

RESEARCH ARTICLE

MeDiA: Mean Distance Association and Its Applications in Nonlinear Gene Set Analysis

Hesen Peng^{1†}, Junjie Ma², Yun Bai³, Jianwei Lu^{4,5*}, Tianwei Yu^{1*}

1 Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, United States of America, **2** Department of Hematology, Yantai Yuhuangding Hospital, Yantai, Shandong, China, **3** Department of Pharmaceutical Sciences, School of Pharmacy, Philadelphia College of Osteopathic Medicine, Suwanee, Georgia, United States of America, **4** School of Software Engineering, Tongji University, Shanghai, China, **5** Advanced Institute of Translational Medicine, Tongji University, Shanghai, China

✉ Current address: Microsoft Corporation, 136 102nd Ave SE Apt 326, Bellevue, Washington, United States of America

* jwlu33@gmail.com (JL); tianwei.yu@emory.edu (TY)



OPEN ACCESS

Citation: Peng H, Ma J, Bai Y, Lu J, Yu T (2015) MeDiA: Mean Distance Association and Its Applications in Nonlinear Gene Set Analysis. PLoS ONE 10(4): e0124620. doi:10.1371/journal.pone.0124620

Academic Editor: Yi Xing, University of California, Los Angeles, UNITED STATES

Received: December 21, 2014

Accepted: March 17, 2015

Published: April 27, 2015

Copyright: © 2015 Peng et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: This study is based on data already published in the Gene Expression Omnibus database. All accession numbers are provided within the paper.

Funding: Tianwei Yu was supported by National Institutes of Health grants P20HL113451 and U19AI090023. Jianwei Lu was supported by China Ministry of Science and Technology 973 Program (No. 2013CB967101), and Shanghai Science Committee Foundation (13PJ1433200). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Probabilistic association discovery aims at identifying the association between random vectors, regardless of number of variables involved or linear/nonlinear functional forms. Recently, applications in high-dimensional data have generated rising interest in probabilistic association discovery. We developed a framework based on functions on the observation graph, named MeDiA (Mean Distance Association). We generalize its property to a group of functions on the observation graph. The group of functions encapsulates major existing methods in association discovery, e.g. mutual information and Brownian Covariance, and can be expanded to more complicated forms. We conducted numerical comparison of the statistical power of related methods under multiple scenarios. We further demonstrated the application of MeDiA as a method of gene set analysis that captures a broader range of responses than traditional gene set analysis methods.

Introduction

In the analysis of high-throughput biological data, such as gene expression data, proteomics data, and metabolomics data, it is often of interest to examine the behavior of groups of variables (genes, proteins or metabolites). The grouping of the variables are commonly pre-determined by functional annotations of the biological units using databases, e.g. Gene Ontology [1] or KEGG pathways [2]. A number of methods were developed in the area of gene set analysis to test for shifts of overall expression levels of genes involved in a gene set under different treatment conditions [3–5]. This approach is commonly referred to as gene set analysis. Besides analyzing the behavior of each gene set in response to certain biological conditions, another class of methods examine the relations between gene sets, both under a single treatment condition [6] and between different treatment conditions [7–9].

Competing Interests: The authors have declared that no competing interests exist.

So far most of the methods developed for the analysis of gene sets are based on linear relations between random variables. However complex and nonlinear relations between genes and between a gene and treatment condition has been documented [10–12]. Utilizing general probabilistic associations beyond linear association could produce more insights into the data.

If we consider each gene set as a random vector consisting of multiple random variables (genes), seeking association between gene sets boils down to finding probabilistic associations between two random vectors. In this manuscript we first propose and generalize new methods to discover probabilistic association between random vectors. Then we demonstrate the utility of such measures in finding the general dependency between gene sets and multi-dimensional clinical outcomes.

Consider two random vectors \mathbf{X} and \mathbf{Y} and n pairs of independent and identically distributed (*i.i.d.*) random samples $\{x_i, y_i\}_{i=1}^n$. We would like to draw inference for the existence of probabilistic association between \mathbf{X} and \mathbf{Y} based on the n pairs of samples. The discussion in this paper will focus on the probabilistic association between continuous random variables defined in the Euclidean space.

Classical association statistics like Pearson's correlation coefficient assume functional forms (for example, piecewise linear, monotonicity) between \mathbf{X} and \mathbf{Y} , which are judged as correlated if $Corr(\mathbf{X}, \mathbf{Y}) \neq 0$. Probabilistic association statistic, as the name suggests, perceives associations from the level of probabilistic dependence. That is, \mathbf{X} and \mathbf{Y} are judged as independent if and only if their joint probability density function can be factored, $F(\mathbf{X}, \mathbf{Y}) = F_X(\mathbf{X})F_Y(\mathbf{Y})$, where $F()$ is the probability density function for the random vector under consideration. Probabilistic association encapsulates a larger group of associations than traditional correlation coefficient. For example, probabilistic association would consider nonlinear interactions involving multiple variables.

There are multiple methods on probabilistic association discovery linked to functions on the observation distance graph. The distance graph consists of nodes representing each observation (x_i, y_i) in the $p+q$ dimensional Euclidean space. Here p and q are the dimensions of \mathbf{X} and \mathbf{Y} , respectively. Edges of the observation graph would connect two nodes (observations) if specific criteria are satisfied. For example, mutual information and its derivatives have been the most popular probabilistic association statistic to date [11,13]. To estimate mutual information, the joint entropy can be approximated using log-transformed K -nearest neighbour distance averaged for each observation [14–16].

Recent breakthrough on distance covariate sheds light on universal association discovery with simplicity and theoretical flexibility [17,18]. Brownian distance covariate was defined as $V_N^2 = \frac{1}{n^2} \sum_{k,l=1}^n D_{kl}^X D_{kl}^Y$, where D_{kl}^X and D_{kl}^Y are linear functions of pairwise distances between sample elements calculated with \mathbf{X} and \mathbf{Y} dimensions, respectively [17]. Given fixed marginal distribution for \mathbf{X} and \mathbf{Y} , large Brownian distance covariate suggests the existence of probabilistic association.

In this manuscript, we first describe a general framework for nonlinear association based on the observation graph. It can incorporate different distance metrics and weighting schemes. We then discuss two specific forms of associations including their properties, and illustrate their performance in simulations. Next we show such metrics can be successfully used in nonlinear gene set analysis, which can incorporate multivariate outcome variables, nonlinear associations, and within-gene set heterogeneity.

Methods

The general form of association based on observation graph

We propose the general form of association as

$$M = \sum_{k=1}^n \sum_{l \neq k} D_{kl}^{(X,Y)} w_{kl}, k = 1, \dots, n, l = 1, \dots, n \tag{1}$$

where n is the total number of data points, $D_{kl}^{(X,Y)}$ is the distance between data points calculated using both X and Y dimensions, and w_{kl} is the weight depending on the specific considerations of the data. This can be seen as a general framework when we consider different distance metrics and weighting schemes can be used. We describe two specific types of association scores in the following sections.

Mean Distance Association (MeDiA) score

We let $w_{kl} = \frac{1}{2} \binom{n}{2}^{-1}$, then the score is the mean distance between all pairs of points in the joint space of (X,Y) . We are interested in testing the existence of Probabilistic association between two random vectors (X,Y) , given n pairs of observations. Consider another pair of random vectors (\tilde{X}, \tilde{Y}) , where \tilde{X} follows independent and identical distribution (*i.i.d.*) as X and \tilde{Y} follows *i.i.d.* distribution as Y . The only difference is that \tilde{X} and \tilde{Y} are mutually independent. As mentioned above, we would like to compare the sample observation distance from (X,Y) against that from (\tilde{X}, \tilde{Y}) . Intuitively, when X and Y are probabilistically associated, the point cloud occupies a smaller space, hence the mean distance tends to be smaller than that from (\tilde{X}, \tilde{Y}) .

Theorem 1. Denote the distance between two independent random samples from (X, Y) as d_{XY} , and the distance between two independent random samples from (\tilde{X}, \tilde{Y}) as $d_{\tilde{X}\tilde{Y}}$. Then we have

$$E(d_{XY}) \leq E(d_{\tilde{X}\tilde{Y}})$$

Proofs of the theoretical results in this section are presented in [S1 File](#). Theorem 1 confirms our intuition: when two random vectors are probabilistically associated, their observations tend to be closer compared with their independent counterparts. Denote distances between two random observations as d_{ij} , where i and j are the indices among the n observations. We have the following property:

Corollary 1. For a given observation i , define its mean peer distance as [Eq 2](#). Also define the mean observation distance for n observations as [Eq 3](#):

$$\bar{d}_i = \frac{1}{n-1} \sum_{j \neq i} d_{ij}, \forall i \tag{2}$$

$$M = \frac{1}{n} \sum_i \bar{d}_i \tag{3}$$

Under the null hypothesis that random vectors X and Y are independent, the mean observation distance follows asymptotic normal distribution as $n \rightarrow \infty$.

Corollary 1 is easily proved using the Central Limit Theorem. Based on Corollary 1, we can approximate the null distribution of mean distance using normal distribution, which alleviate the computational burden of the permutation test when n is reasonably large.

Mean Distance Association using Nearest Neighbor (MeDiANN)

We let $w_{kl} = 1/n$ when the involved elements are nearest neighbors, and $w_{kl} = 0$ otherwise. The association score becomes:

$$\begin{aligned} \tilde{d}_i &= \min(d_{i1}, \dots, d_{im}) \\ M &= \frac{1}{n} \sum_i \tilde{d}_i \end{aligned} \tag{4}$$

MeDiANN also enjoys asymptotic normality following Bickel and Breiman [19]. That is, regardless of the joint distribution of (X, Y) or the norm used to define the distance, the MeDiANN score M follows

$$\frac{1}{\sqrt{n}}(M - E(M)) \rightarrow N(0, \sigma^2) \text{ as } n \rightarrow \infty$$

where $E(M)$ is the expectation of the MeDiANN score M , and σ is the asymptotic standard deviation. This property leads to the proposal of Gaussian plug-in permutation test in the next section, and reduces the computational burden of simulating the score distribution under the null hypothesis of independence between X and Y .

Using Gaussian plug-in permutation test for inference

Following theoretical results from the sub-sections above, we propose a permutation test of probabilistic association using MeDiA or MeDiANN. Given n pairs of observations $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$, the null distribution of the test statistic is generated with the following procedure:

1. Permute relative indices of samples $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$, i.e. generate random pairing between the \mathbf{x} 's and \mathbf{y} 's, and calculate MeDiA or MeDiANN score after permutation.
2. Repeat the above step K times and record all the scores, denoted as $\{\hat{M}_k\}_{k=1, \dots, K}$. Calculate mean and standard deviation of $\{\hat{M}_k\}_{k=1, \dots, K}$, denoted as $(\hat{\mu}, \hat{\sigma})$.
3. Approximate the null distribution using normal distribution with mean and standard deviation equaling to $(\hat{\mu}, \hat{\sigma})$.
4. Compare the score M from the actual data with the approximated null distribution, and generate one-sided p -value of the test, $p = \Phi\left(\frac{M - \hat{\mu}}{\hat{\sigma}}\right)$.

Results

A simple example

We show a simple example in Fig 1. Both plots show 300 samples from two bivariate random vectors. Observations on the left panel were sampled from independent bivariate normal distributions. Observations on the right panel were sampled from a mixture bivariate normal distribution. Half of the samples were from a bivariate normal distribution with variance-covariance

matrix of $\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$, and the other half of the samples were from a bivariate normal distribution

with variance-covariance matrix of $\begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$.

Both distributions in Fig 1A and 1B have standard normal marginal distribution and zero correlation coefficient. However, the two samples differ on a group of metrics defined on the

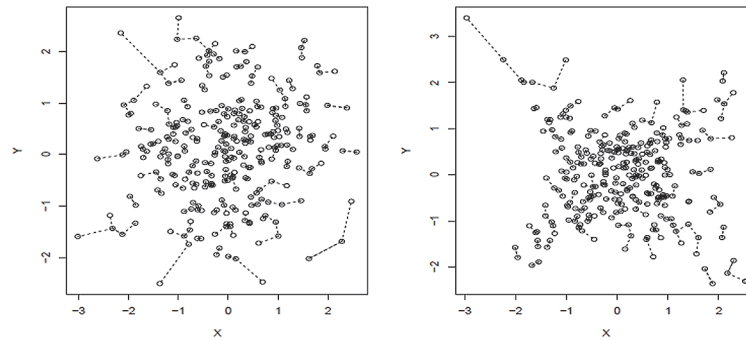


Fig 1. Random samples generated from independent bivariate normal distribution (left), and mixture bivariate normal distribution with ± 0.8 covariates (right). The dashed lines connects two observations if they are nearest neighbors.

doi:10.1371/journal.pone.0124620.g001

observation distances. Table 1 shows the mean distance, mean nearest neighbor distance, and mean log-nearest neighbor distance are all smaller for the dependent case (Fig 1B) compared with independent case (Fig 1A).

Numerical Comparison

For a more systematic assessment of the performance of the proposed method, we conducted a simulation study to examine the size and statistical power of the tests, together with two commonly used non-linear association statistics—the mutual information (MI) and the Brownian covariance (dCov).

The power of the above methods are compared using simulation under scenarios below:

- **Linear association:** X and Y are both p -dimensional random vectors. (X, Y) follow multivariate normal distribution with zero mean, unit variance and covariance of ρ between all pairs of random variables.
- **Variance association:** X follows p -dimensional normal distribution with zero mean, unit variance, and zero covariance. Y is a p -dimensional random vector. y_{ij} is sampled from $N(0, |x_{ij}|)$, $i = 1, \dots, p, j = 1, \dots, n$.
- **Sine curve association:** X is sampled from p -dimensional normal distribution with zero mean, unit variance, and zero covariance. Then x_i is linearly scaled to between 0 and 2π . $y_{ij} = \sin(x_{ij} + \zeta_i) + \epsilon_{ij}$, $i = 1, \dots, p, j = 1, \dots, n$, $\zeta_i \sim \text{Unif}(0, 2\pi)$, $\epsilon_{ij} \sim N(0, \sigma^2)$. Square function association X follows p -dimensional normal distribution with zero mean, unit variance, and zero covariance. $y_{ij} = x_{ij}^2 + \epsilon_{ij}$, $i = 1, \dots, p, j = 1, \dots, n$, $\epsilon_{ij} \sim N(0, \sigma^2)$.

For each of the above scenarios, we generated n pairs of random samples. We tested the existence of association using MeDiA, MeDiANN, MI and dCov. We used $p = 3$ for all cases. The sample size n ranged from 25 to 500. For the linear association case, we used $\rho = 0.5$, which is a

Table 1. Comparison between the independent bivariate normal distribution and mixture normal distribution in Fig 1.

Metric	Left (Independent)	Right (Mixed Normal)
mean distance (MeDiA)	1.81	1.70
mean nearest neighbor distance (MeDiANN)	0.14	0.12
mean log(nearest neighbor distance)	-2.24	-2.43

doi:10.1371/journal.pone.0124620.t001

relatively weak pairwise correlation level in real data. For the sine curve and square function associations, a noise term was involved. We used a noise level such that the signal variance is half that of the noise, which is a relatively high noise level in real data. For each scenario/method/sample size tuple, we repeated the simulation 400 times. The p -value for each simulation was recorded. And finally the power for each method under each scenario and sample size combination is calculated as the percentage of tests with p -values smaller than 0.05.

Power comparison shows differentiated method performance in different scenarios. MeDiANN is defined as average nearest neighbor observation graph edge length. The version of MI used here is estimated as average log-transformed nearest neighbor edge length. They share similar power possibly due to the similar forms of estimation. Nonetheless, there are cases where one is better than the other.

In the linear association scenario, dCov and MeDiA performed almost identically. Both are better than the other two methods (Fig 2, lower-left panel). There is a tight connection between Pearson correlation and Brownian Covariate under multivariate normal distributions [17]. At the same time, the data points generated from the joint density function tend to be closer to the diagonal compared to the independent case, causing the average pairwise distance to be smaller, which allows sufficient power for MeDiA. The average distance to the nearest neighbor is reduced not as much as the average pairwise distance. Thus the MiDiANN and MI showed lower power than dCov and MeDiA when the sample size is not large.

In the square function association case, MeDiA outperformed the other three methods (Fig 2, upper-left panel). To understand the result, we consider a simple case where both X and Y are one-dimensional. Because the data points fall close to a curve $y = x^2$, compared to the independent case, a point (x_i, y_i) is not only closer to other points with similar x values, but also to other points with x values close to $-x_i$. Thus the average distance utilized by MeDiA is substantially reduced. On the other hand, for dCov, two y values that are close may mean the corresponding x values are either close or opposite. Thus dCov showed lower power than MeDiA. Similar to the linear case, the reduction of a point's distance to its nearest neighbor is not very substantial compared to the independent case. Thus the power of MeDiANN and MI trailed the other two methods.

In the variance association case, MeDiA again led the performance, followed by MeDiANN and MI, while dCov trailed in the performance (Fig 2, upper-right panel). As illustrated by the one-dimensional case, the data points tend to stay in a subspace, and the majority of the data points are close to the center. Clearly the average pairwise distance tends to be much smaller than the independent case, favoring MeDiA. At the same time, as the data points are crowded near the origin, the distance of a data point to its nearest neighbor also tends to be small. Thus MeDiANN and MI have reasonably good performance. On the other hand, data points with similar y values may have x values that spread over a big range, causing difficulty for dCov.

In the sine function association case, dCov achieved higher statistical power than the other three methods (Fig 2, lower-right panel). At the high noise level in our simulations, the bulk of the data follows a relationship close to linear. Thus dCov showed the highest power as it is favored by linear relations. MiDiANN showed better power than MI at smaller sample size.

Overall, general dependency between groups of variables can take many forms. No method is uniformly better than other methods. Each method is favored by certain types of associations, while having difficulty in other forms of associations. It appears that when the bulk of the point cloud follows a relation that is close to linear, dCov has the highest statistical power. When the underlying function is far from linear, MeDiA has the highest statistical power. MeDiANN and MI appear to handle both cases reasonably well, but don't excel in either. Between the two, MeDiANN has slightly better power than MI when the relation is close to linear, while MI is better when the underlying function is curved.

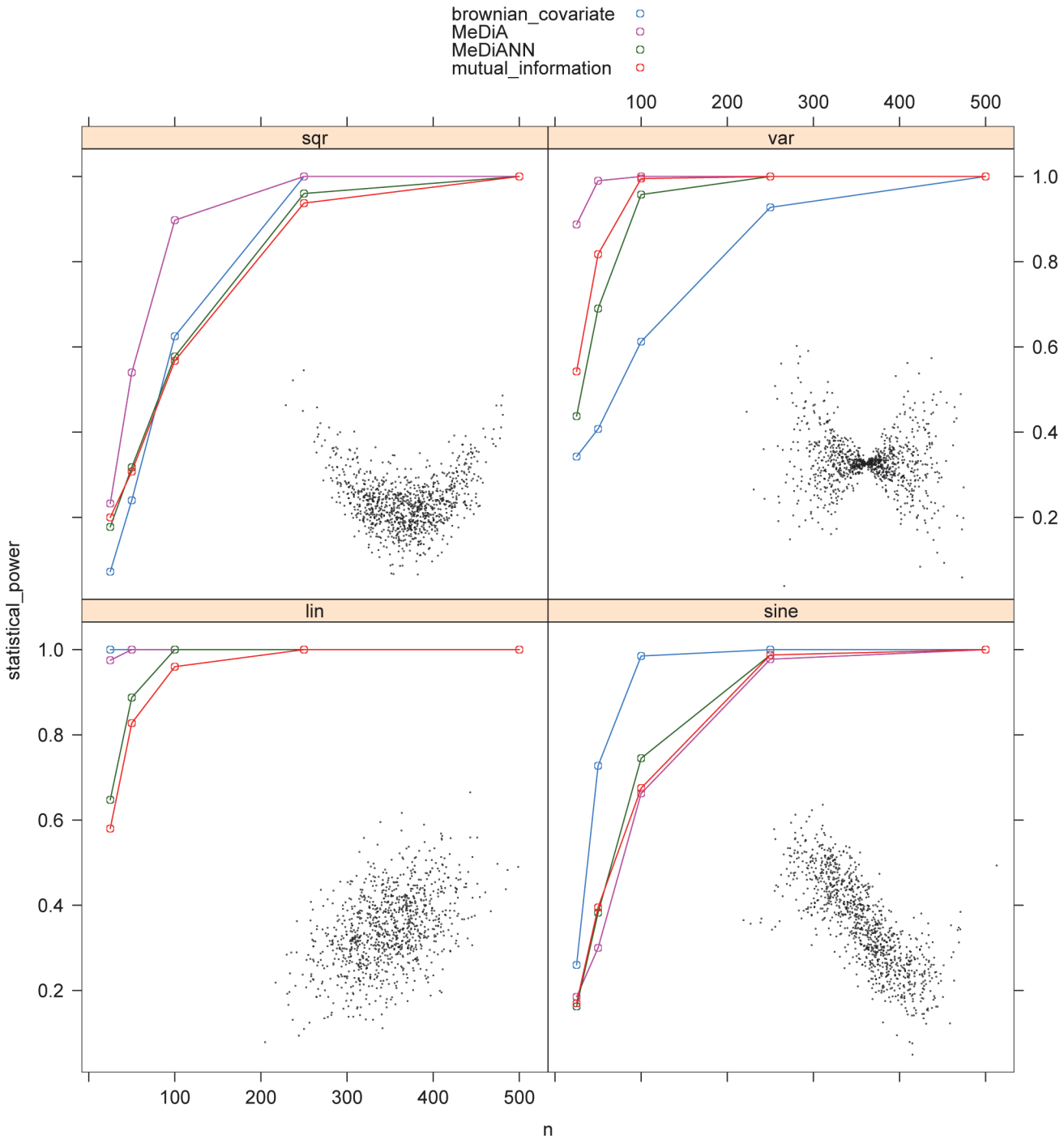


Fig 2. Comparison of statistical power under different scenarios.

doi:10.1371/journal.pone.0124620.g002

Nonlinear gene set analysis using MeDiA

In this section we apply probabilistic association discovery on the analysis of gene sets. Unlike traditional gene set analysis that seeks gene set-level differential expression [3–5], the class of methods we discussed above seeks to find general dependencies between gene sets and the clinical response variable(s). Using this type of methods allows multi-dimensional clinical

outcome, heterogeneity in the genes' behavior within a gene set, as well as non-linear response between genes and the clinical outcome variables. As MeDiA achieved very good overall statistical power in our simulations, in this section, we use MeDiA for the data analysis.

We studied the dataset of gene expression in primary acute lymphoblastic leukemia (ALL) associated with methotrexate (MTX) treatment [20]. The dataset is GSE10255 from the Gene Expression Omnibus [21] (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10255>). The major clinical outcome is the change of circulating leukemia cells after initial MTX treatment. We selected the probesets with known ENTREZ Gene IDs. When a gene was represented by more than one probesets, we merged the corresponding probesets by taking the average expression values of the probesets. The data matrix contained 12704 genes and 161 samples.

We used a two-dimensional outcome—the day 0 white blood cell count, and the day 3 white blood cell count, both on log scale. This type of multi-dimensional outcome is not accommodated by traditional gene set analysis methods. For the gene sets, we used a previously described procedure to reduce redundancy [9,22], and selected a group of Gene Ontology Biological Process (BP) terms that are representative—577 biological process terms that contain a total of 10455 genes. We used MeDiA and its accompanying permutation test procedure to test for association between each gene set and the clinical outcome vector. The p-values were adjusted to False Discovery Rate using the Benjamini and Yekutieli procedure [23].

The top gene sets with $FDR < 0.01$ are shown in Table 2. First we notice a number of immune system GO terms in the table (label 1, Table 2). Besides being an anti-cancer agent, MTX is also used to treat autoimmune diseases. It takes action by inhibiting enzymes in methyltransferase and purine metabolism, hence suppressing immune system function [24]. Because this relation is obvious, we shall skip the discussion of details.

Another observation is that a number of signal transduction pathways are among the top GO terms (label 2, Table 2). Some of the pathways were known to be regulated by MTX. The NF-kappaB activation is suppressed by MTX through the inhibition of IkappaB alpha phosphorylation and degradation [25]. The relation between MTX or leukemia to other pathways are not as clearly documented. But there has been some evidences. For example, PELO, which is involved in both cell-matrix adhesion and integrin-mediated signaling pathway, was found to be differentially expressed in AML [26]. It has been shown that MTX affects human bone cell mechanotransduction by interfering with BMP4, which is involved in both integrin-mediated signaling and regulation of protein kinase activity [27]. CAV1, a gene belonging to the cytokine-mediated signaling pathway, was found to be one of the genes characterizing MTX non-responders in patients with rheumatoid arthritis (RA) [28]. At the same time, CAV1 is considered a general tumor suppressor, the lack of expression of which was implicated in the pathogenesis of many cancers, while the over expression of which has also been associated with tumor progression and metastasis in prostate cancers [29]. With regard to the calcium-mediated signaling pathway, it has been shown that the interaction between CXCR4 and SDF-1 is a key mediator of the resistance to chemotherapy in children with ALL [30]. In mouse experiments, MTX treatment caused reduced CXCR4 and CXCL12 expression [31].

Four metabolism GO terms were among the list (label 3, Table 2), three of which were lipid metabolism pathways. It has recently been shown that MTX potentiated glucose uptake and lipid oxidation in skeletal muscle [32]. MTX was also found to improve lipid parameters and fasting plasma glucose levels in a cross-sectional study of humans [33]. The specific pathways detected by our method could indicate certain mechanistic links to explain the observations.

The GO term “stem cell maintenance” is the most significant in the list (label 4, Table 2). Human hematopoietic stem cell maintenance mediated by the transcriptional coactivator CITED2 contributes to leukemia maintenance [34]. In addition, in AML, the loss of *Leo1* leads to down-regulation of *SOX2* and *SOX4*, potent oncogenes in myeloid transformation [35].

Table 2. Gene sets associated with the two-dimensional clinical outcome based on MeDiA.

GO term*	FDR	Name
⁴ GO:0019827	1.65E-06	stem cell maintenance
¹ GO:0050852	1.47E-05	T cell receptor signaling pathway
³ GO:0006693	0.00042	prostaglandin metabolic process
⁵ GO:0033627	0.00047	cell adhesion mediated by integrin
¹ GO:0030183	0.00051	B cell differentiation
¹ GO:0045058	0.00072	T cell selection
³ GO:0009225	0.0027	nucleotide-sugar metabolic process
¹ GO:0045730	0.0027	respiratory burst
GO:0000122	0.0031	negative regulation of transcription from RNA polymerase II promoter
² GO:0007229	0.0031	integrin-mediated signaling pathway
⁶ GO:0051668	0.0031	localization within membrane
³ GO:0006633	0.0038	fatty acid biosynthetic process
⁶ GO:0008105	0.0038	asymmetric protein localization
¹ GO:0019882	0.0038	antigen processing and presentation
² GO:0043123	0.0038	positive regulation of I-kappaB kinase/NF-kappaB cascade
² GO:0043627	0.0038	response to estrogen stimulus
GO:0001837	0.0047	epithelial to mesenchymal transition
¹ GO:0006959	0.0047	humoral immune response
GO:0044419	0.0047	interspecies interaction between organisms
² GO:0006469	0.0064	negative regulation of protein kinase activity
² GO:0019221	0.0064	cytokine-mediated signaling pathway
² GO:0019722	0.0064	calcium-mediated signaling
⁷ GO:0015012	0.0066	heparan sulfate proteoglycan biosynthetic process
³ GO:0042632	0.0079	cholesterol homeostasis
¹ GO:0050869	0.0079	negative regulation of B cell activation
⁵ GO:0022407	0.0079	regulation of cell-cell adhesion
GO:0046677	0.0087	response to antibiotic
⁸ GO:0006919	0.0094	activation of caspase activity
GO:0006997	0.0099	nucleus organization

* Superscripts by the GO terms are for easy reference from the main text.

doi:10.1371/journal.pone.0124620.t002

Two terms involved in cell adhesion were found in the list (label 5, Table 2). There have been known links between leukemia or MTX to cell adhesion. A few genes in the collagen metabolism pathway are altered with leukemia [36], and the overall expression level of collagen increases with MTX treatment [37]. In addition, several cellular adhesion molecules are known to be influenced by MTX [24]. A number of proteins in the membrane organization process are influenced by leukemia (label 6, Table 2) [36,38]. Overall, almost all the GO biological processes found by MeDiA make biological sense, indicating MeDiA is an effective method for nonlinear gene set analysis.

Differential pathway interaction network discovery

The change of pathway interactions under different cell status is of crucial interest in biomedical study. For example, certain interactions between pathways may be amplified or suppressed in disease state compared with healthy states. These changes in interaction may facilitate the discovery of cell regulatory mechanism. We applied a network reverse engineering procedure

for pathway interaction to celiac disease data (NCBI data set GDS3646) and lung cancer data (NCBI data set GDS2771), in which we were specifically interested in identifying pathway interactions that are amplified or suppressed in the disease state.

The celiac disease data consists of gene expression levels of untouched primary leukocytes from 132 unrelated celiac disease individuals and 22185 probesets [39]. Of the 132 individuals, 110 have sustained celiac disease, and 22 are healthy control individuals. Illumina HumanWGv2 annotation data was used to group probe reads into 214 KEGG pathway groups, covering 5201 genes of the data set. The lung cancer data consists of gene expression levels of large airway epithelial cells from cigarette smokers without cancer, with cancer, and with suspect lung cancer [40]. The probe reads were grouped into 214 KEGG pathways using Affymetrix Human Genome U133A database.

The amplified/suppressed pathway interactions were identified using the following procedure:

1. For each pair of pathways i and j that don't share any gene, a MeDiANN permutation test of association was applied for the disease group and control group, respectively. Denote the test p-value for disease group as p_{ij}^D , and p_{ij}^C for control group.
2. Rank $\{p_{ij}^D\}$ and $\{p_{ij}^C\}$ from all i, j combinations in ascending order, and obtain the ranks $\{r_{ij}^D\}$ and $\{r_{ij}^C\}$. This is done for the two groups separately.
3. Calculate between state rank differences $d_{ij} = r_{ij}^D - r_{ij}^C$.
4. Pathway pairs with rank change d_{ij} smaller than the 1% quantile in all $\{d_{ij}\}$ are identified as amplified in association at the disease state. Pairs with d_{ij} greater than 99% quantile are identified as suppressed in association in disease state.

The identified pathways interactions are then checked for their biological meanings and discussed below.

Celiac Disease Pathway Interaction. The newly developed method was used to analyze the publically available Gene Expression Omnibus (GEO) data set GDS3646 (<http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3646>). GDS3646 record is an expression analysis of untouched primary leukocytes from unrelated celiac disease individuals [39]. In the study, the gene expression in untouched primary leukocytes from individuals with celiac disease were compared with an EBV-transformed HapMap B cell line data. Celiac disease, a multifactorial disorder with complex genetics, is an enteropathy caused by autoimmune response against wheat gluten, the protein component of the cereals wheat, rye and barley in genetically susceptible individuals [41]. Patients with celiac disease have a wide spectrum of gastrointestinal and extraintestinal manifestations, characterized by intestinal malabsorption and atrophy of intestinal villi [42,43]. Celiac patients experience altered carbohydrate, lipid, peptide/protein, metabolism levels. Untreated celiac patients oxidize more carbohydrates as energy substrate compared to treated subject [42].

The amplified pathway pairs are predominantly related to nutrition absorption and metabolism, while a large proportion of the suppressed pathway pairs are between metabolism and signal transduction (Fig 3; S1 Table). Other pathways potentially linked with celiac disease were also identified. For example, the 04062 chemokine signaling pathway has 6 connections. Chemokines are small peptides that provide directional cues for the cell trafficking and thus are vital for protective inflammatory immune response that requires the recruitment of leukocytes to the site of inflammation upon foreign insult. Celiac disease is known to be an inflammation disease caused by dietary gluten. In genetically predisposed people, gliadin peptides (derivatives of gluten) provokes immune response, which leads to the production of pro-inflammatory

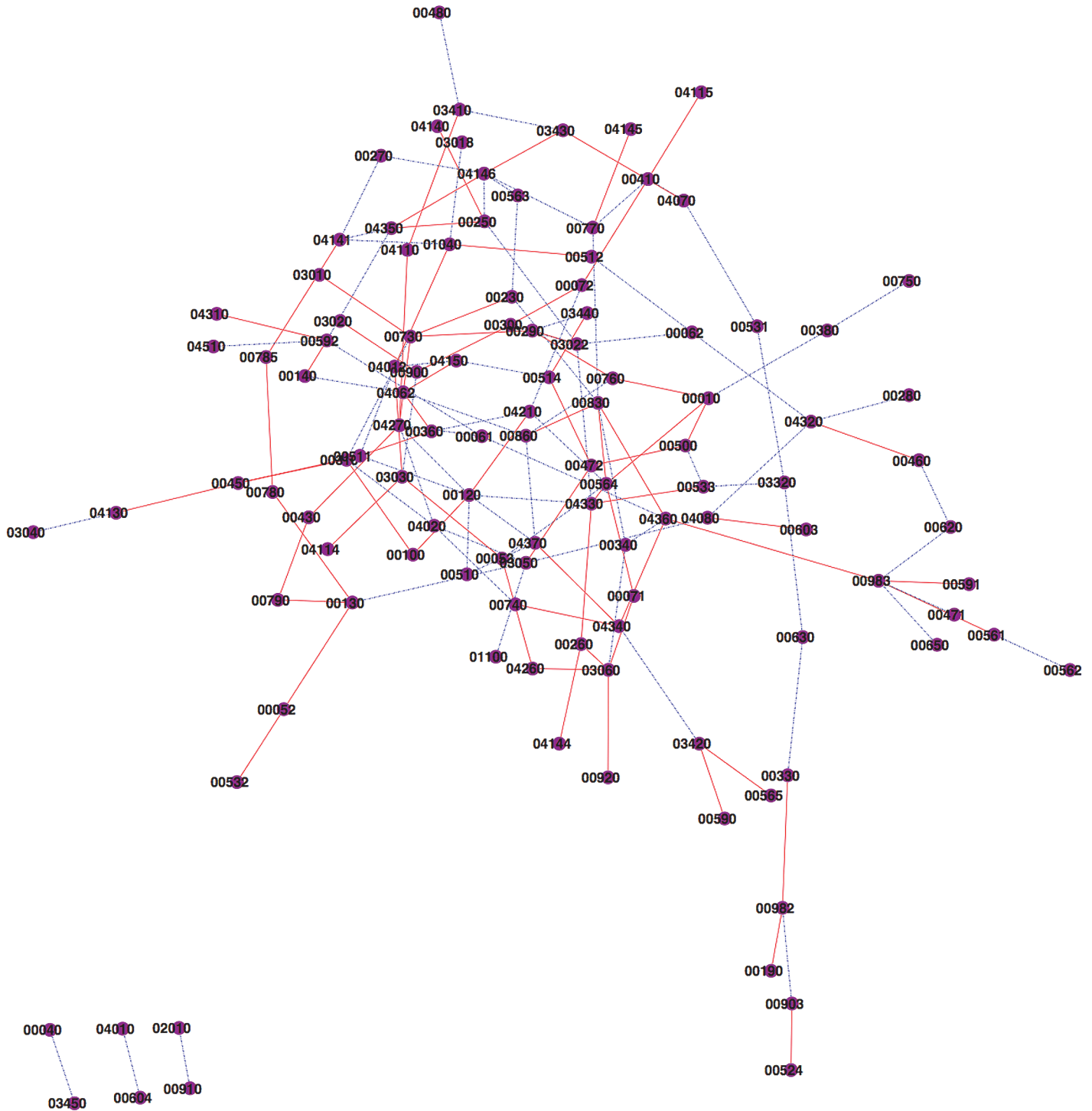


Fig 3. Network interaction for celiac disease pathways. Red edge indicates that the interaction between connected pathways are amplified in disease individuals. Blue edge indicates the interaction suppressed in disease individuals.

doi:10.1371/journal.pone.0124620.g003

cytokines and subsequent damage to, and increased permeability of the intestinal epithelium [44–46].

Another example is the 02010 ABC transporters pathway. The results show that the interaction between ABC transporter pathway and nitrogen metabolism pathway is suppressed in the celiac disease patients. The ATP-binding cassette (ABC) transporters are protein families that couple ATP hydrolysis to activate transport of a wide variety of substrates such as ions, sugars, lipids, sterols, peptides, proteins, and drugs [47,48]. ABC transporters have been confirmed to be related to celiac disease. It has been reported that a close association exists between polymorphism of TAP1 and TAP2 (ABC transporter gene) and disease susceptibility among southern European populations [49]. The products of TAP1 and TAP2 are ABC transporters, which are believed to transport antigenic peptides from the cytoplasm into the endoplasmic reticulum. It was reported that nitrogen balance was modulated in celiac patients [50]. In addition, both nitrate/nitrite are transported by ATP-binding cassette (ABC) transporters [51].

In addition, the relation between 04370 VEGF signaling pathway and several pathways is found to be modulated in celiac patients, including 04340 hedgehog signaling, 510 N-Glycan biosynthesis, 00860 porphyrin and chlorophyll metabolism, 00120 primary bile acid biosynthesis. Vascular endothelial growth factor (VEGF) family and its receptor systems have been demonstrated to be the fundamental regulator in the cell signaling of angiogenesis. Angiogenesis is an essential biological process involved in the progression of a variety of major diseases such as cancer, diabetes and inflammation [52]. It was reported that small-bowel mucosal vascular network was altered in untreated coeliac disease. The study found that on a gluten-containing diet the mucosal vasculature in the small intestine of untreated coeliac disease patients was altered in overall organization as well as in the number and maturity of the vessels when compared to healthy subjects. In patients on a gluten-free diet, the vasculature normalized parallel to mucosal recovery [53]. Angiogenesis is reported to be related to hedgehog signaling [54,55], bile acid [56], glycan biosynthesis [57], porphyrin [58,59]. 04210 apoptosis pathway, the programmed cell death, also shows a number of connections. Much evidence supported the increase of small intestinal apoptosis in celiac disease [60]. Some other study demonstrated that enterocyte apoptosis induced by activated intraepithelial lymphocytes is increased in celiac disease [61].

Lung Cancer Pathway Interaction. The newly developed method is also tested on GDS2771 data set (<http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2771>), which is the microarray data of large airway epithelial cells from cigarette smokers without cancer, with cancer, and with suspect lung cancer. Many studies demonstrated the correlation of altered metabolism with lung cancer, including basal metabolism [62–64], carbohydrate metabolism [65,66], protein metabolism [66–68], lipid metabolism [69], and xenobiotic metabolism [70].

The result show that many relations between metabolism related pathways are regulated (Fig 4; S2 Table). Take the TCA pathway as an example, the citrate cycle (TCA cycle, Krebs cycle) is an important aerobic pathway for the final steps of the oxidation of carbohydrates and fatty acids. Modulation of TCA cycle enzymes have been demonstrated in lung cancer. Decreased activities of TCA cycle key enzymes were observed in lung cancer bearing animals [71].

Some specific pathways related to lung cancer are also caught on the list: It was shown that the correlation between pathway 00072, the synthesis and degradation of ketone bodies pathway and 04145, the phagosome pathway is amplified in lung cancer patients. Phagocytosis is the cellular process of engulfing solid particles by the cell membrane to form an internal phagosome, which is a central mechanism in both immune and apoptosis responses. There is a broadly accepted view that bronchial neoplasms or its products suppress phagocytic functions of alveolar macrophages [72]. The alveolar macrophage is believed to be of central importance in the immune response against infection and tumor. It has been reported that there are type-specific alterations in phagocytosis ability of alveolar macrophage in lung cancer patients, which may result in an inability to stimulate anti-tumor immunity and subsequently cause observed differences between lung cancer subgroups. Altered blood monocyte (BM) phagocytosis

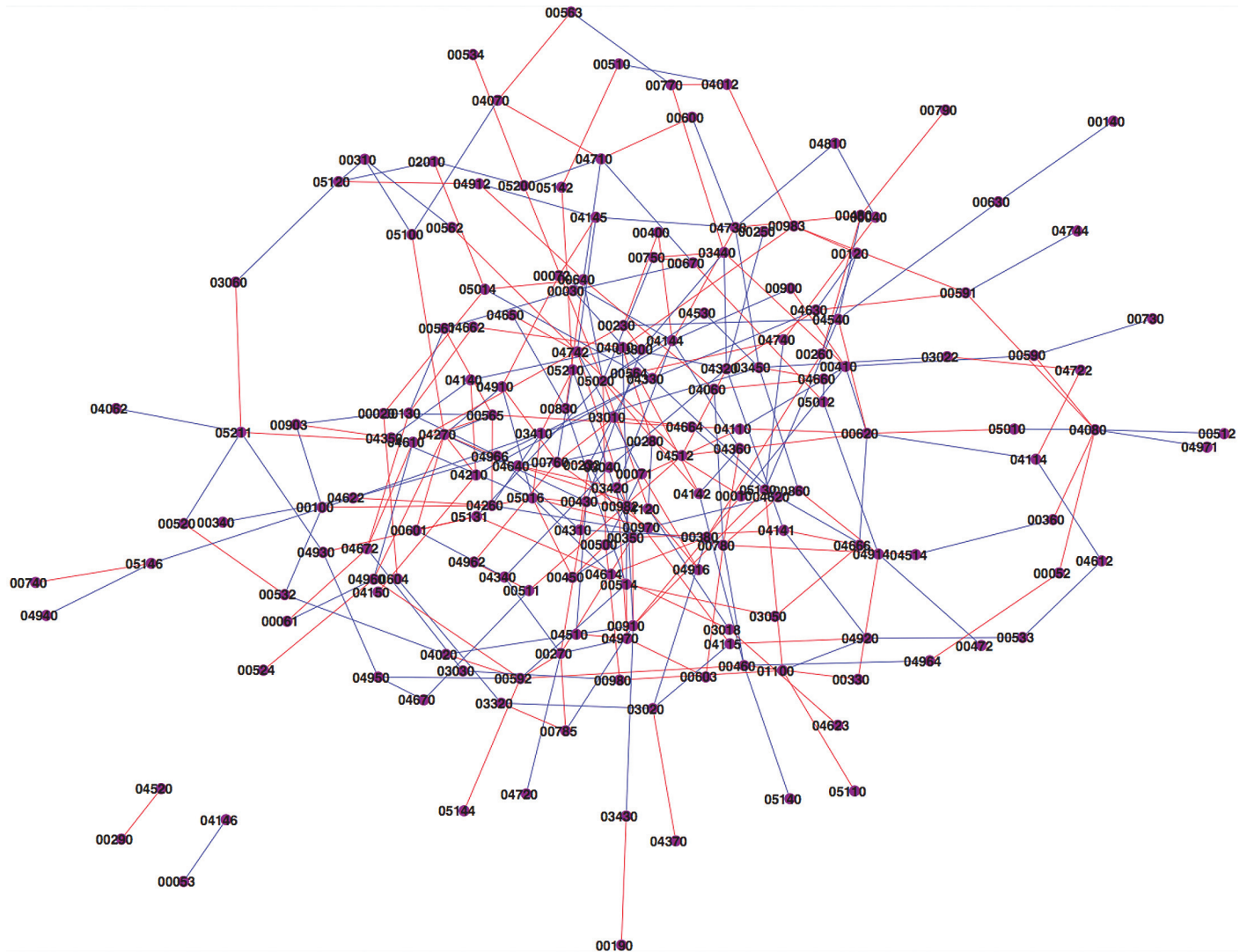


Fig 4. Network interaction for lung cancer pathways. Red edge indicates that the interaction between connected pathways are amplified in disease individuals. Blue edge indicates the interaction suppressed in disease individuals.

doi:10.1371/journal.pone.0124620.g004

ability was also observed in patients with lung cancer [73,74]. More importantly, some studies proved that ketone bodies affect the phagocytic activity of macrophages and leukocytes [75].

Another interesting example is that the correlation between 00232, the pathway of caffeine metabolism, and 00760, the nicotinate and nicotinamide metabolism, amplified in smokers with lung cancer. First of all, both caffeine and nicotine metabolism are generally believed to be related to the risk of lung cancer. Cigarette smoking is a clear risk factor for lung cancer. Even though nicotine, one of the major ingredients and the causative agent for addiction of cigarette smoking, is generally believed not a carcinogen by itself. However, several studies have shown that nicotine can induce cell proliferation and angiogenesis [76]. Nicotine metabolism by cytochrome P450 2A6 (CYP2A6) varies across ethnicity and race, which is indicated to be related to smoking behavior and lung cancer risk [77,78]. The same as smoking, the consumption of coffee is a very old and popular habit. Coffee contains catechins and flavonoids, which exhibit anti-carcinogenic properties. Conversely, caffeine may elevate cancer risk through a variety of

mechanisms [79,80]. Caffeine, an environmentally prominent phosphodiesterase, has been proved to selectively stimulate the growth of pulmonary adenocarcinoma and small airway epithelial cells [81]. Not only are both nicotine and caffeine related to lung cancer, but also many evidences suggested that the metabolism of caffeine and nicotine are closely correlated. Caffeine is mainly metabolized by cytochrome P450 1A2 (CYP1A2). Actually caffeine metabolism has been used as an in vivo marker of CYP1A2 activity, which has been clearly demonstrated to be induced by cigarette smoking [82]. The difference of caffeine intake and plasma concentrations among smokers and nonsmokers was reported [83]. The results from 69 US samples showed that smokers had significantly higher caffeine intake than nonsmokers and the ratio of concentration/dose of caffeine was approximately four-fold higher in nonsmokers than in smokers [83]. In animal studies, nicotine have been proved to induce the activity of several enzymes, including CYP1A2 [84]. It explains why nonsmokers have high plasma caffeine concentration after intake of the same dose of caffeine compared to smokers. Some other research articles reported that the combined NAT2/CYP1A2 status was related to lung adenocarcinoma [78].

Discussion

In this paper we have discussed the general theory and applications of association discovery using functions on the observation graph. Statistics of similar form to Eq 1 are capable of detecting associations between continuous random vectors using permutation test of association (Table 3). However, we would like to point out that Eq 1 is only one of the ways to test probabilistic association. For example, Brownian distance covariate (dCov) utilizes the covariates of between-observation distances calculated using either random vector. Its estimation is derived from estimating the L2 distance between characteristics functions of joint distribution and product of marginal distributions. It arrives at an observation distance product form. This is different from Eq 1. We are confident that there are way more methods to test probabilistic associations to be discovered.

Mutual information estimation is derived from the estimation of joined entropy of variables under consideration. There are two estimation methods generally used. (1) Reverse engineering of joint density. This is the most popular method but is generally not applicable when the dimension under consideration goes too high. (2) Estimation of joint entropy using graph distance. This method as mentioned in the article can circumvent joint density estimation and is more appropriate in high dimensional data analysis.

We have generalized the estimation of mean observation distance (MeDiA), mean nearest neighbor observation distance (MeDiANN), and mutual information (MI) estimate into the same framework of functions on the observation graph. Simulation study showed that the three statistics have different performance in terms of statistical power under different scenarios. In hindsight, we realized that: testing of probabilistic association using observation distance under the framework of Eq 1 actually rests on the testing of observation distance distributions. More specifically, when random vectors X and Y under consideration are associated, their

Table 3. Summary of methods on Probabilistic association discovery discussed in this paper.

Name	Statistic	Inference
MeDiA	Mean distance	Permutation Test
MeDiANN	Mean nearest neighbor distance	Permutation Test
Mutual Information	Mean log nearest neighbor distance	Permutation Test
Brownian Cov	Distance covariate	Permutation Test

doi:10.1371/journal.pone.0124620.t003

distribution of observation distance should be different from their independent counterpart \tilde{X} and \tilde{Y} . In our upcoming work, we will explore an omnibus probabilistic association test based on the observation distance distribution. We would expect this test to be more flexible compared with the methods compared in this paper.

On the applications front, probabilistic dependency discovery methods such as MeDiA and Brownian Covariate can be used to test the hypothesis that a gene set is associated with clinical outcome variables. The association can involve multiple clinical variables, nonlinear interaction, and be heterogeneous within the gene sets. We have also successfully applied the MeDiA for the detection of differential interaction between gene sets under different treatment conditions.

Conclusions

Overall, with the complexity of the biological system, and the documented presence of nonlinear and conditional dependencies, MeDiA and other probabilistic dependency discovery methods based on the observation graph are useful in unraveling high-throughput data to make new biological discoveries.

Supporting Information

S1 File. Mathematical proof of Theorem 1.
(PDF)

S1 Table. Pathway pairs with amplified or suppressed relations in dataset GDS3646.
(XLSX)

S2 Table. Pathway pairs with amplified or suppressed relations in dataset GDS2771.
(XLSX)

Acknowledgments

The authors thank two anonymous reviewers whose comments helped significantly improve the manuscript.

Author Contributions

Conceived and designed the experiments: HP TY. Performed the experiments: HP JM YB JL TY. Analyzed the data: HP JM YB JL TY. Contributed reagents/materials/analysis tools: JL TY. Wrote the paper: HP JM YB TY.

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25–29. PMID: [10802651](#)
2. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 38: D355–360. doi: [10.1093/nar/gkp896](#) PMID: [19880382](#)
3. Efron B, Tibshirani R (2007) On testing the significance of sets of genes. *Ann Appl Stat* 1: 107–129.
4. Goeman JJ, Buhlmann P (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 23: 980–987. PMID: [17303618](#)
5. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545–15550. PMID: [16199517](#)
6. Yu T, Sun W, Yuan S, Li KC (2005) Study of coordinative gene expression at the biological process level. *Bioinformatics* 21: 3651–3657. PMID: [16076891](#)

7. Choi Y, Kendzioriski C (2009) Statistical methods for gene set co-expression analysis. *Bioinformatics* 25: 2780–2786. doi: [10.1093/bioinformatics/btp502](https://doi.org/10.1093/bioinformatics/btp502) PMID: [19689953](https://pubmed.ncbi.nlm.nih.gov/19689953/)
8. Cho SB, Kim J, Kim JH (2009) Identifying set-wise differential co-expression in gene expression microarray data. *BMC Bioinformatics* 10: 109. doi: [10.1186/1471-2105-10-109](https://doi.org/10.1186/1471-2105-10-109) PMID: [19371436](https://pubmed.ncbi.nlm.nih.gov/19371436/)
9. Yu T, Bai Y (2011) Capturing changes in gene expression dynamics by gene set differential coordination analysis. *Genomics* 98: 469–477. doi: [10.1016/j.ygeno.2011.09.001](https://doi.org/10.1016/j.ygeno.2011.09.001) PMID: [21971296](https://pubmed.ncbi.nlm.nih.gov/21971296/)
10. Francesconi M, Lehner B (2014) The effects of genetic variation on gene expression dynamics during development. *Nature* 505: 208–211. doi: [10.1038/nature12772](https://doi.org/10.1038/nature12772) PMID: [24270809](https://pubmed.ncbi.nlm.nih.gov/24270809/)
11. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. (2011) Detecting novel associations in large data sets. *Science* 334: 1518–1524. doi: [10.1126/science.1205438](https://doi.org/10.1126/science.1205438) PMID: [22174245](https://pubmed.ncbi.nlm.nih.gov/22174245/)
12. Li KC (2002) Genome-wide coexpression dynamics: theory and application. *Proc Natl Acad Sci U S A* 99: 16875–16880. PMID: [12486219](https://pubmed.ncbi.nlm.nih.gov/12486219/)
13. Tostevin F, ten Wolde PR (2009) Mutual Information between Input and Output Trajectories of Biochemical Networks. *Physical Review Letters* 102.
14. Gorja MN, Leonenko NN, Mergel VV, Inverardi PLN (2005) A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Journal of Nonparametric Statistics* 17: 277–297.
15. Kraskov A, Stogbauer H, Grassberger P (2004) Estimating mutual information. *Physical Review E* 69.
16. Leonenko N, Pronzat L, Savani V (2008) A Class of Renyi Information Estimators for Multidimensional Densities. *Annals of Statistics* 36: 2153–2182.
17. Szekely GJ, Rizzo ML (2009) Brownian Distance Covariance. *Annals of Applied Statistics* 3: 1236–1265.
18. Szekely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. *Annals of Statistics* 35: 2769–2794.
19. Bickel P, Breiman L (1983) Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *Annals of Probability* 11: 185–214.
20. Sorich MJ, Pottier N, Pei D, Yang W, Kager L, Stocco G, et al. (2008) In vivo response to methotrexate forecasts outcome of acute lymphoblastic leukemia and has a distinct gene expression profile. *PLoS Med* 5: e83. doi: [10.1371/journal.pmed.0050083](https://doi.org/10.1371/journal.pmed.0050083) PMID: [18416598](https://pubmed.ncbi.nlm.nih.gov/18416598/)
21. Barrett T, Edgar R (2006) Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol* 411: 352–369. PMID: [16939800](https://pubmed.ncbi.nlm.nih.gov/16939800/)
22. Yu T, Bai Y (2011) Improving gene expression data interpretation by finding latent factors that co-regulate gene modules with clinical factors. *BMC Genomics* 12: 563. doi: [10.1186/1471-2164-12-563](https://doi.org/10.1186/1471-2164-12-563) PMID: [22087761](https://pubmed.ncbi.nlm.nih.gov/22087761/)
23. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29: 1165–1188.
24. Wessels JA, Huizinga TW, Guchelaar HJ (2008) Recent insights in the pharmacological actions of methotrexate in the treatment of rheumatoid arthritis. *Rheumatology (Oxford)* 47: 249–255. PMID: [18045808](https://pubmed.ncbi.nlm.nih.gov/18045808/)
25. Majumdar S, Aggarwal BB (2001) Methotrexate suppresses NF-kappaB activation through inhibition of I-kappaBalpha phosphorylation and degradation. *J Immunol* 167: 2911–2920. PMID: [11509639](https://pubmed.ncbi.nlm.nih.gov/11509639/)
26. Stirewalt DL, Meshinchi S, Kopecky KJ, Fan W, Pogossova-Agadjanyan EL, Engel JH, et al. (2008) Identification of genes with abnormal expression changes in acute myeloid leukemia. *Genes Chromosomes Cancer* 47: 8–20. PMID: [17910043](https://pubmed.ncbi.nlm.nih.gov/17910043/)
27. Elliot KJ, Millward-Sadler SJ, Wright MO, Robb JE, Wallace WH, Salter DM (2004) Effects of methotrexate on human bone cell responses to mechanical stimulation. *Rheumatology (Oxford)* 43: 1226–1231. PMID: [15238642](https://pubmed.ncbi.nlm.nih.gov/15238642/)
28. Oliveira RD, Fontana V, Junta CM, Marques MM, Macedo C, Rassi DM, et al. (2012) Differential gene expression profiles may differentiate responder and nonresponder patients with rheumatoid arthritis for methotrexate (MTX) monotherapy and MTX plus tumor necrosis factor inhibitor combined therapy. *J Rheumatol* 39: 1524–1532. doi: [10.3899/jrheum.120092](https://doi.org/10.3899/jrheum.120092) PMID: [22753658](https://pubmed.ncbi.nlm.nih.gov/22753658/)
29. Tsuji Y, Nakagawa T, Hatanaka M, Takeuchi T, Matsumoto E, Takenaka H, et al. (2006) Quantification of caveolin isoforms using quantitative real-time RT-PCR, and analysis of promoter CpG methylation of caveolin-1alpha in human T cell leukemia cell lines. *Int J Mol Med* 18: 489–495. PMID: [16865235](https://pubmed.ncbi.nlm.nih.gov/16865235/)
30. Sison EA, Magoon D, Li L, Annesley CE, Rau RE, Small D, et al. (2014) Plerixafor as a chemosensitizing agent in pediatric acute lymphoblastic leukemia: efficacy and potential mechanisms of resistance to CXCR4 inhibition. *Oncotarget* 5: 8947–8958. PMID: [25333254](https://pubmed.ncbi.nlm.nih.gov/25333254/)

31. Georgiou KR, Scherer MA, King TJ, Foster BK, Xian CJ (2012) Deregulation of the CXCL12/CXCR4 axis in methotrexate chemotherapy-induced damage and recovery of the bone marrow microenvironment. *Int J Exp Pathol* 93: 104–114. doi: [10.1111/j.1365-2613.2011.00800.x](https://doi.org/10.1111/j.1365-2613.2011.00800.x) PMID: [22220905](https://pubmed.ncbi.nlm.nih.gov/22220905/)
32. Pirkmajer S, Kulkarni SS, Tom RZ, Ross FA, Hawley SA, Hardie DG, et al. (2014) Methotrexate Promotes Glucose Uptake and Lipid Oxidation in Skeletal Muscle via AMPK Activation. *Diabetes*.
33. Toms TE, Panoulas VF, John H, Douglas KM, Kitas GD (2009) Methotrexate therapy associates with reduced prevalence of the metabolic syndrome in rheumatoid arthritis patients over the age of 60- more than just an anti-inflammatory effect? A cross sectional study. *Arthritis Res Ther* 11: R110. doi: [10.1186/ar2765](https://doi.org/10.1186/ar2765) PMID: [19607680](https://pubmed.ncbi.nlm.nih.gov/19607680/)
34. Korthuis PM, Berger G, Bakker B, Rozenveld-Geugien M, Jaques J, de Haan G, et al. (2014) CITED2-mediated human hematopoietic stem cell maintenance is critical for acute myeloid leukemia. *Leukemia*.
35. Chong PS, Zhou J, Cheong LL, Liu SC, Qian J, Guo T, et al. (2014) LEO1 is regulated by PRL-3 and mediates its oncogenic properties in acute myelogenous leukemia. *Cancer Res* 74: 3043–3053. doi: [10.1158/0008-5472.CAN-13-2321](https://doi.org/10.1158/0008-5472.CAN-13-2321) PMID: [24686170](https://pubmed.ncbi.nlm.nih.gov/24686170/)
36. Shemon AN, Sluyter R, Wiley JS (2007) Rottlerin inhibits P2X(7) receptor-stimulated phospholipase D activity in chronic lymphocytic leukaemia B-lymphocytes. *Immunol Cell Biol* 85: 68–72. PMID: [17130901](https://pubmed.ncbi.nlm.nih.gov/17130901/)
37. Jaskiewicz K, Voigt H, Blakolmer K (1996) Increased matrix proteins, collagen and transforming growth factor are early markers of hepatotoxicity in patients on long-term methotrexate therapy. *J Toxicol Clin Toxicol* 34: 301–305. PMID: [8667468](https://pubmed.ncbi.nlm.nih.gov/8667468/)
38. Dubielecka PM, Jazwiec B, Potoczek S, Wrobel T, Miloszewska J, Haus O, et al. (2005) Changes in spectrin organisation in leukaemic and lymphoid cells upon chemotherapy. *Biochem Pharmacol* 69: 73–85. PMID: [15588716](https://pubmed.ncbi.nlm.nih.gov/15588716/)
39. Heap GA, Trynka G, Jansen RC, Bruinenberg M, Swertz MA, Dinesen LC, et al. (2009) Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med Genomics* 2: 1. doi: [10.1186/1755-8794-2-1](https://doi.org/10.1186/1755-8794-2-1) PMID: [19128478](https://pubmed.ncbi.nlm.nih.gov/19128478/)
40. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, et al. (2007) Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat Med* 13: 361–366. PMID: [17334370](https://pubmed.ncbi.nlm.nih.gov/17334370/)
41. Alaedini A, Green PH (2005) Narrative review: celiac disease: understanding a complex autoimmune disorder. *Ann Intern Med* 142: 289–298. PMID: [15710962](https://pubmed.ncbi.nlm.nih.gov/15710962/)
42. Malandrino N, Capristo E, Farnetti S, Leggio L, Abenavoli L, Addolorato G, et al. (2008) Metabolic and nutritional features in adult celiac patients. *Dig Dis* 26: 128–133. doi: [10.1159/000116770](https://doi.org/10.1159/000116770) PMID: [18431062](https://pubmed.ncbi.nlm.nih.gov/18431062/)
43. Rubio-Tapia A, Murray JA (2010) Celiac disease. *Curr Opin Gastroenterol* 26: 116–122. doi: [10.1097/MOG.0b013e3283365263](https://doi.org/10.1097/MOG.0b013e3283365263) PMID: [20040864](https://pubmed.ncbi.nlm.nih.gov/20040864/)
44. Meresse B, Ripoche J, Heyman M, Cerf-Bensussan N (2009) Celiac disease: from oral tolerance to intestinal inflammation, autoimmunity and lymphomagenesis. *Mucosal Immunol* 2: 8–23. doi: [10.1038/mi.2008.75](https://doi.org/10.1038/mi.2008.75) PMID: [19079330](https://pubmed.ncbi.nlm.nih.gov/19079330/)
45. Bianchi ML (2010) Inflammatory bowel diseases, celiac disease, and bone. *Arch Biochem Biophys* 503: 54–65. doi: [10.1016/j.abb.2010.06.026](https://doi.org/10.1016/j.abb.2010.06.026) PMID: [20599670](https://pubmed.ncbi.nlm.nih.gov/20599670/)
46. De Carolis S, Botta A, Fatigante G, Garofalo S, Ferrazzani S, Gasbarrini A, et al. (2004) Celiac disease and inflammatory bowel disease in pregnancy. *Lupus* 13: 653–658. PMID: [15485096](https://pubmed.ncbi.nlm.nih.gov/15485096/)
47. Linton KJ (2007) Structure and function of ABC transporters. *Physiology (Bethesda)* 22: 122–130. PMID: [17420303](https://pubmed.ncbi.nlm.nih.gov/17420303/)
48. Teodori E, Dei S, Martelli C, Scapecchi S, Gualtieri F (2006) The functions and structure of ABC transporters: implications for the design of new inhibitors of Pgp and MRP1 to control multidrug resistance (MDR). *Curr Drug Targets* 7: 893–909. PMID: [16842220](https://pubmed.ncbi.nlm.nih.gov/16842220/)
49. Tighe MR, Hall MA, Cardi E, Ashkenazi A, Lanchbury JS, Ciclitira PJ (1994) Associations between alleles of the major histocompatibility complex-encoded ABC transporter gene TAP2, HLA class II alleles, and celiac disease susceptibility. *Hum Immunol* 39: 9–16. PMID: [8181966](https://pubmed.ncbi.nlm.nih.gov/8181966/)
50. Caughey RH, Mc CW, Kaye R (1955) Nitrogen balances in a patient with fibrocystic disease of the pancreas and a patient with the celiac syndrome and the effect of aureomycin. *Pediatrics* 16: 174–183. PMID: [13245321](https://pubmed.ncbi.nlm.nih.gov/13245321/)
51. Maeda S, Omata T (2009) Nitrite transport activity of the ABC-type cyanate transporter of the cyanobacterium *Synechococcus elongatus*. *J Bacteriol* 191: 3265–3272. doi: [10.1128/JB.00013-09](https://doi.org/10.1128/JB.00013-09) PMID: [19286804](https://pubmed.ncbi.nlm.nih.gov/19286804/)

52. Shibuya M (2001) Structure and function of VEGF/VEGF-receptor system involved in angiogenesis. *Cell Struct Funct* 26: 25–35. PMID: [11345501](#)
53. Myrsky E, Syrjanen M, Korponay-Szabo IR, Maki M, Kaukinen K, Lindfors K (2009) Altered small-bowel mucosal vascular network in untreated coeliac disease. *Scand J Gastroenterol* 44: 162–167. doi: [10.1080/003665520802400875](#) PMID: [18985542](#)
54. Chen W, Tang T, Eastham-Anderson J, Dunlap D, Alicke B, Nannini M, et al. (2011) Canonical hedgehog signaling augments tumor angiogenesis by induction of VEGF-A in stromal perivascular cells. *Proc Natl Acad Sci U S A* 108: 9589–9594. doi: [10.1073/pnas.1017945108](#) PMID: [21597001](#)
55. Koyama E, Young B, Nagayama M, Shibukawa Y, Enomoto-Iwamoto M, Iwamoto M, et al. (2007) Conditional Kif3a ablation causes abnormal hedgehog signaling topography, growth plate dysfunction, and excessive bone and cartilage formation during mouse skeletogenesis. *Development* 134: 2159–2169. PMID: [17507416](#)
56. Soma T, Kaganoi J, Kawabe A, Kondo K, Tsunoda S, Imamura M, et al. (2006) Chenodeoxycholic acid stimulates the progression of human esophageal cancer cells: A possible mechanism of angiogenesis in patients with esophageal cancer. *Int J Cancer* 119: 771–782. PMID: [16557574](#)
57. Banerjee DK (2007) Requirement of protein kinase type I for cAMP-mediated up-regulation of lipid-linked oligosaccharide for asparagine-linked protein glycosylation. *Cell Mol Biol (Noisy-le-grand)* 53: 55–63. PMID: [17531150](#)
58. Lee JM, Lee WH, Kay HY, Kim ES, Moon A, Kim SG (2011) Hemin, an iron-binding porphyrin, inhibits HIF-1 α induction through its binding with heat shock protein 90. *Int J Cancer*.
59. Aviezer D, Cotton S, David M, Segev A, Khaselev N, Galili N, et al. (2000) Porphyrin analogues as novel antagonists of fibroblast growth factor and vascular endothelial growth factor receptor binding that inhibit endothelial cell proliferation, tumor progression, and metastasis. *Cancer Res* 60: 2973–2980. PMID: [10850445](#)
60. Moss SF, Attia L, Scholes JV, Walters JR, Holt PR (1996) Increased small intestinal apoptosis in coeliac disease. *Gut* 39: 811–817. PMID: [9038662](#)
61. Giovannini C, Sanchez M, Straface E, Scazzocchio B, Silano M, De Vincenzi M (2000) Induction of apoptosis in caco-2 cells by wheat gliadin peptides. *Toxicology* 145: 63–71. PMID: [10771132](#)
62. Kurgan J (1969) [Basal metabolism in patients with lung cancer and the effect of corticosteroid therapy]. *Gruzlica* 37: 297–302. PMID: [5788092](#)
63. Kurgan J (1970) [Basal metabolism in patients with lung cancer and its modification by corticotherapy]. *Z Erkr Atmungsorgane Folia Bronchol* 132: 181–186. PMID: [5203359](#)
64. Tokovoi VA, Matytsin AN (1967) [Basal metabolism in patients with lung cancer]. *Vopr Onkol* 13: 74–77. PMID: [5601552](#)
65. Giovacchini G, Picchio M, Schipani S, Landoni C, Gianolli L, Bettinardi V, et al. (2009) Changes in glucose metabolism during and after radiotherapy in non-small cell lung cancer. *Tumori* 95: 177–184. PMID: [19579863](#)
66. Heber D, Chlebowski RT, Ishibashi DE, Herrold JN, Block JB (1982) Abnormalities in glucose and protein metabolism in noncachectic lung cancer patients. *Cancer Res* 42: 4815–4819. PMID: [7127317](#)
67. Koukourakis MI, Giatromanolaki A, Bougioukas G, Sivridis E (2007) Lung cancer: a comparative study of metabolism related protein expression in cancer cells and tumor associated stroma. *Cancer Biol Ther* 6: 1476–1479. PMID: [17881895](#)
68. Gabazza EC, Taguchi O, Yoshida M, Yamakami T, Kobayashi H, Ibata H, et al. (1995) Neutrophil activation and collagen metabolism in lung cancer. *Clin Chim Acta* 236: 101–108. PMID: [7664461](#)
69. Dessi S, Batetta B, Pulisci D, Spano O, Cherchi R, Lanfranco G, et al. (1992) Altered pattern of lipid metabolism in patients with lung cancer. *Oncology* 49: 436–441. PMID: [1465281](#)
70. Kiyohara C, Shirakawa T, Hopkin JM (2002) Genetic polymorphism of enzymes involved in xenobiotic metabolism and the risk of lung cancer. *Environ Health Prev Med* 7: 47–59. doi: [10.1007/BF02897330](#) PMID: [21432264](#)
71. Senthilnathan P, Padmavathi R, Magesh V, Sakthisekaran D (2006) Modulation of TCA cycle enzymes and electron transport chain systems in experimental lung cancer. *Life Sci* 78: 1010–1014. PMID: [16143346](#)
72. Sulowicz W (1983) Phagocytosis and peroxidase activity in neutrophils from peripheral blood of patients with malignant tumours of lung, stomach and large intestine. *Folia Haematol Int Mag Klin Morphol Blutforsch* 110: 48–54. PMID: [6192054](#)
73. Pouniotis DS, Plebanski M, Apostolopoulos V, McDonald CF (2006) Alveolar macrophage function is altered in patients with lung cancer. *Clin Exp Immunol* 143: 363–372. PMID: [16412062](#)

74. Hosker HS, Corris PA (1991) Alveolar macrophage and blood monocyte function in lung cancer. *Cancer Detect Prev* 15: 103–106. PMID: [2032250](#)
75. Klucinski W, Degorski A, Miernik-Degorska E, Targowski S, Winnicka A (1988) Effect of ketone bodies on the phagocytic activity of bovine milk macrophages and polymorphonuclear leukocytes. *Zentralbl Veterinarmed A* 35: 632–639. PMID: [3142190](#)
76. Puliappadamba VT, Cheriyan VT, Thulasidasan AK, Bava SV, Vinod BS, Prabhu PR, et al. (2010) Nicotine-induced survival signaling in lung cancer cells is dependent on their p53 status while its down-regulation by curcumin is independent. *Mol Cancer* 9: 220. doi: [10.1186/1476-4598-9-220](#) PMID: [20727180](#)
77. Derby KS, Cuthrell K, Caberto C, Carmella SG, Franke AA, Hecht SS, et al. (2008) Nicotine metabolism in three ethnic/racial groups with different risks of lung cancer. *Cancer Epidemiol Biomarkers Prev* 17: 3526–3535. doi: [10.1158/1055-9965.EPI-08-0424](#) PMID: [19029401](#)
78. Murray RP, Connett JE, Zapawa LM (2009) Does nicotine replacement therapy cause cancer? Evidence from the Lung Health Study. *Nicotine Tob Res* 11: 1076–1082. doi: [10.1093/ntr/ntp104](#) PMID: [19571249](#)
79. Martinet Y, Debry G (1992) [Effects of coffee on the respiratory system]. *Rev Mal Respir* 9: 587–592. PMID: [1470751](#)
80. Baker JA, McCann SE, Reid ME, Nowell S, Beehler GP, Moysich KB (2005) Associations between black tea and coffee consumption and risk of lung cancer among current and former smokers. *Nutr Cancer* 52: 15–21. PMID: [16090999](#)
81. Al-Wadei HA, Takahashi T, Schuller HM (2006) Caffeine stimulates the proliferation of human lung adenocarcinoma cells and small airway epithelial cells via activation of PKA, CREB and ERK1/2. *Oncol Rep* 15: 431–435. PMID: [16391865](#)
82. Zevin S, Benowitz NL (1999) Drug interactions with tobacco smoking. An update. *Clin Pharmacokinet* 36: 425–438. PMID: [10427467](#)
83. de Leon J, Diaz FJ, Rogers T, Browne D, Dinsmore L, Ghosheh OH, et al. (2003) A pilot study of plasma caffeine concentrations in a US sample of smoker and nonsmoker volunteers. *Prog Neuropsychopharmacol Biol Psychiatry* 27: 165–171. PMID: [12551740](#)
84. Carrillo JA, Benitez J (1996) CYP1A2 activity, gender and smoking, as variables influencing the toxicity of caffeine. *Br J Clin Pharmacol* 41: 605–608. PMID: [8799528](#)