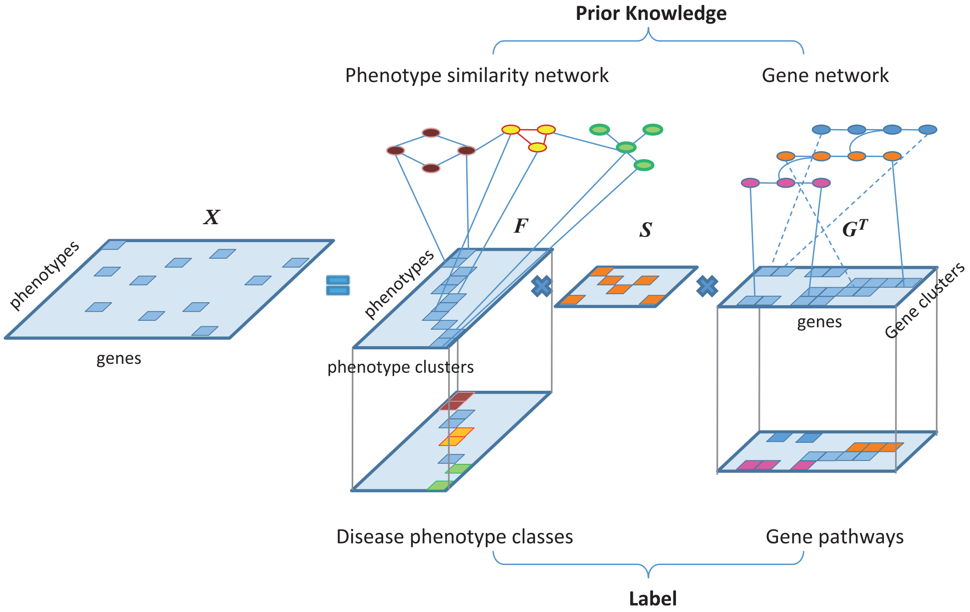

Figure 1. NMTF of disease phenotype–gene associations. The phenotype–gene association matrix X is factorized into products of three matrices, phenotype cluster membership F, gene cluster membership G and phenotype cluster–gene cluster association S for supervised co-clustering of phenotypes and genes. Label information for the disease classes and the pathways are available for a small number of phenotypes and genes. Prior knowledge is also introduced from phenotype similarity network and gene network. For better visualization, different colors are used to distinguish the phenotypes and the genes in different clusters.

Current classification of human disease is mainly based on observational correlation between pathological analysis and clinical syndromes (20), and more recently, by text mining of clinical records and synopsis (21). An accurate classification of human diseases based on its phenotypic and molecular basis will help to establish syndromic patterns for selecting phenotypes to consider in diagnosis. Existing phenotype clustering approaches cluster phenotypes based on only text descriptions and synopsis (22–24) or shared disease genes (25), which do not fully reflect both phenotypic and genetic basis of the disease phenotypes. R-NMTF integrates various sources of phenotypic and genomic data as well as prior knowledge to perform supervised co-clustering of phenotypes and genes simultaneously. R-NMTF is the first of its kind that effectively discovers disease classes based on the molecular underpinnings of the phenotypes and the molecular interactions in a network. This approach implements the philosophy of network-based medicine (26), which is believed to be the promising approach for generating the next generation of disease categorization (20). The R-NMTF-based co-clustering also naturally induces the associations between the phenotype clusters and gene clusters, which provides a global pathway activity view of human disease classes for understanding the unique as well as common underlying molecular mechanisms of diseases.

MATERIALS AND METHODS

In this section, we first describe the notations for the data of disease phenotypes, genes and their associations. We then review NMTF and introduce the framework of R-NMTF for co-clustering phenotypes and genes. We also outline the multiplicative update algorithm for solving the R-NMTF model.

Notations

The notations and definitions used in the article are specified in Table 1. We denote the OMIM phenotype–gene associations by a m by n binary matrix X with 1 for known associations and 0 otherwise. The objective is to derive phenotype clusters (F) and find their association (S) with gene clusters (G) based on X (Figure 1). F and G are non-negative matrices representing the soft memberships of each gene/phenotype against the k_1 phenotype clusters or the k_2 gene clusters. To perform more reliable phenotype clustering in a supervised setting, we use the partial phenotype annotations by (25) represented by a binary matrix F^0 with 1 for the known class memberships. Similarly, KEGG pathways (27) are also included in a binary matrix G^0 to guide gene clustering. Note that, since training samples are not required for each disease category to classify the phenotypes in the model, we use the word 'co-clustering' instead of 'classification' or 'semi-supervised learning' for the learning problem although in the experiments, we only focused on recovering the 21 disease categories with at least one OMIM disease phenotype. Finally, a phenotype similarity network M (21) and the gene interaction network N were also introduced to capture modular relations among phenotypes and genes. M and N contain edges weighted by the degree of similarity between phenotypes or the confidence of interaction between genes, respectively.

Table 1. Notations

Notation	Definition
m	Number of disease phenotypes
n	Number of genes
k_1	Number of phenotype clusters (e.g. classes)
k_2	Number of gene clusters (e.g. pathways)
X	Disease phenotype–gene association matrix ($m \times n$)
F	Phenotype cluster membership ($m \times k_1$)
S	Phenotype cluster–gene cluster association Matrix ($k_1 \times k_2$)
G	Gene cluster membership ($n \times k_2$)
F^0	Annotated phenotype cluster membership ($m \times k_1$)
G^0	Annotated gene cluster membership ($n \times k_2$)
M	Disease phenotype similarity network ($m \times m$)
N	Gene interaction network ($n \times n$)

Non-negative matrix tri-factorization

Non-negative matrix factorization (NMF) was proposed by (28,29) as an alternative to principle component analysis and vector quantization for parts-based decomposition of a data matrix. NMF has been applied to solve various bioinformatics problems such as identifying gene clusters (30–32), bi-clustering (33) and identifying cancer tumor categories (34) in gene expression data analysis, and finding modules in protein–protein interaction (PPI) network (35).

By imposing the orthogonality on the two factorized matrices, (36) proposed a framework to perform NMTF as $X \simeq FSG^T$ under the constraints $F^TF = \mathbf{1}$ and $G^TG = \mathbf{1}$. This framework has the advantage of simultaneously clustering the columns and rows, and finding a condense representation of the data matrix by the row clusters and the column clusters, which can also be considered as associations between row clusters and column clusters. For co-clustering phenotypes and genes, the NMTF approach provides novel insights into the phenotype–gene associations beyond clustering and decomposition.

Regularization by phenotype and gene labels

To cluster phenotypes and genes based on their associations, we adopt supervised NMTF proposed for finding associations between document clusters and word clusters in text categorization (37,38). We use manually labeled phenotype clusters as the phenotype label F^0 and gene clusters from existing pathway database as the gene label G^0, and simultaneously cluster phenotypes and genes with tri-factorization as illustrated in Figure 1. The following optimization framework can be solved to achieve the goal:

$$\begin{aligned}&\min_{F,S,G}\|X - FSG^T\|_F^2\\&\qquad + \alpha\|F - F^0\|_F^2 + \beta\|G - G^0\|_F^2\\&\text{subject to } \sum_{j=1}^{k_1} F_{i,j} = 1, \sum_{j=1}^{k_2} G_{i,j} = 1.\end{aligned} \qquad \textbf{(1)}$$

In equation (1), the first term is the NMTF of X, and the second and the third terms are the fitting penalties to keep the new cluster assignment consistent with the known phenotype and gene cluster labels. These two terms are introduced as a supervised way of minimizing the squared loss between the predicted phenotype cluster assignment F and the initial phenotype cluster assignment F^0, and between the predicted gene cluster assignment G and the initial gene cluster assignment G^0. Specifically, the phenotype clusters are taken from the 21 disease classes manually curated by (25), in which 872 disease phenotypes are assigned to 21 classes. The gene clusters are derived from the genes in KEGG pathways (27). The information of the labeled phenotypes and genes provides the useful guidance to learning more accurate co-clustering.

A limitation of the approach in equation (1) is the low coverage and the sparsity of the disease gene association matrix used to cluster phenotypes and genes. The known disease–gene association only cover a small fraction of phenotypes and genes (one-third of the phenotypes and 5% of the genes), with very few associations between them (less than one association per phenotype/gene). Moreover, the phenotype cluster annotations and KEGG pathways also only provide a low coverage of around 15% phenotypes and one-fourth of the genes. The statistics simply suggest that with this model only a very small fraction of phenotypes and genes could be clustered properly.

Regularization by graph Laplacians

To address the above problem, we design R-NMTF to incorporate the prior knowledge in the phenotype similarity network and the PPI network (Figure 1) to cluster phenotypes and genes with matrix tri-factorization. Given the phenotype similar network M and the PPI network N, the following optimization problem is formulated for the purpose:

$$\begin{aligned}&\min_{F,S,G}\|X - FSG^T\|_F^2\\&\qquad + \alpha\|F - F^0\|_F^2 + \beta\|G - G^0\|_F^2\\&\qquad + \gamma\,\mathrm{tr}(F^T(D_M - M)F)\\&\qquad + \lambda\,\mathrm{tr}(G^T(D_N - N)G)\\&\text{subject to } \sum_{j=1}^{k_1} F_{i,j} = 1, \sum_{j=1}^{k_2} G_{i,j} = 1,\end{aligned} \qquad \textbf{(2)}$$

where D_M is the diagonal matrix with the row summation of matrix M on the diagonal and D_N is similarly defined from N. In this equation, the first three terms are identical to those in equation (1). The fourth and fifth terms introduce the phenotype similarity network and the PPI network as prior knowledge to guide the clustering of the phenotypes and the genes. These two terms are called smoothness terms, which encourage the connected nodes (phenotypes/genes) in a graph to be assigned to the same cluster. Specifically, the term $\mathrm{tr}(F^T(D_M - M)F)$ requires that the phenotype clusters identified by NMTF are also densely connected in the phenotype network, and similarly for $\mathrm{tr}(G^T(D_N - N)G)$. $D_M - M$ and $D_N - N$ are known as the Laplacian matrices of the graphs, which are positive semi-definite (39).

Algorithm 1

```
Regularized Non-negative Matrix Tri-factorization
INPUT: $X$, $F^0$, $G^0$, $L_M$, $L_N$, parameters $\alpha$, $\beta$, $\gamma$, and $\lambda$, maximum interation $T$
OUTPUT: $F$, $G$, $S$
while not converged and $t \leq T$ do
    (1) Update $F_{ij} \leftarrow F_{ij}\sqrt{\frac{(XGS^T+\alpha F^0+\gamma MF)_{ij}}{(FSG^TGS^T+\alpha F+\gamma D_M F)_{ij}}}$.
    (2) Normalize $F_{i.} \leftarrow \frac{F_{i.}}{\sum_{j=1}^{k_1} F_{ij}}$
    (3) Update $G_{ij} \leftarrow G_{ij}\sqrt{\frac{(X^TFS+\beta G^0+\lambda NG)_{ij}}{(GS^TF^TFS^S+\beta G+\lambda D_N G)_{ij}}}$.
    (4) Normalize $G_{i.} \leftarrow \frac{G_{i.}}{\sum_{j=1}^{k_2} G_{ij}}$.
    (5) Compute $S_{ij} \leftarrow S_{ij}\sqrt{\frac{(F^TXG)_{ij}}{(F^TFSG^TG)_{ij}}}$.
end while
```

Multiplicative update algorithms

We extend the optimization algorithms for the original NMTF to handle the four additional penalty terms in equation (2). The alternative iterative scheme to solve the problem with respect to one variable while fixing the other variables are described.

Computation of F

If we fix variables S and G, solving equation (2) with respect to F is equivalent to minimizing the following function:

$$L(F) = \|X - FSG^T\|_F^2 + \alpha\|F - F^0\|_F^2 + \gamma \mathrm{tr}(F^T L_M F)$$

subject to $\sum_{j=1}^{k_1} F_{i,j} = 1$, where L_M is $D_M - M$.

The differentiation of L with respect to F is

$$\frac{\partial L(F)}{\partial F} = -2XGS^T + 2FSG^TGS^T + 2\alpha(F - F^0) + 2\gamma L_M F.$$

The multiplicative update rule is

$$F_{ij} \leftarrow F_{ij}\sqrt{\frac{(XGS^T + \alpha F^0 + \gamma MF)_{ij}}{(FSG^TGS^T + \alpha F + \gamma D_M F)_{ij}}}.$$

To satisfy the equality constrain, we normalize F as

$$F_{i.} \leftarrow \frac{F_{i.}}{\sum_{j=1}^{k_1} F_{ij}}.$$

Computation of G

If we fix variables S and F, solving equation (2) with respect to G is equivalent to minimizing the function,

$$L(G) = \|X - FSG^T\|_F^2 + \alpha\|G - G^0\|_F^2 + \gamma \mathrm{tr}(G^T L_N G)$$

subject to $\sum_{j=1}^{k_2} G_{ij} = 1$, where L_N is $D_N - N$.

The differentiation of L with respect to G is

$$\frac{\partial L(G)}{\partial G} = -2X^TFS + 2GS^TF^TFS^S + 2\beta(G - G^0) + 2\lambda L_N G.$$

The multiplicative update rule is

$$G_{ij} \leftarrow G_{ij}\sqrt{\frac{(X^TFS + \beta G^0 + \lambda NG)_{ij}}{(GS^TF^TFS^S + \beta G + \lambda D_N G)_{ij}}}.$$

To satisfy the equality constrain, we normalize G as

$$G_{i.} \leftarrow \frac{G_{i.}}{\sum_{j=1}^{k_2} G_{ij}}.$$

Computation of S

After F and G are computed, solving equation (2) with respect to S is equivalent to minimizing the following function:

$$L(S) = \|X - FSG^T\|_F^2.$$

The differentiation of L with respect to S is

$$\frac{\partial L(S)}{\partial S} = -2F^TXG + 2F^TFSG^TG.$$

The multiplicative update rule is

$$S_{ij} = \leftarrow S_{ij}\sqrt{\frac{(F^TXG)_{ij}}{(F^TFSG^TG)_{ij}}}.$$

The complete R-NMTF algorithm is outlined in Algorithm 1. Since the updating steps for F, S and G are non-increasing, the objective function will decrease until a lower bound is reached. Empirically, the algorithm converges fast within 100 iterations in the experiments.

EXPERIMENTS

To evaluate the performance of supervised co-clustering of phenotypes and genes, R-NMTF was applied to classifying OMIM human disease phenotypes and KEGG disease pathway genes with leave-one-out cross-validation. R-NMTF was compared with several baseline methods, including support vector machines (SVMs), Label Propagation (LP) and a NMTF model without network regularization defined in equation (1). R-NMTF was then applied to classify unannotated OMIM disease phenotypes and identify new member genes of KEGG disease pathways. The predictions were verified and analyzed by comparison with human phenotype ontology (HPO) and literature survey.

Data preparation

We collected the disease phenotype–gene associations in OMIM, which consist of the associations between 1284 disease phenotypes and 1777 disease genes. We also collected 200 KEGG pathways, which contain 4128 genes in total, from molecular signature database (40).

Table 2. Performance of phenotype classification in leave-one-out cross-validation

Compared methods	Avg. rank	win/draw/loss (*P*-value)
R-NMTF versus NMTF	**3.124** versus 5.590	300/154/136 (4.617e−13)
versus SVM-linear	versus 6.103	308/154/128 (3.693e−12)
versus SVM-rbf	versus 5.037	268/213/109 (1.497e−4)
versus LP	versus 3.700	161/388/41 (9.145e−05)

This table reports the average rank of the target class out of the 20 classes, and the pairwise 'win/draw/loss' comparisons of each leave-one-out case between R-NMTF and the baselines, SVMs with linear and rbf kernels, NMTF and LP. The last column reports the statistical significance of the ranking results using Wilcoxon rank sum test.

We obtained the human protein-protein interaction (PPI) network from HPRD (41). The PPI network contains 76232 binary undirected interactions between 9667 genes. We obtained the phenotype similarity network from (21). The phenotype similarity network is an undirected graph with 5080 vertices representing OMIM disease phenotypes, and edges weighted by a number in [0,1]. The edge weights measure the similarity between phenotypes by their overlap in the text and the clinical synopsis in OMIM records, calculated by text mining (21).

In the leave-one-out cross-validation, after preprocessing (removing the phenotypes classified as multiple and unclassified, removing disease phenotypes not present in both the disease phenotype–gene associations and the phenotype similarity network and removing genes not present in both the disease phenotype–gene associations and the PPI network), we generated a dataset containing 590 disease phenotypes in 20 disease classes (25) and 7997 genes in 200 gene pathways. This dataset was used in leave-one-out cross-validation on disease phenotype classification and disease pathway gene discovery.

To further evaluate R-NMTF with more phenotypes and other independent phenotype annotations, we generated another larger dataset containing 1325 disease phenotypes with at least one known causal gene in OMIM. Among the 1325 disease phenotypes, 501 disease phenotypes intersect with the labeled disease phenotypes in the first dataset and the rest 824 disease phenotypes are unlabeled. Our task in this experiment is to perform a supervised clustering to assign the 824 unannotated disease phenotypes to the 20 disease classes.

Baselines and parameter tuning

Four baselines were introduced for comparison with R-NMTF, SVMs with linear kernel and radial basis kernel, LP and the NMTF model defined in equation (1) without the prior knowledge from the phenotype network and the PPI network (named NMTF). The SVMs used a binary vector representing the disease genes of each phenotype as the features for classification (25). We also tested SVMs with the similarity scores in the phenotype similarities network as features for classification. Since the results are close to random, we did not report them in the experiments. We also compared R-NMTF with a semi-supervised learning method, LP, which uses the disease similarity network and the PPI network for disease phenotype classification and disease gene discovery, respectively (42). The hyper-parameters (α and β for NMTF; α, β, γ and λ for R-NMTF and C and σ for SVMs) were chosen by a grid search in $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. The hyper-parameter α for LP was chosen by a grid search in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. More analysis of parameter tuning is described in the supplementary Table S1 and S2.

In the leave-one-out cross-validation on the 590 labeled phenotypes in disease phenotype classification, we held out one phenotype as the test case to be classified by all the compared methods. The performance is measured by the rank of the true disease class among the 20 target classes ranked by the corresponding classification scores generated by a classification method. Similarly, in the leave-one-out cross-validation for disease gene discovery on the same data, we held out one gene in a KEGG disease pathway as the test case to be classified by all the compared methods. Since one gene could belong to multiple disease pathways, the performance is measured by the area under the curve of receiver operating characteristic (AUC). Since leave-one-out cross-validation usually gives less overfitting bias, we reported the results with the best parameters for all the methods in the experiments on both disease phenotype classification and disease gene discovery.

Performance of disease phenotype classification in leave-one-out cross-validation

The average ranking performance of the compared methods are reported in Table 2 and Figure 2. On average, R-NMTF were able to rank the target class at around third out of the 20 classes, while the other methods performed worse. To further assess the statistical significance of the difference in the performance between R-NMTF and the baselines, we also report the pairwise comparison of each test case and performed a Wilconsin test on the difference of the ranks in Table 2. The *P*-values suggest that R-NMTF performed significantly better than the baselines. Supplementary Figure S1 visualizes the pairwise comparison between R-NMTF and the baselines by scatter plot. Many more cases appeared in the top left triangle indicating a better ranking by R-NMTF. LP performed worse than R-NMTF but better than SVMs and NMTF. The observation indicates that the global structural information in the phenotype similarity network provides substantial information on phenotype classes. To further understand the classification performance in each disease class, we show in Table 3 the classification performance for the phenotypes by disease classes. R-NMTF outperformed all the baseline methods in 11 disease classes. In some of the small classes such as 'ear, nose, throat', 'nutritional' and 'respiratory', less relations among the training points are available for R-NMTF to improve classification.

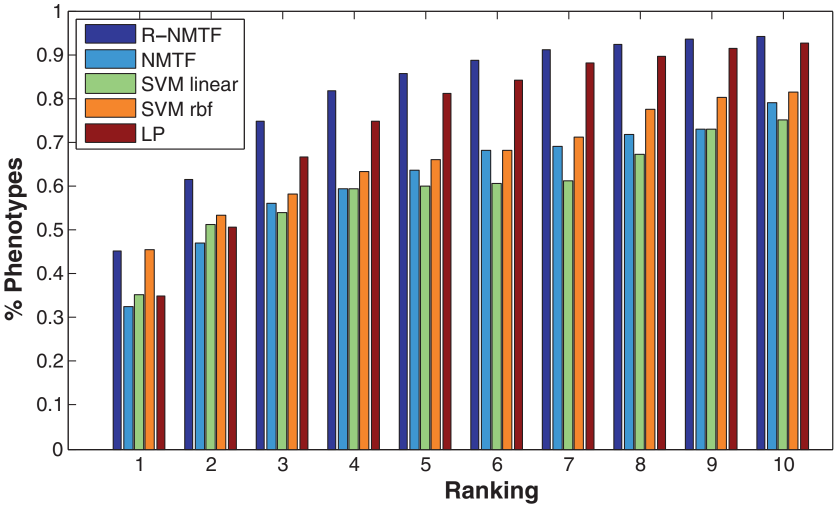

Figure 2. Performance of phenotype classification in leave-one-out cross-validation. In this plot, the *x*-axis represents the cutoffs of the rank of the target disease class out of the 20 classes. The *y*-axis represents the faction of phenotypes with their target disease class ranked within a certain cutoff. For example, R-NMTF ranked the target class of >60% of the phenotypes within Rank 2, while the other methods only ranked around or <50% within the same rank cutoff.

Table 3. Disease phenotype classification results by disease classes

Disease classes (No)	Avg. rank				
	R-NMTF	NMTF	SVM-linear	SVM-rbf	LP
Bone (23)	**3.3**	8.5	4.7	7.6	4.7
Cancer (53)	**1.6**	5.0	4.2	2.0	1.9
Cardiovascular (28)	**3.8**	10.1	10.0	6.0	4.3
Connective tissue (16)	**8.5**	8.9	10.6	11.4	11.1
Dermatological (32)	**2.0**	4.4	3.0	4.0	2.5
Developmental (28)	5.7	**2.5**	9.6	9.2	6.5
Ear,Nose,Throat (3)	20.0	20.0	**14.7**	15.0	16.7
Endocrine (30)	**4.2**	5.4	13.4	5.4	4.9
Gastrointestinal (12)	9.7	7.8	**7.8**	9.7	11.7
Hematological (30)	3.5	9.5	**2.3**	6.9	3.8
Immunological (31)	**2.6**	10.0	8.1	5.2	2.8
Metabolic (84)	**1.0**	2.2	4.1	2.2	**1.0**
Muscular (18)	5.7	**5.3**	12.2	9.1	7. 3
Neurological (80)	**1.4**	6.2	5.8	2.7	1.4
Nutritional (2)	16.0	3.0	19.0	**2.0**	20
Ophthamological (35)	**1.9**	4.2	2.5	2.9	2.5
Psychiatric (9)	7.9	**6.1**	8.0	11.4	14.8
Renal (23)	4.1	**3.5**	4.4	6.8	4.9
Respiratory (7)	15.4	**10.4**	**10.4**	14.1	15.7
Skeletal (46)	**1.5**	3.3	4.8	5.2	1.8

This table reports the ranking performance by R-NMTF, SVM with linear and rbf kernels, NMTF and LP in each disease class in the leave-one-out cross-validation. The number of phenotypes in each disease class is reported in the parentheses.

Performance of disease gene discovery in leave-one-out cross-validation

In the experiment of disease gene discovery, we collected the member genes in the 200 pathways from KEGG. In the preprocessed data, there are 590 member genes in 27 KEGG disease pathways such as Alzheimer, diabetes and cancer-related pathways. In the leave-one-out cross-validation, each of the 590 member gene was held out and then classified into the 200 pathways as a multi-label classification problem since some of the disease genes are members of multiple pathways. The higher the target pathways in the ranking of the 200 pathways, the better the performance. We measured the performance by the AUC. LP was applied on the PPI network to predict the disease genes as the baseline. The other 589 member genes was used as the initialization of label propagations to classify the held-out gene. The average AUC across the 590 member genes by all the methods are reported in Table 4 and Figure 3. The results clearly show that by integration of phenotype similarity, phenotype class annotation and phenotype–gene associations with PPI network R-NMTF more accurately classified the disease genes compared with LP, which only uses the PPI network for disease gene discovery. R-NMTF performed better on >500 cases with an average AUC 0.930 compared with 0.73 by LP.

Analysis of phenotype clusters with HPO

To bette characterize the discovered phenotype clusters for the 824 unannotated disease phenotypes, we compared the phenotype clusters with HPO (43). HPO describes human phenomic abnormalities with a controlled hierarchical vocabulary. Since the vocabulary in the HPO was developed independently of the disease classification by (25), it is an external resources for the validation of the phenotype clusters discovered by R-NMTF. Each OMIM phenotype was mapped to the hierarchy of HPO to retrieve the matched HPO terms. Then, a new HPO similarity is calculated for each pair of phenotypes by Jaccard similarity coefficient

$$Sim_{\text{HPO}} = \frac{|P_1 \cap P_2|}{|P_1 \cup P_2|},$$

where P_1 and P_2 are the set of the matched HPO terms of the two phenotypes, respectively. We arranged the phenotypes into the 20 disease classes (clusters) based on the

Table 4. Performance of disease gene discovery in leave-one-out cross-validation

Compared methods	Avg. AUC	win/draw/loss (*P*-value)
R-NMTF versus LP	**0.930** versus 0.730	526/1/63 (5.4482e−113)

This table reports the average AUC for disease gene classification, and the pairwise 'win/draw/loss' comparisons of each leave-one-out case between R-NMTF and LP. The last column reports the statistical significance of ranking results using Wilcoxon rank sum test.

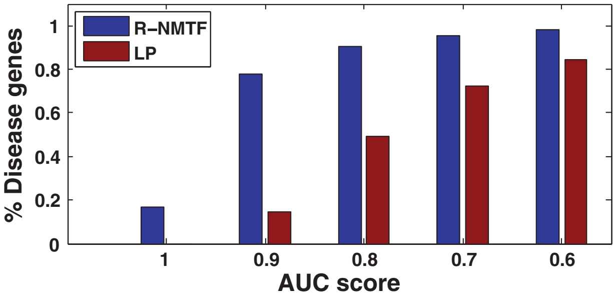

Figure 3. Performance of disease gene discovery in leave-one-out cross-validation. In the plot, the *x*-axis represents AUC cutoffs. The *y*-axis represents the faction of disease genes with a AUC score above the cutoffs. For example, R-NMTF achieved AUCs above 0.9 for >80% of the genes, while LP only achieved the same level of AUC for 20% of the genes.

R-NMTF clustering, and show their HPO similarity by a heat map in Figure 4. There are clearly block structures among the predicted 20 clusters. Most of the phenotypes in the same cluster also share strong HPO similarity. The consistency between the predicted disease clusters and HPO similarities suggest that R-NMTF produced a phenotype clustering supported by HPO annotations. Another interesting observation is that there are also strong HPO similarities between different clusters (i.e. different disease classes share HPO similarities). This may imply that some of the disease classes may share common molecular mechanisms such as skeletal diseases and developmental diseases.

Analysis of new phenotypes in disease classes

Table 5 lists the newly predicted disease phenotypes in the 20 disease classes. Our survey identified supporting literatures for many of the predictions. One interesting finding is faconi anemia (FA) (OMIM:227650), a rare, inherited blood disorder, predicted as a cancer-related disease. Surprisingly, a recent study found that FA could share a common pathogenesis with diseases related with chromosomal instability including cancers, and suggested a possible use of cancer treatment for patients with FA (48). R-NMTF also predicted Proteus syndrome (OMIM:176920) as a cancer-related disease. PTEN, a well-known tumor suppressor gene, is a known causative gene for Proteus syndrome, which may indicate that cancer risk accompanying Proteus syndrome could be increased (49–52). Other interesting newly predicted disease phenotypes are Amyotrophic lateral sclerosis (ALS) (OMIM:105400), also known as Lou Gehrig's disease in neurological disease class, and Gambling, pathologic (OMIM:606349) in psychiatric disease class. ALS is a disease of the nerve cells in the brain and causes unstable muscle movement and Gambling, pathologic is a disabling disorder to fail to resist impulses to gamble, known for frequently co-occur with other psychiatric disorders (85,86). R-NMTF also accurately predicted a few disease phenotypes including juvenile myelomonocytic leukemia and breast cancer which were previously missed in the annotation of the cancer disease class (25). These findings suggest that R-NMTF could correctly classify complex and rare disease phenotypes into their relevant disease classes, which could be used to guide clinical decisions.

Analysis of new member genes in disease pathways

KEGG provides a list of manually curated disease pathways. However, the current knowledge of biological pathways related with diseases is still incomplete and inaccurate, and there are many missing member genes in the disease-related pathways. Table 6 lists the newly predicted member genes in the KEGG disease pathways. Our literature review also identified supporting evidences for many of the predictions. Interesting examples include TMED10 and PRND, which are newly predicted member genes in Alzheimer's pathway and Prion disease pathway, respectively. TMED10 inhibits production of amyloid beta peptides, which is a critical feature of Alzheimers disease and RPND (prion protein 2) is known for that mutations in this gene may lead to neurological disorders. Other examples include EXO1 and ADIPOR1 in colorectal cancer pathway and FGFR3 and FGFR4 in melanoma pathway. Single nucleotide polymorphisms in EXO1 increases risk of colorectal cancer (106,107), and expression of ADIPOR1 is known for involving cancer progression in colorectal cancer (108,109). Mutations in FGFR3 and FGFR4 were previously described in melanoma (121).

We also provide a network view of three examples of disease pathways with addition of the newly predicted member genes in Figure 5. These examples demonstrate that, while KEGG disease pathways were manually curated, there are still missing member genes in the pathways. One example is WNT5A, a newly predicted member gene in the colorectal cancer pathway in Figure 5A. Recent study showed that WNT5A is a potential biomarker for colorectal cancer and could act as tumor suppressor for colorectal cancer by antagonizing the WNT signaling pathway (135). Another example is FGFR3 gene, the newly predicted member gene in the melanoma pathway, in Figure 5B. It has been shown that mutation and overexpression in FGFR3 are associated with survival of melanoma patients (136). However, FGFR3 was not annotated in the melanoma pathway although it is interacting with several members in the pathway. The network views of all the 27 expanded KEGG disease pathways with newly predicted member genes are available at the article's Supplementary Web. These results support that R-NMTF correctly predicted

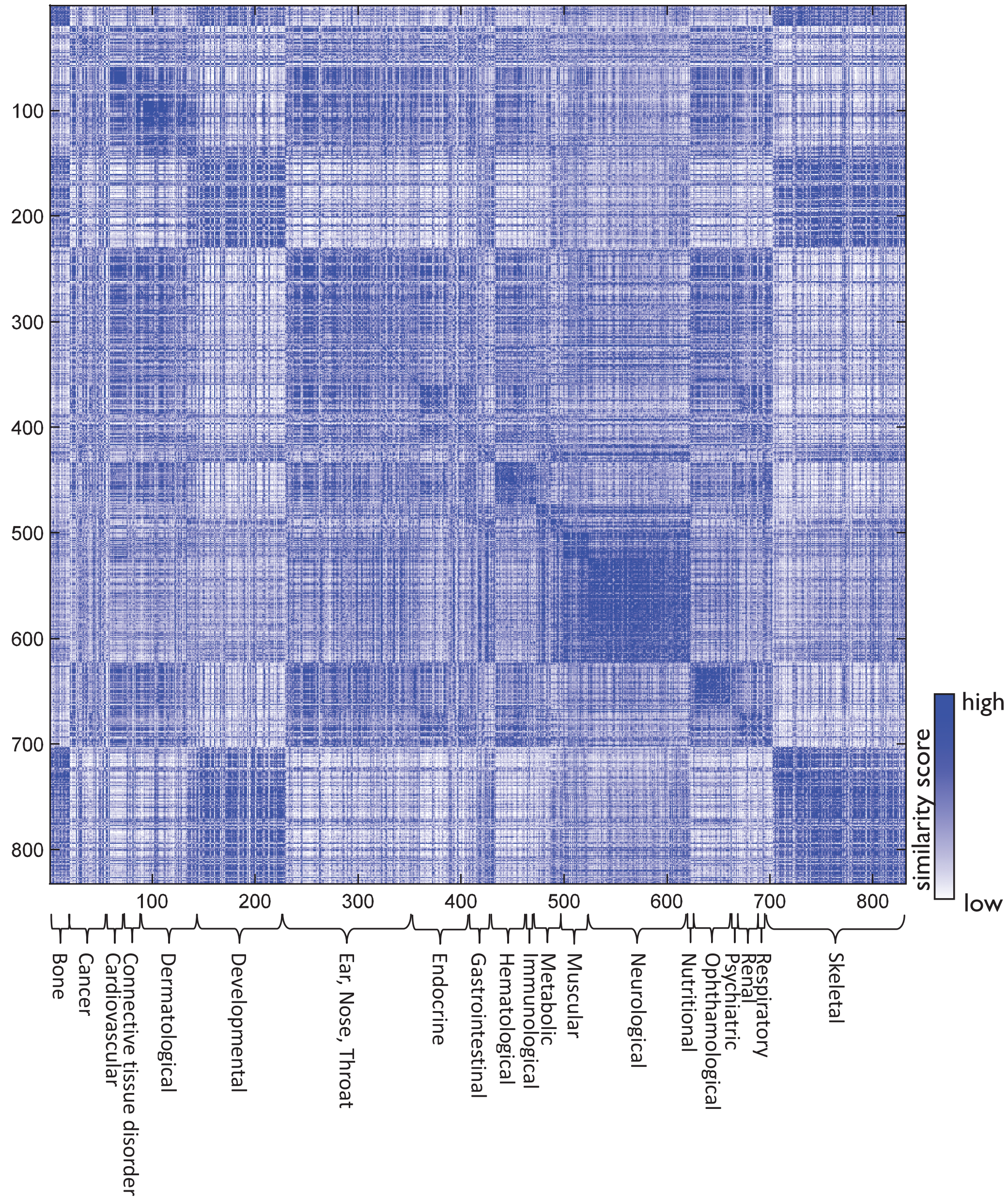

Figure 4. HPO phenotype similarities by clusters. The HPO similarity matrix of the phenotypes are display as a heap map. The phenotypes are grouped into 20 clusters with the disease classes annotated below.

new member genes in several disease-related pathways, and these novel disease genes could play important roles in the disease pathways.

Analysis of predicted disease phenotype cluster–gene cluster associations

We evaluated the predicted disease phenotype cluster–gene cluster associations by a literature survey. We performed two-way hierarchical clustering for the predicted disease phenotype cluster–gene cluster associations. Figure 6 shows the predicted associations between 20 disease phenotype clusters and 200 gene clusters (pathways). Interesting examples are the manually curated KEGG disease pathways. These disease-related pathways include pathways related to cancers, neurological diseases and psychiatric diseases. R-NMTF accurately predicted association between many of these disease-related pathways to the related disease classes. For example, many cancer-related pathways including colorectal, pancreatic, bladder, non-small cell lung, glioma and prostate cancer were correctly identified as cancer pathways. We also identified a set of biological pathways such as apoptosis, p53 signaling and ERBB signaling, hedgehog signaling which are previously known to contribute to tumorigenesis, as well as targets of many anti-cancer drugs (137–141). Other interesting examples are the pathways predicted to be associated to neurological and psychiatric disease classes. Prion disease is one of the well-known rare progressive neurodegenerative

Table 5. New disease phenotypes in 20 disease classes

Disease classes	New disease phenotypes				
Bone	Achondrogenesis, Type III (44)	Canine Teeth (Omim:114600)	Dens Evaginatus (45)	Dental Noneruption (46)	Dentin Dysplasia, Type I(47)
Cancer	Fanconi Anemia (48)	Juvenile Myelomonocytic Leukemia	Breast Cancer	Proteus Syndrome (49,50,51,52)	Bannayan-Riley-Ruvalcaba Syndrome (53,54)
Cardiovascular	Cardiomyopathy (Omim:192600)	Atrial Standstill (55)	Cardiomyopathy, Dilated, 1E	Long Qt Syndrome 3 (56,57)	Sudden Infant Death Syndrome (58)
Connective tissue	Arthritis, Sacroiliac (59)	Spondyloarthropathy (Omim:183840)	Slipped Femoral Capital Epiphyses (60)	Facial Asymmetry (61)	Cervical Rib
Dermatological	Deafness; Dfna3 (62)	Epidermolysis Bullosa (Omim:131800)	Pachyonychia Congenita, Type 1 (63)	Epidermolysis Bullosa Herpetiformis (64)	Epidermolysis Bullosa Simplex, Koebner Type (64)
Developmental	Leucine Transport, High	Uterine Anomalies (65)	Testes, Rudimentary (66)	Oligosynaptic Infertility	Hypospadias, Autosomal (67)
Ear,Nose,Throat	Otosclerosis 3 (68)	Otosclerosis 2 (68)	Otosclerosis 5 (68)	Periodontitis, Aggressive, 2	Red Cell Permeability Defect
Endocrine	Diabetes Mellitus	Hypoglycemia (Omim:601820) (69)	Polycystic Ovary Syndrome 1 (70)	Diabetes Mellitus, Transient Neonatal	Goiter, Multinodular 2 (71)
Gastrointestinal	Cholestasis2 (Omim:605479) (72)	Bile Acid, Synthetic Defect Of	Cholestasis; Pfic2 (Omim:601847) (72)	Cholestasis; Pfic3 (Omim:602347) (72)	Pancreatitis, Hereditary (73)
Hematological	Anemia (74)	Hyperheparinemia	Sideroblastic Anemia, Autosomal (75)	Platelet Groups–ko System	Anemia, Familial Pyridoxine-Responsive (76)
Immunological	Herpesvirus Sensitivity (77)	Interleukin (Omim:243110) (78)	Panbronchiolitis, Diffuse (79)	Immune Deficiency Disease	Allergic Bronchopulmonary Aspergillosis (80)
Metabolic	Immunoglobulin D Level In Plasma	Magnesium, Elevated Red Cell	Flood Factor Deficiency	Citrulline Transport Defect	Amobarbital, Deficient N-Hydroxylation of
Muscular	Palmomental Reflex	Myopathy (Omim:255100)	Muscular Hypoplasia	Pleoconial Myopathy With Salt Craving	Myopathy, Congenital
Neurological	Amyotrophic Lateral Sclerosis 1	Amyotrophic Lateral Sclerosis 2	Alzheimer Disease 2	Prion Disease (Omim:603218)	Frontotemporal Dementia (Omim:607485)
Nutritional	Bulimia Nervosa	Red Cell Permeability Defect	Labia Minora (Omim:149600) (81)	Schizophrenia 9 (82)	Amyotrophic Lateral Sclerosis 6 (83)
Ophthamological	Cone Dystrophy 3	Cone-Rod Dystrophy 3	Leber Congenital Amaurosis	Cone-Rod Dystrophy 6	Retinitis Pigmentosa 19
Psychiatric	Fg Syndrome 2 (86)	Fg Syndrome 3 (84)	Schizophrenia 5	Cerebral Angiopathy, Dysphoric (85,86)	Gambling, Pathologic
Renal	Nephrotic Syndrome, Type 2 (87,88)	Hypertensive Nephropathy (89)	Enuresis, Nocturnal, 2 (90)	Enuresis, Nocturnal, 1 (90)	Blue Diaper Syndrome
Respiratory	Hemangiomatosis	Respiratory Underresponsiveness	Emphysema (Omim:130700)	Asthma, Short Stature, and Elevated Iga	Asthma-Related Traits, Susceptibility To, 1
Skeletal	Brachydactyly, Mononen Type	Tibial Hemimelia (91)	Acropectoral Syndrome	Syndactyly, Type IV	Spondyloepimetaphyseal Dysplasia, Irapa Type

The 5 most confident predictions of phenotypes in each disease class are reported.

Table 6. New member genes of KEGG disease pathways

Kegg disease pathways	New member genes									
Hsa04930: Type II Diabetes Mellitus	KCNJ8 (92)	EFHC1	ADIPOR2 (93)	ABCC9	LDHA	CDH13 (94)	ENSA	CRYBB1	CASR	KCNJ2
Hsa04940: Type I Diabetes Mellitus	CKAP5	SPTBN4	PTPRT	SNX19	CD74	LILRB1 (95)	LILRB2	GAST	LRRC23	CTLA4 (96)
Hsa04950: Maturity Onset Diabetes of the Young	OLIG2	EN2	PCSK1	PNRC1	PCSK2	GATA5	GATA6	PNRC2	OTX2	RAMP2
Hsa05010: Alzheimers Disease	TMED10 (97)	BRI3	PTX3	APH1B (98)	TFCP2 (99)	HRG	C1R	FKBP2	KHSRP	NEDD8 (100)
Hsa05020: Parkinsons Disease	ARIH1 (101)	AMFR	AGXT	TRIM25	CCNB1IP1	GAN	TMCC2	STUB1	SH2D3C	SLC6A1
Hsa05030: ALS	SSR3	JUB	ALS2CL (102)	APBA1	MTMR2	ABL2	HOXB2	RAB37	PKN1 (103)	CHML
Hsa05040: Huntingtons Disease	HIP1R (104)	SNX5	IFT20	PICALM	RPS10	PQBP1	NECAP1	ARF1	KPNA4	MBTPS1
Hsa05050: Dentatorubropallidoluysian Atrophy	ALG13	TRIM22	CLCN5	ECM1	MYST3	NET1	SYNPO	EFEMP1	CPSF6	NDFIP2
Hsa05060: Prion Disease	PRND (106)	CHD6	LAMA2	RPS21	EIF2AK3	KEAP1	ADAM23	DPP6	MOG	OPCML
Hsa05110: Cholera Infection	SERP1	SEC63	ARFIP2	APOB	ARFIP1	PIP5K1A	FLAD1	TRAM1	ETHE1	AP1B1
Hsa05120: Epithelial Cell Signaling in Helicobacter Pylori Infection	GRLF1	ETHE1	HBA1	EFNA2	TOMM34	DARC	ADD2	SH3D19	PFKM	ANG
Hsa05130: Pathogenic Escherichia Coli Infection Ehec	ARPC4	GRM7	HS1BP3	CGN	PLA2G7	KIAA1543	LAPTM4A	NOX4	ACTR2	SSB
Hsa05210: Colorectal Cancer	EXO1 (106,107)	ADIPOR1 (108,109)	MUTYH (111)	PMS2	CDCA8	ROR2	PMS1	MAZ	WNT5A	WNT7A
Hsa05211: Renal Cell Carcinoma	HIF3A	OS9	EGLN2	ING4	ARNTL2	SIM1	ASB8	LRRC41	SENP6	SIM2
Hsa05212: Pancreatic Cancer	REPS1	REPS2	PLCD1	SHFM1	EXOC1	RAD51AP1	RAD54L	RALGPS1	EXOC5	EXOC3
Hsa05213: Endometrial Cancer	MSR1	BRCA2 (112)	NF1	MXI1 (113)	RNASEL	FH	MSH2	ELAC2	MAD1L1	CHEK2
Hsa05214: Glioma	PDAP1	KIAA1683	RHBDF1	RPS18	ART1	BRD2	NKD2	MYO10	TFDP2	SETD8
Hsa05215: Prostate Cancer	KRT27	MTTP	ATF6 (114)	PTHLH (115)	SEMG1	ATF2 (116)	G6PC	NFIL3 (117)	ASGR1	MALL
Hsa05216: Thyroid Cancer	TSSK2	TMOD2	RNF14	TRIM25	PPP4C	IFI16	CNN1	TMOD1	S100A2	NUP98
Hsa05217: Basal Cell Carcinoma	IHH	DHH	ZIC1	ZIC2	PORCN	SFRP1	ROR2	FRMPD4	GPC3	GAS1
Hsa05218: Melanoma	FGFR4 (118)	FGFR2 (119)	PHEX	FGFR3 (120,121)	SCN8A	EBNA1BP2	RPS2	MAPK8IP2	TFEB	PDAP1
Hsa05219: Bladder Cancer	MLC1	UNC5B (122)	UNC5A	PAWR	AATF	TNXB	CAMK2A	RECK	HIST3H2A	ATF4 (123)
Hsa05220: Chronic Myeloid Leukemia	APBA3	MAP4K5 (124)	BAZ2B	KLF3	TDGF1	MAPK4	FMOD	RAI2	ELF2	SPRY2 (125)
Hsa05221 Acute Myeloid Leukemia	RPL21	NDUFB8 (126)	FBXO18	GATA2 (127)	CEBPD (128)	GFI1 (129)	TAF9B	MYST3 (130)	CBFA2T3	NFATC1
Hsa05222: Small Cell Lung Cancer	CKS2	BCKDK	TBC1D8	TNFRSF19	DUSP1	TNFRSF4	TNFRSF12A	NGFRAP1	LTBR	MAP6
Hsa05223: Non Small Cell Lung Cancer	FDXR (131)	LATS1 (132)	MAP6	NR1H2 (133)	PRKRIR	CSN1S1	NR1H3	CNKSR1	FOXG1 (134)	PNRC1

The 10 most confident predictions of member genes in KEGG disease pathways are reported.

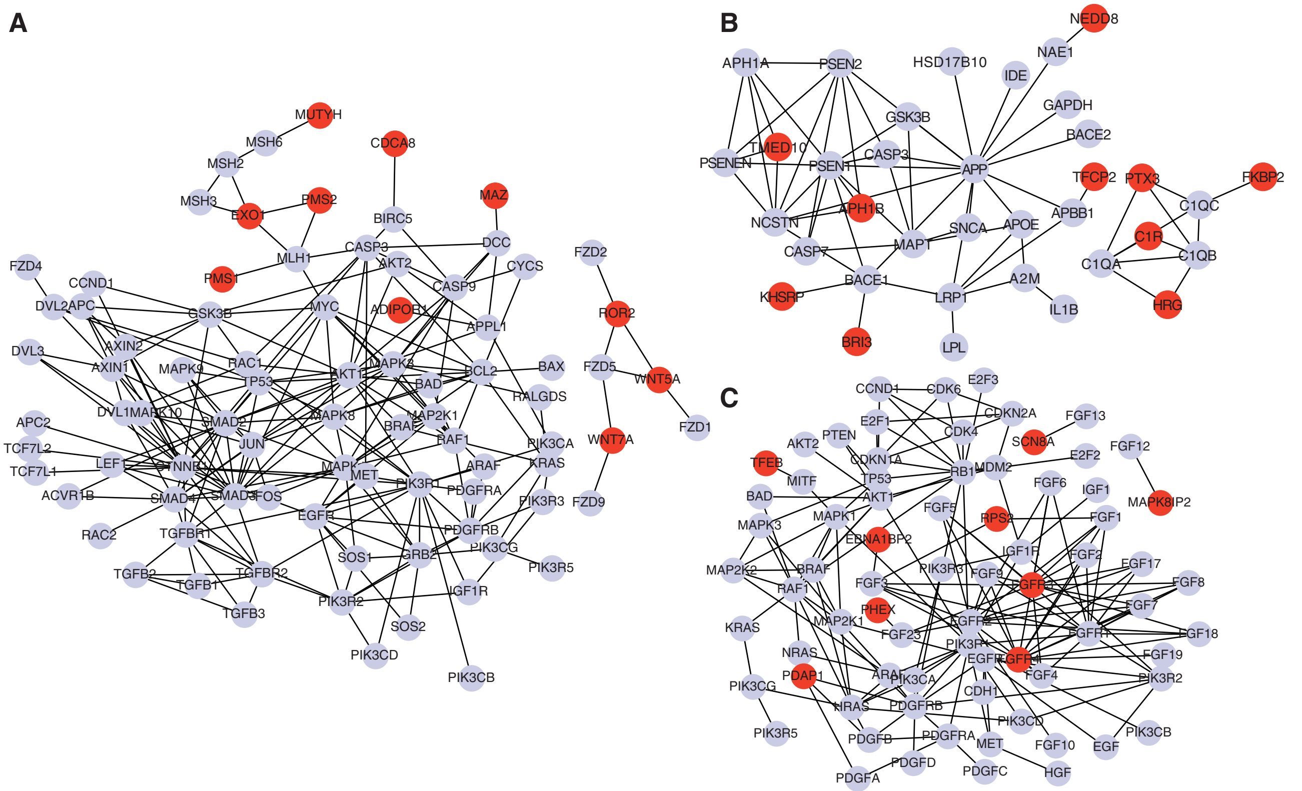

Figure 5. PPI subnetworks of the extended disease pathways. In each pathway, gray nodes are known member genes in the disease pathways and red nodes are newly predicted member genes. Edges represent PPI between two genes. Note that if a known or a newly predicted member gene is not interacting with any other member genes in the pathway, the gene is not included. (**A**) Colorectal cancer pathway. The predicted colorectal cancer genes EXO1 and ADIPOR1 are interacting with many other genes in the colorectal cancer pathway. (**B**) Alzheimer pathway. Over-expression of C1R is known for involving alzheimer disease. (**C**) Melanoma pathway. Mutation and copy number changes in new member gene FGFR3 were recently discovered in melanoma.

disorders that affect both humans and animals. R-NMTF accurately predicted the prion disease pathway as one of the pathways associated with neurological disease class. MAPK pathway is predicted to be associated with both neurological and psychiatric disease classes. Recent study reported that activation of MAPK pathway could play a role in alzheimer and psychiatric disorders such as increasing anxiety and depression and schizophrenia etc. (142,143). R-NMTF also correctly predicted Huntington's disease pathway to be associated with neurological and psychiatric diseases.

DISCUSSION

The number of documented disease phenotypes and phenotype–gene associations increases quickly. Since 2007, the number of OMIM disease–gene associations is nearly doubled. These determined associations provide valuable resources not only for predicting novel associations but also for understanding disease phenotypes. Our research work in the article explored this possibility and reported promising results. Recently, phenotype databases have been proposed and in the progress of becoming comprehensive and systematic for many species. R-NMTF will be a useful model for analyzing the new 'phenomes'. Moreover, R-NMTF also identifies pathways associated with disease phenotype clusters. Since many drugs are developed to target proteins that act in disease-related pathways, precise identification of members of disease pathways could accelerate the development of more efficient targeted therapies, as well as improve understanding of the molecular mechanisms underlying complex human diseases. More recently, cross-species phenotype–gene association analysis based on ortholog genes and similar phenotypes has been performed (144). An interesting future direction is to extend R-NMTF to perform cross-species phenome–genome co-clustering. It is also possible to apply other advanced machine learning models to integrate the phenotype similarity network and the PPI network with phenotype–gene association data for co-clustering phenotypes and genes. More refined modeling might lead to further improvement in phenotype classification and disease–gene discovery.

Previously, regularized NMTF models were only proposed for applications in image and document classification. Gu and Zhou (145) introduced a dual regularized co-clustering (DRCC), which extended NMTF by incorporating the graph Laplacian as additional regularizations in the objective function. DRCC was applied to classify images, documents and newsgroups. Zh vang *et al.* (38) introduced a matrix tri-factorization-based classification framework (MTrick) for transfer learning. MTrick first learns an association matrix from source domain by

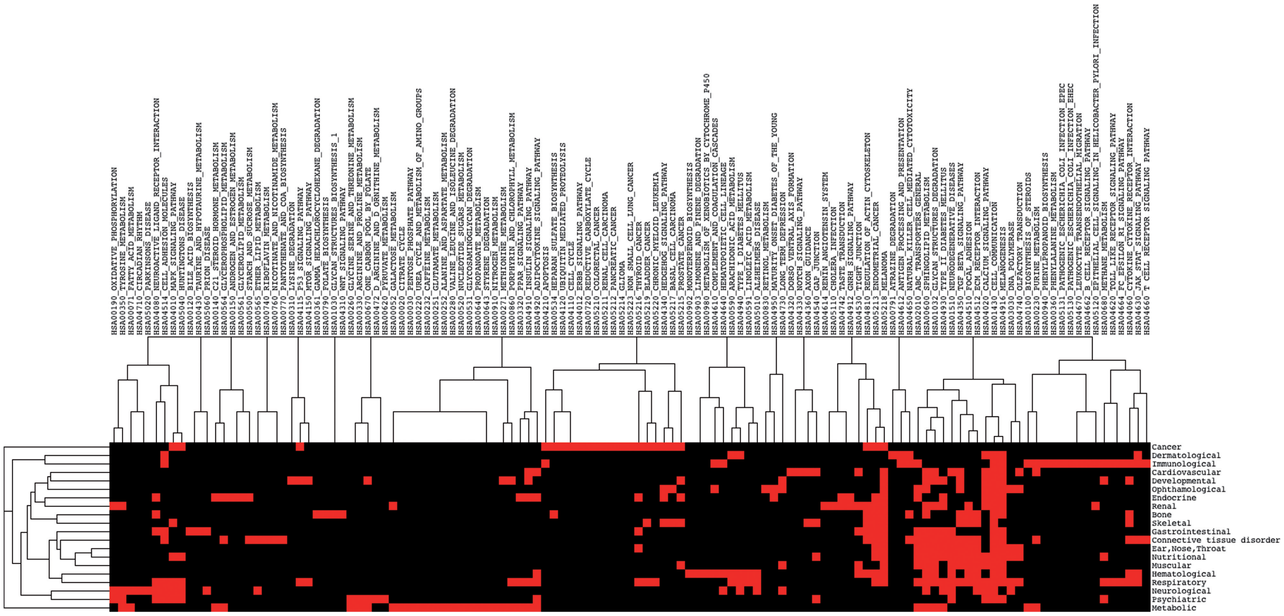

Figure 6. Predicted associations between disease classes and pathways. Each red entry represents a predicted association between 20 disease classes and 200 KEGG pathways.

performing non-negative tri-factorization and use incorporates inferred association matrix S from source domain into non-negative tri-factorization for target domain classification. R-NMTF introduces regularization terms for label information from both phenotype and gene clusters, and thus R-NMTF is a supervised co-clustering method while DRCC is unsupervised. Compared with MTrick, which only uses label information, R-NMTF incorporates the prior knowledge in phenotype similarity network and PPI networks to cluster phenotypes and genes with tri-matrix factorization. To our best knowledge, no previous NMF-based model has been applied to clustering phenotypes or analyzing disease phenotype–gene associations. R-NMTF is an advanced model which integrates phenome, genome and interactome information for both problems.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2 and Supplementary Figure 1.

FUNDING

Funding for open access charge: National Science Foundation [III-1117153].

Conflict of interest statement. None declared.

REFERENCES

1. McKusick,V. (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.
2. Groth,P. and Weiss,B. (2006) Phenotype Data: a Neglected Resource in Biomedical Research? *Curr. Bioinformatics*, **1**, 347–358.
3. Sawyers,C. (2008) The cancer biomarker problem. *Nature*, **452**, 548–552.
4. Rubin,E. (2008) Genomics of cellulosic biofuels. *Nature*, **454**, 841–845.
5. Edwards,D. and Batley,J. (2004) Plant bioinformatics: from genome to phenome. *Trends Biotechnol*, **22**, 232–237.
6. The Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
7. van't Veer,L. and Bernards,R. (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, **452**, 564–570.
8. Johnson,A. and O'Donnell,C. (2009) An open access database of genome-wide association resutls. *BMC Med. Gent.*, **6**
9. Shlien,A. and Malkin,D. (2009) Copy number variations and cancer. *Genome Med.*, **1**, 62.
10. Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotech.*, **26**, 1135–1145.
11. Rothberg,J. and Leamon,J. (2008) The development and impact of 454 sequencing. *Nat. Biotechnol.*, **26**, 1117–1124.
12. Franke,L., van Bakel,H., Fokkens,L., de Jong,E., Egmont-Petersen,M. and Wijmenga,C. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.
13. Kohler,S., Bauer,S., Horn,D. and Robinson,P. (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.
14. Wu,X., Jiang,R., Zhang,M. and Li,S. (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol.*, **4**, 189.
15. Linghu,B., Snitkin,E., Hu,Z., Xia,Y. and Delisi,C. (2009) Genome-wide prioritization of disease genes and identification of disease–disease associations from an integrated human functional linkage network. *Genome Biol.*, **10**, R91.
16. Hwang,T. and Kuang,R. (2010) A Heterogeneous Label Propagation Algorithm for Disease Gene Discovery. *Proc of SIAM International Conference on Data Mining*, pp. 583–594.
17. Vanunu,O., Magger,O., Ruppin,E., Shlomi,T. and Sharan,R. (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.*, **6**, e1000641.
18. Li,Y. and Patra,J. (2010) Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, **26**, 1219–1224.

19. Navlakha,S. and Kingsford,C. (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, **26**, 1057–1063.
20. Loscalzo,J., Kohane,I. and Barabasi,A. (2007) Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol. Syst. Biol*., **3**, 124.
21. van Driel,M., Bruggeman,J., Vriend,G., Brunner,H. and Leunissen,J. (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet*., **14**, 535–542.
22. Freimer,N. and Sabatti,C. (2003) The human phenome project. *Nat. Genet*., **34**, 15–21.
23. Scriver,C. (2004) After the genome–the phenome? *J. Inherit. Metab. Dis*., **27**, 305–317.
24. Groth,P., Kalev,I., Kirov,I., Traikov,B., Leser,U. and Weiss,B. (2010) Phenoclustering: online mining of cross-species phenotypes. *Bioinformatics*, **26**, 1924–1925.
25. Goh,K.-I., Cusick,M.E., Valle,D., Childs,B., Vidal,M. and Barabási,A.-L. (2007) The human disease network. *Proc Natl Acad. Sci. USA*, **104**, 8685–8690.
26. Barabasi,A., Gulbahce,N. and Loscalzo,J. (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet*., **12**, 56–68.
27. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*, **32**, D277–D280.
28. Lee,D. and Seung,H. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.
29. Lee,D.D. and Seung,H.S. (2000) *The proceeding of Neural Information Processing Systems*. MIT Press, Denver, CO, USA, pp. 556–562.
30. Brunet,J., Tamayo,P., Golub,T. and Mesirov,J. (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, **101**, 4164–4169.
31. Kim,H., Park,H. and Drake,B. (2007) Extracting unrecognized gene relationships from the biomedical literature via matrix factorizations. *BMC Bioinformatics*, **8(Suppl 9)**, S6.
32. Schachtner,R., Lutter,D., Knollmuller,P., Tome,A., Theis,F., Schmitz,G., Stetter,M., Vilda,P. and Lang,E. (2008) Knowledge-based gene expression classification via matrix factorization. *Bioinformatics*, **24**, 1688–1697.
33. Kim,H. and Park,H. (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, **23**, 1495–1502.
34. Inamura,K., Fujiwara,T., Hoshida,Y., Isagawa,T., Jones,M., Virtanen,C., Shimane,M., Satoh,Y., Okumura,S., Nakagawa,K. *et al*. (2005) Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. *Oncogene*, **24**, 7105–7113.
35. Greene,D., Cagney,G., Krogan,N. and Cunningham,P. (2008) Ensemble non-negative matrix factorization methods for clustering protein–protein interactions. *Bioinformatics*, **24**, 1722–1728.
36. Ding,C., Li,T., Peng,W. and Park,H. (2006) *Proceeding of the 12th ACM International Conference on Knowledge Discovery and Data Mining*. ACM (Association for Computing Machinery), Philadelphia, PA, USA, pp. 126–135.
37. Li,T., Sindhwani,V., Ding,C. and Zhang,Y. (2010) *SIAM Conference on Data Mining*. SIAM (Society for Industrial and Applied Mathematics), Columbus, Ohio, USA, pp. 293–302.
38. Zhuang,F., Luo,P., Xiong,H., He,Q., Xiong,Y. and Shi,Z. (2010) *SIAM Conference on Data Mining*. Press, pp. 13–24.
39. Chung,F. (1997) Spectral graph theory. In: *Regional Conference Series in Mathematics*, Vol. 92. American Mathematical Society, Providence, RI, Ann Arbor, MI.
40. Higgins,M., Claremont,M., Major,J., Sander,C. and Lash,A. (2007) CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res*, **35**, D721–D726.
41. Peri,S., Navarro,J., Amanchy,R., Kristiansen,T., Jonnalagadda,C., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T., Gronborg,M. *et al*. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, **13**, 2363–2371.
42. Zhou,D., Bousquet,O., Lal,T.N., Weston,J. and Schlkopf,B. (2004) In *Advances in Neural Information Processing Systems 16*. MIT Press, pp. 321–328.
43. Robinson,P. and Mundlos,S. (2010) The human phenotype ontology. *Clin. Genet*., **77**, 525–534.
44. Hall,C.M. (2002) International nosology and classification of constitutional disorders of bone (2001). *Am. J. Med. Genet*., **113**, 65–77.
45. McCulloch,K.J., Mills,C.M., Greenfeld,R.S. and Coil,J.M. (1997) Dens evaginatus from an orthodontic perspective: report of several clinical cases and review of the literature. *Am. J. Orthod. Dentofacial Orthop*., **112**, 670–675.
46. Marks,S.C. Jr and Cahill,D.R. (1987) Regional control by the dental follicle of alterations in alveolar bone metabolism during tooth eruption. *J. Oral Pathol. Med*., **16**, 164–169.
47. Morris,M.E. and Augsburger,R.H. (1977) Dentine dysplasia with sclerotic bone and skeletal anomalies inherited as an autosomal dominant trait: a new syndrome. *Oral Surg. Oral Med. Oral Patho*., **43**, 267–283.
48. D'Andrea,A.D. (2010) Susceptibility pathways in fanconi's anemia and breast cancer. *N. Engl J. Med*., **362**, 1909–1919.
49. Yilmaz,A., Hamel,N., Schwartz,C.E., Houlston,R.S., Harper,J.I. and Foulkes,W.D. (2010) A genome-wide analysis of loss of heterozygosity and chromosomal copy number variation in Proteus syndrome using high-density SNP microarrays. *J. Hum. Genet*., **55**, 627–630.
50. Bahcall,O. (2011) Proteus syndrome exomes. *Nat. Genet*., **43**, 824–824.
51. Lindhurst,M.J., Sapp,J.C., Teer,J.K., Johnston,J.J., Finn,E.M., Peters,K., Turner,J., Cannons,J.L., Bick,D., Blakemore,L. *et al*. (2011) A Mosaic Activating Mutation in AKT1 Associated with the Proteus Syndrome. *New Engl. Jo. Med*., **365**, 611–619.
52. Zbuk,K.M. and Eng,C. (2007) Cancer phenomics: RET and PTEN as illustrative models. *Nat. Rev. Cancer*., **7**, 35–45.
53. Pilarski,R., Stephens,J.A., Noss,R., Fisher,J.L. and Prior,T.W. (2011) Predicting PTEN mutations: an evaluation of Cowden syndrome and bannayanrileyruvalcaba syndrome clinical features. *J. Med. Genet*., **48**, 505–512.
54. Patnaik,M.M., Raza,S.S., Khambatta,S., Stanich,P.P. and Goetz,M.P. (2010) Oncophenotypic Review and Clinical Correlates of Phosphatase and Tensin Homolog on Chromosome 10 Hamartoma Tumor Syndrome. *J. Clini. Oncol*., **28**, e767–e768.
55. Groenewegen,W.A., Firouzi,M., Bezzina,C.R., Vliex,S., van Langen,I.M., Sandkuijl,L., Smits,J.P., Hulsbeek,M., Rook,M.B., Jongsma,H.J. *et al*. (2003) A Cardiac Sodium Channel Mutation Cosegregates With a Rare Connexin40 Genotype in Familial Atrial Standstill. *Circ. Res*., **92**, 14–22.
56. Roden,D.M. (2008) Long-QT Syndrome. *New Engl. J. Med*., **358**, 169–176.
57. Crotti,L., Monti,M.C., Insolia,R., Peljto,A., Goosen,A., Brink,P.A., Greenberg,D.A., Schwartz,P.J. and George,A.L. (2009) NOS1AP Is a Genetic Modifier of the Long-QT Syndrome. *Circulation*, **120**, 1657–1663.
58. Horne,R.S.C., Witcombe,N.B., Yiallourou,S.R., Scaillet,S., Thiriez,G. and Franco,P. (2010) Cardiovascular control during sleep in infants: implications for Sudden Infant Death Syndrome. *Sleep Medicine*, **11**, 615–621.
59. Bennett,R.M. and O'Connell,D.J. (1978) The arthritis of mixed connective tissue disease. *Ann. Rheumatic Diseases*, **37**, 397–403.
60. Mcafee,P.C. and Cady,R.B. (1983) Endocrinologic and Metabolic Factors in Atypical Presentations of Slipped Capital Femoral Epiphysis: Report of Four Cases and Review of the Literature. *Clin. Orthopa. Relat Res*., **180**.
61. Kale,A., Sah,K., Rastogi,P., Awasthi,S. and Chandra,S. (2011) Traumatic pseudolipoma causing facial asymmetry: an uncommon pathology and review of its pathogenesis. *J. Oral Maxillofacial Pathol*., **15**, 113–115.
62. Lai-Cheong,J.E., Arita,K. and McGrath,J.A. Genetic Diseases of Junctions. *J. Invest Dermatol*, **127**, 2713–2725.
63. Leachman,S.A., Kaspar,R.L., Fleckman,P., Florell,S.R., Smith,F.J.D., McLean,W.H.I., Lunny,D.P., Milstone,L.M., van Steensel,M.A.M., Munro,C.S. *et al*. (2005) Clinical and Pathological Features of Pachyonychia Congenita. *J. Investig. Dermatol. Symp. Proc*., **10**, 3–17.

64. Yaoita,H., Briggaman,R.A., Lawley,T.J., Provost,T.T. and Katz,S.I. (1981) Epidermolysis Bullosa Acquisita: Ultrastructural and Immunological Studies. *J. Investig. Dermatol.*, **76**, 288–292.
65. Saravelos,S.H., Cocksedge,K.A. and Li,T.-C. (2008) Prevalence and diagnosis of congenital uterine anomalies in women with reproductive failure: a critical appraisal. *Hum. Reproduct. Update*, **14**, 415–429.
66. Bergada,C., Cleveland,W.W., Jones,H.W. and Wilkins,L. (1962) Variants of embryonic testicular dysgenesis: Bilateral anorchia and the syndrome of rudimentary testes. *Acta Endocrinologica.*, **40**, 521–536.
67. Wang,M.-H. and Baskin,L.S. (2008) Endocrine Disruptors, Genital Development, and Hypospadias. *J. Androl.*, **29**, 499–505.
68. Ruedi,L. (1963) Pathogenesis of Otosclerosis. *Arch. Otolaryngol.*, **78**, 469–477.
69. Gerich,J.E., Mokan,M., Veneman,T., Korytkowski,M. and Mitrakou,A. (1991) Hypoglycemia Unawareness. *Endocrine Rev.*, **12**, 356–371.
70. Franks,S. (1995) Polycystic Ovary Syndrome. *New Eng. J. Med.*, **333**, 853–861.
71. Krohn,K., Fhrer,D., Bayer,Y., Eszlinger,M., Brauer,V., Neumann,S. and Paschke,R. (2005) Molecular Pathogenesis of Euthyroid and Toxic Multinodular Goiter. *Endocrine Rev.*, **26**, 504–524.
72. Mcrae,C.A., Prince,M.I., Hudson,M., Day,C.P., James,O.F.W. and Jones,D.E.J. (2003) Pain as a complication of use of opiate antagonists for symptom control in cholestasis. *Gastroenterology*, **125**, 591–596.
73. Lowenfels,A.B., Maisonneuve,P., DiMagno,E.P., Elitsur,Y., Gates,L.K., Perrault,J. and Whitcomb,D.C. (1997) Hereditary Pancreatitis and the Risk of Pancreatic Cancer. *J. Nat. Cancer Instit.*, **89**, 442–446.
74. Jayabose,S., Tugal,O., Sandoval,C., Patel,P., Puder,D., Lin,T. and Visintainer,P. (1996) Clinical and hematologic effects of hydroxyurea in children with sickle cell anemia. *The Journal of Pediatrics*, **129**, 559–565.
75. Fiske,D.N., McCoy,H.E. and Kitchens,C.S. (1994) Zinc-induced sideroblastic anemia: report of a case, review of the literature, and description of the hematologic syndrome. *Ame. J. Hematol.*, **46**, 147–150.
76. Bourne,M.S., Elves,M.W. and Israls,M.C.G. (1965) Familial Pyridoxine-Responsive Anaemia. *Brit. J. Haematol.*, **11**, 1–10.
77. Enbom,M., Wang,F.-Z., Fredrikson,S., Martin,C., Dahl,H. and Linde,A. (1999) Similar Humoral and Cellular Immunological Reactivities to Human Herpesvirus 6 in Patients with Multiple Sclerosis and Controls. *Clin. Diagn. Lab. Immunol.*, **6**, 545–549.
78. Durum,S.K., Schmidt,J.A. and Oppenheim,J.J. (1985) Interleukin 1: an Immunological Perspective. *Annu. Rev. Immunol.*, **3**, 263–287.
79. Iwata,M., Colby,T.V. and Kitaichi,M. (1994) Diffuse panbronchiolitis: diagnosis and distinction from various pulmonary diseases with centrilobular interstitial foam cell accumulations. *Human Pathol.*, **25**, 357–363.
80. Rosenberg,M., Patterson,R., Mintzer,R., Cooper,B.J., Roberts,M. and Harris,K.E. (1977) Clinical and Immunologic Criteria for the Diagnosis of Allergic Bronchopulmonary Aspergillosis. *Ann. Inter. Medi.*, **86**, 405–414.
81. Fiorillo,L. (2004) Therapy of pediatric genital diseases. *Dermatologic Therapy*, **17**, 117–128.
82. Malcolm,P. (2004) Nutrition and schizophrenia: beyond omega-3 fatty acids. *Prostaglandins Leukot. Essent. Fatty Acids*, **70**, 417–422.
83. Desport,J.C., Preux,P.M., Magy,L., Boirie,Y., Vallat,J.M., Beaufrre,B. and Couratier,P. (2001) Factors correlated with hypermetabolism in patients with amyotrophic lateral sclerosis. *Am. J. Clin. Nutr.*, **74**, 328–334.
84. Graham,J.M., Clark,R.D., Moeschler,J.B. and Rogers,R.C. (2010) Behavioral features in young adults with FG syndrome (OpitzKaveggia syndrome). *Am. J. Med. Genet. C Seminar Med. Genet.*, **154C**, 477–485.
85. Petry,N.M., Stinson,F.S. and Grant,B.F. (2005) Comorbidity of DSM-IV Pathological Gambling and Other Psychiatric Disorders: results From the National Epidemiologic Survey on Alcohol and Related Conditions. *J. Clin. Psychiatr.*, **66**, 564–574.
86. Blaszczynski,A. and Nower,L. (2002) A pathways model of problem and pathological gambling. *Addiction*, **97**, 487–499.
87. Blainey,J., Brewer,D., Hardwicke,J. and Soothill,J. (1960) The nephrotic syndrome. Diagnosis by renal biopsy and biochemical and immunological analyses related to the response to steroid therapy. *Quart. j. Med.*, **29**, 235–256.
88. Boute,N., Gribouval,O., Roselli,S., Benessy,F., Lee,H., Fuchshuber,A., Dahan,K., Gubler,M.-C., Niaudet,P. and Antignac,C. (2000) NPHS2, encoding the glomerular protein podocin, is mutated in autosomal recessive steroid-resistant nephrotic syndrome. *Nat. Genet.*, **24**, 349–354.
89. Brenner,B.M., Cooper,M.E., de Zeeuw,D., Keane,W.F., Mitch,W.E., Parving,H.-H., Remuzzi,G., Snapinn,S.M., Zhang,Z. and Shahinfar,S. (2001) Effects of Losartan on Renal and Cardiovascular Outcomes in Patients with Type 2 Diabetes and Nephropathy. *New Eng. J. Med.*, **345**, 861–869.
90. Natochin,Y.V. and Kuznetsova,A.A. (2000) Nocturnal enuresis: correction of renal function by desmopressin and diclofenac. *Pediatr. Nephrol.*, **14**, 42–47, 10.1007/s004670050011.
91. McKay,M., Clarren,S.K., Zorn,R. and Opitz,J.M. (1984) Isolated tibial hemimelia in sibs: an autosomal-recessive disorder? *Am. J. Med. Genet.*, **17**, 603–607.
92. Koehn,J., Fountoulakis,M. and Krapfenbauer,K. (2008) Multiple drug resistance associated with function of ABC-transporters in diabetes mellitus: molecular mechanism and clinical relevance. *Infect. Disorders Drug Targets (Formerly Current Drug Targets-Infectious*, **8**, 109–118.
93. Heilbronn,L., Smith,S. and Ravussin,E. (2004) Failure of fat cell proliferation, mitochondrial function and fat oxidation results in ectopic fat storage, insulin resistance and type II diabetes mellitus. *Int. J. Obe.*, **28**, S12–S21.
94. Andersson,L., Petersen,G. and Ståhl,F. (2009) Ranking candidate genes in rat models of type 2 diabetes. *Theor. Biol. Med. Mode.*, **6**, 12.
95. Planas,R., Carrillo,J., Sanchez,A., Ruiz de Villa,M., Nuñez,F., Verdaguer,J., James,R., Pujol-Borrell,R. and Vives-Pi,M. (2010) Gene expression profiles for the human pancreas and purified islets in Type 1 diabetes: new findings at clinical onset and in long-standing diabetes. *Clini. Exp. Immunol.*, **159**, 23–44.
96. Donner,H., Rau,H., Walfish,P., Braun,J., Siegmund,T., Finke,R., Herwig,J., Usadel,K. and Badenhoop,K. (1997) CTLA4 alanine-17 confers genetic susceptibility to Graves' disease and to type 1 diabetes mellitus. *J. Clin. Endocrinol. Metab.*, **82**, 143–146.
97. Haygood,R., Fedrigo,O., Hanson,B., Yokoyama,K. and Wray,G. (2007) Promoter regions of many neural-and nutrition-related genes have experienced positive selection during human evolution. *Nat. Genet.*, **39**, 1140–1144.
98. Tanzi,R. and Bertram,L. (2005) Twenty years of the Alzheimers disease amyloid hypothesis: a genetic perspective. *Cell*, **120**, 545–555.
99. Carter,C. (2007) Convergence of genes implicated in Alzheimer's disease on the cerebral cholesterol shuttle: app, cholesterol, lipoproteins, and atherosclerosis. *Neurochem. Int.*, **50**, 12–38.
100. De Vrij,F., Fischer,D., Van Leeuwen,F. and Hol,E. (2004) Protein quality control in Alzheimer's disease by the ubiquitin proteasome system. *Prog. neurobiol.*, **74**, 249–270.
101. Parelkar,S., Cadena,J., Kim,C., Wang,Z., Sugal,R., Bentley,B., Moral,L., Ardley,H. and Schwartz,L. (2012) The Parkin-Like Human Homolog of Drosophila Ariadne-1 (HHARI) Can Induce Aggresome Formation in Mammalian Cells and Is Immunologically Detectable in Lewy Bodies. *J. Mol. Neurosci.*, **46**, 109–121, 10.1007/s12031-011-9535-1.
102. Hadano,S., Otomo,A., Suzuki-Utsunomiya,K., Kunita,R., Yanagisawa,Y., Showguchi-Miyata,J., Mizumura,H. and Ikeda,J.-E. (2004) ALS2CL, the novel protein highly homologous to the carboxy-terminal half of ALS2, binds to Rab5 and modulates endosome dynamics. *FEBS Lett.*, **575**, 64–70.
103. Manser,C., Stevenson,A., Banner,S., Davies,J., Tudor,E.L., Ono,Y., Nigel˜Leigh,P., McLoughlin,D.M., Shaw,C.E. and Miller,C.C.J. (2008) Deregulation of PKN1 activity disrupts neurofilament organisation and axonal transport. *FEBS Lett.*, **582**, 2303–2308.

104. Borrell-Pags,M., Zala,D., Humbert,S. and Saudou,F. (2006) Huntingtons disease: from huntingtin function and dysfunction to therapeutic strategies. *Cell. Mol. Life Sci.*, **63**, 2642–2660, 10.1007/s00018-006-6242-0.

105. Peoc'h,K., Guérin,C., Brandel,J., Launay,J. and Laplanche,J. (2000) First report of polymorphisms in the prion-like protein gene (PRND): implications for human prion diseases. *Neurosc. lett.*, **286**, 144–148.

106. Wu,Y., Berends,M., Post,J., Mensink,R., Verlind,E., Van Der Sluis,T., Kempinga,C., Sijmons,R., Van Der Zee,A., Hollema,H. *et al.* (2001) Germline mutations of EXO1 gene in patients with hereditary nonpolyposis colorectal cancer (HNPCC) and atypical HNPCC forms. *Gastroenterology*, **120**, 1580–1587.

107. Yamamoto,H., Hanafusa,H., Ouchida,M., Yano,M., Suzuki,H., Murakami,M., Aoe,M., Shimizu,N., Nakachi,K. and Shimizu,K. (2005) Single nucleotide polymorphisms in the EXO1 gene and risk of colorectal cancer in a Japanese population. *Carcinogenesis*, **26**, 411–416.

108. Kaklamani,V., Wisinski,K., Sadim,M., Gulden,C., Do,A., Offit,K., Baron,J., Ahsan,H., Mantzoros,C. and Pasche,B. (2008) Variants of the adiponectin (ADIPOQ) and adiponectin receptor 1 (ADIPOR1) genes and colorectal cancer risk. *JAMA*, **300**, 1523.

109. Byeon,J.-S., Jeong,J.-Y., Kim,M.J., Lee,S.-M., Nam,W.-H., Myung,S.-J., Kim,J.G., Yang,S.-K., Kim,J.-H. and Suh,D.J. (2010) Adiponectin and adiponectin receptor in relation to colorectal cancer progression. *Inte. J. Cancer*, **127**, 2758–2767.

110. Sieber,O., Lipton,L., Crabtree,M., Heinimann,K., Fidalgo,P., Phillips,R., Bisgaard,M., Orntoft,T., Aaltonen,L., Hodgson,S. *et al.* (2003) Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. *New Engl. J. Med.*, **348**, 791–799.

111. De Jong,A., Van Puijenbroek,M., Hendriks,Y., Tops,C., Wijnen,J., Ausems,M., Meijers-Heijboer,H., Wagner,A., Van Os,T., Bröcker-Vriends,A. *et al.* (2004) Microsatellite instability, immunohistochemistry, and additional PMS2 staining in suspected hereditary nonpolyposis colorectal cancer. *Clin. Cancer Res.*, **10**, 972–980.

112. Beiner,M.E., Finch,A., Rosen,B., Lubinski,J., Moller,P., Ghadirian,P., Lynch,H.T., Friedman,E., Sun,P. and Narod,S.A. (2007) The risk of endometrial cancer in women with brca1 and brca2 mutations. a prospective study. *Gynecol. Oncol.*, **104**, 7–10.

113. Yoo,K.H., Sung,Y.H., Yang,M.H., Jeon,J.O., Yook,Y.J., Woo,Y.M., Lee,H.-W. and Park,J.H. (2007) Inactivation of Mxi1 induces Il-8 secretion activation in polycystic kidney. *Biochem. Biophys. Res. Commun.*, **356**, 85–90.

114. Zu,K., Bihani,T., Lin,A., Park,Y.-M., Mori,K. and Ip,C. (2005) Enhanced selenium effect on growth arrest by BiP//GRP78 knockdown in p53-null human prostate cancer cells. *Oncogene*, **25**, 546–554.

115. Wu,Y.-M., Robinson,D.R. and Kung,H.-J. (2004) Signal Pathways in Up-regulation of Chemokines by Tyrosine Kinase MER/NYK in Prostate Cancer Cells. *Cancer Res.*, **64**, 7311–7320.

116. Deng,X., Liu,H., Huang,J., Cheng,L., Keller,E.T., Parsons,S.J. and Hu,C.-D. (2008) Ionizing Radiation Induces Prostate Cancer Neuroendocrine Differentiation through Interplay of CREB and ATF2: Implications for Disease Progression. *Cancer Res.*, **68**, 9663–9670.

117. Zhao,H., Whitfield,M.L., Xu,T., Botstein,D. and Brooks,J.D. (2004) Diverse Effects of Methylseleninic Acid on the Transcriptional Program of Human Prostate Cancer Cells. *Mol. Biol. Cell.*, **15**, 506–519.

118. Streit,S., Mestel,D., Schmidt,M., Ullrich,A. and Berking,C. (2006) FGFR4 Arg388 allele correlates with tumour thickness and FGFR4 protein expression with survival of melanoma patients. *Brit. Jo. cancer*, **94**, 1879–1886.

119. Gartside,M., Curtis,A., Yudt,L., Harper,U., Bengston,A., Pavey,S., Tuthill,R., Bastian,B., Meltzer,P. and Pollock,P. (2005) *AACR Meeting Abstracts*, **2005**, 608.

120. Bloethner,S., Mould,A., Stark,M. and Hayward,N. (2008) Identification of arhgef17, dennd2d, fgfr3, and rb1 mutations in melanoma by inhibition of nonsense-mediated mRNA decay. *Genes, Chromosomes and Cancer*, **47**, 1076–1085.

121. Bloethner,S., Mould,A., Stark,M. and Hayward,N.K. (2008) Identification of arhgef17, dennd2d, fgfr3, and rb1 mutations in melanoma by inhibition of nonsense-mediated mRNA decay. *Genes, Chromosomes and Cancer*, **47**, 1076–1085.

122. Koed,K., Wiuf,C., Christensen,L., Wikman,F., Zieger,K., Møller,K., von˜der Maase,H. and Ørntoft,T. (2005) High-density single nucleotide polymorphism array defines novel stage and location-dependent allelic imbalances in human bladder tumors. *Cancer Res.*, **65**, 34.

123. Buytaert,E., Matroule,J.Y., Durinck,S., Close,P., Kocanova,S., Vandenheede,J.R., de Witte,P.A., Piette,J. and Agostinis,P. (2007) Molecular effectors and modulators of hypericin-mediated cell death in bladder cancer cells. *Oncogene*, **27**, 1916–1929.

124. Diaz-Blanco,E., Bruns,I., Neumann,F., Fischer,J.C., Graef,T., Rosskopf,M., Brors,B., Pechtel,S., Bork,S., Koch,A. *et al.* (2007) Molecular signature of CD34+ hematopoietic stem and progenitor cells of patients with CML in chronic phase. *Leukemia*, **21**, 494–504.

125. Steelman,L.S., Franklin,R.A., Abrams,S.L., Chappell,W., Kempf,C.R., Basecke,J., Stivala,F., Donia,M., Fagone,P., Nicoletti,F. *et al.* (2011) Roles of the Ras/Raf/MEK/ERK pathway in leukemia therapy. *Leukemia*, **25**, 1080–1094.

126. Schmidt,S. and Wolf,D. (2009) Role of gene-expression profiling in chronic myeloid leukemia. *Expert Rev. Hematol.*, **2**, 93–103.

127. Zhang,S., Ma,L., Huang,Q., Li,G., Gu,B., Gao,X., Shi,J., Wang,Y., Gao,L., Cai,X. *et al.* (2008) Gain-of-function mutation of GATA-2 in acute myeloid transformation of chronic myeloid leukemia. *Proc. Nat. Acad. Sci.*, **105**, 2076.

128. Forestier,E., Izraeli,S., Beverloo,B., Haas,O., Pession,A., Michalová,K., Stark,B., Harrison,C., Teigler-Schlegel,A. and Johansson,B. (2008) Cytogenetic features of acute lymphoblastic and myeloid leukemias in pediatric patients with Down syndrome: an iBFM-SG study. *Blood*, **111**, 1575.

129. Valk,P., Verhaak,R., Beijen,M., Erpelinck,C., van Doorn-Khosrovani,S., Boer,J., Beverloo,H., Moorhouse,M., van˜der Spek,P., Löwenberg,B. *et al.* (2004) Prognostically useful gene-expression profiles in acute myeloid leukemia. *New Engl. J. Med.*, **350**, 1617–1628.

130. Rozman,M., Camós,M., Colomer,D., Villamor,N., Esteve,J., Costa,D., Carrió,A., Aymerich,M., Aguilar,J., Domingo,A. *et al.* (2004) Type I MOZ/CBP (MYST3/CREBBP) is the most common chimeric transcript in acute myeloid leukemia with t (8; 16)(p11; p13) translocation. *Genes Chromosomes Cancer*, **40**, 140–145.

131. Sun,Y. (2006) p53 and its downstream proteins as molecular targets of cancer. *Mol. Carcinogenesis*, **45**, 409–415.

132. Xia,H., Qi,H., Li,Y., Pei,J., Barton,J., Blackstad,M., Xu,T. and Tao,W. (2002) LATS1 tumor suppressor regulates G2/M transition and apoptosis. *Oncogene*, **21**, 1233–1241.

133. Altucci,L., Leibowitz,M.D., Ogilvie,K.M., de Lera,A.R. and Gronemeyer,H. (2007) RAR and RXR modulation in cancer and metabolic disease. *Nat Rev Drug Discov*, **6**, 793–810.

134. Massague,J. (2004) G1 cell-cycle control and cancer. *Nature*, **432**, 298–306.

135. Ying,J., Li,H., Yu,J., Ng,K.M., Poon,F.F., Wong,S.C.C., Chan,A.T., Sung,J.J. and Tao,Q. (2008) WNT5A exhibits tumor-suppressive activity through antagonizing the Wnt/beta-catenin signaling, and is frequently methylated in colorectal cancer. *Clin Cancer Res.*, **14**, 55–61.

136. Streit,S., Mestel,D.S., Schmidt,M., Ullrich,A. and Berking,C. (2006) FGFR4 Arg388 allele correlates with tumour thickness and FGFR4 protein expression with survival of melanoma patients. *Br J Cancer*, **94**, 1879–1886.

137. Vogelstein,B. and Kinzler,K.W. (2004) Cancer genes and the pathways they control. *Nat. Med.*, **10**, 789–799.

138. Brown,C.J., Lain,S., Verma,C.S., Fersht,A.R. and Lane,D.P. (2009) Awakening guardian angels: drugging the p53 pathway. *Nat. Rev. Cancer*, **9**, 862–873.

139. Citri,A. and Yarden,Y. (2006) EGF-ERBB signalling: towards the systems level. *Nat. Rev. Mole. Cell Biol.*, **7**, 505–516.

140. (2006) Targeting the Hedgehog pathway in cancer. *Nature Reviews Drug Discovery*, **5**, 1026–1033.
141. (2005) Promoting apoptosis as a strategy for cancer drug discovery. *Nat. Rev. Cancer*, **5**, 876–885.
142. Tsankova,N., Renthal,W., Kumar,A. and Nestler,E.J. (2007) *Epigenetic regulation in psychiatric disorders*, **8**, 355–367.
143. Zarubin,T. and Han,J. (2005) Activation and signaling of the p38 MAP kinase pathway. *Cell Res.*, **15**, 11–18.
144. Hoehndorf,R., Schofield,P. and Gkoutos,G. (2011) PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res*, **39**, e119.
145. Gu,Q. and Zhou,J. (2009) *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, pp. 359–368.