Bioinformatics, 38(6), 2022, 1560–1567
https://doi.org/10.1093/bioinformatics/btab851
Advance Access Publication Date: 22 December 2021
Original Paper

Gene expression

Gene set analysis with graph-embedded kernel association test

Jialin Qu and Yuehua Cui *

Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA

*To whom correspondence should be addressed.

Associate Editor: Teresa Przytycka

Received on September 24, 2021; revised on November 20, 2021; editorial decision on December 13, 2021; accepted on December 16, 2021

Abstract

Motivation: Kernel-based association test (KAT) has been a popular approach to evaluate the association of expressions of a gene set (e.g. pathway) with a phenotypic trait. KATs rely on kernel functions which capture the sample similarity across multiple features, to capture potential linear or non-linear relationship among features in a gene set. When calculating the kernel functions, no network graphical information about the features is considered. While genes in a functional group (e.g. a pathway) are not independent in general due to regulatory interactions, incorporating regulatory network (or graph) information can potentially increase the power of KAT. In this work, we propose a graph-embedded kernel association test, termed gKAT. gKAT incorporates prior pathway knowledge when constructing a kernel function into hypothesis testing.

Results: We apply a diffusion kernel to capture any graph structures in a gene set, then incorporate such information to build a kernel function for further association test. We illustrate the geometric meaning of the approach. Through extensive simulation studies, we show that the proposed gKAT algorithm can improve testing power compared to the one without considering graph structures. Application to a real dataset further demonstrate the utility of the method.

Availability and implementation: The R code used for the analysis can be accessed at https://github.com/JialinQu/gKAT.

Contact: cuiy@msu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Significant developments in gene expression analysis in the past decades have drastically promoted our comprehension of genomic aspect of diverse diseases. There is a paradigm shift of gene expression analysis from the single gene-level analysis to analyses focusing on gene sets, with the hope to gain better biological insights into the molecular mechanisms of various phenotypic traits, in particular disease traits. The resultant gene sets are analyzed as a whole to determine which of these properties are relevant to the phenotype of interest (Mathur *et al.*, 2018). This global view of gene-set analysis (e.g. pathway analysis) features a number of advantages when compared with a single-gene analysis. First, identifying pathways and processes can have more explanatory power. Gene sets tend to be more interpretable and more reducible than a simple list of different genes. Second, genes within a pathway are likely to interact with each other. Gene-set analysis can boost the signal-to-noise ratio and make it possible to detect modest changes in individual genes when there exists strong cross-correlation between the members of a gene set (Subramanian *et al.*, 2005). Moreover, it reduces the number of tests that need to be performed and further reduces computational complexity by grouping thousands of genes or proteins sharing biological, functional or other characteristics. Gene-set analysis approach can also provide valuable insight into the collaboration of particular biological pathways or cellular functions of complex diseases by considering functionally associated gene sets simultaneously.

With well over a decade of development of gene set analysis, various methods are available. Subramanian *et al.* (2005) proposed gene set enrichment analysis, ranking all genes in a gene set depending on differential expressions and calculating enrichment score from a ranked list to test gene set significance. In a study by Liu *et al.* (2007), a semiparametric regression model called least squares kernel machines for assessing pathway effects on a continuous outcome is presented, where the kernel machines is utilized to handle interactions between expressions of several genes. This model established a close connection between kernel machine methods and a linear mixed model. A similar idea was applied to develop a logistic kernel machine regression model for binary outcome, establishing close relationship between logistic kernel machine regression and logistic mixed model (Liu *et al.*, 2008). Similar ideas were also generalized to SNP data. For example, Wu *et al.* (2010) proposed to assess the association of a group of SNPs with a binary disease trait using a logistic kernel machine. The method was later on extended to incorporate rare variants, called sequence kernel association test

(SKAT), to analyze the joint influence of region-based multiple variants on a disease phenotype (Wu *et al.*, 2011).

During the last decade, the abundance of biological knowledge or pathway information from multitudinous scientific research has made it possible to incorporate biological information, particularly molecular interactions networks, in the analysis of gene expression data. Networks or graphs are popular ways of characterizing these biological messages which contain valuable information. With rapid development of well-known pathway databases containing tens of thousands of reactions and interactions, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2004), Reactome (Joshi-Tope *et al.*, 2005), STRING v10.0 (Szklarczyk *et al.*, 2015), BioCyc (Karp *et al.*, 2005) and BioCarta (Nishimura, 2001), how to utilize these knowledge to supplement the standard experimental data raises interesting statistical challenges.

A lot of published literatures have been focused on taking advantages of the prior network knowledge to solve classification or clustering problems. For instance, Liu *et al.* (2014) developed a 'network-assisted co-clustering for the identification of cancer subtypes' (NCIS) algorithm that incorporated gene network information to simultaneously cluster both samples and features. Rapaport *et al.* (2007) showed how to derive supervised and unsupervised classification algorithms based on spectral decomposition incorporating gene network information. Li and Li (2008) and Gao *et al.* (2019) introduced a network-constrained regularization procedure for gene selection that integrated graphical information. Cun and Fröhlich (2012) compared 14 published gene prior knowledge-based selection methods, while Manica *et al.* (2019) proposed a pathway-induced multiple kernel learning method and compared their method with all approaches mentioned in Cun's paper. These works underscore the importance of incorporating prior gene network information in genomic analysis, focusing either on gene selection or prediction. However, none of these methods are focused on hypothesis testing by borrowing prior network information. It is thus the purpose of this work to develop a new testing procedure that evaluate the joint association of multiple genes with a phenotypic trait, while incorporating prior network/graph knowledge.

In a recent work by Liu *et al.* (2007), the author presented a kernel association test (KAT) framework, a regression approach to test for association between multiple genes and a phenotype, while also taking covariates into consideration. Individual similarity is taken into account to construct a kernel matrix which is then used to build a score test to assess associations. In fact, all the KAT-based testing methods do not consider feature similarities when establishing associations. Instead of depicting correlations between individuals, feature similarity focuses on connectivity between gene features which are further utilized to improve the power of an association test. The advancement of pathway knowledge enables us to integrate biological information and construct such feature similarities. In this article, building upon the KAT framework, we propose a graph-embedded kernel association test (gKAT), a method that exploits prior network information and tests for association between a gene set and a trait. Based on the prior network information, we can compute a diffusion kernel to describe relationships between genes in a set. Such information is regarded as a weight matrix when computing the gKAT test statistic. The simulation study indicates that this method is effective and has higher power than the original KAT test that do not borrow the graph information. We apply our method to a liver enzyme gene expression dataset combined with pathway information extracted from the KEGG database. Our approach achieves meaningful result and provides a new tool for gene set analysis.

2 Materials and methods

2.1 Association test with kernel machine regression

To make the article self-contained, we briefly introduce the KAT-based method under the kernel machine regression framework. Kernel machine method is a powerful tool for association analysis. Association test using the kernel machine regression method was initially introduced by Liu *et al.* (2007), where the authors introduced a score test statistic to evaluate the joint effect of a gene set on a phenotype, instead of assessing individual gene effect. Some notations need to be made before we introduce the KAT method. Consider N samples measured on P genes; for the ith sample, y_i refers to the phenotypic response; $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iM})^T$ denotes the ith sample measured on M covariates; and $X_i = (X_{i1}, X_{i2}, \ldots, X_{iP})^T$ denotes P gene variables. The phenotype y_i depends on Z_i and X_i through the following semi-parametric model:

$$y_i = \alpha_0 + \alpha' Z_i + h(X_i) + \epsilon_i \tag{1}$$

where $h(X_i)$ is an unknown centered smooth function subjected to gene expression X_i. If we assume $h(X_i)$ is linear in X_i, then (1) can be written as,

$$y_i = \alpha_0 + \alpha' Z_i + \beta' X_i + \epsilon_i,$$

when y_i is continuous and

$$\operatorname{logit} P(y_i = 1) = \alpha_0 + \alpha' Z_i + \beta' X_i,$$

when y_i is a binary outcome taking values either 0 or 1. Then, one can test association between P gene variables and Y by $H_0 : \beta = 0$. When the dimension of gene variables P is large, the above linear or logistic regression model will be under-powered due to large degrees of freedom. To overcome this issue, KAT assumes the regression coefficients β are random and follow a normal distribution with mean 0 and variance $\tau^2 \boldsymbol{K}$, where $\boldsymbol{K}$ is a kernel matrix whose (i, j)th element is $K(X_i, X_j)$. Therefore, testing $H_0 : \beta = 0$ is equivalent to test $H_0 : \tau^2 = 0$. For a continuous outcome, the KAT test statistic Q under the null is defined as,

$$Q = (y - \hat{\mu})' \boldsymbol{K} (y - \hat{\mu}) \tag{2}$$

where $\hat{\mu} = \hat{\alpha}_0 + Z\hat{\alpha}$ is the predicted mean of y under the null hypothesis. Some popular choices of kernel functions include linear kernel, polynomial kernel, Gaussian kernel and sigmoid kernel. When the gene variables are SNPs, the kernel matrix can be the IBS kernel (Kwee *et al.*, 2008; Wu *et al.*, 2011).

In all the KAT-based methods, the kernel matrix $\boldsymbol{K}$ is calculated assuming genes are independent. Genes function in networks to fulfill their joint tasks. When such network information is available, incorporating such network information can improve the gene selection performance (e.g. Li and Li, 2008; Gao *et al.*, 2019). However, no testing work has been developed by borrowing such network information when assessing associations. Here, we develop a graph-embedded KAT-based test (gKAT) which can incorporate prior network/graph information to improve association signals.

We take gene expression data to illustrate the idea. Extension to SNP variables is similar as long as such network/graph information is available. Figure 1 gives a summary of the proposed gKAT approach. Figure 1a shows that a subset of gene variables belonging to a particular gene set (e.g. a KEGG pathway or a GO term) may form a network which can be expressed as an adjacency matrix. This set of genes are highlighted with red color and their graph information can be extracted from a database. Figure 1b shows the extracted sub-network of genes with which a diffusion kernel can be defined to extract the gene similarity information in the network. Such information is then incorporated into step (c) to form the gKAT statistic for further association evaluation.

2.2 Constructing the diffusion kernel from gene expression pathway

To consider the regulatory interaction network information in a gene set, we first introduce feature similarity which can be described by a diffusion kernel. Kernel-based algorithms are suitable for capturing similarities between data points induced by graph-like structures. In kernel methods, a symmetric function $K : \boldsymbol{\chi} \times \boldsymbol{\chi} \to \mathbb{R}$, where $\boldsymbol{\chi}$ denotes the input space, is called a kernel function if it satisfies the Mercer's condition (Schölkopf *et al.*, 2002). Kernel function

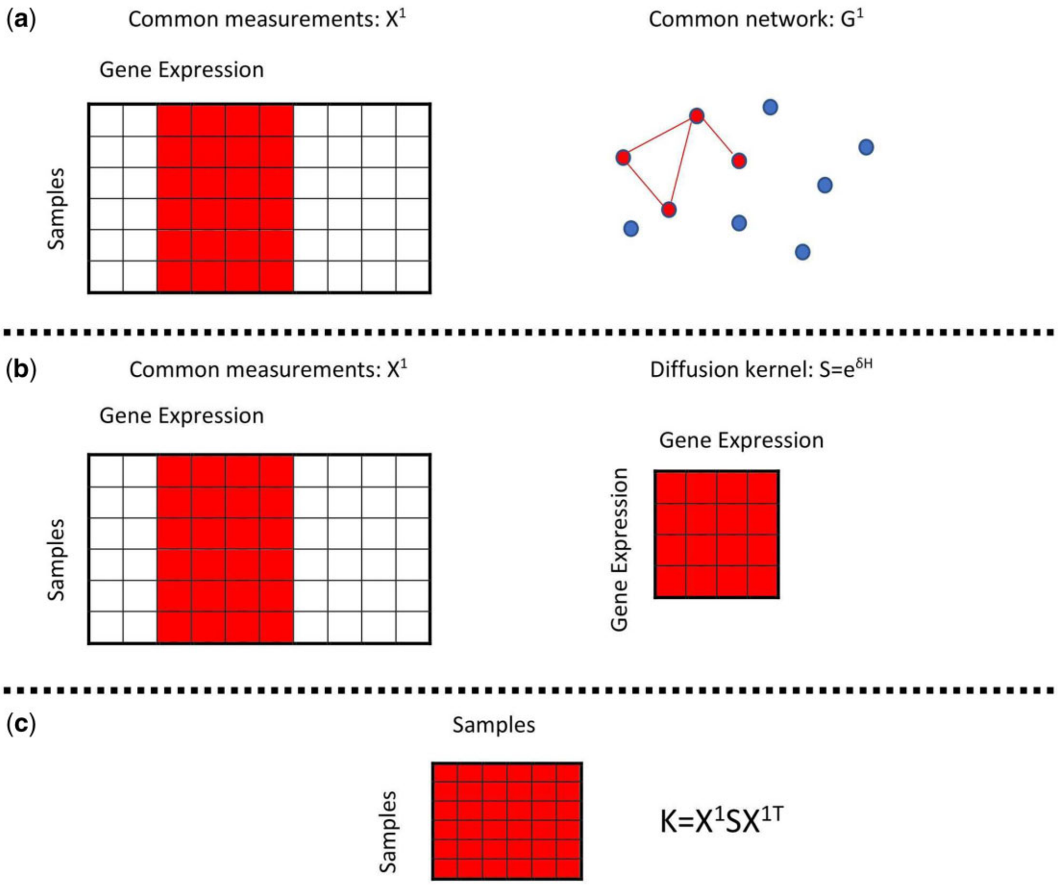


Fig. 1. Step (**a**): a subset of gene variables (red color) is extracted to form a sub-network based on some prior knowledge (e.g. a KEGG pathway or a GO term); (**b**) compute a diffusion kernel based on the sub-network graph information extracted from the database; and (**c**) calculate the final kernel that contains the graph information for further association test

implicitly constructs a mapping from the input space to a Hilbert space $\boldsymbol{H}_K$, i.e. $\phi : \boldsymbol{\chi} \to \boldsymbol{H}_K$, which is equipped with the inner product,

$$K(x, x') = \langle \phi(x), \phi(x') \rangle.$$

Based on matrix exponential idea, a diffusion kernel is proposed to construct kernels on graphs, which can be formally described as:

$$S = e^{\delta H} = \lim_{n \to \infty} \left(1 + \frac{\delta H}{n}\right)^n, \tag{3}$$

where δ is the bandwidth parameter that controls the extent of diffusion and H is a negative Laplacian matrix. The limit of diffusion kernel always exists and is equivalent to (Kondor and Lafferty, 2002),

$$e^{\delta H} = I + \delta H + \frac{\delta^2}{2} H^2 + \frac{\delta^3}{3} H^3 + \cdots.$$

Consider a pathway that is represented by an undirected and unweighted graph $\mathcal{G} = (V, E)$, where V denotes a set of nodes corresponding to genes and $E = \{u \sim v\}$ denotes edges indicating connections (or interactions) between genes. Such a graph is defined by a generator H as follows,

$$H = \begin{cases} 1 & \text{for } u \sim v \\ -d_u & \text{for } u = v \\ 0 & \text{otherwise} \end{cases}$$

where the degree of vertex u is defined as $d_u = \sum_{u \sim v} 1$. For an isolated vertex u, we set $d_u = 0$. It is obvious to see that the negative of H is the Laplacian matrix. Since H is symmetric, it turns out that $e^{\delta H}$ is always positive semi-definite for any symmetric matrix H (Kondor and Lafferty, 2002).

The parameter δ in the diffusion kernel controls the extent of diffusion and it has similar effect as the scaling parameter in a Gaussian kernel (Sun *et al.*, 2008). Selecting appropriate δ values is a key step. In a prediction analysis, one can simply pick the optimal one by checking the prediction accuracy using cross-validation. However, things become complicated when the focus is to evaluate associations in a hypothesis testing situation. In this work, we consider a sequence of δ values centered at 0 (the diffusion kernel S degenerates to an identity matrix if $\delta = 0$) based on which we can construct different diffusion kernels. Under each S, we can conduct a hypothesis test and get a P-value. Then, we borrow the omnibus testing idea to get an aggregated P-value. The details are given in Section 2.4.

2.3 Combing data information with graph similarity

Given a diffusion kernel which represents relationships between genes, it can be incorporated into the calculation of the sample similarities to get the sample kernel function. The final kernel function, thus, considers both the sample and feature similarity information. For simplicity, we consider a linear kernel to illustrate the idea.

Recall that for a symmetric Laplacian matrix $L \in \mathbb{R}^{P \times P}$, there exists an eigen-decomposition, $L = R \Lambda R^T$, where columns of $R \in \mathbb{R}^{P \times P}$ contains orthogonal and normalized eigenvectors of L, and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_P)$ is a diagonal matrix whose entries are the eigenvalues of L. As stated by explanation of the function of matrices (Golub and Van Loan, 2012), a diffusion kernel has decomposition as follows,

$$S = e^{\delta H} = e^{-\delta L} = I - \delta L + \frac{\delta^2}{2}L^2 - \frac{\delta^3}{6}L^3 + \cdots = RDR^T$$

where $D = \text{diag}(e^{-\delta\lambda_1}, \ldots, e^{-\delta\lambda_P})$. Note that, when $\delta = 0$, S is degenerated to an identity matrix. In this case, no graph information is considered.

The linear kernel is the simplest kernel function which is given by the inner product, i.e. $K_L(x, y) = x^T y$, where x and $y \in \mathbb{R}^{P\times 1}$ are two samples with P measurements. Gene expression pathway information is regarded as prior information gathered over biomedical researches from which we can construct diffusion kernel to describe similarities between genes, which is shown in Section 2.2. Using this prior knowledge, a diffusion kernel matrix $S \in \mathbb{R}^{P\times P}$ can be built, which leads to the final kernel function given as,

$$K_L(x, y) = x^T S y = x^T RDR^T y = (x^T RD^{\frac{1}{2}})(D^{\frac{1}{2}} R^T y) = \Phi(x)^T \Phi(y) \tag{4}$$

Once we calculate K_L, it can be plugged into Equation (2) to assess the association between a gene set and a phenotypic trait.

One can also consider a weighted graph $G = (V, E, W)$, where W denotes weights of the edges with $w(u, v)$ being the weight of edge between node u and v. In this case, the degree of vertex can be calculated as $d_u = \sum_{u\sim v} w(u, v)$. A normalized Laplacian matrix of a weighted graph can be defined as,

$$L = \begin{cases} -w(u,v)/\sqrt{d_u d_v} & \text{for } u \sim v \\ 1 - w(u,v)/d_u & \text{for } u = v \\ 0 & \text{otherwise} \end{cases}$$

which is symmetric and positive semi-definite. Comparable eigen-decomposition can be applied to L as described above.

Here, we use a toy example to illustrate the idea (see Fig. 2), similar as what illustrated in Manica *et al.* (2019). For two samples with seven gene variables, an adjacency matrix can be extracted based on the sub-network information extracted from certain database (Fig. 2a). Applying the graph induction steps described above, given any fixed tuning parameter δ, Figure 2b suggests that the original data can be mapped from the input space to a new space defined by the newly calculated kernel function specified by the parameter δ.

The idea is illustrated with the linear kernel. It can be generalized to other kernels, such as the Gaussian and polynomial kernel, or a combination of different kernels to fully utilize the advantages of different kernel functions. For example, when a Gaussian kernel is applied, the final kernel function that incorporates the diffusion kernel S can be given as $K_G(x, y) = \exp(-\sigma\kappa)$, where σ is the bandwidth parameter and $\kappa = (x-y)^T S(x-y) = (x-y)^T RDR^T (x-y) = \left\{(x-y)^T RD^{\frac{1}{2}}\right\} \left\{D^{\frac{1}{2}} R^T (x-y)\right\} = \Phi(x,y)^T \Phi(x,y)$. So the final diffusion-based Gaussian kernel is given as $K_G(x, y) = \exp\{-\sigma\Phi(x, y)^T \Phi(x, y)\}$.

2.4 *P*-value aggregation with Cauchy transformation

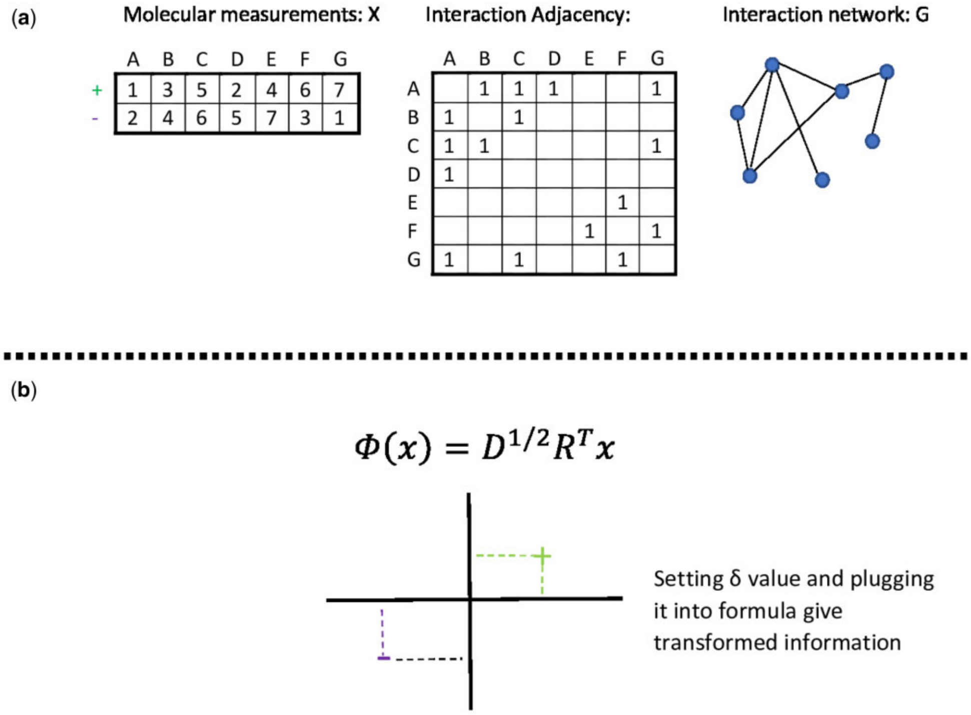
Since we do not know the optimal tuning parameter δ when calculating the diffusion kernel function, we define a sequence of J δ values centered at 0. Note that when $\delta = 0$, the diffusion kernel is degenerated to an identity matrix and the final kernel is the same as the regular kernel assuming independence between features. For a given δ value, we can get a P-value denoted as $p_j, j = 1, \ldots, J$. Since these J P-values are calculated based on the same dataset, they are correlated. Borrowing the P-value aggregation idea termed as ACAT by Liu *et al.* (2019), we can do a Cauchy transformation of the jth P-value by $w_j \tan\{(0.5 - p_j)\pi\}$, where w_j is a non-negative weight. ACAT defines the final test statistic as a linear combination of Cauchy transformed P-values by,

$$T_{ACAT} = \sum_{j=1}^{J} w_j \tan\{(0.5 - p_j)\pi\}$$

Then the final aggregated P-value can be calculated by,

$$P - \text{value} \approx 0.5 - \{arctan(T_{ACAT}/w)\}/\pi$$

based on the cumulative density function of the Cauchy distribution. If no information is available for the weight function w, it can be simply taken as $1/J$.


(a)
Molecular measurements: X
Interaction Adjacency:
Interaction network: G
(b)
Φ(x) = D^{1/2}R^T x
Setting δ value and plugging it into formula give transformed information

	A	B	C	D	E	F	G
+	1	3	5	2	4	6	7
-	2	4	6	5	7	3	1

	A	B	C	D	E	F	G
A		1	1	1			1
B	1		1				
C	1	1					1
D	1						
E						1	
F					1		1
G	1		1			1	

Fig. 2. Illustration of the pathway induction idea. (**a**) Prior graph knowledge is converted to an adjacency matrix; and (**b**) the geometric representation of the samples in the mapped space defined by the new kernel function specified by the parameter δ. The two samples in the mapped space can be considered as projections from the original measurement space to the mapped space defined by the new kernel function incorporating the diffusion kernel

3 Simulation studies

We conducted simulations under various settings to evaluate the performance of the gKAT method and compared it with the regular KAT method without considering the graphical information. We fist simulated two gene networks assuming different P (50 and 100). We randomly generated two adjacency matrices with different dimensions ($P = 50$ and $P = 100$), based on which the R igraph package was used to depict relationships between nodes. See Figure 3 for the two gene network structures. We regarded the network structures as known prior knowledge for the two gene sets. A Laplacian matrix was then defined according to the graph network information. Next, we computed the Jaccard similarity based on each network that would be treated as the variance-covariance matrix Σ of a gene set. The Jaccard similarity coefficient of two vertices is the number of common neighbors divided by the number of vertices that are neighbors of at least one of the two vertices being considered. The network information and matrix Σ were then used in all the simulations.

To simulate gene variables X, we generated a multivariate normal distribution with mean 0 and variance-covariance matrix Σ. Table 1 reported the type I error rates of the test under different data dimension ($P = 50$, 100) and different sample sizes ($N = 200$, 400, 1000). We analyzed the data using linear kernel (denoted as L) and Gaussian kernel (denoted as G). The corresponding ones by incorporating the graph information with a diffusion kernel were denoted as L_g and G_g. We also applied the ACAT idea to integrate P-values obtained with the linear and Gaussian kernel and the results were denoted by C. The one by integrating P-values incorporating graph information was denoted by C_g. Since the true gene effect is unknown in practice, this omnibus testing idea can take advantage of both linear and non-linear effect in which a group of genes may have on a phenotype, hence improves the power. In both the type I error and power simulation studies, the significance level was set as 0.05.

Under different data dimensions ($P = 50$, 100), the type I error rates are all under-estimated if the graph information is ignored. Typically the KAT-based method, such as the SKAT method for sequencing data analysis, gives conservative type I error rate (Lee et al., 2012). In contrast, the type I error rate is reasonably estimated when the graph information is incorporated. For example, the type I error rate is 0.038 for the combined result when $N = 200$ without considering the graph information, while the estimate is 0.05 when prior graph information is considered. Interestingly, the Gaussian kernel gives quite conservative error rate compared to the linear kernel, in particular when no graph information is considered. By integrating the two P-values obtained under the linear and Gaussian kernel, the type I error rate is reasonably controlled. Overall, the test by considering the graph information can reasonably control the type I error.

Table 1. Type I error (T1E) under different sample sizes ($N = 200$, 400, 1000) and different data dimensions ($P = 50$, 100) analyzed with a linear kernel (L), a Gaussian kernel (G) and integrating linear and Gaussian kernel (C) with and without incorporating the graph information

P	N	T1E without graph			T1E with graph		
		L	G	C	L_g	G_g	C_g
50	200	0.035	0.025	0.038	0.043	0.037	0.050
	400	0.045	0.035	0.038	0.047	0.046	0.052
	1000	0.041	0.035	0.039	0.050	0.048	0.055
100	200	0.034	0.011	0.022	0.044	0.038	0.039
	400	0.045	0.019	0.026	0.052	0.042	0.047
	1000	0.048	0.034	0.038	0.051	0.050	0.050

To evaluate the testing power, three simulation scenarios (A, B and C) were considered:

$$\mathrm{A}: Y_i = X_i\beta$$

$$\begin{aligned}\mathrm{B}: Y_i = {} & 0.15X_{i1} + 0.15X_{i3} + 0.15X_{i5} + 0.15X_{i7} + 0.15X_{i9} \\ & + 0.6X_{i1}X_{i3} + 0.6X_{i5}X_{i7} + 0.6X_{i9}X_{i11} + 0.6X_{i13}X_{i15} \\ & + 0.6X_{i17}X_{i19}\end{aligned}$$

$$\mathrm{C}: Y_i = \sum_{j=1}^{16}\{0.01X_{ij}^2 + 0.015X_{ij}^3 + \exp(-X_{ij}^2/10)\}$$

In scenario A, we let 25 genes have effects and the rest have no effect and considered two different settings: (A1) weak effect size, i.e. $\beta = \{0, 0.035, 0, 0.035, \ldots, 0, 0.035, 0, 0.035\}$; and (A2) relatively strong effect size, i.e. $\beta = \{0, 0.05, 0, 0.05, \ldots, 0, 0.05\}$. In scenario B, both main and interaction effects were considered and in scenario C, only non-linear effect were considered. The purpose was to check the testing performance under different situations.

It is worth mentioning that the choice of the bandwidth parameter δ is a challenging issue when computing the diffusion kernel. Empirical evidence suggests that considering an interval $\delta \in [-1, 1]$ is sufficient to cover a good range of δ values centered at 0. In our analysis, we chose δ values between –1 and 1 with an increment 0.1. A total of 21 diffusion kernels were computed under each δ value. When $\delta = 0$, the result is the same as the one without incorporating network information since the diffusion kernel S degenerates to an identity matrix I. The ACAT

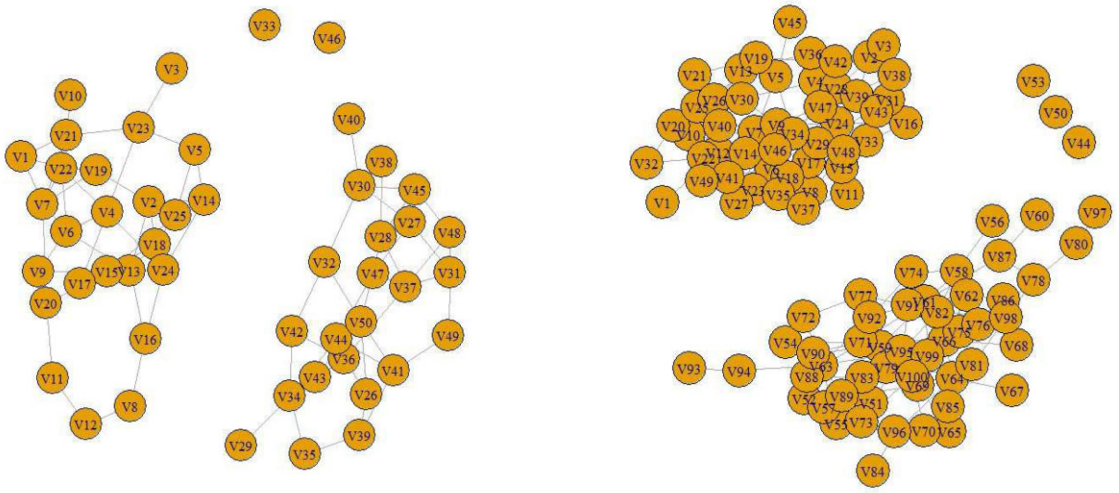


Fig. 3. The simulated network structures under different data dimensions ($P = 50$ for the left figure and $P = 100$ for the right figure)

Table 2. Testing power under different scenarios with and without incorporating graph information

P	*N*	Scenario	*β*	Power without graph			Power with graph		
				L	*G*	*C*	L_g	G_g	C_g
50	200	A1	(0, 0.035)	0.479	0.383	0.452	0.563	0.462	0.516
		A2	(0, 0.05)	0.842	0.776	0.808	0.901	0.829	0.865
		B		0.386	0.483	0.437	0.392	0.628	0.555
		C		0.496	0.506	0.504	0.577	0.585	0.582
	400	A1	(0, 0.035)	0.863	0.839	0.840	0.906	0.873	0.883
		B		0.717	0.859	0.847	0.724	0.957	0.925
		C		0.855	0.924	0.906	0.900	0.945	0.939
100	200	A1	(0, 0.035)	0.332	0.252	0.325	0.492	0.420	0.483
		A2	(0, 0.05)	0.741	0.635	0.643	0.873	0.801	0.838
		B		0.178	0.120	0.151	0.199	0.179	0.187
		C		0.215	0.198	0.199	0.332	0.282	0.300
	400	A1	(0, 0.035)	0.783	0.699	0.766	0.886	0.836	0.885
		B		0.368	0.358	0.361	0.378	0.422	0.417
		C		0.576	0.557	0.565	0.669	0.657	0.663

method was then applied to combine the 21 *P*-values obtained under the 21 kernels to get the final aggregated *P*-value. In each scenario, 200 or 400 individuals along with 50 or 100 genes were generated to estimate the empirical power, each with 1000 simulation replicates.

Table 2 shows the empirical power under different scenarios. In general, the empirical power improves with increasing sample size and gene effect. Under scenarios A1 and A2, the linear kernel always outperforms the Gaussian kernel since the effects are all linear. By further integrating different *P*-values obtained under the linear and Gaussian kernel, the integrated power is close to the best one. The same results were observed in other scenarios. This shows that in practice, one can integrate *P*-values obtained under different kernels to get a robust one when the underlying true gene function is unknown. For fixed *N* and *P*, the power increases as the effect size increases. Under the same *P* and effect size, the power increases as the sample size increases. For example, the power increases from 0.563 to 0.906 when increasing sample size from 200 to 400, analyzed with the linear kernel incorporating the graph information. Under the same sample size and effect size, the power decreases as the data dimension *P* increases as we expected, since increasing the null variables adds more noise to the model. For example, under scenario A1, the power decreases from 0.563 to 0.492 when *P* increases from 50 to 100 under the L_g model.

Under scenario B and C, the Gaussian kernel performs better than the linear kernel when $P = 50$. When *P* increases to 100, more noises were added to the model and the Gaussian kernel suffered a little bit from power loss. This could be due to the fact that the introduced noises were linear noises having zero effect, which makes the Gaussian kernel under-performed. In any case, the gKAT model considering prior graph information always outperforms the KAT method without incorporating the graph information. In addition, integrating *P*-values analyzed with different kernels can be done in practice in order to utilize the strength of different kernels.

In the above-presented simulation studies, we incorporated the correct graph information when implementing the gKAT method. However, as the true graph information may not always available in practice, it is interesting to investigate the situation where the prior network knowledge may not be correctly specified. Due to space limit, we provided the simulations in Supplementary File. The results showed that the proposed gKAT method does not suffer too much from power loss even the network information was not correctly specified, indicating the robustness of the method.

4 Real data analysis

We applied the gKAT method to a human liver cohort dataset to demonstrate the utility of our approach. Liver tissues of 466 Caucasian samples including 213 females and 253 males whose age ranging from 0 to 93 with an average of 50 and a standard deviation of 18 were used. The dataset contains 466 individuals along with 40 638 gene expression measurements and 10 enzyme activity phenotypes, which can be downloaded from the Sage Bionetworks' synapse platform using Synapse ID syn4499 (Gao *et al.*, 2019). Schadt *et al.* (2008) and Yang *et al.* (2010) gave detail description of the dataset. We picked enzyme activity CYP2E1 as our phenotype which was further log-transformed to pass the normality test. We then extracted three gene expression pathways from the KEGG database, including two CYP2E1-related pathways, hsa00982 and hsa00980, and one CYP2E1-unrelated pathway hsa00730. Pathway hsa00982 and hsa0090 contain gene *CYP2E1* and served as positive controls. After matching genes in the dataset with the ones in the pathways, 66 and 70 genes were, respectively, remained in pathway hsa00982 and hsa00980, whereas only 12 genes were mapped to the hsa00730 pathway. Subjects whose information was not complete (e.g. having missing gene expressions or phenotypic measurement) were removed, resulting in 379 individuals. Each gene expression was standardized to have mean 0 and standard deviation 1.

Figure 4 showed the subgraph of (a) hsa00982, (b) hsa00980 and (c) hsa00730 with the matched genes. Based on the subgraph, we constructed diffusion kernels to quantify the graphical structure of each pathway. For pathway hsa00982 containing 66 genes (e.g. 66 nodes), the diffusion kernel based on this subgraph was a $\mathbb{R}^{66\times 66}$ matrix. Similarly for pathway hsa00980 and hsa00730, we constructed $\mathbb{R}^{70\times 70}$ and $\mathbb{R}^{12\times 12}$ kernel matrices, respectively. We first considered the single kernel situation when calculating the Gaussian and linear kernel while incorporating the graph information to get the corresponding *P*-value. The results were compared with those obtained with the corresponding Gaussian and linear kernel without incorporating prior graph information. Finally, we integrated the *P*-values obtained under the Gaussian and linear kernel to obtain a final aggregated *P*-value with the ACAT method. In total, we compared three different cases: 1) results analyzed with the linear kernel with and without incorporating the diffusion kernel; 2) results analyzed with the Gaussian kernel with and without incorporating the diffusion kernel; and 3) results after integrating the linear and Gaussian kernel with and without considering the graph information.

As we described in Section 2.4, we selected a sequence of values for *δ*, i.e. $\delta = \{-1.0, -0.9, -0.8, \ldots, 0.8, 0.9, 1.0\}$. This ended up with 21 different *δ* values. Under each *δ* value, a diffusion kernel was computed and further incorporated into the Gaussian or linear kernel calculation to get a *P*-value. The 21 *P*-values were then aggregated with the ACAT method to get the final *P*-value for the pathway. Figure 5 shows the distribution of *P*-values under three different pathways. When $\delta = 0$, the diffusion kernel is degenerated to an identity matrix; thus, the diffusion combined Gaussian or linear kernel is the same as the original Gaussian or linear kernel. The red dot in the figures shows the minimum *P*-value. For pathway hsa00982 and hsa00730, the minimum *P*-values were obtained when $\delta = 0$, whereas for pathway hsa00980, the minimum *P*-value was obtained when $\delta = -0.1$.

Table 3 shows *P*-values computed using the original Gaussian and linear kernel, the Cauchy combined *P*-value and the minimum *P*-value along with the optimal bandwidth *δ* value for pathway hsa00982, hsa00980 and hsa00730. In the table, P_L, P_G and P_C refer to *P*-values calculated using the linear, Gaussian and combination of the linear and Gaussian kernel, respectively. Similarly, P_{DL}, P_{DG} and P_{DC} refer to the corresponding *P*-values after incorporating the prior graph information with the diffusion kernel. The data in the corresponding second line shows the minimum *P*-value along with the bandwidth *δ* value in the parenthesis.

For pathway hsa00982 and hsa00730, the minimum *P*-values were obtained when $\delta = 0$. There is no power gain by incorporating prior graph information for these two pathways. Therefore, it is not surprising to see that the aggregated *P*-values under different kernels incorporating the prior graph information are larger than the ones without considering the graph information. For pathway hsa00980, the minimum *P*-value was observed when $\delta = -0.1$. The final

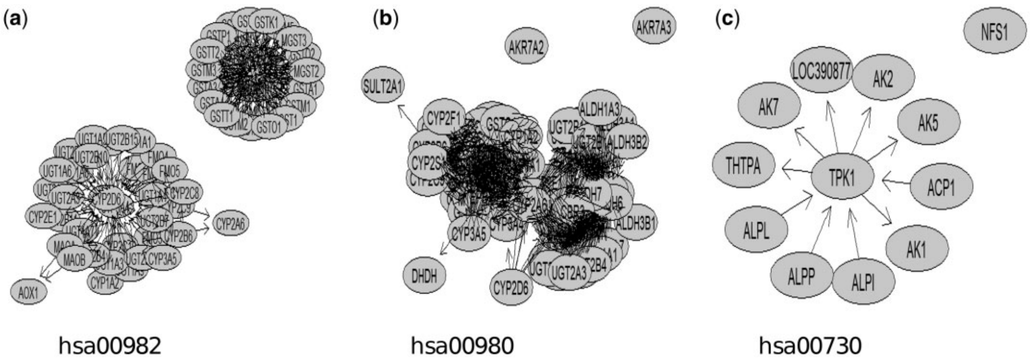


Fig. 4. The graphical representation of the three pathways analyzed in this study

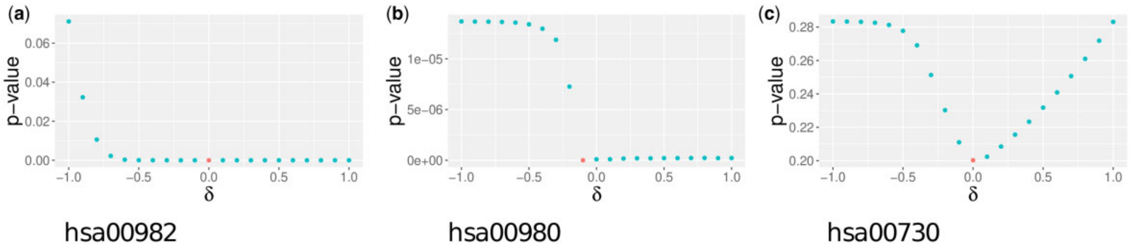


Fig. 5. Plot of P-values under a sequence of δ values in different pathways

Table 3. List of the final aggregated P-values with and without incorporating prior graph information in different pathways

Pathway	P_L	P_G	P_C	P_{DL}	P_{DG}	P_{DC}
hsa00982	2.68E-9	9.52E-9	4.18E-9	1.23E-8 *2.68E-9(0)*	1.10E-7 *9.52E-9(0)*	2.21E-8
hsa00980	9.68E-8	1.31E-7	1.11E-7	5.65E-8 *3.25E-9(–0.1)*	2.30E-9 *1.10E-10(–0.1)*	4.42E-9
hsa00730	0.20	0.06	0.09	0.25 *0.20(0)*	0.11 *0.06(0)*	0.16

Note: The minimum P-value along with the δ value (in the parenthesis) is shown with the italic font.

aggregated P-values under different kernels incorporating the graph information are thus smaller compared to the ones without considering the graph information, showing the improvement of our method. For example, the Cauchy aggregated P-value by integrating the two kernel results is 1.11E-7 without considering the prior graph information, whereas the P-value reduces to 4.42E-9 when prior graph information is incorporated. This shows the power gain by using considering the prior graph information.

For pathway hsa00730, since it is not related to CYP2E1, it serves as a negative control. Both the Gaussian and linear kernel gave P-values larger than 0.05, though the Gaussian kernel generated a smaller P-value than the linear kernel. Gao *et al.* (2019) recently analyzed this pathway associated with CYP2E1 enzyme activity using a high-dimensional variable selection model considering prior graph information. They found no genes with a selection rate larger than 60% in this pathway. Their result is consistent with the result obtained in this analysis.

5 Conclusion and discussion

Although it is well-known that genes function in networks to fulfill their joint task, few association studies considered gene network information when assessing association between a gene set and a phenotype. Building upon the KAT framework, we proposed a new method, termed gKAT that incorporated prior network graph information to test for association between a gene set and a trait. We demonstrated the advantage of our proposed method in simulation studies when compared with the regular KAT-based method without considering graph information. The simulation results suggested that integrating network information can improve the testing power. In addition, gKAT can correctly control the type I error rate at the $\alpha = 0.05$ level. In an application to a liver enzyme dataset, the analysis results showed that our proposed method could achieve better performance with a smaller P-value for pathway hsa00980. For pathway hsa00982, though it is related to CYP2E1 activity, testing results showed that the smallest P-value was obtained when the diffusion kernel bandwidth parameter is 0. Thus, incorporating network information did not provide power gain for this pathway. Both simulation and real data analysis demonstrated that the proposed method could be a powerful tool for gene set association analysis by incorporating prior network knowledge.

It should be noted that choosing optimal bandwidth parameter δ is a challenging task when we compute diffusion kernel to quantify gene connectivity in a network. Instead of choosing an optimal δ value which is difficult to obtain, we tried a range of different δ values and applied the ACAT P-value combination method to combine all P-values to get an aggregated P-value. The ACAT method works well under arbitrary correlation structures and performs similarly as the minimum P-value approach. Since no resampling is involved, it is computationally faster than the minimum P-value approach does. We did try to enlarge the range of δ to [-2,2] and found no difference compared to the range [-1,1]. On the other hand, the interval size cannot be too small to miss the point where the minimum P-value can be obtained. From the empirical evidence, we found that $\delta \in [-1, 1]$ worked reasonably well.

In the simulation study, when the number of features increased from 50 to 100, the model noises increased as the signals were fixed. Such increased noises had an impact on the testing power. Since the noises were considered as linear noises with zero effects, they affected the performance of the Gaussian kernel more than the linear kernel. This was the reason that we observed better performance of the linear kernel compared to the Gaussian one in scenario B and C, even though the relationship between Y and X was non-linear under $P = 100$. With the increase of noises, sometimes the testing size may not be well controlled. In this case, feature selection needs to be done to improve the testing performance. To avoid using the same dataset to do the feature selection and testing, one can split the data into two halves, with one half being used for feature selection and the other half for testing (Meinshausen *et al.*, 2009). This will be investigated in our future study.

Our method is illustrated with the gene expression data. It can be applied to other genetic or genomic data with the goal to assess the association of a gene set with a phenotypic trait, as long as the graph information is available for the variants in a set. Under the ACAT framework, kernels other than the linear and Gaussian kernels, such as the polynomial kernel, can be applied to achieve a more robust omnibus testing power. Our method provides a general quantitative framework to assess gene set associations while considering gene network graph information.

Acknowledgement

The authors acknowledge the support from the National Institutes of Health [1R21HG010073, 1R01GM131398].

Funding

This work was supported by the National Institutes of Health [1R21HG010073, 1R01GM131398].

Conflict of Interest: none declared.

References

Cun,Y. and Fröhlich,H. (2012) Prognostic gene signatures for patient stratification in breast cancer-accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC Bioinformatics*, **13**, 69.

Gao,B. *et al.* (2019) Integrative analysis of genetical genomics data incorporating network structures. *Biometrics*, **75**, 1063–1075.

Golub,G.H. and Van Loan,C.F. (2012) *Matrix Computations*, **Vol. 3**. JHU Press, Baltimore.

Joshi-Tope,G. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.

Kanehisa,M. *et al.* (2004) The kegg resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Karp,P.D. *et al.* (2005) Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.*, **33**, 6083–6089.

Kondor,R.I. and Lafferty,J. (2002) Diffusion kernels on graphs and other discrete structures. In: *Proceedings of the 19th International Conference on Machine Learning*, Vol. 2002. pp. 315–322.

Kwee,L.C. *et al.* (2008) A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.*, **82**, 386–397.

Lee,S. *et al.*; NHLBI GO Exome Sequencing Project—ESP Lung Project Team. (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.*, **91**, 224–237.

Li,C. and Li,H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**, 1175–1182.

Liu,D. *et al.* (2007) Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, **63**, 1079–1088.

Liu,D. *et al.* (2008) Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, **9**, 292.

Liu,Y. *et al.* (2014) A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC Bioinformatics*, **15**, 37.

Liu,Y. *et al.* (2019) Acat: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.*, **104**, 410–421.

Manica,M. *et al.* (2019) Pimkl: pathway-induced multiple kernel learning. *NPJ Syst. Biol. Appl.*, **5**, 8–8.

Mathur,R. *et al.* (2018) Gene set analysis methods: a systematic comparison. *BioData Min.*, **11**, 8.

Meinshausen,N. *et al.* (2009) P-values for high-dimensional regression. *J. Am. Stat. Assoc.*, **104**, 1671–1681.

Nishimura,D. (2001) Biocarta. Biotech software & internet report. *Comput. Softw. J. Sci.*, **2**, 117–120.

Rapaport,F. *et al.* (2007) Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**, 35.

Schadt,E.E. *et al.* (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, **6**, e107.

Schölkopf,B. *et al.* (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Sun,L. *et al.* (2008) Adaptive diffusion kernel learning from biological networks for protein function prediction. *BMC Bioinformatics*, **9**, 162.

Szklarczyk,D. *et al.* (2015) String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.

Wu,M.C. *et al.* (2010) Powerful snp-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.*, **86**, 929–942.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Yang,X. *et al.* (2010) Systematic genetic and genomic analysis of cytochrome p450 enzyme activities in human liver. *Genome Res.*, **20**, 1020–1036.