Title: PaIRKAT: A pathway integrated regression-based kernel association test with applications to metabolomics and COPD phenotypes

Authors: Charlie M. Carpenter, Weiming Zhang, Lucas Gillenwater, Cameron Severn, Tusharkanti Ghosh, Russel Bowler, Katerina Kechris, Debasish Ghosh

Affiliations:
1. Department of Biostatistics and Informatics, University of Colorado Denver, Anschutz medical campus, Denver, CO. USA.
2. Syneos Health, Morrisville, NC. USA
3. Computational Bioscience Program, University of Colorado Denver, Anschutz medical campus, Denver, CO. USA.
4. Department of Medicine, National Jewish Health, Denver; University of Colorado Denver, Anschutz Medical Campus, Denver, CO.

† Previously affiliated with Anschutz Medical Campus, Denver, CO. USA

* Corresponding author: charles.carpenter@cuanschutz.edu

Author Contributions:
CMC wrote the manuscript, performed the simulation studies, analyzed the COPDGene data, and assisted in coding the method’s Shiny app. WZ performed initial methods development. CS performed the majority of coding and visualizations for the Shiny app. LG conducted the COPDGene data generation and cleaning. RB led the team responsible for the COPDGene data
generation, management, and cleaning. TG assisted with COPDGene data cleaning and preprocessing. KK and DG provided guidance for methods development, data analysis, and writing the manuscript.

Keywords: Kernel machine regression, Pathway analysis, Metabolome, Metabolic pathways, Graph Laplacian, Graph Regularization

Abstract

High-throughput data such as metabolomics, genomics, transcriptomics, and proteomics have become familiar data types within the “-omics” family. For this work, we focus on subsets that interact with one another and represent these “pathways” as graphs. Observed pathways often have disjoint components, i.e. nodes or sets of nodes (metabolites, etc.) not connected to any other within the pathway which notably lessens testing power. In this paper we propose the Pathway Integrated Regression-based Kernel Association Test (PaIRKAT), a new kernel machine regression method for incorporating known pathway information into the semi-parametric kernel regression framework. This paper also contributes an application of a graph kernel regularization method for overcoming disconnected pathways. By incorporating a regularized or “smoothed” graph into a score test, PaIRKAT is capable of providing more powerful tests for associations between biological pathways and phenotypes of interest and will be helpful in identifying novel pathways for targeted clinical research. We evaluate this method through several simulation studies and an application to real metabolomics data from the COPDGene study. Our simulation studies illustrate the robustness of this method to incorrect and incomplete pathway knowledge, and the real data analysis shows meaningful improvements of
testing power in pathways. PaIRKAT was developed for application to metabolomic pathway
data, but the techniques are easily generalizable to other data sources with a graph-like structure.

Author Summary

PaIRKAT is a tool for improving testing power on high dimensional data by including
graph topography in the kernel machine regression setting. Studies on high dimensional data can
struggle to include the complex relationships between variables. The semi-parametric kernel
machine regression model is a powerful tool for capturing these types of relationships. They
provide a framework for testing for relationships between outcomes of interest and high
dimensional data such as metabolomic, genomic, or proteomic pathways. Our paper proposes
PaIRKAT, a method for including known biological connections between high dimensional
variables by representing them as edges of ‘graphs’ or ‘networks.’ It is common for nodes (e.g.
metabolites) to be disconnected from all others within the graph, which leads to meaningful
decreases in testing power whether or not the graph information is included. We include a graph
regularization or ‘smoothing’ approach for managing this issue. We demonstrate the benefits of
this approach through simulation studies and an application to the metabolomic data from the
COPDGene study.

Introduction

Metabolomics is the study of the metabolite composition of a cell, tissue, or biological
fluid. Leading metabolomic experimental techniques such as liquid or gas chromatography
coupled with mass spectrometry (LC-MS or GC-MS) and nuclear magnetic resonance (NMR)
spectroscopy are able to capture the abundance of all metabolites within a cell (the metabolome).
These technologies provide high-throughput data similar to other familiar -omics datatypes such as genomics, transcriptomics, and proteomics. An important advantage of metabolomics over other -omics data is its proximity to biological phenotypes. While genomic or proteomic data are vital pieces for understanding the progression from DNA to phenotype, the metabolites are the end products of the enzymatic reactions of a cell. The metabolome is comprised of exogenous (environmentally derived) and endogenous (genetically regulated) metabolites which can be used as biomarkers for the current phenotypic state of a cell or organism.

Like other -omics data, careful considerations of the metabolome’s unique characteristics are required to fully leverage it for biological insights. Specifically, metabolites are known to be related directly and indirectly by enzymatic reactions within a metabolomic pathway. These connections are sometimes handled separately through a secondary pathway enrichment analysis. Clustering methods have been developed to incorporate this connectivity into the primary analysis in order to avoid this two-step approach. These include Bayesian methods for metabolite clustering based on peak detection and ad hoc methods based on singleton metabolite presence. For this work, we choose to group subsets of metabolites that interact with one another and represent these pathways as graphs or networks. Throughout this paper we will use the term graph and network interchangeably. Open source databases with metabolomic pathway documentation such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), the Human Metabolome Database (HMDB), Reactome, OmniPath, and WikiPathways are growing resources, and the pathways within these databases are easily translated to graphs to be used in downstream analyses.

The semiparametric kernel machine regression method has gained popularity in many areas of biomedical research such as genomics, microbiome analysis, and
neuroimaging (14–16). One reason for its popularity is that it provides a computationally scalable method of classification and regression through the introduction of a *kernel* function. Another is that it provides a setting for formal statistical estimation and testing procedures for high-dimensional data sources, often through the use of a score statistic. Formal statistical tests are useful for metabolomic research, as a goal is often identifying specific metabolites and pathways for further inquiry. At a high level, kernel machines test for relationships between an outcome and a set of predictors by testing if variation between the two correspond with one another.

A hurdle more unique to metabolomics is the high levels of sparsity in individual metabolites and pathway connectivity. While metabolomic databases (e.g. KEGG, HMDB) are growing, none are considered complete. Data generating techniques like LC-MS and GC-MS are also imperfect technologies that may miss metabolite abundances that are too low (17). Thus, pathway representations of metabolomic data are often sparse and disconnected, i.e. nodes or sets of nodes (metabolites, etc.) not connected to any other within the pathway.

Disjoint nodes are of concern for graph-structured data since many techniques for inference on graph-structured data involve spectral decomposition and require full rank matrices from a fully connected graph. Techniques that force graphs to be fully connected by making small, uniform changes to the structure have been suggested for handling this issue (18,19). However, it is understood that these alterations impose new challenges by changing the subspaces spanned by the graph. Freytag et al. developed a network-based kernel for genomic pathway analysis in which they impose “as much noise as necessary” to ensure positive semidefinite matrices (20). This method and others like it are tailored to genomic data and not applicable to other omics data. The *PIMKL* method works with pathways within the metabolome by combining them through a weighted summed kernel (21). These weights provide insight into
the importance of each sub-pathway, but this does not surmount to the level of evidence gathered
from a direct comparison between specific pathways and phenotype.

In this paper we propose the Pathway Integrated Regression-based Kernel Association
Test (PaIRKAT), a new kernel machine regression method for incorporating known pathway
information into the semi-parametric kernel regression framework. In addition, PaIRKAT
contributes an application of a graph kernel regularization method for overcoming sparse
connectivity and disjoint pathways. To our knowledge, this is the first method to incorporate
graph regularization into a kernel regression test. PaIRKAT allows for tests of association with
phenotypes and the specific pathways while integrating pathway structure, and, instead of adding
small amounts of noise, this approach dampens noisy components of a pathway while preserving
biologically relevant signals. This leads to improved testing power and better overall biomarker
detection. We evaluate these methods through several simulation studies and an application to
real metabolomics data from the COPDGene(22) study.

Results
Method Overview

Here we provide the main steps of PaIRKAT and provide an overview of the ideas
behind them. The method is described in full in Methods and Models. The primary goal of
PaIRKAT is to include the topographical information of graph structured data into the kernel
machine regression model. We use the semiparametric kernel machine model to test for
relationships between the phenotype of interest, $Y$, and high dimensional data, $Z$, while
controlling for important covariates, $X$, in the model $g(Y) = XB + h(Z) + e$. In this model $h(\cdot)$
is a positive semidefinite kernel function that transforms $Z$ to a feature space.
Omics data (metabolomics, genomics, etc.) can often be represented as a graph (or network) with edges representing biological interactions between the nodes (metabolites, etc.). The topography of a graph is often captured through its normalized Laplacian, \( \bar{L} \). We explored the utility of incorporating the Laplacian directly into the kernel machine but found it to be ineffective. Instead, we transform \( \bar{L} \) using methods designed to dampen noisy aspects of a graph while preserving its biologically relevant features. The PaIRKAT method is to include this regularized Laplacian, \( \bar{L}_R \), in the model through the kernel function thusly \( g(Y) = X\beta + h(Z\bar{L}_R) + \epsilon \). Tests for relationships between \( Y \) and \( h(Z\bar{L}_R) \) are accomplished using a score test based on a Satterthwaite approximation of a \( \chi^2 \) distribution.

Simulation Results

A complete description of our simulation study can be found in \textbf{Methods and Models}, but we give a brief synopsis of the simulation scheme. We first randomly generated a graph. Second, we randomly generated features, \( Z \), from multivariate normal distribution with a covariance structure derived from the graph. Lastly, we randomly generated a normally distributed outcome, \( Y \), with a mean based on a linear relationship between the columns of \( Z \). We performed tests ignoring graph topography, including graph topography via the normalized Laplacian (\( \bar{L} \)), and our proposed method PaIRKAT of including graph topography via the regularized Laplacian (\( \bar{L}_R \)). Our simulations aimed to assess how sensitive our method is to incomplete and/or incorrect graph information.

Type I error rates for PaIRKAT are summarized in Tables 1, 2, 3, and 4. The type I error rates for tests using a graph’s normalized Laplacian, \( \bar{L} \) (see Methods section for definition), are
summarized in Tables S1, S2, S3, and S4. The type I error rate of $\approx 0.05$ is maintained throughout all simulation scenarios.

| Pathway size       | 15   | 30   | 45   |
|--------------------|------|------|------|
| Perfect            | 0.073| 0.051| 0.067|
| Partial 10%        | 0.051| 0.049| 0.059|
| Partial 40%        | 0.054| 0.061| 0.055|
| Partial 70%        | 0.054| 0.061| 0.055|
| Complete Mismatch  | 0.058| 0.047| 0.054|
| No Pathway         | 0.059| 0.058| 0.055|

**Table 1**: Type 1 error rates using complete pathways. Error rates were calculated from score tests on 1000 simulated data sets. All simulations used graphs with 15, 30, or 45 nodes. No nodes or edges were dropped for these simulations.

| Pathway size       | 15   | 30   | 45   |
|--------------------|------|------|------|
| Perfect Network    | 0.052| 0.056| 0.046|
| Partial 10% Mismatch | 0.046| 0.063| 0.047|
| Partial 40% Mismatch | 0.051| 0.068| 0.052|
| Partial 70% Mismatch | 0.051| 0.068| 0.052|
| Complete Mismatch  | 0.052| 0.060| 0.064|
| No Network         | 0.051| 0.048| 0.044|

**Table 2**: Type 1 error rates using pathways with 5% missing edges. Error rates were calculated from score tests on 1000 simulated data sets using graphs with 15, 30, or 45 nodes. The graph used to simulate $Z$ and $Y$ was of medium edge density, while the graph used to test was of low density. The low-density graphs are drawn from the Barabasi-Albert model with edge density 0.13, 0.07, and 0.04 for graphs with 15, 30, and 45 nodes, respectively. Medium edge density graphs are created by giving any 2 nodes without a direct edge between them a 5% chance of becoming directly connected. This creates graphs with an average edge density of 0.18, 0.12, and 0.09 for graphs with 15, 30, and 45 nodes, respectively.
Table 3: Type 1 error rates using pathways with 15% missing edges. Error rates were calculated from score tests on 1000 simulated data sets using graphs with 15, 30, or 45 nodes. The graph used to simulate $Z$ and $Y$ was of high edge density, while the graph used to test was of low density. The low density graphs are drawn from the Barabasi-Albert model with edge density 0.13, 0.07, and 0.04 for graphs with 15, 30, and 45 nodes, respectively. High edge density graphs are created by giving any 2 nodes without a direct edge between them a 15% chance of becoming directly connected. This creates graphs with an average edge density of 0.26, 0.21, and 0.19 for graphs with 15, 30, and 45 nodes, respectively.

| Pathway size     | 15   | 30   | 45   |
|------------------|------|------|------|
| Perfect Network  | 0.061| 0.059| 0.064|
| Partial 10 Mismatch | 0.039| 0.057| 0.059|
| Partial 40 Mismatch | 0.045| 0.065| 0.058|
| Partial 70 Mismatch | 0.045| 0.065| 0.058|
| Complete Mismatch | 0.062| 0.046| 0.054|
| No Network       | 0.048| 0.055| 0.053|

Table 4: Type 1 error rates using pathways with dropped nodes. Error rates were calculated from score tests on 1000 simulated data sets using graphs 15, 30, or 45 nodes initially. The graph used to simulate $Z$ and $Y$ contained all nodes. Nodes with degree below the 25th percentile within a graph had a 25% chance of being dropped before testing.

| Pathway size     | 15   | 30   | 45   |
|------------------|------|------|------|
| Perfect Network  | 0.041| 0.046| 0.062|
| Partial 10 Mismatch | 0.050| 0.054| 0.057|
| Partial 40 Mismatch | 0.049| 0.059| 0.051|
| Partial 70 Mismatch | 0.049| 0.059| 0.051|
| Complete Mismatch | 0.063| 0.064| 0.057|
| No Network       | 0.042| 0.050| 0.056|

The power curves for all pathway structures while simulating complete knowledge, missing edges, and missing nodes are displayed in Figure 1. Having a perfect pathway structure provides the most power. Relationships between an outcome and pathway are easier to detect in larger pathways. The more incorrect direct edges in the pathway, the lower the overall power. We also see that increasing the overall signal to noise ratio improves power for all pathway structures (Figure 2). PaIRKAT ($\tilde{L}_R$) achieves approximately 80% power at a signal to noise ratio around 0.23, whereas ignoring network information requires a signal to noise ratio nearly
three times large, about 0.60, and only including the Laplacian never achieves 80% power (Figure 2).

Fig 1: Power curves from the four pathway knowledge and 6 pathway structure simulation scenarios. Power curves were all calculated from score tests on 1000 simulated data sets using graphs with 15, 30, or 45 nodes. Power curves assuming complete pathway knowledge with no dropped edges or nodes are displayed in a). For (b) and (c), the graph used to simulate $Z$ and $Y$ was of medium or high density, respectively, while the graph used to test was of low density. Medium and high edge density graphs used for data generation had ~5% and ~15% more edges, respectively, than the low density graph used for testing. The power curve generated assuming missing nodes (d) used all graph nodes to generate $Z$ and $Y$. Then nodes (and corresponding columns of $Z$) with degree below the 25th percentile within a graph had a 25% chance of being dropped before testing.
Fig 2: Signal to Noise Ratio. Power curves from increasing the signal to noise ratio while assuming complete pathway knowledge. The signal to noise ratio was calculated as the ratio between the overall variance in $\mathbf{Y}$, $\text{Var}(\mathbf{\beta}_0 + \sum_{i=1}^{p} \mathbf{\beta}_i \mathbf{Z}_i)$, and the overall residual variance, $\text{Var}[Y_i - (\mathbf{\beta}_0 + \sum_{i=1}^{p} \mathbf{\beta}_i \mathbf{Z}_i)]$. Each power calculation comes from score tests on 1000 simulated data sets using graphs with 30 nodes.

COPDGene Analysis Results

A complete description of these analyses can be found in Methods and Models, but here we give a brief description of the outcome variables we analyzed. We create models for two phenotypes from the COPDGene study(22): (1) percent emphysema and (2) the ratio of post-bronchodilator forced expiratory volume at one second divided by forced vital capacity (FEV$_1$/FVC). To normalize FEV$_1$/FVC, we use the following log ratio transformation, $\log \left( \frac{\text{FEV}_1/\text{FVC}}{1-\text{FEV}_1/\text{FVC}} \right)$. This is referred to as the “log FEV$_1$/FVC ratio” for simplicity. We test for associations between 28 pathways and each outcome under the same three conditions in the simulation study: ignoring graph topography, including graph topography via the normalized
Laplacian ($\mathbf{L}$), and our proposed method PaIRKAT of including graph topography via the regularized Laplacian ($\tilde{\mathbf{L}}_R$).

Including the metabolites’ regularized graphs had large impacts on the associations between the log FEV$_1$/FVC ratio and several subsets of metabolites. For the 28 pathways tested, power was improved for 18 pathways when using PaIRKAT vs. using $\tilde{\mathbf{L}}$ or ignoring pathway information. Of note, the strength of the associations between the log FEV$_1$/FVC ratio and the $ABC$ transporters, the arginine and proline metabolism, the cysteine and methionine metabolism, the pyrimidine metabolism, the glycine, serine, and threonine metabolism, and the neuroactive ligand-receptor interaction metabolite subsets increased dramatically. The average p-value was also lower for 5 pathways with using $\tilde{\mathbf{L}}$ vs. ignoring pathway information. However, there was not a significant result from any method for 4 of these pathways, and PaIRKAT provided similar power for the fifth. Figure S1 displays the p-values from the kernel regression tests for associations between the log FEV$_1$/FVC and the 28 pathways of interest for each subsample size.

Including the metabolites’ regularized graphs also had impacts on the associations between percent emphysema and several subsets of metabolites. For the 28 subsets of metabolites tested, power was improved for 17 pathways when using PaIRKAT vs. including $\tilde{\mathbf{L}}$ or ignoring pathway information. Of note, the strength of the associations between percent emphysema and the $ABC$ transporters, the $\beta$-alanine metabolism, the neuroactive ligand-receptor interaction, the glycine, serine and threonine metabolism, and the histidine metabolism metabolite subsets increased dramatically when using PaIRKAT vs. ignoring pathway information. The average p-value was again lower for the same 5 pathways with using $\tilde{\mathbf{L}}$ vs. ignoring pathway information. However, there was still not a significant result from any method for 4 of these pathways, and PaIRKAT provided similar power for the fifth. Figure S2 displays
the p-values from the kernel regression tests for associations between percent emphysema and
the 28 pathways of interest for each subsample size.

Figure 3 displays results from 3 pathways selected to illustrate PaIRKAT’s impact on
power for fully connected (left column), partially disconnected (middle column), and sparse
(right column) graphs. For the *steroid hormone biosynthesis* pathway, an almost completely
sparse pathway, we see virtually no differences between PaIRKAT and ignoring pathway
connectivity. We also see relatively small differences between all three methods for the fully
connected *aminoacyl-tRNA biosynthesis* pathway. The major impacts from PaIRKAT come
when there are a few nodes or node subsets disjoint from the rest of the graph, as we see in the
*cysteine and methionine metabolism*. 

263
264
Figure 3: Selected results from COPDGene subset analysis. Average p-values from kernel regressing tests that do not include pathway information (No Laplacian, red circles), include pathway information through a normalized Laplacian ($\mathbf{L}$, green triangles), and include pathway information through a regularized normalized Laplacian ($\mathbf{L}_R = (I + \tau \mathbf{L})^{-1}$, blue squares) are displayed. P-values were averaged over 100 random subsets of size 100, 200, 300, 400, and 500 from the COPDGene dataset. $\tau$ was set to 1 for all tests that used $\mathbf{L}_R$. The 3 pathways selected illustrate the expected results in fully connected (left), partially disconnected (middle), and sparse (right) graphs.

Discussion

We developed PaIRKAT, a method for incorporating pathway information under a kernel regression framework. Other methods to incorporate pathway connectivity via graph operations have been developed (20, 21, 23–26). PaIRKAT enables the researcher to test on specified pathways instead of aggregating all pathways through a weighted kernel as in (21, 24). It can also handle disjointed pathways without adding in artificial noise to the network as in (18–20). This
allows the investigator to compile information from multiple sources, e.g. KEGG and HMDB. The regression framework also expands upon a method developed for classification (23). Finally, this method provides a one-step testing procedure to simplify workflow compared to the popular two-step enrichment analysis procedure.

Pathway misspecifications from incomplete data collection or imperfect canonical pathways within databases are common hurdles in -omics studies. We explored the sensitivity of the method by simulating data assuming incorrect pathway structures and incomplete pathway knowledge. These studies show that our method is highly robust to pathway misspecifications. In smaller pathways, we see that the partially mismatched structure with ~10% of direct edges being incorrect does as well as the perfect network structure. This is likely due to the very small change from the perfect structure in these cases, as a graph with only 15 nodes could easily be unchanged with only a 10% chance to change an edge. Furthermore, even with incorrect or incomplete pathway information, our method provides significantly improved power over ignoring pathway information while maintaining an appropriate type I error rate. We believe this is because many indirect connections between nodes are preserved, and these connections still provide more accurate information than incorrectly assuming independence among nodes.

One benefit of using PaIRKAT is improved power to identify pathways that are associated with clinical phenotypes. For example, an application to the COPDGene dataset using KEGG’s database of metabolic pathways also illustrated PaIRKAT’s ability to improve testing power over simply treating metabolites as independent (Figure 3, S1, S2). The regularization technique was also able to handle pathways with few metabolites and/or disjoint components. Several tests had a dramatic boost in power from including pathway connectivity for both percent emphysema and the log FEV₁/FVC ratio, and most pathways have been previously
associated with COPD and lung function. Huang et al. linked environmental exposures, COPD risk, and metabolomic pathways, and found associations between COPD and the histidine metabolism, cysteine and methionine metabolism, and β-alanine metabolism pathways (27). The glycine, serine and threonine metabolism, aminoacyl-tRNA biosynthesis, pyrimidine metabolism, pantothenate and CoA biosynthesis pathways have all previously been associated with asthma (28). The β-Alanine metabolism, ABC transporters, purine metabolism, pantothenate and CoA biosynthesis pathways were all differentially associated with COPD subclasses for patients with lung cancer (29). Another study of the COPDGene dataset (22) using a two-step pathway enrichment approach found that the purine metabolism, mineral absorption, arginine biosynthesis, aminoacyl-tRNA biosynthesis, ABC transporters and glycine, serine and threonine metabolism pathways were all associated with various measures of lung function and increased COPD exacerbations (30). The three ABC transporters have also been shown to be related to COPD in several murine knockout and human studies (see Chai et al. (31)). Finally, the arginine biosynthesis pathway has also been associated with COPD in multiple studies (32, 33).

Impacts of Regularization

In simulation studies and real data analyses we saw meaningful improvements in power by including pathway information through a graph’s regularized normalized Laplacian, (PaIRKAT) when compared to ignoring the pathway information or using \( \tilde{L} \). PaIRKAT was essential to maintaining testing power when graphs had disjoint nodes or sub-graphs. Using the normalized Laplacian, \( \tilde{L} \), hindered testing performance compared to using PaIRKAT or ignoring the pathway information when a graph was disconnected. In connected graphs PaIRKAT, using
\( \mathbf{L} \), and ignoring the pathway information all performed similarly in the real data analyses (Figure 3).

It is well established that \( \mathbf{L} \) is a symmetric and positive semidefinite matrix with eigen values \( 0 \leq \lambda_1, \lambda_2, \ldots, \lambda_p \leq 2 \), where the number of \( \lambda_i = 0 \) is the number of disjoint components of the undirected graph \( G \) (see Methods and Models). Therefore, graphs with very low connectivity, meaning many \( \lambda_i = 0 \), will not be as impacted by regularization since all \( r^{-1}(\lambda_i = 0) = a \) for some scalar \( a \). In words, there is no extra information from a graph when most nodes are disconnected from one another (e.g., Figure 3, right column).

In summary, our proposed method serves as a framework for including pathway information into a kernel machine regression test. We developed this method for application to metabolomic pathway data, but the techniques are easily generalizable to other data sources with a graph-like structure. It is important to examine the structure of a graph before applying a regularization step. Unique challenges arose from the sparsity present in many metabolomic pathways which can greatly hinder performance. We implement a graph regularization kernel to handle disconnected pathways. This regularization step is novel in the application of graph-based kernel machine regression to biological data. Our simulation studies illustrate the robustness of this method to improper and incomplete pathway knowledge. The method presented is capable of providing powerful tests for associations between biological pathways and phenotypes of interest and will be helpful in identifying novel pathways for targeted clinical research. The R code for simulations and an example workflow with source code is available at https://github.com/CharlieCarpenter/PaIRKAT, and an R shiny app is currently hosted at https://csevern.shinyapps.io/pairkat/.
The kernel machine model:

We will assume that the data are properly filtered, imputed, and normalized for the methods described in this paper. Consider a dataset with observations from \( n \) subjects. Let \( Y \) be an \( n \times 1 \) vector representing a continuous or discrete phenotype of interest. Also let \( X \) be a \( n \times q \) matrix of clinical covariates and \( Z \) be an \( n \times p \) matrix of metabolite abundances. The phenotype can then be modeled through the following semiparametric model

\[
g(Y) = X\beta + h(Z) + \epsilon,
\]

where \( g \) is any standard link function, \( \beta \) is a \( q \times 1 \) vector of regression coefficients, \( \epsilon \) is an \( n \times 1 \) vector of normally distributed error terms, and \( h \) is a kernel function. There are no parametric assumptions placed on \( h \) except that it lies in some feature space. This more relaxed requirement from the kernel regression provides flexibility and robustness to model misspecification. Another key advantage of introducing the kernel function is its ability to capture nonlinear relationships between the phenotype \( (Y) \) and the metabolome \( (Z) \) in a computationally tractable manner. These relationships are assumed to exist in some feature space, denoted \( \mathcal{H}_K \), that is generated by a positive definite kernel function \( K(\cdot,\cdot) \). Under this condition, the representer theorem then allows \( h(Z) \) to be represented through the kernel function \( K(\cdot,\cdot) \) as \( h(\cdot) = \sum_{i=1}^{n} \alpha_i K(z_i, \cdot, \rho) \) for some coefficients \( \alpha_i \in \mathbb{R} \). More detailed derivations can be found in texts by Schölkopf and Smola(34) as well as Cristianini and Shawe-Taylor(35).

The kernel function \( K \) can be thought of as a measurement of similarity of the metabolome between two individuals. Consequently, the choice of a kernel function defines the notion of similarity and covariance between the metabolome and the phenotype of interest. Two common choices for kernel functions are the \( dth \) Polynomial Kernel: \( K(z_i, z_j, \rho) = (z_i^T z_j + \rho)^d \)
and the Gaussian Kernel: \( K(z_i, z_j, \rho) = \exp\left\{ -\|z_i - z_j\|^2 / \rho \right\} \), where \( \| \cdot \| \) is the Euclidean (\( L_2 \)) norm. Both kernels contain a tuning parameter \( \rho \). It is common in the machine learning literature for this parameter to be fixed to some ad-hoc value. For this work, we employ the Gaussian kernel and use the median of all pairwise Euclidean distances between metabolites as an empirical estimate of \( \rho \). We choose to work with the Gaussian kernel since it is a \textit{characteristic} kernel, a desirable property meaning that probability measures embedded onto \( \mathcal{H}_K \) through the kernel function are unique. These familiar choices are used to illustrate our new method, but, in principle, several kernel functions and choices for estimating \( \rho \) could be applied.

\textit{Kernel-based Score Test}

Liu et al. show a connection between kernel machine regression and linear mixed models for semiparametric modeling of high dimensional data\(^{12,13}\). The parameters \( \beta \) and \( h(Z) \) can be estimated by maximizing the scale penalized likelihood

\[
L(\beta, h) = -\frac{1}{2} \sum_{i=1}^{n} \left[ y_i - x_i^T \beta - h(z_i) \right]^2 - \frac{1}{2} \lambda \| h \|^2
\]

\[
= -\frac{1}{2} \sum_{i=1}^{n} \left[ y_i - x_i^T \beta - \sum_{j=1}^{n} \alpha_j K(z_i, z_j) \right]^2 + \lambda \sum_{i=1}^{n} \alpha_i^2 \] (2)

where \( K = K(z_i, z_j, \rho) \) is the semi-positive definite kernel function of choice. \( h(Z) \) can then be understood as subject specific random effects with mean 0 and variance \( \tau K \). Testing for an association between phenotype and pathway is then equivalent to testing the null hypothesis
\( H_0: \tau = 0 \text{ vs } H_1: \tau > 0 \). Liu et al follow Zhang and Lin in the formulation of a score statistic for this variance component test (12, 36). Specifically, the score statistic for \( \tau \) is 
\[
Q(\widehat{\beta}, \widehat{\phi}, \rho) = \frac{1}{2\phi} (Y - X\beta)^T K (Y - X\beta).
\]

Since \( Q \) is a quadratic function of the residuals, an approximation for the null distribution can be obtained using Satterthwaite’s method of matching the mean and variance of \( Q \) and a scaled chi-squared distribution, \( \kappa \chi^2 \). This results in \( \kappa = \tilde{I}_{\tau \tau} / (2\bar{e}) \) and \( \eta = 2\bar{e}^2 / \tilde{I}_{\tau \tau} \) where \( \bar{e} = \)
\[
tr(P_0K)/2 \text{ and } \tilde{I}_{\tau \tau} \text{ is the efficient information } I_{\tau \tau} = I_{\tau \tau} - I_{\tau \sigma^2} I_{\sigma^2 \sigma^2}^{-1} I_{\tau \sigma^2}. \]

Here \( I_{\tau \tau} = \)
\[
tr[(P_0K)^2]/2, \quad I_{\tau \sigma^2} = tr[P_0KP_0], \quad \text{and } I_{\sigma^2 \sigma^2} = tr(P_0^2). \]

\( \phi \) serves as a scaling parameter and \( P_0 \) is the projection matrix under the null model: \( g(Y) = X\beta + \epsilon \). For continuous outcomes the projection matrix is 
\( P_0 = I - X(X^TX)^{-1}X^T \) and \( \phi \) is the residual variance, \( \sigma^2 \). For dichotomous outcomes the projection matrix is 
\( P_0 = D_0 - D_0X(X^TD_0X)^{-1}X^TD_0 \) and \( \phi = 1 \), with \( D_0 = diag(\hat{\mu}_i(1 - \hat{\mu}_i)) \) and \( \hat{\mu}_i \) being the fitted probability \( P(y_i = 1) \) under the null model. A key advantage of working with the score statistic is the need to only fit the null model for testing, meaning \( \widehat{\beta} \) and \( \widehat{\phi} \) are their respective maximum likelihood estimates under the null model. Liu et al also demonstrated that this method is robust to the choice of \( \rho \)(12).

**Graph Laplacian:**

A network or graph, \( G = \{V, E\} \), is a mathematical representation of any interconnected structure through a set \( V \) of \( p \) nodes (or vertices) and a set \( E \) of edges, where the elements of \( E \) are pairs \( \{u, v\} \) of distinct vertices, \( u, v \in V \). When applied to metabolomic pathways, nodes
represent individual metabolites within the pathway and edges represent direct biochemical interactions/reactions between metabolites.

Two important features of any graph are its adjacency matrix, $A$, and degree matrix, $D$. $A$ is a $p \times p$ matrix that is non-zero when an edge exists between two vertices. $D$ is a $p \times p$ diagonal matrix with $D_{[i,i]}$ representing the number of nodes connected to node $i$. For this work, we represent pathways using undirected unweighted graphs, i.e. there is no ordering to the vertices defining an edge and $\{u, v\} = \{v, u\} \in E$. This means $A$ will be a symmetric matrix with all entries either 1 or 0. Using these features, we can calculate a graph’s Laplacian $L := D - A$ and its normalized Laplacian $\bar{L} := D^{-1/2}LD^{-1/2} = I - D^{-1/2}AD^{-1/2}$, where $I$ is a $p \times p$ identity matrix.

Both $L$ and $\bar{L}$ can be regarded as linear operators of functions $f: V \rightarrow \mathbb{R}$ that induce a semi-norm $\|f\|_L = \langle f, Lf \rangle = f^T Lf$. This semi-norm can be interpreted as a measure of “smoothness” or how much $f$ varies over its domain. Standardizing $L$ by the number of connections per node to obtain $\bar{L}$ is a common approach in graph theory since $\bar{L}$ has several well-known and desirable properties. In particular, $\bar{L}$ is symmetric and positive semidefinite, and its eigenvalues, $\lambda_i$, are bounded such that they satisfy $0 \leq \lambda_i \leq 2$ for $i \in 1, 2, \ldots p$.

Graph Regularization

A key component of PaIRKAT is the ability to handle disjoint components within a graph. Pathway databases may not be complete, and untargeted data generating techniques may not be able capture all components within a pathway. This leaves some pathways with low connectivity and others with completely disconnected nodes. An interesting feature of a graph’s normalized Laplacian, $\bar{L}$, is that the number of disjoint pieces within a graph is captured by the
number of \( \mathbf{L} \)'s eigenvalues equal to 0. When \( G \) has disjoint nodes and disjoint nodes \( \mathbf{L} \) becomes less than full rank. This can lead to a decrease in our power to detect associations between phenotypes and metabolomic pathways. One proposed solution is to simply manipulate the adjacency matrix by adding a small constant to all entries \((18,19)\), i.e. working with a modified adjacency matrix \( \mathbf{\tilde{A}} = \mathbf{A} + t \mathbf{e} \mathbf{e}^T \), where \( t \) is a nonnegative tuning parameter and \( \mathbf{e} \) is a vector of 1s. This yields a full rank matrix as desired, but we know that the subspace spanned by \( \mathbf{\tilde{A}} \) is not the correct subspace on which our graph lies.

A more elegant solution can be drawn from Smola and Kondor’s work on regularization of graphs \((38)\) in which they draw on parallels between the standard Laplacian operator \( (\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \cdots + \frac{\partial^2}{\partial x_m^2}) \) and the graph Laplacian to design regularization kernels for graphs.

Rapaport, et al. \((23)\) took a similar approach to graph smoothing, though this work was done in the context of classification not hypothesis testing. These ideas can be generalized further to represent any metric on a space. That is, for any two observations \( i \) and \( j \), the inner product can be expressed as \( \langle \mathbf{z}_i, \mathbf{z}_j \rangle_M = \mathbf{z}_i^T \mathbf{M} \mathbf{z}_j \), where \( \mathbf{M} \) defines the metric on the vector space based on \( \| \mathbf{z}_i - \mathbf{z}_j \|_M \). Purdom \((39)\) presents this argument in the context of a “generalized” principal component analysis using a general metric \( \mathbf{M} \). This can be seen as an application of a linear kernel on any metric space, whereas we apply the Gaussian kernels for hypothesis testing and, like Rapaport, focus on graph Laplacians for our metric.

For this work, we apply a regularization kernel to obtain a \textit{regularized} normalized Laplacian: \( r(\mathbf{L}) \equiv \mathbf{\bar{L}}_R \). Regularizations of the Laplacian can be seen as regularizations of the eigenvalues of \( \mathbf{L} \), \( r(\lambda) \). There are many possible choices for \( r \); the only requirement is that \( r^{-1}(\lambda) > 0 \) for \( \lambda \in [0,2] \) to ensure \( r(\mathbf{L}) \geq 0 \). In classical Fourier analysis the size of \( \lambda_i \in [0,2] \)
is directly proportional to the frequency of component $i$ within Fourier space, which translates to
the degree of noise within the system. This intuition tells us to limit $r^{-1}(\lambda)$ to monotonically
increasing functions in order to impose higher penalties to more uneven portions of the graph
while preserving the lower frequency components, which we assume translate to the prevalent
biological signals. Smola and Kondor recommend further limiting choices of $r$ to functions
expressible by power series such as a diffusion kernel (40), $r(L) = e^{-\tau/2L}$. See their paper for
complete details on the derivation of different regularization functions.

PaIRKAT implements a “linear” regularization function

$$\bar{L}_R = (I + \tau\bar{L})^{-1},$$  

where $\tau > 0$ is a bandwidth parameter and $I$ is a $p \times p$ identity matrix. We choose this
regularization for its simplicity and interpretability of $\tau$. Increasing $\tau$ linearly increases the
amount of smoothing performed in $r^{-1}(\lambda) = 1 + \tau\lambda$. With the graph’s sparse connectivity
accounted for through the linear regularization function, we can now conduct a kernel machine
test while incorporating connectivity within a pathway through $\bar{L}_R$ into (1) thusly

$$g(Y) = X\beta + h(Z\bar{L}_R) + \epsilon,$$

where $h$ is a kernel function applied to $Z\bar{L}_R$ and the other model components are as described in
(1). The kernel-based score test can then be applied to obtain powerful tests for associations
between connected or disconnected pathways and a phenotype of interest.
We conducted multiple simulation studies to assess whether or not the proposed method is robust to imperfect pathway information. We assumed 3 different “pathway knowledge” scenarios and 4 different “pathway structure” scenarios (Figure 4). Different pathway knowledge scenarios refer to different types of missing information, whereas pathway structure scenarios refer to different configurations of the “known” nodes and edges. We simulate using both the normalized Laplacian, $\mathbf{L}$, and PaIRKAT’s regularized normalized Laplacian, $\mathbf{L}_R$, as well as ignoring the pathway information.

**Fig 4:** Flowchart of simulation procedure. We (1) simulate a graph $G$, (2) generate $Z$ and $Y$ from $G$, (3) drop nodes or edges from $G$ to give a smaller graph $G_s$ (drop corresponding columns of $Z$ when dropping edges to create $Z_s$), (4) permute edges to create an improperly structured graph $G_s^*$, (5) calculate the regularized normalized Laplacian $\mathbf{L}_R^*$ from $G_s^*$, and finally (5) test for an association between $h(Z \mathbf{L}_R^*)$ (or $h(Z_s \mathbf{L}_R^*)$) and $Y$ in the model $Y = \beta_0 + h(Z \mathbf{L}_R^*)$. For the “no network” simulations, we only use step (1), step (2) and step (5) without including $\mathbf{L}_R^*$. 
We simulated three different knowledge scenarios to represent incomplete pathway database information and/or incomplete data collection.

1. **No missing**: Assuming the nodes measured (metabolites, genes, etc.) and edges connecting them are a perfect representation of the biological pathway of interest.

2. **Missing edges**: Assuming that some biological interactions (edges) are missing from the documented pathway. Here we generate a graph $G = \{V, E\}$ according to the Barabasi-Albert model for a “low” edge density. We then give every set $\{u, v\} \notin E$ a 5% or 15% percent chance of being added to $E$ for a “medium” or “high” edge density graph, respectively. $Z$ and $Y$ are then generated from the medium or high edge density graph, but $\mathbf{L}$ or $\tilde{\mathbf{L}}_R$ is calculated from the original “low” edge density graph. Examples of these graphs are shown in Figure S3.

3. **Missing nodes**: Assuming that some of the nodes (and hence their edges) are missing from the documented pathway. Here a graph is used to generate $Z$ and $Y$. Then nodes with degree below the 25\textsuperscript{th} percentile have a 25% chance of being removed before calculating $\mathbf{L}$ or $\tilde{\mathbf{L}}_R$. The corresponding columns and rows of $Z$ and $\tilde{\beta}$ are removed as well. Examples of these graphs are shown in Figure S4.

**Pathway Structures**

After we simulate a pathway knowledge scenario, we alter the pathway structure to represent incorrect edge connections within a database. Examples of structures 1, 2, and 3 are displayed in Figure 5.
1. **No mismatch**: No alterations to graph edges. The graph used to simulate $Z$ and $Y$ is the same graph used to calculate $\bar{L}$ or $\bar{L}_R$ (Figure 5, left).

2. **Partial Mismatch**: a graph, $G_1 = \{V_1, E_1\}$, is used to simulate $Z$ and $Y$. This graph’s edges are permuted such that any edge $\{u, v\} \in E_1$ has a 10%, 40%, or 70% chance of being changed to some $\{u, w\} \notin E_1$; i.e. approximately 10%, 40%, or 70% of direct edges are incorrect before calculating $\bar{L}$ or $\bar{L}_R$ (Figure 5, middle).

3. **Complete Mismatch**: a network $G_1 = \{V_1, E_1\}$ is used to simulate $Z$ and $Y$. A new random graph, $G_2$, is then draw and forced to have no edges that match $G_1$, i.e. $V_1 = V_2$ but if $\{u, v\} \in E_1$ then $\{u, v\} \notin E_2$. We then calculate $\bar{L}$ or $\bar{L}_R$ from $G_2$ (Figure 5, right).

4. **No Pathway**: a graph is used to simulate $Z$ and $Y$. This connectivity is ignored while testing by not including $\bar{L}$ or $\bar{L}_R$ in the kernel function.

---

**Fig 5**: Examples of the three different pathway structures. Nodes 6 and 7 are highlighted in red to help display the effects of different pathway structures. (Left) The “true” pathway or graph that is used to simulated $Z$ and $Y$. This is the graph used for tests under a “perfect pathway structure” scenario. (Middle) A graph with approximately 40% of the edges from the “true” graph directly connecting the wrong nodes. This is used for tests under a “partial mismatch (40)” structure scenario. (Right) A graph with 0 shared edges with the “true” graph. This is the graph used for tests under a “complete mismatch structure” scenario.

All pathway structures were considered under each different pathway knowledge scenario. The different pathway structures were imposed after simulating under different
pathway knowledge assumptions. Each simulated pathway structure and knowledge combination
followed 5 steps: (1) simulate a graph $G$, (2) generate $Z$ and $Y$ from $G$, (3) drop nodes and/or
edges (based on knowledge assumption) from $G$ and $Z$ to give a smaller graph and node set $G_s$
and $Z_s$, (4) alter $G_s$ (based on structure assumption) to create a graph $G_s^*$ with improper edge
connections, (5) calculate $\mathbf{L}^*$ or $\mathbf{L}_R^*$ from $G_s^*$, and (5) test for an association between $h(Z_s \mathbf{L}_R^*)$
and $Y$ in the model $Y = \beta_0 + h(Z_s \mathbf{L}_R^*)$. See Figure 4 for a flowchart of these simulation
scenarios.

Simulated Data

To evaluate PaIRKAT’s overall testing performance and robustness to incorrect pathway
information, we simulate data and tests assuming various different types of misspecified
pathways. All simulations were accomplished using R(41). Random graphs were generated using
the igraph(42) package according to the Barabasi-Albert model(43) with $p$ nodes representing $p$
metabolites within a pathway. The graph’s adjacency matrix was converted into a positive
definite precision matrix, $\mathbf{\Omega}$, using an approach developed by Danaher, et al.(44) and also
applied by Shaddox, et al(45). An $n$ by $p$ matrix of metabolite abundances, $Z$, was then
simulated from a multivariate normal distribution with mean $\bar{\mu}$ and covariance $\mathbf{\Omega}^{-1}$, where $\bar{\mu}$ is a
vector of $p$ 0s. In this way, metabolite connectivity is captured by $\mathbf{\Omega}$. A continuous outcome $Y_i$
was then simulated from a normal distribution with mean $\beta_0 + \sum_{j=1}^{p} \beta_j Z_{ij}$ and variance $\sigma^2$,
where $\beta_0 = 0.2644$ and $\sigma^2 = 1.3688^2$. These values for $\beta_0$ and $\sigma^2$ are drawn from observed
metabolomics data. The regularization parameter $\tau$ is set to 1 for all simulations. All $\beta_j$ were set
to 0 to assess Type I error rates or set to 0.1 to assess power for the twelve different pathway
information scenarios described below. Each used 1000 simulations of graphs of size $p =$
assuming a sample size of \( n = 160 \), and a testing level of \( \alpha = 0.05 \) was used for all simulations.

We also conducted a simulation study to assess the impacts of increasing the overall signal to noise ratio on the method’s power. The signal to noise ratio was calculated as the ratio between the overall variance in the mean component, \( \text{Var}(\beta_0 + \sum_{j=1}^{p} \beta_j Z_{ij}) \), and the overall residual variance, \( \text{Var}[Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j Z_{ij})] \). Here \( \beta_j \) is drawn from a normal distribution with a mean and standard deviation of 0.1. Then from normal distributions with gradually increasing means to increase the signal to noise ratio. For simplicity, we only assumed perfect pathway knowledge and simulated under each of the 4 different pathway structures.

COPDGene Data

We analyzed data collected from the COPDGene study (22), a multicenter observational study that collected genetic data as well as multiple measures of lung function to study chronic obstructive pulmonary disease (COPD). Between 2007 and 2011, 10,198 participants with and without chronic obstructive pulmonary disease (COPD) enrolled (Visit 1). A five-year follow up visit took place between 2013 and 2017 (Visit 2). Blood samples were also obtained for -omics analyses from participants who provided consent. In total, 1136 subjects (1040 non-Hispanic white, 96 African American) participated in a metabolomics ancillary study in which they provide fresh frozen plasma collected using an 8.5 mL p100 tube (Becton Dickson) at Visit 2.

Metabolomics and Data Processing

P100 plasma was profiled using the Metabolon (Durham, NC, USA) Global Metabolomics platform. Briefly, untargeted liquid chromatography–tandem mass spectrometry
(LC–MS/MS) was used to quantify 1392 metabolites and described in (46,47). A data normalization step was performed to correct variation resulting from instrument inter-day tuning differences: metabolite intensities were divided by the metabolite run day median, then multiplied by the overall metabolite median. It was determined that no further normalization was necessary based on the reduction in the significance of association between the top PCs and sample run day after normalization. Subjects with aggregate metabolite median z-scores greater than 3.5 standard deviation from the mean (n = 6) of the cohort were removed. Metabolites were excluded if >20% of samples were missing values (48). For the 995 remaining metabolites, missing values were imputed across metabolites with k-nearest neighbors imputation (k = 10) using the R package `impute` (49). As a final step, metabolomic data was natural log transformed and standardized. Linear regression models were fit to each metabolite controlling for white blood cell count, percent eosinophil, percent lymphocytes, percent monocytes, percent neutrophils, and hemoglobin. The partial residuals were then used as the observed metabolomics data. These data are available at Metabolomics Workbench with identifier PR000907.

Four hundred and thirty six of these metabolites had an id in the KEGG database of human pathways, which was accessed using the `keggLink` function from the `KEGGREST` package (50). These 436 metabolites appear in 161 KEGG pathways, and 28 of these 161 KEGG pathways contained 10 or more metabolites. Edges in a pathway’s graph were defined by connections within a pathway from the KEGG reaction database. Note that our filtered dataset did not contain every metabolite within the 28 KEGG pathways selected, and therefore some of the analyzed pathways have less that 10 metabolites.
Clinical Variables

We focus on two COPD phenotypes: (1) percent emphysema and (2) the ratio of post-bronchodilator forced expiratory volume at one second divided by forced vital capacity (FEV$_1$/FVC). Emphysema, a measure of erosion of the distal airspaces, has been linked with the clinical severity of COPD(51). It is an imaging-based phenotype defined as the 15th percentile lung voxel density in Hounsfield units adjusted for total lung capacity from quantitative CT imaging analyses. FEV$_1$/FVC is a measure of airflow obstruction. To normalize FEV$_1$/FVC, we use the following log ratio transformation, $\log \left( \frac{FEV_1/FVC}{1-FEV_1/FVC} \right)$. After removing incomplete cases, we were left with 1,113 complete cases for the FEV$_1$/FVC analysis and 1,065 complete cases for the percent emphysema analysis.

Analysis

We compared results from tests that included pathway connectivity via $L$, $L_R$, and tests that ignored pathway connectivity for the 28 pathways that had measurements on at least 10 of the metabolites in the pathway. P-values were calculated from a score test as described Section 2 with $\tau = 1$. P-values from each method were indistinguishable from one another for both data sets with over 1,000 observations. However, many data sets may not be that large. To demonstrate the differences in performance, 100 random subsets of sizes 100, 200, 300, 400, and 500 were taken from both the log FEV$_1$/FVC ratio and the percent emphysema data sets. All three methods were used to test for associations between phenotype and metabolites within a pathway. The 100 p-values were then averaged to measure the performance of each method. All null models included subject age, sex, BMI, smoking status (current, former, never), pack-years of smoking, and the clinical center as covariates.
Acknowledgements:
The project described was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute, and U01 CA235488 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute, National Cancer Institute, or the National Institutes of Health.

COPDGene is also supported by the COPD Foundation through contributions made to an Industry Advisory Board that has included AstraZeneca, Bayer Pharmaceuticals, Boehringer-Ingeheim, Genentech, GlaxoSmithKline, Novartis, Pfizer, and Sunovion.

References:
1. Fiehn O. Metabolomics — the link between genotypes and phenotypes. In: Town C, editor. Functional Genomics [Internet]. Dordrecht: Springer Netherlands; 2002 [cited 2021 Apr 6]. p. 155–71. Available from: http://link.springer.com/10.1007/978-94-010-0448-0_11
2. Alonso A, Marsal S, Julià A. Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. Front Bioeng Biotechnol [Internet]. 2015 Mar 5 [cited 2020 Aug 3];3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4355044/
3. López-Ibáñez J, Pazos F, Chagoyen M. MBROLE 2.0—functional enrichment of chemical compounds. Nucleic Acids Res. 2016 Jul 8;44(Web Server issue):W201–4.
4. Suvitaival T, Rogers S, Kaski S. Stronger findings from mass spectral data through multiplex modeling. BMC Bioinformatics. 2014 Jun 19;15(1):208.
5. Suvitaival T, Rogers S, Kaski S. Stronger findings for metabolomics through Bayesian modeling of multiple peaks and compound correlations. Bioinformatics. 2014 Sep 1;30(17):i461–7.
6. Zhan X, Patterson AD, Ghosh D. Kernel approaches for differential expression analysis of mass spectrometry-based metabolomics data. BMC Bioinformatics [Internet]. 2015 Mar 11 [cited 2020 Jun 27];16. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4359587/
7. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27–30.

8. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. HMDB: the Human Metabolome Database. Nucleic Acids Research. 2007 Jan 1;35(suppl_1):D521–6.

9. Croft D, O’Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011 Jan;39(Database issue):D691–7.

10. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nature Methods. 2016 Dec 1;13(12):966–7.

11. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2018 Jan 4;46(Database issue):D661–7.

12. Liu D, Lin X, Ghosh D. Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. Biometrics. 2007;63(4):1079–88.

13. Liu D, Ghosh D, Lin X. 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. 2008 Jun 24;9:292.

14. Broadaway KA, Cutler DJ, Duncan R, Moore JL, Ware EB, Jhun MA, et al. A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants. The American Journal of Human Genetics. 2016 Mar 3;98(3):525–40.

15. Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, et al. Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. The American Journal of Human Genetics. 2015 May;96(5):797–807.

16. Jensen AM, Tregellas JR, Sutton B, Xing F, Ghosh D. Kernel machine tests of association between brain networks and phenotypes. PLoS One [Internet]. 2019 Mar 21 [cited 2020 Jun 27];14(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6428401/

17. Chaleckis R, Meister I, Zhang P, Wheelock CE. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. Current Opinion in Biotechnology. 2019 Feb 1;55:44–50.

18. Arash A. Amini, Aiyou Chen, Peter J. Bickel, Elizaveta Levina. Pseudo-Likelihood Methods for Community Detection in Large Sparse Networks. Ann Stat. 2013;41(4).

19. Le CM, Levina E, Vershynin R. Concentration and regularization of random graphs. Random Structures & Algorithms. 2017;51(3):538–61.
20. Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, et al. A Network-Based Kernel Machine Test for the Identification of Risk Pathways in Genome-Wide Association Studies. Hum Hered. 2013;76(2):64–75.

21. Manica M, Cadow J, Mathis R, Rodríguez Martínez M. PIMKL: Pathway-Induced Multiple Kernel Learning. npj Systems Biology and Applications. 2019 Mar 5;5(1):1–8.

22. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, et al. Genetic epidemiology of COPD (COPDGene) study design. COPD. 2010 Feb;7(1):32–43.

23. Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert J-P. Classification of microarray data using gene networks. BMC Bioinformatics. 2007 Feb 1;8(1):

24. Ha SS, Kim I, Wang Y, Xuan J. Applications of Different Weighting Schemes to Improve Pathway-Based Analysis. Comp Funct Genomics [Internet]. 2011 [cited 2020 Sep 1];2011. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3114410/

25. Kim I, Pang H, Zhao H. Bayesian semiparametric regression models for evaluating pathway effects on continuous and binary clinical outcomes. Stat Med. 2012 Jul 10;31(15):1633–51.

26. Kim I, Pang H, Zhao H. Statistical properties on semiparametric regression for evaluating pathway effects. J Stat Plan Inference. 2013 Apr;143(4):745–63.

27. Huang Q, Hu D, Wang X, Chen Y, Wu Y, Pan L, et al. The modification of indoor PM2.5 exposure to chronic obstructive pulmonary disease in Chinese elderly people: A meet-in-metabolite analysis. Environment International. 2018 Dec 1;121:1243–52.

28. Kelly RS, Virkud Y, Giorgio R, Celedón JC, Weiss ST, Lasky-Su J. Metabolomic profiling of lung function in Costa-Rican children with asthma. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2017 Jun 1;1863(6):1590–5.

29. Li X, Cheng J, Shen Y, Chen J, Wang T, Wen F, et al. Metabolomic analysis of lung cancer patients with chronic obstructive pulmonary disease using gas chromatography-mass spectrometry. Journal of Pharmaceutical and Biomedical Analysis. 2020 Oct 25;190:113524.

30. Cruickshank-Quinn CI, Jacobson S, Hughes G, Powell RL, Petrache I, Kechris K, et al. Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD. Sci Rep [Internet]. 2018 Nov 20 [cited 2020 Oct 1];8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6244246/

31. Chai AB, Ammit AJ, Gelissen IC. Examining the role of ABC lipid transporters in pulmonary lipid homeostasis and inflammation. Respir Res [Internet]. 2017 [cited 2020 Oct 1];18. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5330150/

32. Ruzsics I, Nagy L, Keki S, Sarosi V, Illes B, Illes Z, et al. L-Arginine Pathway in COPD Patients with Acute Exacerbation: A New Potential Biomarker. COPD: Journal of Chronic Obstructive Pulmonary Disease. 2016 Mar 3;13(2):139–45.
33. Scott JA, Duongh M, Young AW, Subbarao P, Gauvreau GM, Grasemann H. Asymmetric Dimethylarginine in Chronic Obstructive Pulmonary Disease (ADMA in COPD). Int J Mol Sci. 2014 Apr 10;15(4):6062–71.

34. Bernhard Schölkopf, Alexander J. Smola. Learning with Kernels. Massachusetts Institute of Technology; 2002.

35. Nello Cristianini, John Shawe-Taylor. An Introduction to Support Vector Machines [Internet]. Cambridge University Press; 2000. Available from: Http://www.cambridge.org

36. Zhang D. Hypothesis testing in semiparametric additive mixed models. Biostatistics. 2003 Jan 1;4(1):57–74.

37. Fan Chung, Graham. Spectral Graph Theory. 1997. (CBMS Regional Conference Series in Mathematics).

38. Smola AJ, Kondor R. Kernels and Regularization on Graphs. In: Schölkopf B, Warmuth MK, editors. Learning Theory and Kernel Machines [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003 [cited 2020 May 15]. p. 144–58. (Goos G, Hartmanis J, van Leeuwen J, editors. Lecture Notes in Computer Science; vol. 2777). Available from: http://link.springer.com/10.1007/978-3-540-45167-9_12

39. Purdom E. Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree. Ann Appl Stat. 2011 Dec;5(4):2326–58.

40. Kondor RI, Lafferty J. Diffusion Kernels on Graphs and Other Discrete Input Spaces. :8.

41. R Core Team. R: A language and environment for statistical computing. 2019; Available from: https://www.R-project.org/

42. Csardi G, Nepusz T. The igraph software package for complex network research. :9.

43. Barabási A-L, Albert R. Emergence of Scaling in Random Networks. Science. 1999 Oct 15;286(5439):509–12.

44. Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. J R Stat Soc Series B Stat Methodol. 2014 Mar;76(2):373–97.

45. Shaddox E, Peterson CB, Stingo FC, Hanania NA, Cruickshank-Quinn C, Kechriss K, et al. Bayesian inference of networks across multiple sample groups and data types. Biostatistics. 2020 Jul 1;21(3):561–76.

46. Gillenwater LA, Pratte KA, Hobbs BD, Cho MH, Zhuang Y, Halper-Stromberg E, et al. Plasma Metabolomic Signatures of Chronic Obstructive Pulmonary Disease and the Impact of Genetic Variants on Phenotype-Driven Modules. Network and Systems Medicine. 2020 Dec 1;3(1):159–81.
47. Gillenwater LA, Kechris KJ, Pratte KA, Reisdorph N, Petrache I, Labaki WW, et al. Metabolomic Profiling Reveals Sex Specific Associations with Chronic Obstructive Pulmonary Disease and Emphysema. Metabolites [Internet]. 2021 Mar 11 [cited 2021 Apr 21];11(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7999201/

48. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, et al. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. Anal Chem. 2006 Jan 15;78(2):567–74.

49. Trevor Hastie, Robert Tibshirani, Balasubramanian Narasimhan, Gilbert Chu. impute: Imputation for microarray data [Internet]. Available from: https://www.bioconductor.org/packages/release/bioc/html/impute.html

50. Tenenbaum D. KEGGREST: Client-side REST access to KEGG [Internet]. Available from: https://bioconductor.riken.jp/packages/3.0/bioc/html/KEGGREST.html

51. Li K, Gao Y, Pan Z, Jia X, Yan Y, Min X, et al. Influence of Emphysema and Air Trapping Heterogeneity on Pulmonary Function in Patients with COPD. Int J Chron Obstruct Pulmon Dis. 2019 Dec 6;14:2863–72.