Yuan *et al. BMC Systems Biology* 2016, **10**(Suppl 1):8
DOI 10.1186/s12918-015-0245-0

BMC Systems Biology

PROCEEDINGS Open Access

A network based covariance test for detecting multivariate eQTL in saccharomyces cerevisiae

Huili Yuan[1], Zhenye Li[1], Nelson L. S. Tang[2] and Minghua Deng[1,3,4*]

From The Fourteenth Asia Pacific Bioinformatics Conference (APBC 2016)
San Francisco, CA, USA. 11 - 13 January 2016

Abstract

Background: Expression quantitative trait locus (eQTL) analysis has been widely used to understand how genetic variations affect gene expressions in the biological systems. Traditional eQTL is investigated in a pair-wise manner in which one SNP affects the expression of one gene. In this way, some associated markers found in GWAS have been related to disease mechanism by eQTL study. However, in real life, biological process is usually performed by a group of genes. Although some methods have been proposed to identify a group of SNPs that affect the mean of gene expressions in the network, the change of co-expression pattern has not been considered. So we propose a process and algorithm to identify the marker which affects the co-expression pattern of a pathway. Considering two genes may have different correlations under different isoforms which is hard to detect by the linear test, we also consider the nonlinear test.

Results: When we applied our method to yeast eQTL dataset profiled under both the glucose and ethanol conditions, we identified a total of 166 modules, with each module consisting of a group of genes and one eQTL where the eQTL regulate the co-expression patterns of the group of genes. We found that many of these modules have biological significance.

Conclusions: We propose a network based covariance test to identify the SNP which affects the structure of a pathway. We also consider the nonlinear test as considering two genes may have different correlations under different isoforms which is hard to detect by linear test.

Keywords: eQTL, Pathway, Isoform

Background

GWAS aims to detect the association between genetic variation and complex diseases. Recent years, GWAS has found 2000 loci associated to complex diseases by statistical methods [1]. As the development of the next-generation sequencing and other high-throughput technology, various types of genome-scale datasets have been collected, providing opportunity to find the mechanism of genetic variation leading to complex diseases by connect the high-throughout data to GWAS. The eQTL study is one of them, which aims to uncover the genetic effects to gene expression and have been conducted in many organisms [2–5]. A common approach in eQTL data analysis is to consider association between each expression trait and each genetic marker through regression analysis. Despite great success with this approach, some regulatory signals may not be detected due to complex interaction between SNPs like epistasis.

Although most eQTL studies considered the expression levels of individual genes as response (single outcome variable), the change of correlation between genes under different genetic status still contains some biological information. For example, post-transcriptional

*Correspondence: dengmh@math.pku.edu.cn
[1] LMAM, School of Mathematical Sciences, Peking University, Yiheyuan Road, 100871 Beijing, China
[3] Center for Quantitative Biology, Peking University, Yiheyuan Road, 100871 Beijing, China
Full list of author information is available at the end of the article

regulations such as phosphorylations and dephosphorylations often affect the activities of transcriptional factors (TFs), which further affect the correlation among TF genes and TF target genes, also the co-expression patterns of the targets of TFs. However, such regulations are hard to be detected if only individual gene considered because there may be little change at the expression levels of individual TF genes. The approach considering "liquid association" (LA) between a pair of genes proposed by [6] is a method to identify such loci, which is later introduced into eQTL study [7]. Subsequently, conditional bi-variate normal model has been developed to capture the change of correlation between a pair of genes [8–10].

However, a biological process is usually performed by a group of genes (more than two genes as in the bi-variate model). Network approaches should be used to study these interactions [11–13]. If we want to see the effects of a cellular change to the organism, it is better for us to consider the change in a functional gene-set such as a pathway. Therefore, some papers has considered the multivariate circumstances by applying CCA to gene expressions and SNP (or CNV) data [14–16]. However, these methods do not consider the network structure when finding the association between gene sets and genetic variant, which will miss the information contained in the network. Li et al. [17], Kim and Xing [18], Zhang and Kim [19], Casale et al. [20] have considered pathway structure when studied the association between genetic variation and gene expression. However, they assume the network structure is the same (static) under different genetic variant. In fact, network structure may be dynamic and biologists have realized that differential network analysis will become a standard mode in network analysis and insightful discoveries could be made with differential network analysis [21]. For example, [22] identified a cancer point mutation in the kinase domain of RET, which causes multiple endocrine neoplasia type 2B by leading to a switch in peptide specificity and then altering the network structure.

So we propose a method to test whether the co-expression pattern in a pathway is affected by a SNP. Our goal is to test for a global change in covariance structure in each pathway, which is different from other network-based methods, which tries to detect non-zero edges from all pairs of genes. When we applied our method to a yeast eQTL dataset, we were able to find some pathway-SNP modules that have biological significance.

Methods

Let $(X_1, X_2, \ldots, X_p)$ be the expression levels of a group of genes and $(Z_1, Z_2, \ldots, Z_m)$ be the set of SNPs. Suppose that there are n independent samples and let $(x_{1i}, x_{2i}, \ldots, x_{pi})_{i=1,\ldots,n}$ denote the expression level of $(X_1, X_2, \ldots, X_p)$ in the ith sample and $(Z_{1i}, Z_{2i}, \ldots, Z_{mi})$ denote the SNP types of the SNP set in the ith sample. Since the mean expression levels of $(X_1, X_2, \ldots, X_p)$ are also possibly affected by some SNPs in $(Z_1, Z_2, \ldots, Z_m)$, we can imitate the procedure in [9] that we first perform regression analysis or penalized regression analysis such as Lasso [23] or SCAD [24] to adjust the effects of $(Z_1, Z_2, \ldots, Z_m)$ on the means and then model the residuals. We assume that the covariate-adjusted expression levels are appropriately centered to have mean values of zero and our interest is to test whether the covariate-adjusted covariance of expression levels is changed under each SNP. In our analysis, the group of genes are a pathway in KEGG [25]. Figure 1 describes our strategy to detect pathway-SNP associations. In this manuscript, we define a module as the collection of a SNP and a pathway, and our objective is to find pathway-SNP modules where the SNP

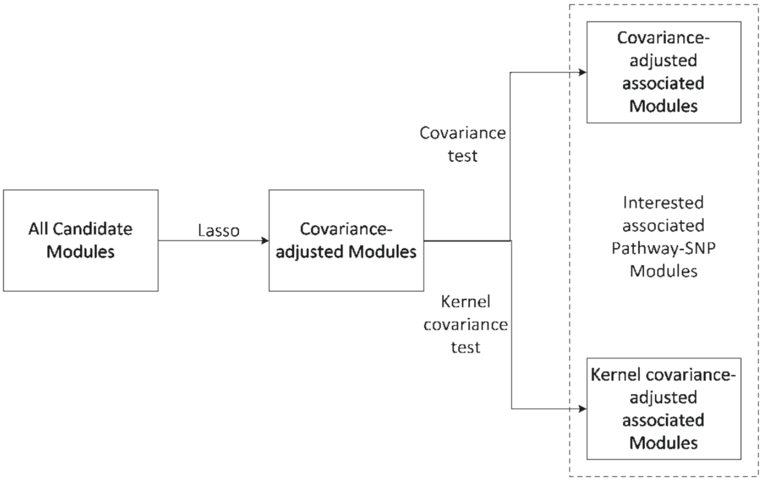
All Candidate Modules
Lasso
Covariance-adjusted Modules
Covariance test
Kernel covariance test
Covariance-adjusted associated Modules
Interested associated Pathway-SNP Modules
Kernel covariance-adjusted associated Modules

Fig. 1 Flowchart of our strategy. Flowchart of our strategy for detecting pathway-associated SNP. We first perform Lasso to adjust the effects of SNPs on the means of gene expression. Then we use covariance test and kernel covariance test to select candidate pathway-SNP modules

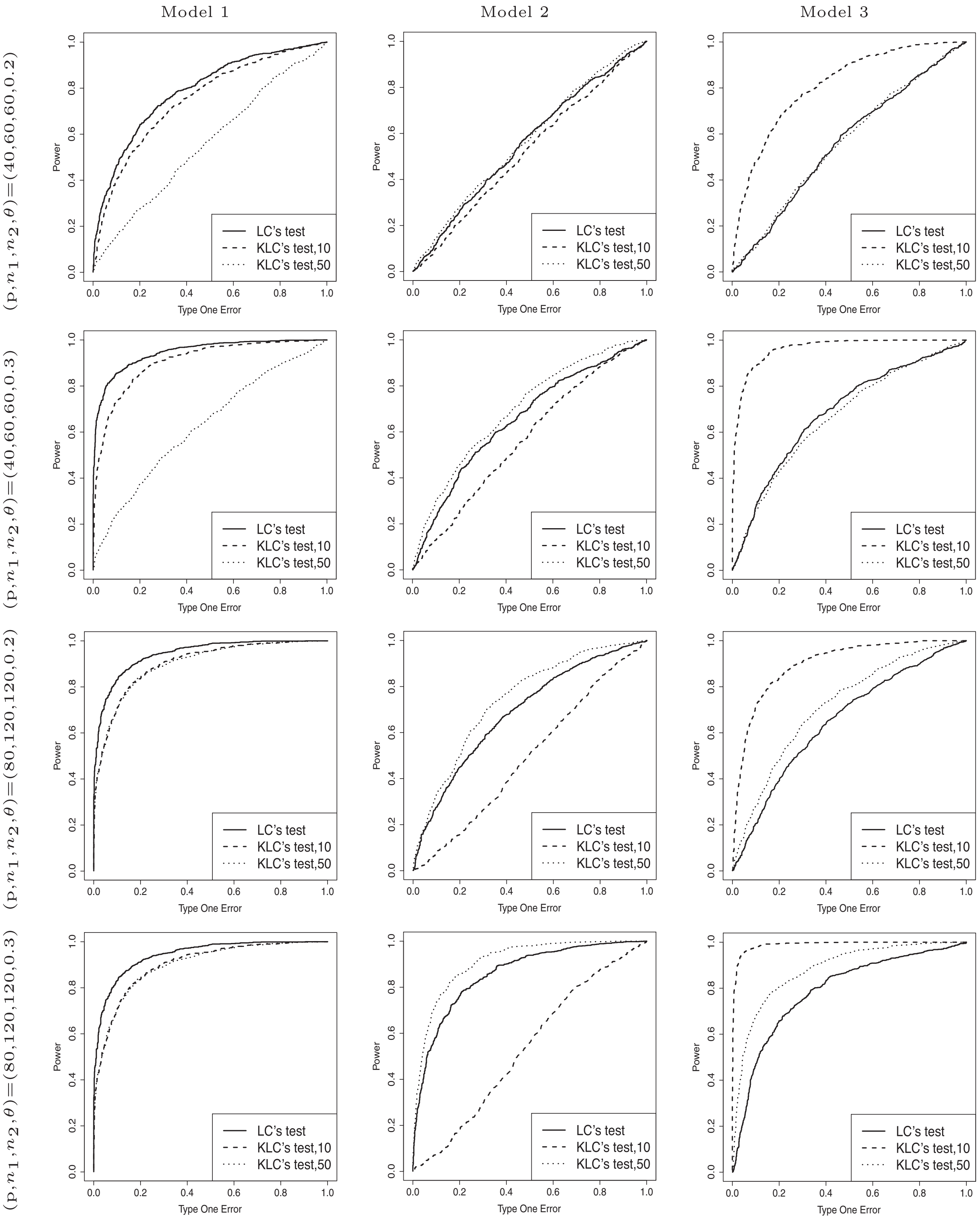

Fig. 2 Comparison between linear method and kernel method. Simulations under different setups. Setup of the first column is under model 1, the second column is under model 2 and the third column is under model 3. First row: (p, n_1, n_2, θ) = (40, 60, 60, 0.2); Second row: (p, n_1, n_2, θ) = (40, 60, 60, 0.3); Third row: (p, n_1, n_2, θ) = (80, 120, 120, 0.2); Fourth row: (p, n_1, n_2, θ) = (80,120,120, 0.3)

affect the co-expression patterns among the genes in the pathway.

Model

We use covariance test to find the pathway-SNP modules. There are three key elements of covariance test for a given gene set S. We consider the strategy similar to [26].

- **Calculation of T statistics.** We calculate a T statistics that reflects the difference of the covariance matrix of the two classes of samples. The statistics is calculated by estimating the Frobenius norm of the difference of the covariance matrix. We first perform the method by [27] to do the test:

  $$H_0 : \Sigma_1 = \Sigma_2,\ \ H_1 : \Sigma_1 \neq \Sigma_2 \tag{1}$$

  where Σ_1 is the covariance matrix of gene expression under one genotype and Σ_2 is that of gene expression under the other genotype. Then we consider the nonlinear relationship between gene expressions by applying kernel method.
- **Estimation of significance level of T statistics.** We estimate the statistical significance (nominal P value) of the T statistics by using an empirical SNP-based permutation test procedure that preserves the complex correlation structure of the gene expression data. Specifically, we permute the SNP labels and recompute the T statistics of the gene set for the permuted data, which generates a null distribution for the T statistics. The empirical, nominal P value of the observed T statistics is then calculated relative to this null distribution. Importantly, the permutation of class labels preserves gene-gene correlations and, thus, provides a more biologically reasonable assessment of significance than would be obtained by permuting genes.

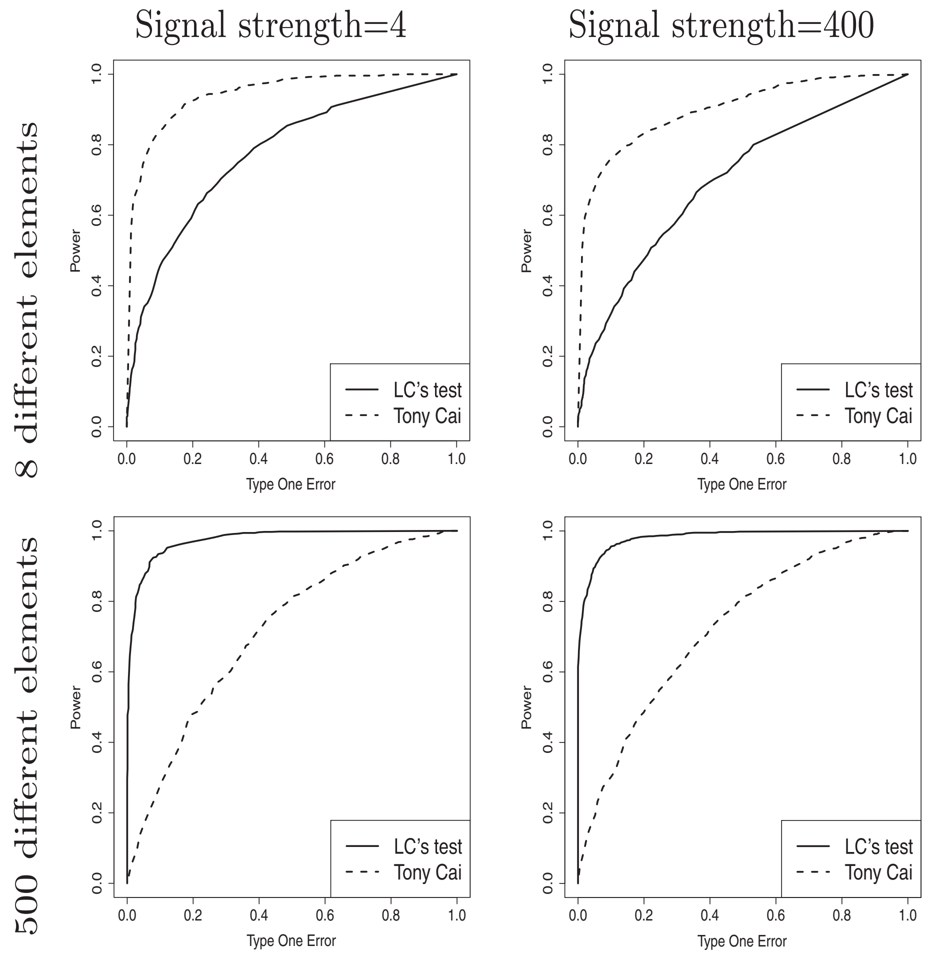

Fig. 3 Comparison between Chen's linear method and other method. *Topleft*: The two covariance matrices have eight different elements, each with a magnitude generated from $Unif(0, 4) * \max_{1\leq j\leq p} \sigma_{jj}$; *Topright*: The two covariance matrices have eight different elements, each with a magnitude generated from $Unif(0, 400) * \max_{1\leq j\leq p} \sigma_{jj}$; *Bottomleft*: The two covariance matrices have 500t different elements, each with a magnitude generated from $Unif(0, 4) * \max_{1\leq j\leq p} \sigma_{jj}$; *Bottomright*: The two covariance matrices have 500 different elements, each with a magnitude generated from $Unif(0, 400) * \max_{1\leq j\leq p} \sigma_{jj}$

- **Adjustment for multiple hypothesis testing.** When an entire database of gene sets is evaluated, we adjust the estimated significance level to account for multiple hypothesis testing. We first normalize the T statistics for each gene set to account for the size of the set, yielding a normalized T statistics. We then control the proportion of false positives by calculating the false discovery rate (FDR) corresponding to each NT statistics. The FDR is the estimated probability that a set with a given NT statistics represents a false positive finding; it is computed by comparing the tails of the observed and null distributions for the NT statistics. To capture the change of the structure of the gene network, we consider the covariance of the gene expression.

Test for high-dimensional covariance matrices

To simplify the problem, we just consider there are two possible values of each SNP. Covariance matrices under two genotypes of the SNP are denoted as Σ_1 and Σ_2, respectively. The primary interest is to test

$$H_0 : \Sigma_1 = \Sigma_2, \; H_1 : \Sigma_1 \neq \Sigma_2$$

which is a nontrivial statistical problem because the number of genes is greater than the number of samples sometimes. The test statistic for the hypothesis is formulated by targeting on $\mathrm{tr}(\Sigma_1 - \Sigma_2)^2$, the squared Frobenius norm of $\Sigma_1 - \Sigma_2$ [27]. Specifically, the test statistic is

$$\begin{aligned}
T_{n_1,n_2} &= A_{n_1} + A_{n_2} - 2C_{n_1 n_2} \\
A_{n_h} &= \frac{1}{n_h(n_h-1)} \sum_{i \neq j} (X'_{hi} X_{hj})^2 \\
&\quad - \frac{2}{n_h(n_h-1)(n_h-2)} \sum_{i,j,k}^{*} X'_{hi} X_{hj} X'_{hj} X_{hk} \\
&\quad + \frac{1}{n_h(n_h-1)(n_h-2)(n_h-3)} \sum_{i,j,k,l}^{*} X'_{hi} X_{hj} X'_{hk} X_{hl} \\
C_{n_1 n_2} &= \frac{1}{n_1(n_2)} \sum_{i} \sum_{j} \left(X'_{1i} X_{2j}\right)^2 \\
&\quad - \frac{1}{n_1 n_2 (n_1-1)} \sum_{i,k}^{*} \sum_{j} X'_{1i} X_{2j} X'_{2j} X_{1k} \\
&\quad - \frac{1}{n_1 n_2 (n_2-1)} \sum_{i,k}^{*} \sum_{j} X'_{2i} X_{1j} X'_{1j} X_{2k} \\
&\quad + \frac{1}{n_1 n_2 (n_1-1)(n_2-1)} \sum_{i,k}^{*} \sum_{j,l}^{*} X'_{1i} X_{2j} X'_{1k} X_{2l}
\end{aligned}$$

where h refers to a subpopulation with a particular SNP.

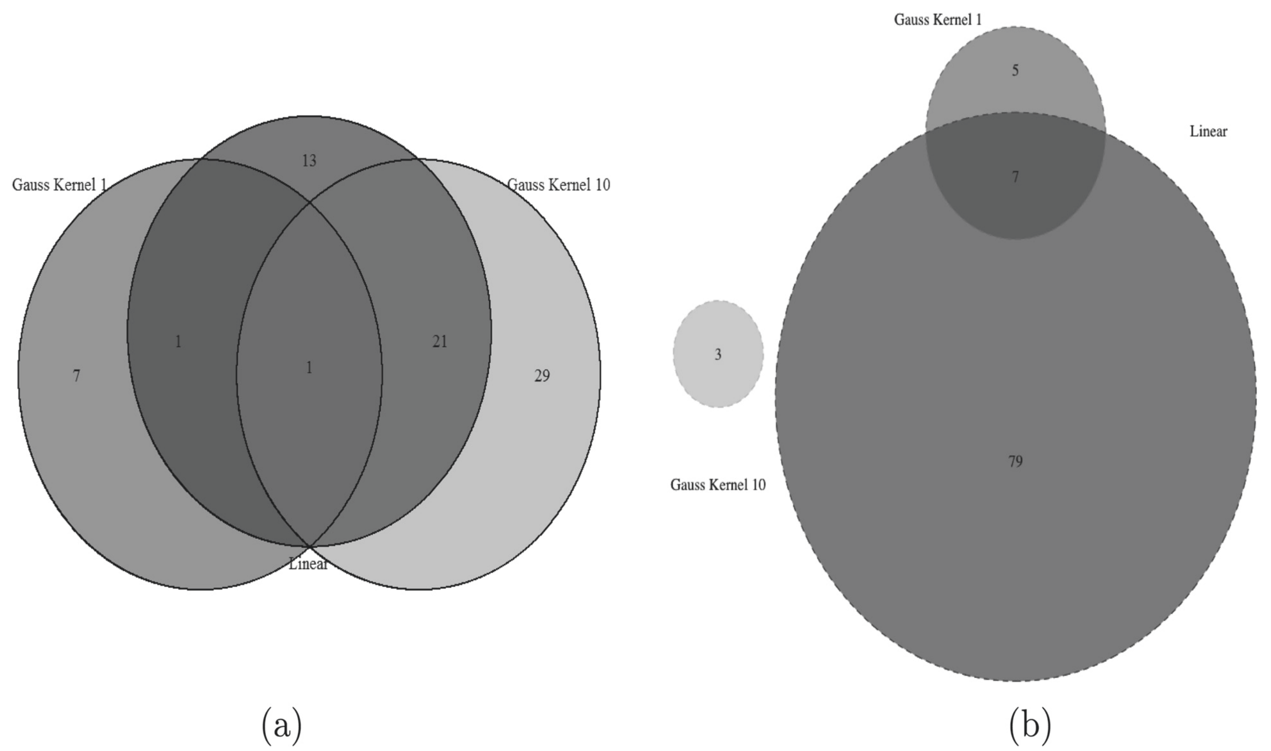

Fig. 4 Different modules detected under different conditions by different method. **a** Detected modules under ethanol condition. We found 36 modules by covariance test, 9 modules by kernel covariance test with parameter 1 and 51 modules by kernel covariance test with parameter 10; **b** Detected modules under glucose condition. We found 86 modules by covariance test, 3 modules by kernel covariance test with parameter 10 and 12 modules by kernel covariance test with parameter 1

Table 1 New associated pathways and SNPs under ethanol condition

Pathways	Associated markers
Glycolysis/Gluconeogenesis	*gOL02*$^{(10)}$
Synthesis and degradation of ketone bodies	*YLR257W*$^{(10)}$, *YLR261C*$^{(10)}$
Steroid biosynthesis	*YEL021W*^{L}, *YFR035C*^{L}, *YJL001W*^{L}, *YJR006W*$^{L,(10)}$, *YJL007C*$^{L,(10)}$,
Valine, leucine and isoleucine degradation	*YOR006C*^{L}, *NOR005W*$^{L,(10)}$, *YOR051C*$^{L,(10)}$, *YOR076C*$^{L,(10)}$
Valine, leucine and isoleucine biosynthesis	*NLR116W*$^{L,(10)}$, *YOR076C*^{L}, *YCL023C*$^{(10)}$, *YLR257W*$^{(10)}$
Histidine metabolism	*gOL02*$^{L,(10)}$, *YOR025W*$^{(10)}$
Tyrosine metabolism	*YFL019C*^{L}, *gOL02*$^{L,(10)}$
Phenylalanine metabolism	*gOL02*$^{L,(10)}$
beta-Alanine metabolism	*gOL02*$^{L,(10)}$
Taurine and hypotaurine metabolism	*gOL02*$^{(10)}$
Selenocompound metabolism	*YOR006C*$^{L,(10)}$, *YOR019W*$^{L,(10)}$, *YOR025W*$^{L,(10)}$, *NOR005W*$^{L,(10)}$
Purine metabolism	*NNL035W*$^{(1)}$
Cyanoamino acid metabolism	*YLR027C*$^{(1)}$
Arachidonic acid metabolism	*gPL09*$^{(1)}$
Linoleic acid metabolism	*YFL029C*$^{L,(1),(10)}$, *YFL019C*$^{(1)}$
Glyoxylate and dicarboxylate metabolism	*NNL035W*$^{(1)}$, *YNL074C*$^{(1)}$
Porphyrin and chlorophyll metabolism	*NBR008W*$^{(1)}$
Sphingolipid metabolism	*YHL047C*^{L}
Pantothenate and CoA biosynthesis	*YGL053W*^{L}, *NLR116W*$^{L,(10)}$
Terpenoid backbone biosynthesis	*YJL007C*$^{L,(10)}$, *YJL001W*$^{L,(10)}$, *YJR006W*$^{L,(10)}$, *NJR006C*$^{L,(10)}$
Sesquiterpenoid and triterpenoid biosynthesis	*YOR334W*^{L}, *YOR343C*^{L}, *YLR261C*$^{(10)}$, *NLR116W*$^{(10)}$, *YLR257W*$^{(10)}$
Metabolic pathways	*YIL078W*$^{(10)}$, *YLR257W*$^{(10)}$, *YLR308W*$^{(10)}$, *NNL035W*$^{(10)}$, *gOL02*$^{(10)}$, *YOR006C*$^{(10)}$, *YOR051C*$^{(10)}$, *YOR019W*$^{(10)}$, *YNL066W*$^{(10)}$, *YLR261C*$^{(10)}$
Biosynthesis of secondary metabolites	*YOR025W*$^{(10)}$, *YOR063W*$^{(10)}$
Carbon metabolism	*YOR019W*$^{(10)}$
2-Oxocarboxylic acid metabolism	*YLR261C*$^{L,(10)}$, *YLR308W*$^{L,(10)}$, *YCL022C*$^{(10)}$, *YLR265C*$^{(10)}$, *NLR116W*$^{(10)}$, *YLR322W*^{L}
mRNA surveillance pathway	*YOR072W*^{L}
Mismatch repair	*gKR08*^{L}
Non-homologous end	*YGR006W*^{L}
Biosynthesis of amino acids	*gOL02*$^{(10)}$
MAPK signaling pathway	*YDR164C*$^{(10)}$, *gDR10*$^{(10)}$

L means detected by covariance test, (1) means detected by kernel covariance test with parameter 1 and (10) means detected by kernel covariance test with parameter 10. The FDR of the covariance test, kernel covariance test with parameter 1 and kernel covariance test with parameter 10 are 0.25, 0.33 and 0.25 respectively. The FWER of the test by Tony Cai is 0.2

For test
$H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3, H_1 : \Sigma_1 \neq \Sigma_2$ or $\Sigma_2 \neq \Sigma_3$

We consider $\mathrm{tr}(\Sigma_1 - \Sigma_2)^2 + \mathrm{tr}(\Sigma_2 - \Sigma_3)^2$. Specifically, the test statistic is

$$T_{n_1,n_2} + T_{n_2,n_3}$$

where T_{n_2,n_3} is defined similar to T_{n_1,n_2}.

Kernel method

We generalize the method of [27] to the kernel space inspired by the method of [28]. We give the similar definition of Frobenius norm and covariance matrix. Let p_x and p_y be Borel probability measures defined on a domain Ω. Given observations $X := \{x_1, \ldots, x_m\}$ and $Y := \{y_1, \ldots, y_n\}$, drawn independently and identically distributed(i.i.d.) from p_x and p_y, respectively.

Definition (HSDCC) Given separable reproducing kernel Hilbert space (RKHS) $\mathcal{F}$, and measures p_x, p_y over $(\mathcal{X}, \Gamma)$, we define the Hilbert-Schmidt Different Covariance Criterion(HSDCC) as the squared HS-norm of the difference of covariance Σ_{xx} and Σ_{yy}:

$$HSDCC(p_x, p_y, \mathcal{F}) :=\| \Sigma_{xx} - \Sigma_{yy} \|_{HS}^2$$

The detailed computation of above norm can be found in text of the Additional file 1. We give the unbiased statistics to $HSDCC(P_x, P_y, \mathcal{F})$ like [27]

$$\begin{aligned} A_{n_h} = & \frac{1}{n_h(n_h-1)} \sum_{i \neq j} k(X_{hi}, X_{hj})^2 \\ & - \frac{2}{n_h(n_h-1)(n_h-2)} \sum_{i,j,k}^{*} k(X_{hi}, X_{hj}) k(X_{hj}, X_{hk}) \\ & + \frac{1}{n_h(n_h-1)(n_h-2)(n_h-3)} \sum_{i,j,k,l}^{*} k(X_{hi}, X_{hj}) k(X_{hk}, X_{hl}) \end{aligned}$$

Table 2 New associated pathways and SNPs under glucose condition

Pathways	Associated markers
Synthesis and degradation of ketone bodies	$gJL07^{(10)}$
Inositol phosphate metabolism	$YBR259W^{(10)}$
Riboflavin metabolism	$YML056C^{(10)}$
Fatty acid degradation	$YBR045C^{(1)}$
Cysteine and methionine metabolism	$YGL195W^{(1)}$
Valine, leucine and isoleucine biosynthesis	$YCL025C^{L}$, $NGR093C^{L}$
	$YOR253W^{L}$, $YOR274W^{L}$
	$YOR326W^{L}$, $YOR334W^{L}$
	$YOR343C^{L}$, $YCL022C^{(1)}$
Phenylalanine metabolism	$YJR040W^{L}$, $YOL123W^{L}$
	$YOL118C^{L}$, $YOL106W^{L}$
	$YOL093W^{L}$, $YOL088C^{L}$
	$gOL02^{L,(1)}$
beta-Alanine metabolism	$YBR271W^{L}$, $gOL02^{L,(1)}$
	$NJR007C^{L}$, $YOL106W^{L}$
Arachidonic acid metabolism	$YIR022W^{L,(1)}$
Vitamin B6 metabolism	$YKL118W^{(1)}$
Porphyrin and chlorophyll metabolism	$YML071C^{(1)}$, $gFL02^{L}$
Degradation of aromatic compounds	$YMR316C^{L,(1)}$, $YMR316C^{L}$
ABC transporters	$YBR131W^{L,(1)}$, $YBR137W^{L}$
Glycolysis/Gluconeogenesis	$YJR071W^{L}$
Pentose phosphate pathway	$NOL043W^{L}$, $YOL151W^{L}$
	$YOL123W^{L}$, $YOL094C^{L}$
	$YOL093W^{L}$, $YOL088C^{L}$
	$gOL02^{L}$
Pentose and glucuronate interconversions	$YGL263W^{L}$
Purine metabolism	$YLR140W^{L}$
Pyrimidine metabolism	$YBL010C^{L}$, $YGL217C^{L}$
Glycine, serine and threonine metabolism	$YCL065W^{L}$, $YJR038C^{L}$
Lysine biosynthesis	$YBR087W^{L}$
Histidine metabolism	$YBR271W^{L}$, $NJR007C^{L}$
	$YJR040W^{L}$, $YJR057W^{L}$
	$YOL106W^{L}$, $YOL093W^{L}$,
	$gOL02^{L,(1)}$
Tyrosine metabolism	$YOL123W^{L}$, $YOL106W^{L}$
	$YOL094C^{L}$, $gOL02^{L,(1)}$
Cyanoamino acid metabolism	$YDR351W^{L}$
Starch and sucrose metabolism	$YER095W^{L}$, $YER116C^{L}$
Linoleic acid metabolism	$NDR174C^{L}$
Butanoate metabolism	$YBR271W^{L}$
Pantothenate and CoA biosynthesis	$YOR274W^{L}$
Lipoic acid metabolism	$gLL01^{L}$, $YNL158W^{L}$
Folate biosynthesis	$NML013W^{L}$, $YNL066W^{L}$, $YNL050C^{L}$
Sesquiterpenoid and triterpenoid biosynthesis	$YMR084W^{L}$
Aminoacyl-tRNA biosynthesis	$YCL065W^{L}$, $YCL047C^{L}$
	$YCL039W^{L}$, $NJR007C^{L}$, $YNL010W^{L}$
Biosynthesis of unsaturated fatty acids	$YFL029C^{L}$
Metabolic pathways	$YCL065W^{L}$, $YJR071W^{L}$
Biosynthesis of secondary metabolites	*YJR038C6L*
Biosynthesis of amino acids	$YJR071W^{L}$, $YOL123W^{L}$
	$YOL118C^{L}$, $YOL106W^{L}$
	$YOL094C^{L}$, $YOL093W^{L}$
	$YOL088C^{L}$, $gOL02^{L}$
Ribosome	$YAR035W^{L}$, $YJL026W^{L}$
RNA transport	$YBL010C^{L}$
RNA polymerase	$YLR140W^{L}$
Proteasome	$YBL010C^{L}$
Phosphatidylinositol signaling system	$YBR045C^{L}$
Meiosis - yeast	$YOL106W^{L}$

L means detected by covariance test, (1) means detected by kernel covariance test with parameter 1 and (10) means detected by kernel covariance test with parameter 10. The FDR of the covariance test, kernel covariance test with parameter 1 and kernel covariance test with parameter 10 are 0.20, 0.24 and 0.33 respectively. The FWER of the test by Tony Cai is 0.2

$$\begin{aligned}C_{n_1n_2} = {} & \frac{1}{n_1(n_2)}\sum_i\sum_j k(X_{1i},X_{2j})^2 \\ & - \frac{1}{n_1n_2(n_1-1)}\sum_{i,k}^{*}\sum_j k(X_{1i},X_{2j})k(X_{2j},X_{1k}) \\ & - \frac{1}{n_1n_2(n_2-1)}\sum_{i,k}^{*}\sum_j k(X_{2i},X_{1j})k(X_{1j},X_{2k}) \\ & + \frac{1}{n_1n_2(n_1-1)(n_2-1)}\sum_{i,k}^{*}\sum_{j,l}^{*} k(X_{1i},X_{2j})k(X_{1k},X_{2l})\end{aligned}$$

$$T_{n_1,n_2} = A_{n_1} + A_{n_2} - 2C_{n_1n_2}$$

For test $H_0 : \Sigma_{xx} = \Sigma_{yy} = \Sigma_{zz}, H_1 : \Sigma_{xx} \neq \Sigma_{yy}$ or $\Sigma_{yy} \neq \Sigma_{zz}$ We consider $\parallel \Sigma_{xx} - \Sigma_{yy} \parallel_{HS}^2 + \parallel \Sigma_{yy} - \Sigma_{zz} \parallel_{HS}^2$. Specifically, the test statistic is

$$T_{n_1,n_2} + T_{n_2,n_3}$$

where T_{n_2,n_3} is defined similar to T_{n_1,n_2}.

Results

Simulation

Comparison between linear method and kernel method

We performed a simulation study to evaluate the power of the proposed kernel methods, and compared the results with the primary method by [27]. Three models have been considered, as below.

Model 1: $X_{ijk} = Z_{ijk} + \theta Z_{ijk+1}$, where Z_{ijk} were i.i.d. standard normally distributed, and $\theta = 0.5$ in the null hypothesis while 0.2 or 0.3 in the alternative hypothesis.

Model 2: $X_{ijk} = Z_{ijk}^3 + \theta Z_{ijk+1}^3$, where Z_{ijk} and θ were defined the same as that in Model 1.

Model 3: $X_{ijk} = e^{Z_{ijk}} + \theta e^{Z_{ijk+1}}$, where Z_{ijk} and θ were defined the same as that in Model 1.

The correlation between variables are linear in model 1, while the correlation between variables are nonlinear in model 2 and 3.

We chose (p, n_1, n_2)=(40, 60, 60) and (80, 120, 120) respectively. The power of the tests are shown by ROC curves (Fig. 2). All the simulation results reported were based on 1000 simulations. We can see from the simulation that kernel methods with some parameters have higher power than the linear test when the true relationships between variables are nonlinear (Model 2 and Model 3). A similar simulation results with different setup of parameters can be found in Additional file 1: Figure S3.

Comparison between Chen et al.'s linear method and other method

We conducted a simulation to compare the power of Chen et al.'s method [27] and Tony Cai et al.'s method [29]. We consider four simulation setups represented different signal quantities and strength, the first of which is the same as the model 2 in [29].

Model 1: Let

$$\Sigma^* = (\sigma_{ij}^*),\ where\ \omega_{ij}^* = 0.5^{|i-j|}\ for\ 1 \le i,j \le p.$$

$$\Sigma = D^{1/2}\Sigma^* D^{1/2},\ where\ D = (d_{ij}),\ d_{ii} = Unif(0.5, 2.5), 1 \le i \le p$$

$$\Sigma_1 = \Sigma + \delta I,\ \Sigma_2 = \Sigma + U + \delta I,\ where\ \delta$$

$$= |\min\{\lambda_{min}(\Sigma + U), \lambda_{min}(\Sigma)\}| + 0.05,$$

$U = (u_{kl})$ be a matrix with eight random nonzero entries, each with a magnitude generated from $Unif(0,4) * \max_{1 \le j \le p} \sigma_{jj}$. The number of each class samples is 50 and the number of variables is 50.

Model 2: $U = (u_{kl})$ be a matrix with eight random nonzero entries, each with a magnitude generated from $Unif(0,400) * \max_{1 \le j \le p} \sigma_{jj}$.

Model 3: $U = (u_{kl})$ be a matrix with 500 random nonzero entries, each with a magnitude generated from $Unif(0,4) * \max_{1 \le j \le p} \sigma_{jj}$.

Model 4: $U = (u_{kl})$ be a matrix with 500 random nonzero entries, each with a magnitude generated from $Unif(0,400) * \max_{1 \le j \le p} \sigma_{jj}$.

As shown in Fig. 3, under the sparse setups (Model 1 and 2), the results of Tony Cai et al.'s method is much better than those of Chen et al.'s method. Chen et al.'s method is better than Tony Cai et al.'s method when the setups are not sparse (Model 3 and 4). Since Tony Cai et al.'s method corresponds to testing each element in the covariance matrix by Hoteling's test and then give the judgement according to the maximum statistic of all of the Hoteling's tests, so Chen et al.'s

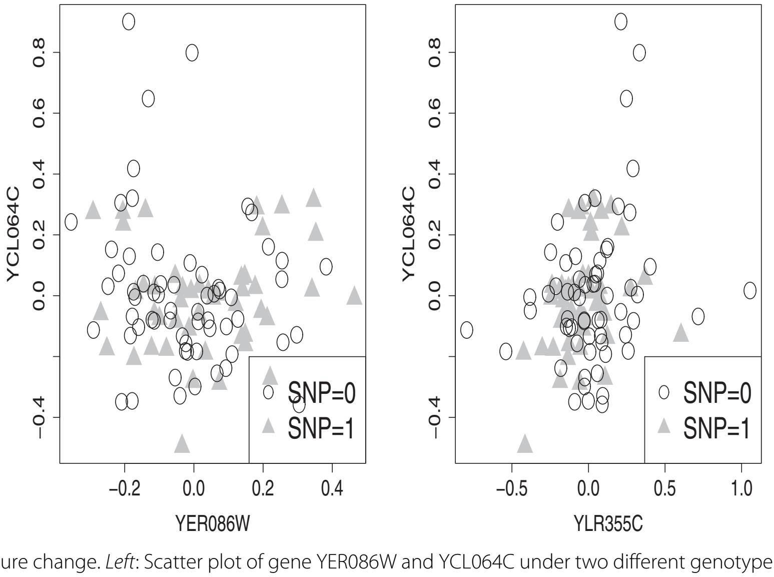

Fig. 5 Isoform-specific structure change. *Left*: Scatter plot of gene YER086W and YCL064C under two different genotypes of YCL023C; *Right*: Scatter plot of gene YLR355C and YCL064C under two different genotypes of YCL023C. We found the associated pathway-SNP modules only by kernel covariance test. The scatter figures show that YER086W-YCL064C and YLR355C-YCL064C were nonlinear correlated under genotypes of marker YCL023C

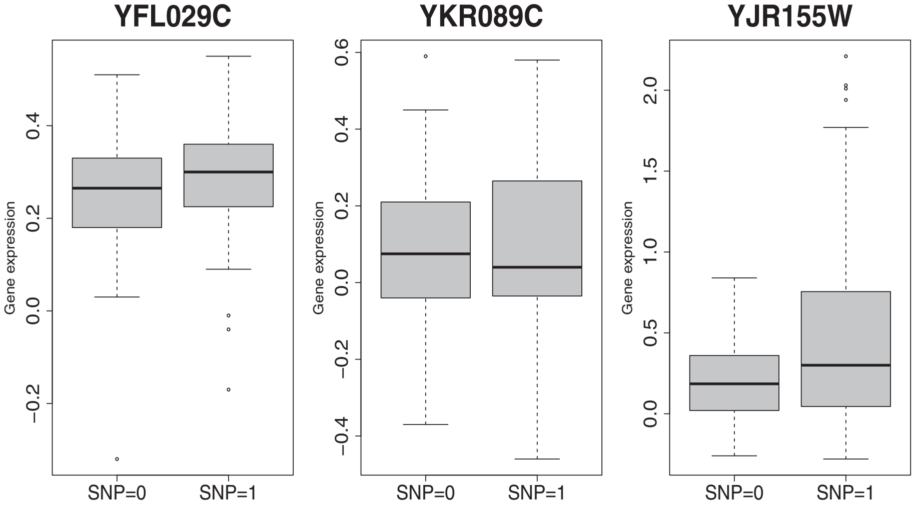

Fig. 6 Failure to detect correlation between single gene expression level and genotype of YFL029C. *Left*: Boxplot of expression level of YFL029C; *middle*, Boxplot of expression level of YKR089C; *right*: Boxplot of expression level of YJR155W. We can see that the means of YKR089C and YJR155W expressions do not change significantly

linear method has higher power than bi-variate model when the setups are not sparse. A similar simulation results with different number of samples can be found in Additional file 1: Figure S2.

Real data results

Associated SNP and pathways

We analyzed the yeast dataset collected by Kruglyak and colleagues [30]. The expression data were downloaded from http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0060083, with 4482 genes measured in 109 segregants derived from a cross between BY and RM. The experiments were performed under two conditions, glucose and ethanol. We did the pre-processing like [10], after which 4419 genes and 820 merged markers remained. We mapped 4419 genes to 103 pathways and analyzed the effect of each SNP to each pathway. Therefore, we tested 103*820 times. The algorithm was implemented in R, which can be found at http://www.math.pku.edu.cn/teachers/dengmh/NetworkBiomarker.

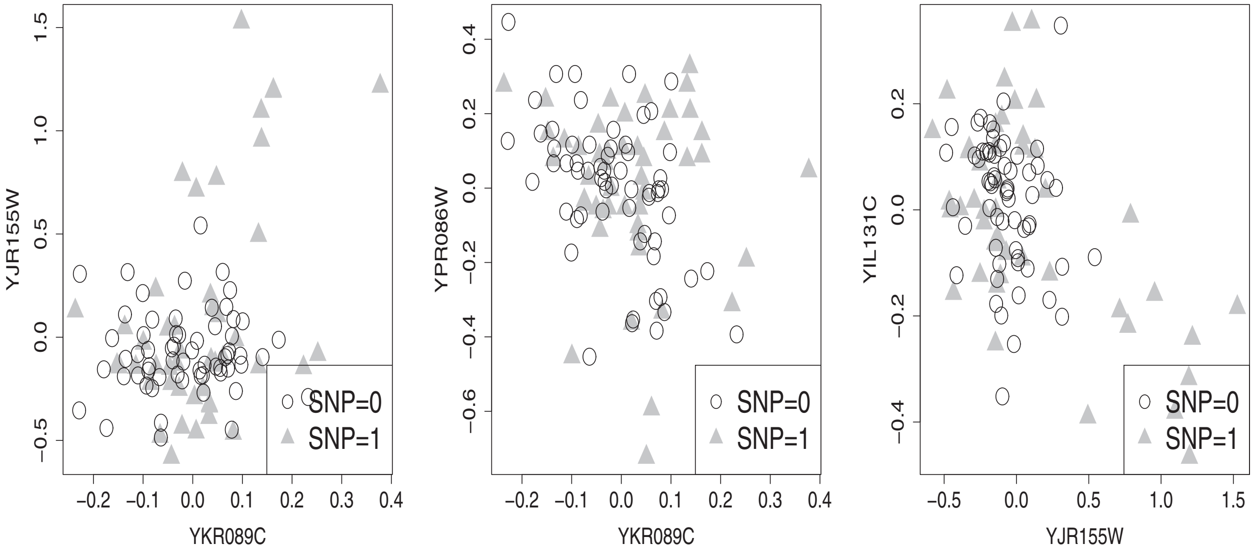

Fig. 7 Three examples of differential coexpression patterns of 2 genes due to genotype of YFL029C. *Left*: the co-expression patterns between the two genes YKR089C and YJR155W depend on the genotype of YFL029C. For samples with genotype 1, the co-expression correlation is different from the other samples. *Middle*: the co-expression patterns between the two genes YKR089C and YPR086W depend on the genotype of YFL029C. *Right*: the co-expression patterns between the two genes YJR155W and YIL131C depend on the genotype of YFL029C

We performed covariance test and kernel covariance with parameter 1 and 10 respectively to the pathway-SNP adjusted modules. We consider 103 pathways in KEGG [25] (the number of genes in each pathway can be found in the Additional file 1: Figure S1) and 820 merged markers under ethanol and glucose condition respectively. We

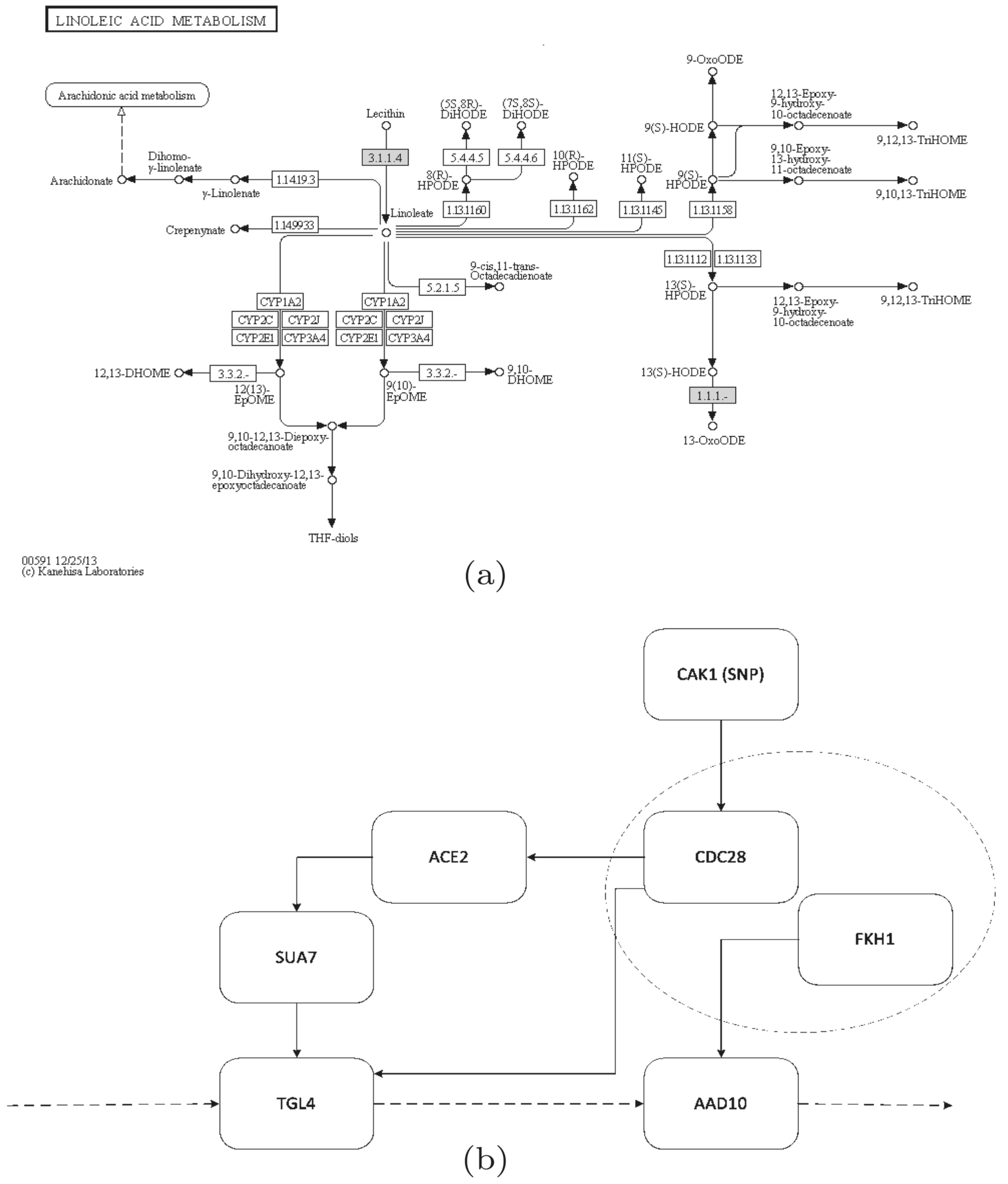

Fig. 8 The possible regulatory mechanism of the marker YFL029C to Linoleic acid metabolism pathway. **a** KEGG pathway of Linoleic acid metabolism pathway; **b** The potential regulatory relationship between marker YFL029C and Linoleic acid metabolism pathway

found 72 pathway-SNP modules under ethanol condition and 94 modules under glucose condition. Specifically, we found 36 modules by covariance test, 9 modules by kernel covariance test with parameter 1 and 51 modules by kernel covariance test with parameter 10 under ethanol condition, while 86 modules by covariance test, 3 modules by kernel covariance test with parameter 10 and 12 modules by kernel covariance test with parameter 1 under glucose condition (Fig. 4). Table 1 showed the associated pathways and SNPs under ethanol condition while Table 2 showed the associated pathways and SNPs under glucose condition.

Kernel Method found isoform-specific structure change

In our result, we found Valine, leucine and isoleucine biosynthesis pathway was associated with YCL023C marker only by kernel method under ethanol condition. Figure 5 shows the non-linear correlation between two pairs of genes, YER086W-YCL064C and YLR355C-YCL064C were nonlinear correlated with genotypes of YCL023C. And more than 10 isoforms of YER086W and 6 isoforms of YLR355C have been found (Saccharomyces Genome Database, http://www.yeastgenome.org/). The nonlinear correlation between two pairs of genes might be caused by samples in different isoforms. Specifically, two genes may be positive correlated under one isoform while negetive correlated under another isoform. However, the correlation of two genes might be missed if when we only considered linear correlation.

Linoleic acid metabolism is associated to cell cycle

Our method found YFL029C is associated with linoleic acid metabolism pathway under ethanol condition. With single gene correlation analysis, both of the mean of expression levels of YKR089C (TGL4) and YJR155W (AAD10) were not associated with YFL029C (Fig. 6 middle and right). Specifically, with p-value 0.5174 and 0.002804 (not significant for multiple test). However, the scatterplot after correction shows the correlation of two genes change apparently under YFL029C (Fig. 7 left). Under one status of SNP, the two genes are positive correlated while under the other status of SNP, the two are nearly independent. To understand this from the biological meaning which was showned in Fig. 8b, we found that marker locates in gene CAK1 (The expression of CAK1 is slightly different under two SNP status which was shown in Fig. 6 left.), the product of which can increase the activity of CDC28 [31]. CDC28 plays an important role in cell cycle. It can control the progress of cell cycle by phosphorylate different transcription factor. In our case, CDC28 phosphorylate ACE2 [32] which can increase the activity of transcription factor, SUA7 [33]. SUA7 is the transcription factor of TGL4, which is a lipase in linoleic acid metabolism pathway. Meanwhile, CDC28 and FKH1 can form complex [34] and FKH1 is the transcription factor of AAD10, which is another enzyme in linoleic acid metabolism pathway. The correlation between YKR089C and its TF was shown in Fig. 7 middle and the correlation between YJR155W and its TF was shown in Fig. 7 right. From the structure of the pathway in KEGG [25] as shown in Fig. 8a, the different status of the SNP YFL029C might lead to different amounts of intermediate product in the pathway.

Discussion and conclusion

We propose a network based covariance test to identify the marker which affects the structure of a pathway. It has an advantage that a static network structure is not assumed. The biomarker we defined is the SNP associated to the structure of genes in the pathway. Considering two genes may have different correlations under different isoforms which is hard to detect by linear test, so we also consider the nonlinear test. We identified a total of 166 modules, with each module consisting of a group of genes and one eQTL where the eQTL regulate the co-expression patterns of the group of genes. We found that many of these modules have biological interpretations. Till now, we consider the difference of two networks by covariance matrix and covariance operators. We will focus on difference of precision matrix in the future research.

Additional file

Additional file 1: Supplementary materials for the computation of HSDCC and additional figures Figures S1–S3. (PDF 1474 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MD and NT conceived of the project. The flowchart of the strategy was designed by MD and HY. HY and ZL developed the kernel method. HY conducted the simulation and real data analysis. HY wrote the first draft of the manuscript, which all authors revised and approved.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Nos. 31171262, 31428012, 31471246), and the National Key Basic Research Project of China (No. 2015CB910303). We thank Dr. Lin Wang for helpful feedback during the saccharomyces cerevisiae data pre-processing and Dr. Minping Qian and Dr. Kai Song's discussions. The publication costs for this article are from the National Natural Science Foundation of China (No. 31471246).

Declarations

This article has been published as part of BMC Systems Biology Volume 10 Supplement 1, 2016: Selected articles from the Fourteenth Asia Pacific Bioinformatics Conference (APBC 2016): Systems Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/10/S1.

Author details

[1]LMAM, School of Mathematical Sciences, Peking University, Yiheyuan Road, 100871 Beijing, China. [2]Department of Chemical Pathology, Prince of Wales Hospital, Faculty of Medicine, The Chinese University of Hong Kong, Shatin,

Hong Kong, China. [3]Center for Quantitative Biology, Peking University, Yiheyuan Road, 100871 Beijing, China. [4]Center for Statistical Sciences, Peking University, Yiheyuan Road, 100871 Beijing, China.

Published: 11 January 2016

References

1. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of gwas discovery. Am J Hum Genet. 2012;90(1):7–24.
2. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015;16:197–212. doi:10.1038/nrg3891.
3. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nat Rev Genet. 2009;10(3):184–94.
4. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. Phil Trans R Soc B: Biol Sci. 2013;368(1620):20120362.
5. Wang P, Dawson JA, Keller MP, Yandell BS, Thornberry NA, Zhang BB, et al. A model selection approach for expression quantitative trait loci (eqtl) mapping. Genetics. 2011;187(2):611–21.
6. Li KC. Genome-wide coexpression dynamics: theory and application. Proc Natl Acad Sci. 2002;99(26):16875–80.
7. Sun W, Yuan S, Li KC. Trait-trait dynamic interaction: 2d-trait eqtl mapping for genetic variation study. BMC Genomics. 2008;9(1):242.
8. Ho YY, Parmigiani G, Louis TA, Cope LM. Modeling liquid association. Biometrics. 2011;67(1):133–41.
9. Chen J, Xie J, Li H. A penalized likelihood approach for bivariate conditional normal models for dynamic co-expression analysis. Biometrics. 2011;67(1):299–308.
10. Wang L, Zheng W, Zhao H, Deng M. Statistical analysis reveals co-expression patterns of many pairs of genes in yeast are jointly regulated by interacting loci. PLoS Genet. 2013;9(3):1003414.
11. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human b cells. Nat Genet. 2005;37(4):382–90.
12. Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, et al. A predictive model for transcriptional control of physiology in a free living cell. Cell. 2007;131(7):1354–65.
13. Pereira-Leal JB, Enright AJ, Ouzounis CA. Detection of functional modules from protein interaction networks. Proteins Struct Function Bioinforma. 2004;54(1):49–57.
14. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009;10(3):515–34. doi:10.1093/biostatistics/kxp008.
15. Naylor MG, Lin X, Weiss ST, Raby BA, Lange C. Using canonical correlation analysis to discover genetic regulatory variants. PloS ONE. 2010;5(5):10395.
16. Lin D, Zhang J, Li J, Calhoun VD, Deng HW, Wang YP. Group sparse canonical correlation analysis for genomic data integration. BMC Bioinformatics. 2013;14(1):245.
17. Li Y, Nan B, Zhu J. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. Biometrics. 2015.
18. Kim S, Xing EP. Statistical estimation of correlated genome associations to a quantitative trait network. PLoS Genet. 2009;5(8):1000587.
19. Zhang L, Kim S. Learning gene networks under snp perturbations using eqtl datasets. PLoS Comput Biol. 2014;10(2):1003420.
20. Casale FP, Rakitsch B, Lippert C, Stegle O. Efficient set tests for the genetic analysis of correlated traits. Nat Meth. 2015.
21. Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol. 2012;8.
22. Zhou S, Carraway KL, Eck MJ, Harrison SC, Feldman RA, Mohammadi M, et al. Catalytic specificity of protein-tyrosine kinases is critical for selective signalling. Nature. 1995;373(6514):536–9.
23. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc. Series B (Methodological). 1996;58(1):267–88.
24. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96(456):1348–60.
25. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 1999;27:29–34.
26. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci of the USA. 2005;102(43):15545–50.
27. Li J, Chen SX. Two sample tests for high-dimensional covariance matrices. Ann Stat. 2012;40(2):908–40.
28. Jain S, Simon HU, Tomita E. Algorithmic Learning Theory. 16th International Conference, ALT 2005, Singapore, October 8–11, 2005, Proceedings. Springer-Verlag Berlin Heidelberg 2005.
29. Cai T, Liu W, Xia Y. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. J Am Stat Assoc. 2013;108(501):265–77.
30. Smith EN, Kruglyak L. Gene–environment interaction in yeast gene expression. PLoS Biol. 2008;6(4):83.
31. Kurat CF, Wolinski H, Petschnigg J, Kaluarachchi S, Andrews B, Natter K, et al. Cdk1/cdc28-dependent activation of the major triacylglycerol lipase tgl4 in yeast links lipolysis to cell-cycle progression. Mol Cell. 2009;33(1): 53–63.
32. Ubersax JA, Woodbury EL, Quang PN, Paraz M, Blethrow JD, Shah K, et al. Targets of the cyclin-dependent kinase cdk1. Nature. 2003;425(6960):859–64.
33. Zhang DY, Dorsey MJ, Voth WP, Carson DJ, Zeng X, Stillman DJ, et al. Intramolecular interaction of yeast tfiib in transcription control. Nucleic Acids Res. 2000;28(9):1913–20.
34. Venters BJ, Wachi S, Mavrich TN, Andersen BE, Jena P, Sinnamon AJ, et al. A comprehensive genomic binding map of gene and chromatin regulatory proteins in saccharomyces. Mol Cell. 2011;41(4):480–92.