Published in final edited form as:

IEEE Trans Med Imaging. 2018 April ; 37(4): 860–870. doi:10.1109/TMI.2017.2783244.

Fast and Accurate Detection of Complex Imaging Genetics Associations Based on Greedy Projected Distance Correlation

Jian Fang,

Department of Biomedical Engineering, Tulane University, New Orleans, LA, 70118

Chao Xu,

Department of Biostatistics and Bioinformatics, Tulane University, New Orleans, LA, 70112

Pascal Zille,

Department of Biomedical Engineering, Tulane University, New Orleans, LA, 70118

Dongdong Lin,

The Mind Research Network, Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM, 87131

Hong-Wen Deng,

Department of Biostatistics and Bioinformatics, Tulane University, New Orleans, LA, 70112

Vince D. Calhoun [Fellow, IEEE], and

The Mind Research Network, Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM, 87131

Yu-Ping Wang [Senior Member, IEEE]

Department of Biomedical Engineering, Tulane University, New Orleans, LA, 70118

Abstract

Recent advances in imaging genetics produce large amounts of data including functional MRI images, single nucleotide polymorphisms (SNPs), and cognitive assessments. Understanding the complex interactions among these heterogeneous and complementary data has the potential to help with diagnosis and prevention of mental disorders. However, limited efforts have been made due to the high dimensionality, group structure, and mixed type of these data. In this paper we present a novel method to detect conditional associations between imaging genetics data. We use projected distance correlation to build a conditional dependency graph among high-dimensional mixed data, then use multiple testing to detect significant group level associations (e.g., ROI-gene). In addition, we introduce a scalable algorithm based on orthogonal greedy algorithm, yielding the greedy projected distance correlation (G-PDC). This can reduce the computational cost, which is critical for analyzing large-volume of imaging genomics data. The results from our simulations demonstrate a higher degree of accuracy with G-PDC than distance correlation, Pearson's correlation and partial correlation, especially when the correlation is nonlinear. Finally, we apply our method to the Philadelphia Neurodevelopmental data cohort with 866 samples including fMRI images and SNP profiles. The results uncover several statistically significant and biologically

interesting interactions, which are further validated with many existing studies. The Matlab code is available at <https://sites.google.com/site/jianfang86/gPDC>.

Index Terms

Imaging genetics; fMRI; SNP; computerized neural-cognitive battery; distance correlation; projected distance correlation; orthogonal greedy algorithm

I. Introduction

With the rapid development of techniques to detect the structural, functional, and genetic factors of brain disorders, the integration of neuroimaging and genetic biomarkers has received increasing attention [1] [2] for the comprehensive and systematic study of mental illnesses. With many applications, imaging genetics aims to discover genetic variants that explain brain activities, in order to better understand neurogenetic mechanisms for the treatment and prevention of complex psychiatric diseases. As the collection of larger datasets becomes increasingly frequent, statistical methods to detect imaging genetics associations with greater accuracy while reducing computational cost are in high demand [3].

Mass-univariate linear modeling (MULM) is a conventional method to detect potential imaging genetics relationships. It does so by first calculating all possible univariate associations and then determines the important connections by multiple testing. However, the massive number of tests leads to limited power. To overcome this problem, several multivariate approaches, such as canonical correlation analysis (CCA) [4], partial least squares (PLS) [5] and independent component analysis (ICA) [6] [7], were developed. The basic idea is to maximize the correlation between linear combinations of variables from different data types to find interrelated patterns. To further avoid overfitting for high-dimensional cases, methods based on sparse penalties were introduced to select a small number of features [8], [9], e.g., sparse CCA [10], [11], sparse PLS [12] and sparse reduced rank regression [13]. The penalized multivariate methods have the advantage of detecting complicated associations by estimating all the variables together. However, as compared to the MULM, the testing statistics are much more difficult to obtain.

Further, the complex structures of imaging genetics data, as shown in Fig. 1, have not been fully accounted in these methods. Firstly, the data are multi-modal and mixed types. Imaging genetics datasets always contain, but are not limited to neuroimaging data (e.g., task fMRI) and genetic profiles (e.g., SNPs). The variables consist of different data types, which are not necessarily Gaussian, hence hinder the direct applications of most conventional statistical methods. Secondly, the data are group structured [14]. For example, in brain network studies, the fMRI voxels can be clustered to regions of interest (ROI) [15]. Similar cases exist for SNPs, which can be grouped into genes. Usually, a group level analysis can utilize local structures to infer complex associations, and facilitate interpretations, hence is more advantageous. Thirdly, the data are high dimensional. Neuroimaging data can have tens of thousands of voxels while the number of SNPs can even reach millions, which greatly increase the risk of overfitting, the difficulty in statistical testing, and the computational cost.

Finally, complex interactions can exist. Widely studied brain and gene networks [16] [17] indicate that strong and dense interactions exist within each modality. In contrast, several researches found that the interactions between fMRI and gene are weak [9]. As a result, the intra-modal interactions implicate many indirect interactions, which make the detection of imaging genetic associations even more challenging [18] [19].

In this paper, we propose a novel method to detect associations among multi-modal data based on greedy projected distance correlation (G-PDC). Distance correlation [20]–[22] was recently developed to measure the dependence of two random vectors. It was shown that distance correlation of two random vectors equals to zero if and only if these two random vectors are independent. Projected distance correlation [23] is an extension of distance correlation, which is able to capture conditional dependence for high dimensional data. These remarkable properties of distance correlation and projected distance correlation motivate us to use the projected distance correlation to detect complex imaging genetics associations. The method has the potential to detect more complex relationships and reduce the detections of indirect interactions. Meanwhile, multivariate and nonlinear dependence are detected straightforwardly on the original high-dimensional data without dimension reduction, e.g., averaging.

As shown in Fig. 1, when applying to imaging genetic data, the main step is to perform a group level conditional independence test to detect significant edges, where each node corresponds to a group of voxels (ROI) or SNPs (gene). We adopt the t-test procedure proposed in [21] to calculate the p -value, and used the Benjamini-Hochberg procedure [24] to control the FDR. However, the voxel dependency of fMRI data can cause the failure of FDR control. To overcome this problem, an effective decorrelation method was introduced without additional computational cost.

In addition, when computing projected distance correlation we apply orthogonal matching pursuit with high dimensional information criterion [25], instead of penalized least square. This greedy projected distance correlation can improve the algorithm efficiency, yet without sacrificing theoretical guarantees. For validation, we first apply the proposed method to a dataset containing real fMRI images and SNP profiles, but simulated across-modal correlations. Through a comprehensive comparison, we demonstrate the higher power and accuracy of the proposed method, as compared to classical univariate linear methods including Pearson's correlation and partial correlation, multivariate linear method CCA, and other distance correlation based methods. Furthermore, the efficiency is also improved over conventional projected distance correlation. Finally, we apply the method to a neurodevelopmental dataset with 866 subjects, with ages of 8–22 years. The data are multi-modal including fMRI and SNPs data, collected by the Philadelphia Neurodevelopmental Cohort [26]. We found a number of statistically significant interactions among the two types of data.

The rest of the paper is organized as follows. Section II introduces the graph construction and edge detection method. The performance of the proposed method is evaluated through both simulations and real data analysis in Section III, followed by some discussions and concluding remarks in the last section.

II. Material and Methods

Suppose we have p_f voxels from imaging data $F \in R^{n \times p_f}$, and p_g SNPs from genetic data $G \in R^{n \times p_g}$, where n is the sample size. Let's assume the voxels can be grouped into C_f regions of interest (ROIs) and the SNPs can be grouped into C_g genes, say, $F = [F_1, F_2, \dots, F_{C_f}]$, $G = [G_1, G_2, \dots, G_{C_g}]$. The goal of this paper is to test the conditional independence of all ROI and gene pairs. For example, given the i -th ROI and the j -th gene, we test the independence of F_i and G_j by controlling all the other SNPs and voxels,

$$F_i \perp\!\!\!\perp G_j \mid \{F \cup G\} \setminus \{F_i \cup G_j\} \quad (1)$$

Then based on the testing statistics, we can construct a bipartite graph. Each node is a gene or ROI and each edge corresponds to a significant conditional association. Specifically, Eqn. (1) is different from conventional univariate tests, e.g., those based on partial correlation, that both F_i and G_j have multiple features. In the rest of this section, we will describe how we can obtain the testing statistics and the graph both effectively and efficiently.

A. Distance Correlation

Distance correlation is a measure that characterizes statistical dependence between two random vectors [20]. More specifically, given two random vectors $x \in \mathcal{R}^p$ and $y \in \mathcal{R}^q$ with characteristic functions ϕ_x and ϕ_y , we can define the distance covariance by the non-negative square root of

$$V^2(x, y) = \int_{\mathcal{R}^{p+q}} |\phi_{x,y}(t, s) - \phi_x(t)\phi_y(s)|^2 w(t, s) dt ds \quad (2)$$

where $w(t, s) = (c_p c_q |t|_p^1 + |s|_q^1 + q)^{-1}$, $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$ and $\phi_{x,y}$ is the joint characteristic function. We can further define the standardized measure, i.e., the distance correlation as the non-negative square root of

$$R^2(x, y) = \begin{cases} \frac{V^2(x, y)}{\sqrt{V^2(x, x)V^2(y, y)}}, & V^2(x, x)V^2(y, y) > 0 \\ 0, & V^2(x, x)V^2(y, y) = 0 \end{cases} \quad (3)$$

A key feature of distance correlation is that it is an equivalent measure of statistical independence, that is, $R^2(x, y) = 0$ if and only if x and y are independent. In contrast, the conventional Pearson's correlation coefficient only measures linear dependency. It is possible that the Pearson's correlation of two random variables is zero but they are nonlinearly dependent. Therefore, the distance correlation can capture more complex relationships that Pearson's correlation cannot.

Given n paired samples $\{x_i\}_{i=1}^n, \{y_j\}_{j=1}^n$, we follow the steps described in [21] to derive the unbiased sample distance correlation:

1. Compute the Euclidean distance matrices A, B where
 $a_{i,j} = \|x_i - x_j\|, b_{i,j} = \|y_i - y_j\|$.
2. Get U-centered distance matrices $\hat{A}$ with

$$\hat{a}_{i,j} = \begin{cases} a_{i,j} - \frac{\sum_{l=1}^n a_{i,l} + \sum_{k=1}^n a_{k,j}}{n-2} + \frac{\sum_{k,l=1}^n a_{k,l}}{(n-1)(n-2)} & i \neq j \\ 0 & i = j \end{cases} \quad (4)$$

and $\hat{B}$ accordingly.

3. Define sample squared distance covariance

$$V_n^2(x, y) = \frac{1}{n(n-3)} \sum_{i \neq j} \hat{a}_{i,j} \hat{b}_{i,j} \quad (5)$$

4. Calculate sample squared distance correlation by Eqn. (3).

To test whether the distance correlation R_n^2 is equal to zero, we can use the test statistics proposed in [21]. This test was based on the student's t -test and designed for high dimensional problems. Since every ROI and gene contains a number of voxels and SNPs, the test statistics is suitable for our problem. In particular, Theorem 1 of [21] proved that under the null hypothesis that x and y are independent and $p, q \rightarrow \infty$, it holds

$$P(T = \sqrt{v-1} \frac{R_n^2}{\sqrt{1 - (R_n^2)^2}} < t) \rightarrow P(t_{v-1} < t) \quad (6)$$

where t_{v-1} denotes for student t distribution with $v-1$ degrees of freedom, $v = \frac{n(n-3)}{2}$.

Therefore, with predetermined significance level α , we can get the p -value

$$p_t = 1 - t_{v-1}(T) \quad (7)$$

and reject the null hypothesis when $p_t < \alpha$.

Via these procedures, the distance correlation provides a concise and rigorous way to test statistical independence. Note that the test can be applied generally to two random vectors with different dimensions. This exclusive feature helps us to deal with the group structure difficulty (e.g., multiple SNPs in a gene, multiple voxels in a ROI) for the proposed method.

B. Greedy Projected Distance Correlation

Projected distance correlation is a conditional dependence measure based on distance correlation [23]. It aims at testing whether x and y are independent given a controlling random vector $z \in \mathcal{R}^r$:

$$x \perp\!\!\!\perp y | z \quad (8)$$

In general, it assumes that the conditional dependency of x and y on z is expressed in a linear form of z :

$$x = B_x z + \varepsilon_x; y = B_y z + \varepsilon_y; \quad (9)$$

where $B_x \in \mathcal{R}^{p \times r}$, $B_y \in \mathcal{R}^{q \times r}$ are coefficient matrices. It is worthy to note that this linear projection assumption does not always hold. In the case of (1), z is a high dimensional vector, since it consists of all the variables except those in x and y . Although the true dependence between z and x, y may not be linear, given limited samples, x and y can usually be linearly expressed by z . This is very similar to the kernel method when applying to nonlinear regression. In fact, the main idea of kernel regression is to transform the features to a very high dimensional space to make the response (e.g. x and y) linearly representable [27] [28]. Hence, we think the assumption (9) is acceptable to some extent. Potential extensions to nonlinear projections will be discussed in the Conclusion.

Under the linear projection assumption, it is easy to see that the test of (8) is equivalent to the test of $\varepsilon_x \perp\!\!\!\perp \varepsilon_y$. Motivated by this observation, the projected distance correlation generally contains the following two steps:

1. Estimate the residues $\hat{\varepsilon}_x, \hat{\varepsilon}_y$ through linear regression, e.g., ordinary least square in low-dimensional case and penalized sparse regression like LASSO in high-dimensional case.
2. Test the independence between $\hat{\varepsilon}_x$ and $\hat{\varepsilon}_y$ using distance correlation described in the previous section.

In imaging genetics studies, both fMRI and SNP are in high-dimensions. Although we can adopt penalized sparse regression suggested in [23], the parameter tuning, especially for large scale problems, remains a challenging task. Here, we will instead use orthogonal greedy algorithm with high-dimensional information criterion (HDIC) to estimate the residues. The orthogonal greedy algorithm (OGA) has been developed as a computationally efficient alternative for sparse learning [25]. It starts from a null model and fits the model based on a series of expanding subspaces, hence can avoid the computation in the original high-dimensional linear space. In particular, considering a linear regression problem

$$x_n = z_n \beta + \varepsilon \quad (10)$$

where $x_n \in R^{n \times I}$, $z_n \in R^{n \times r}$, the algorithm starts from the active set $T_0 = \emptyset$. Given $T_{k-1} \in \{1, 2, \dots, r\}$ and the estimated residue $r_{k-1} = x_n - z_n \beta_{k-1}$, it updates the active set

$T_k = T_{k-1} \cup z_n^{t_k}$ where

$$t_k = \operatorname{argmax}_{j \in \{1, \dots, q\}} \left| \langle r_{k-1}, z_n^j \rangle \right|$$

, and then solves a least square problem restricting on T_k ,

$$\beta_k = \operatorname{argmin}_{\beta^i = 0 \text{ } i \notin T_k} \|x_n - z_n \beta\|_2^2$$

It is easy to see that the new estimate $f_k = z_n \beta_k$ is the orthogonal projection of x_n onto the space of T_k , hence the residue r_k is orthogonal to T_k . The algorithm finally achieves a path of sparse solutions when the iteration terminates. To select an appropriate model along the OGA path, we adopt the HDIC [25], which is a consistent model selection procedure for high-dimensional problems. This can be achieved by finding

$$k^* = \operatorname{argmin}_k n \log \left(\frac{\|r_k\|_2^2}{n} \right) + 2k \log n \log r$$

We summarize the algorithm in Algorithm 1. It is worthy noting that in Step 5 of Algorithm 1, the projection can be updated efficiently in an iterative way (see [29] for details). As compared with penalized least square, the OGA with HDIC is able to select the tuning parameter during a single run of the algorithm, hence can greatly reduce the computational time. To promote the application of the proposed method, we make the Matlab code available at <https://sites.google.com/site/jianfang86/gPDC>

Algorithm 1

Algorithm for projected distance correlation with OGA

Require: Observed random sample vectors $x_n \in R^{n \times p}$, $y_n \in R^{n \times q}$, $z_n \in R^{n \times r}$.

1: **for** $i = 1$ to p **do**

2:

Initialize $r_0 = x_n^i$, $T_0 = \emptyset$.

3:

for $k = 1$ to $\left(1 + \log q\right) \sqrt{\frac{n}{\log q}}$ **do**

4:

Update the active set by $T_k = T_{k-1} \cup z_n^{t_k}$, where $t_k = \operatorname{argmax}_{j \in \{1, \dots, q\}} \left| \langle r_{k-1}, z_n^j \rangle \right|$.

5:

Update the estimate by $f_k = P_{T_k} x_n^i$, where P_{T_k} is the orthogonal projection onto T_k .

```

6:      Update the residue  $r_k = x_n^i - f_k$ .
7:
      Record  $\text{HDIC}_k = n \log\left(\frac{\|r_k\|_2^2}{n}\right) + 2k \log r$ .
8:   end for
9:   Find  $k^* = \text{argmin}_k \text{HDIC}_k$ .
10:  Select  $\varepsilon_x^i = r_{k^*}$ .
11: end for
12:  Calculate squared sample projected distance correlation  $R_n^2(\varepsilon_x, \varepsilon_y)$  and the corresponding  $p$ -value.

```

We now provide some theoretical justifications on using Algorithm 1. Based on the results in [23], [25], we can get the following result.

Theorem 1—If $E\|e_x\| < \infty$, $E\|e_y\| < \infty$, $E\|e_x\|^2 < \infty$, $E\|e_y\|^2 < \infty$, there exist constants $C_1 > 1$, $r > 0$ and $C_2 > 1$ such that

$$\|(B_x - \hat{B}_x)z_n\|_{2,\infty} \leq C_1 a_n, \|(B_y - \hat{B}_y)z_n\|_{2,\infty} \leq C_2 a_n \quad (11)$$

where the sequence $a_n = o\{(n^{1+r} \log n)^{-1/3}\}$, then the following holds

$$V_n^2(\hat{\varepsilon}_x, \hat{\varepsilon}_y) \rightarrow V_n^2(\varepsilon_x, \varepsilon_y) \quad (12)$$

The proof is given in the supplementary document. Supplementary materials are available in the supplementary files/multimedia tab. Theorem 1 shows that the sample distance covariance between the estimated residual vectors converges to the sample distance covariance between the true error vectors, which is an unbiased estimate of the population error vectors [21]. It enables us to use the distance covariance of the estimated residual vectors to construct the conditional independence test as described in the previous section. In addition, from Theorem 1 of [30], when B_x and B_y are truly sparse, we can easily check that (11) holds for OGA estimator in high probability. More details are included in the supplementary material.

C. Construct the graph based on G-PDC

Based on the greedy projected distance correlation described above, we can apply Algorithm 1 to (1) and get $C_g p$ -values. To deal with the problem of multiple testing, we control the proportion of false positives by calculating the false discovery rate (FDR) using the Benjamini-Hochberg procedure [24]. Finally, we can obtain a set of significantly correlated pairs of ROIs and genes. However, as shown in Fig. 2, the naive implementation of G-PDC failed to control type 1 error rate in real imaging genetic data. More specifically, in Fig. 2(a),

the quantile-quantile (QQ) plot under null distribution shows that the false positive rate(FPR) was highly inflated. Fig. 3 displays the QQ plots for G-PDC when applying to independent multivariate normal vectors. We can see that the FPR was controlled well when the dimension is high enough. In contrast, inflated FPR was observed in Fig. 3(a), since the dimension is too low that the null distribution no longer follows student t distribution in (6). Although the dimensionality of imaging genetic data is high enough, Fig. 2(a) still exhibits similar behavior as Fig. 3(a). One important feature of the distance correlation is that it is invariant under shift and orthogonal transformations [20], say, $R^2(x, y) = R^2(\Phi_x x + c_x, \Phi_y y + c_y)$, where Φ_x, Φ_y are any orthogonal transformations. More specifically, suppose Φ_x is the transformation derived by PCA, and if there are some dominant principal components in x , the intrinsic dimension is then very small, which can explain the failure of FPR control. In fact, this is true for the fMRI data. The adjacent voxels in a ROI are highly correlated, and the resulting redundancy leads to some dominant features. The intra-modal correlation may not be relevant to genes, but can yield a high FPR and a high FDR. To overcome this problem, we slightly modified Algorithm 1 to reduce the correlation in each ROI. More specifically, motivated by the works in [31] [32], we assume that a voxel x_j in the i th ROI follows an approximate model:

$$x_j = h_j + u_j \quad (13)$$

We expect h_j to be a highly correlated component, and u_j to be weakly correlated across voxels. It is reasonable to assume that the local correlation contributes most to the correlation among voxels; hence, we suppose that h_j is a sparse linear combination of nearby voxels:

$$h_j = x_{(F_i \setminus j)} \gamma \quad (14)$$

where F_i is the set of voxels in the i -th ROI as in (1). Then (13) becomes

$$x_j = x_{(F_i \setminus j)} \gamma + u_j \quad (15)$$

By solving the above problem with sparse constraint on γ , we can obtain an estimate of $\hat{\gamma}$ and the residue $\hat{u}_j = x_j - x_{(F_i \setminus j)} \hat{\gamma}$, which is expected to be weakly correlated. Then, to avoid

an additional model selection for this preprocessing step, we combine the implementation with (9). That is to say, instead of regressing x_j sparsely over $F_i \setminus j$ and then over z (see (1) and (8) for definition), we perform the regression once over $F_i \setminus (j \cup z)$. In practice, we found the difference between the two configurations is ignorable. And in this way, we can simultaneously obtain the residue and decorrelate the data without additional computations. As shown in Fig. 2 (b), this procedure can significantly reduce the FPR, providing a reliable testing statistic. We also include the comparison with another popular decorrelation method in the supplementary material.

Through the procedures introduced above, we construct a graph that can overcome the four difficulties mentioned at the beginning of this paper. First, the projected distance correlation is not limited to Gaussian variables and linear dependence. The general idea of (1) is to test the independence of two vectors by controlling the complementary set of variables, which can be straightforwardly extended to three datasets or more. Second, the use of greedy projected distance correlation is not limited by the number of voxels or SNPs, leading to a consistent way to compute multivariate conditional dependence. Third, by clustering the variables into groups, the number of tests can be reduced from $p_p p_g$ to $C_l C_g$, leading to a moderate scale of multiple testing. Finally, by adopting conditional dependence measures, the method is able to distinguish direct associations from indirect ones, which can potentially increase the accuracy of the detection.

III. Results

In this section, we evaluate the proposed method using both simulated data and a real imaging genetics study.

A. Simulation

In a series of simulations, we aim at evaluating the potential power of the proposed method. In particular, we compare the graphs constructed by different dependence measures with respect to the power, false discovery rate control, and computational cost in both linear and nonlinear cases.

We simulated a paired dataset containing fMRI images and SNPs to match with a realistic data structure. To this end, we carefully select a subset of fMRI images and another subset of SNP profiles from the real data. There are no overlaps between the two subsets so as to minimize data dependencies. (Further details regarding real data will be introduced in Sec III.B). Then, a pre-defined relationship was added between ROIs and genes to make them correlated.

More specifically, the fMRI images were generated by randomly selecting 50 ROIs with voxel size between 100 and 200 (The mean and median voxel sizes of all the ROIs are 123.35 and 116, respectively). The SNP data were generated by directly collecting the chromosome 1 data from random subjects that have no fMRI scans. This leads to 177 genes and 4428 SNPs.

To simulate correlated patterns between ROIs and genes, we first randomly selected 100 pairs of ROIs and genes. For each pair, we randomly activated 10 voxels and 10 SNPs. Then for each activated voxel f , we added a correlation with the activated SNPs G in two cases: 1) The correlation is linear. To this end, we draw a multivariate normal vector $\beta \sim \mathcal{N}(0, I_{10})$ and let $\tilde{f} = f + 0.1G\beta$. 2) The correlation is non-linear. To do this, we similarly draw a multivariate uniform vector $\beta \sim \mathcal{U}^{10}(0, 1)$ and let $\tilde{f} = f + \sin(\pi G\beta)$. In this way, we can get a sparse and weak relationship between ROIs and genes in both linear and nonlinear cases, and keep the intra-modal interactions.

We used the true positive rate (TPR) or power, false positive rate (FPR), false discovery rate (FDR), and computational time to evaluate the performance of the model. 7 dependence measures were compared, including four multivariate nonlinear methods, two univariate linear methods, and one multivariate linear method shown in Table I. They are briefly described as follows: 1) Distance correlation [20], 2) Partial distance correlation [22], 3) Projected distance correlation with LASSO [23]. The LASSO is solved via Matlab function "lasso" and the regularization parameter is tuned by HDIC. 4) Greedy Projected distance correlation proposed in this paper, 5) Pearson's correlation on each pair of voxel and SNP. 6) Partial correlation on each pair of voxel and SNP, which was calculated using the same procedures as in [33]. The significance of a ROI-gene interaction of the two univariate methods was defined by whether there was a significant voxel-SNP pair within the ROI and gene. 7) Canonical correlation analysis on each pair of ROI and gene. We used the Matlab function 'canoncorr' and got the p -value under the null hypothesis that all the canonical correlations are zero.

In particular, the partial distance correlation is an alternative approach to detect conditional dependency between x and y over z , which is defined as

$$pR^2(x, y) = \begin{cases} 0 & \text{if } (R^2(x, z) - 1)(R^2(y, z) - 1) = 0 \\ \frac{R^2(x, y) - R^2(x, z)R^2(y, z)}{\sqrt{(1 - R^4(x, z))(1 - R^4(y, z))}} & \text{otherwise} \end{cases} \quad (16)$$

Different from the projected distance correlation, the partial distance correlation assumes that the dependence of x and y is expressed linearly over z in the Hilbert space generated by U-centered distance matrices [22]. The performance of partial distance correlation when z is in high-dimension is not well studied, which motivated us to compare it with projected distance correlation. In each simulation, the statistics were averaged over 100 replications.

In all the following simulations, we generated the data with sample size from 100 to 800. Fig. 4 displays the TPR against FPR in the linear case. The curves were drawn by varying the significance threshold. As can be seen in the figure, the projected distance correlation performs much better than the distance correlation and partial distance correlation. There is no obvious improvement of partial distance correlation over distance correlation, which indicates that the partial distance correlation is not suitable for high-dimensional data. CCA performs well given sufficient samples (e.g. 500). When the sample size is small, it fails to control the FPR. This is mainly because CCA is unavailable to get p -value when the correlation matrix is singular. The projected distance correlation is also more accurate than the two univariate linear methods, Pearson's correlation and partial correlation. In addition, the great improvement of partial correlation over Pearson's correlation further demonstrates the necessity of conditional dependency measure in imaging genetic studies.

Fig. 5 displays the TPR against FPR in nonlinear cases. In this case, the projected distance correlation performs the best among all the 7 methods, and much better than CCA. This

demonstrates the exclusive advantage of distance correlation in detecting nonlinear associations.

Fig. 4 and 5 also show no significant difference between OGA and LASSO when computing the projected distance correlation. But from the comparison of the computational time in Fig. 6, we can see that the OGA is much faster, demonstrating its efficiency. This reduction of computational time is highly desirable, because large volume of multiscale data are ubiquitous in imaging genomics studies.

We further compare the performance of the competitive methods in combination with FDR control. Fig. 7 displays the false discovery rate versus the sample size when we applied the Benjamini-Hochberg procedure with two different target FDR rates, 0.05 and 0.01. All the methods reach a higher FDR than the target threshold. Nevertheless, with target FDR rate 0.01, the G-PDC with decorrelation, Pearson's correlation and partial correlation perform quite well, and the true FDR is around 0.1, which is reasonable in many problems [34]. However, the FDR reaches 0.2 given the target FDR 0.05. So we will use a target FDR rate 0.01 in our real data analysis. The distance correlation and CCA have a very high FDR, which may hinder the application to real data. We also compared the power in Fig. 8. The projected distance correlation usually obtains the highest power except CCA. Considering the failed FDR control of CCA, the projected distance correlation appears to be a better choice. Moreover, in the case of nonlinear correlation, the power of CCA degrades as the sample size increases. This may be due to the fact that the SNP data are sparse category data, and the nonlinearity can be easily fitted by linear functions in low sample size cases. Nevertheless, as the sample size increases, the nonlinearity of the interactions becomes stronger, posing difficulties for linear methods to capture.

All these simulations results demonstrate that the projected distance correlation can achieve comparable accuracy with conventional linear methods to detect linear interactions, but works much better for nonlinear ones. Meanwhile, the FDR is well controlled, which makes it a reliable method to detect complex imaging genetic associations.

B. Real data results

The data were acquired from the Philadelphia Neurodevelopmental Cohort [26] (PNC), which is a collaboration between the Brain Behaviour Laboratory at the University of Pennsylvania and the Children's Hospital of Philadelphia. It is available in the dbGaP database [35], and contains fMRI and SNP for children aged from 8 to 22 years. The data were preprocessed as follows.

1) fMRI data—The data were collected during an emotion task from 929 patients. All MRI scans were acquired on a single 3T Siemens TIM Trio whole-body scanner. In this task, participants were asked to identify happy, angry, sad, fearful and neutral faces. The data were acquired with a single-shot, interleaved multi-slice, gradient-echo, echo planar imaging (GE-EPI) sequence. The repetition time and echo time are 3000ms and 32ms, respectively. The total scan duration for this task was 10.5 min [26].

Standard preprocessing steps were applied using SPM12¹, including motion correction, spatial normalization to standard MNI space, spatial smoothing with a 3mm FWHM Gaussian kernel. The functional time series were band-pass filtered using a 0.01Hz to 0.1Hz frequency range. 246 ROI were extracted based on the Brainnetome Atlas [36], which contains information on both anatomical and functional connections. Then, multiple regression considering the influence of motion was performed using the SPM software and the stimulus on-off contrast maps for each subject were collected. The voxels with missing rate higher than 5% were deleted and missing values were imputed by the mean of that variable from all other samples. Within each ROI, the voxels with correlation higher than 0.9 were pruned to reduce the intra-ROI correlation. All these procedures result in 27168 voxels in 221 ROIs.

2) SNP data—We selected the genomic data from 3 platforms, including the Illumina HumanHap 610 array, the Illumina Human Omni Express array. There are totally 7863 samples and each platforms has 620k, 561k, 731k SNPs, respectively [26]. The common sets of SNPs(313k) were first extracted, and then PLINK [37] was used to perform standard quality controls, including the Hardy-Weinberg equilibrium test with $p\text{-value} < 1e-5$ for genotyping errors, extraction of common SNPs(MAF > 5%), Linkage disequilibrium pruning threshold of 0.9 for highly correlated SNPs. SNPs with missing call rates > 10% and samples with missing SNPs > 5% were removed. The missing values were further imputed by Minimac3 [38] using the reference genome from 1000 Genome Project Phase 1 data². Each SNP was recoded to discrete numbers, which equal to the number of alleles. Finally, the SNPs within gene bodies and the gene with SNP count greater than 10 were kept and 63010 SNPs in 2035 genes were left.

Then, the overlapping samples were extracted from the fMRI, SNP described above, leading to 866 subjects.

According to the preprocessing steps, the voxels, SNPs were grouped into 221 ROIs and 2035 genes. For each pair of ROI and gene, we calculated the projected distance correlation based on Algorithm 1 and the decorrelation of fMRI data. Then, we applied the student's- t test (6) to get a collection of p -values. Under FDR control with target FDR 0.01, 455 ROI-Gene pairs were detected. As a comparison, we randomly permuted the subjects and repeated the same procedures under null hypothesis. The average number of discoveries is 17.6 over 100 independent runs, which is far less than 455. Meanwhile, we also performed the same procedure on subsets of the data. For each of the 100 runs, we randomly subsampled 750 subjects. The average number of discoveries is 287.2 and the mean overlap with that of the full data is 207.5. The above results demonstrate that our findings are neither from randomness nor from false positives. In addition, the power of G-PDC in real data relies heavily on the sample size.

Table II further compared different methods in the number of discoveries. As shown in the table, there are few edges detected by Pearson's correlation and partial correlation. This is

¹<http://www.fil.ion.ucl.ac.uk/spm/>

²<http://www.1000genomes.org>

because there are nearly 1.6×10^9 edges between voxels and SNPs, which makes multiple testing extremely challenging. As a result, the power of the univariate methods was too low to provide useful information. In contrast, there were only 4×10^5 gene-ROI pairs, which greatly reduced the difficulties in multiple testing. Nevertheless, the conventional distance correlation and CCA detected too many associations. Since the imaging genetic associations are believed to be weak, it is reasonable to suspect that these discoveries contain a great amount of indirect associations. The partial distance correlation seems to reduce a great amount of false positives in the real data, but the power may be also low according to the simulation results. We also count the number of overlap in findings with G-PDC. Through hypergeometric test, we found that the overlaps were not due to randomness. It is worth to note that only 25% of the edges detected by G-PDC were overlapped with CCA, which may infer that only a small amount of imaging genetic associations are linear.

Table III then lists the top 20 significantly correlated gene and ROI pairs³. Each association shows significantly strong projected distance correlation. Interestingly, most of these genes are connected to the prefrontal gyrus, especially the bilateral superior frontal gyrus and middle frontal gyrus. The medial frontal network was recognized to be one of the most successful networks in individual subject identification [39], and exhibits delayed stabilization and individualization between healthy adolescent and those with increased clinical symptoms [40]. The activity of superior frontal gyrus is sensitive in emotion recognition [41], which also plays important role in the emotion task. The q -value of these pairs derived from some other methods were also reported, including distance correlation, partial distance correlation and CCA. The Pearson's correlation and partial correlation were excluded because of their limited power. From Table II, found that there were quite a lot of overlaps between G-PDC and distance correlation. For the partial distance correlation, many of the p -values were below 0.05, although not small enough to pass the multiple testing threshold. There were pairs shared by CCA and G-PDC. It is interesting that these edges also reached very small q -value in distance correlation and partial distance correlation. This observation may further indicate that the detection of nonlinear interactions is more challenging.

Fig. 9 displays the degrees of ROIs calculated from the ROI-Gene network and drawn by BrainNet Viewer [42], where the degree denotes the number of edges. From the figure, we can find several highly connected nodes, i.e., the hubs. Here we define hubs as nodes with degrees at least two standard deviation higher than the mean degrees. In this way we identified 5 ROI hubs, which were listed in Table IV. All of these ROIs are from the frontal lobe, and have much larger connections than average.

Following the same definition of hubs, we detected 24 hub genes, which were listed in Fig. 10. By checking the Genecard⁴, we found 13 genes, including ABCC8, ADCY9, BCAS3, BCR, CCDC88C, DAB1, DLGAP1, GLDN, GRXCR1, HKDC1, MACROD2, SLC6A11 and SCARA5, to be overexpressed in brain. In addition, 13 genes, including ACSS3, ADCY9, DAB1, DCDC2, DLGAP1, FBLN1, GRXCR1, HKDC1, MACROD2, MGAT5,

³The full list is given in the supplementary data

⁴<http://www.genecards.org>

NMNAT3, SLC6A11, SYNE1, are related to mental disorders, including depression, schizophrenia, autism, intellectual disability and etc. All these results support the potential interactions between these genes and the brain activities.

We further explored the biological implications of the detected interactions. Specifically, among the ROI-gene pairs in Table III, five connections were supported by existing literatures. Specifically, the DAB1 mRNAs for autistic subjects were significantly reduced in superior frontal cortex than controls [43]. The gene DCDC2 was important role in gray matter distribution in middle and superior frontal gyrus in healthy individuals [44]. Both regions are detected in Table III. LYRM4 was down-regulated in the prefrontal cortex of mice with microdeletions in the locus syntenic to human 22q11.2 patients affected by schizophrenia [45]. Although there were no direct evidences on the remaining pairs, we found potential links between three pairs. Firstly, the gene GRXCR1 was identified as a cause of hearing impairment [46]. The corresponding ROI in Table III, right postcentral gyrus, displayed a decreased connectivity in networks responding to emotional sounds in patients with hearing loss [47]. Secondly, the gene MACROD2 was identified to be associated with autism spectrum disorders by different research groups [48] [49]. A functional connectivity study of autism spectrum disorders showed altered connectivity with respect to superior frontal gyrus and temporal lobe [50]. Both regions were identified to interact with MACROD2 by our method. Finally, the gene SLC6A11 is one major GABA transporter in brain, the microdeletion of this gene was found to cause intellectual disability, epilepsy, ataxia, and stereotypic behavior [51]. The precentral gyrus was identified to be one origin of the epilepsy [52] and was related to several types of ataxia [53] [54]. In addition, since 3 out of the 4 overlaps between our method and CCA were supported or indirectly supported by literatures, we are also interested in the relationship between DLGAP1 and right superior frontal gyrus, which is also shared by CCA and G-PDC. The gene DLGAP1 is located in a chromosome region linked to schizophrenia, and was found to be a candidate gene to the risk [55]. It also encodes the protein SAPAP1, which is involved in maintaining normal brain function and development [56]. The superior frontal gyrus was well known to be associated with schizophrenia. All these results reveal some potential genetic factors of brain activities.

IV. Conclusion

Imaging genetics is an emerging area in brain research. However, the detection of imaging genetics associations remains a challenging problem due to the group structure, multi-model, high-dimension, and complex structure of the data. Our aim in this paper was to find a way to detect complex imaging genetic relationships. The main contributions of our work can be summarized as follows. First, we proposed a method based on projected distance correlation that can detect conditional dependence between imaging and genetics data in high-dimensional cases. Second, we combined orthogonal greedy algorithm to accelerate the calculation of projected distance correlation, which is highly demanded for imaging genomics studies, where large scale of multiple data are ubiquitous. Third, we studied the numerical performance via a series of simulations. Our results show that the projected distance correlation can achieve comparable accuracy with classical linear approaches in detecting linear correlation, but is much better in finding non-linear correlations. Finally, we

applied the proposed method to the analysis of a neurodevelopment dataset. We have discovered some novel interactions among ROIs and genes. These interactions hold the promise for shedding light on how dysfunction of gene-brain interactions can imply the risk of brain disorders.

There is still room for further improvement. First, the linear projection assumption (8) can be relaxed to nonlinear projections. One possible direction is the sparse additive model [57] [23]. It combines sparse linear modeling with additive nonparametric regression, which works for a class of nonlinear additive models. It is worth to design fast algorithm for high-dimensional data and study theoretical properties in the application of projected distance correlation. Second, a limitation of the proposed approach is that it may lead to inflated false positives, especially when the prescribed FDR is 0.05. On the other hand, we showed that the use of more conservative target rates leads to more reasonable FDR control. In practice, we recommend to use our approach with a prescribed FDR of 0.01 or lower. To further address this important problem, we plan to work on different FDR control algorithms while improving the accuracy of sparse regression in the meantime. Finally, it is necessary to validate our results with replication experiments with additional biological evidences. Our dedication to tackling these interesting issues remains strong and further discoveries will be shared with the research community in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank the NIH (R01GM109068, R01MH104680, R01MH107354, R01AR059781, R01EB006841, R01EB005846, P20GM103472), and NSF (#1539067) for their partial support.

References

1. Hariri AR, Drabant EM, Weinberger DR. Imaging genetics: perspectives from studies of genetically driven variation in serotonin function and corticolimbic affective processing. *Biological psychiatry*. 2006; 59(10):888–897. [PubMed: 16442081]
2. Meyer-Lindenberg A. The future of fMRI and genetics research. *NeuroImage*. 2012; 62(2):1286–1292. [PubMed: 22051224]
3. Pearlson GD, Calhoun VD, Liu J. An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders. *Frontiers in genetics*. 2015; 6:276. [PubMed: 26442095]
4. Hotelling H. Relations between two sets of variates. *Biometrika*. 1936:321–377.
5. Wold H. Partial least squares. *Encyclopedia of statistical sciences*. 1985
6. Comon P. Independent component analysis, a new concept? *Signal processing*. 1994; 36(3):287–314.
7. Calhoun VD, Liu J, Adalı T. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage*. 2009; 45(1):S163–S172. [PubMed: 19059344]
8. Liu J, Calhoun VD. A review of multivariate analyses in imaging genetics. *Frontiers in neuroinformatics*. 2014; 8

9. Grellmann C, Bitzer S, Neumann J, Westlye LT, Andreassen OA, Villringer A, Horstmann A. Comparison of variants of canonical correlation analysis and partial least squares for combined analysis of MRI and genetic data. *NeuroImage*. 2015; 107:289–310. [PubMed: 25527238]
10. Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*. 2009; 8(1): 1–34.
11. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*. 2009; 8(1):1–27.
12. Chun H, Kele S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010; 72(1):3–25. [PubMed: 20107611]
13. Vounou M, Nichols TE, Montana G, A. D. N. Initiative. et al. Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. *Neuroimage*. 2010; 53(3):1147–1159. [PubMed: 20624472]
14. Lin D, Zhang J, Li J, Calhoun VD, Deng H-W, Wang Y-P. Group sparse canonical correlation analysis for genomic data integration. *BMC bioinformatics*. 2013; 14(1):245. [PubMed: 23937249]
15. Achard S, Salvador R, Whitcher B, Suckling J, Bullmore E. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *Journal of Neuroscience*. 2006; 26(1):63–72. [PubMed: 16399673]
16. Greicius MD, Krasnow B, Reiss AL, Menon V. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*. 2003; 100(1):253–258.
17. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*. 2008; 9(1):559. [PubMed: 19114008]
18. Calhoun VD, Sui J. Multimodal fusion of brain imaging data: A key to finding the missing link (s) in complex mental illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. 2016; 1(3):230–244. [PubMed: 27347565]
19. Yang E, Baker Y, Ravikumar P, Allen GI, Liu Z. Mixed graphical models via exponential families. *AISTATS*. 2014; 2012:1042–1050.
20. Székely GJ, Rizzo ML, Bakirov NK, et al. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*. 2007; 35(6):2769–2794.
21. Székely GJ, Rizzo ML. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*. 2013; 117:193–213.
22. Székely GJ, Rizzo ML, et al. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*. 2014; 42(6):2382–2412.
23. Fan J, Feng Y, Xia L. A projection based conditional dependence measure with applications to high-dimensional undirected graphical models. *ArXiv e-prints*. Jan.2015
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society Series B (Methodological)*. 1995:289–300.
25. Ing C-K, Lai TL. A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica*. 2011:1473–1513.
26. Satterthwaite TD, Elliott MA, Ruparel K, Loughhead J, Prabhakaran K, Calkins ME, Hopson R, Jackson C, Keefe J, Riley M, et al. Neuroimaging of the Philadelphia neurodevelopmental cohort. *Neuroimage*. 2014; 86:544–553. [PubMed: 23921101]
27. Rahimi A, Recht B. Random features for large-scale kernel machines. in *Advances in neural information processing systems*. 2008:1177–1184.
28. Scholkopf B, Smola AJ. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press; 2001.
29. Sturm BL, Groesb M, et al. Comparison of orthogonal matching pursuit implementations. *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European IEEE*. 2012:220–224.
30. Zhang T. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*. Mar.2009 10:555–568.

31. Fan J, Ke Y, Wang K. Decorrelation of covariates for high dimensional sparse regression. arXiv preprint arXiv:1612.08490. 2016
32. Kneip A, Sarda P, et al. Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics*. 2011; 39(5):2410–2447.
33. Wang T, Ren Z, Ding Y, Fang Z, Sun Z, MacDonald ML, Sweet RA, Wang J, Chen W. FastGGM: an efficient algorithm for the inference of gaussian graphical model in biological networks. *PLoS computational biology*. 2016; 12(2):e1004755. [PubMed: 26872036]
34. Genovese CR, Lazar NA, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*. 2002; 15(4):870–878. [PubMed: 11906227]
35. Satterthwaite TD, Connolly JJ, Ruparel K, Calkins ME, Jackson C, Elliott MA, Roalf DR, Hopson R, Prabhakaran K, Behr M, et al. The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth. *Neuroimage*. 2016; 124:1115–1119. [PubMed: 25840117]
36. Fan L, Li H, Zhuo J, Zhang Y, Wang J, Chen L, Yang Z, Chu C, Xie S, Laird AR, et al. The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cerebral Cortex*. 2016; 26(8):3508–3526. [PubMed: 27230218]
37. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*. 2007; 81(3):559–575. [PubMed: 17701901]
38. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, et al. Next-generation genotype imputation service and methods. *Nature Genetics*. 2016; 48(10):1284–1287. [PubMed: 27571263]
39. Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, Papademetris X, Constable RT. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*. 2015; 18(11):1664–1671. [PubMed: 26457551]
40. Kaufmann T, Alnæs D, Doan NT, Brandt CL, Andreassen OA, Westlye LT. Delayed stabilization and individualization in connectome development are related to psychiatric disorders. *Nature neuroscience*. 2017; 20(4):513–515. [PubMed: 28218917]
41. McLellan T, Wilcke J, Johnston L, Watts R, Miles L. Sensitivity to posed and genuine displays of happiness and sadness: A fMRI study. *Neuroscience letters*. 2012; 531(2):149–154. [PubMed: 23123788]
42. Xia M, Wang J, He Y. BrainNet Viewer: a network visualization tool for human brain connectomics. *PloS one*. 2013; 8(7):e68910. [PubMed: 23861951]
43. Fatemi SH, Snow AV, Strydom JM, Araghi-Niknam M, Reutiman TJ, Lee S, Brooks AI, Pearce DA. Reelin signaling is impaired in autism. *Biological psychiatry*. 2005; 57(7):777–787. [PubMed: 15820235]
44. Meda SA, Gelernter J, Gruen JR, Calhoun VD, Meng H, Cope NA, Pearlson GD. Polymorphism of DCDC2 reveals differences in cortical morphology of healthy individuals: a preliminary voxel based morphometry study. *Brain imaging and behavior*. 2008; 2(1):21–26. [PubMed: 19096528]
45. Bozza M, Bernardini L, Novelli A, Brovedani P, Moretti E, Canapicchi R, Doccini V, Filippi T, Battaglia A. 6p25 interstitial deletion in two dizygotic twins with gyral pattern anomaly and speech and language disorder. *European Journal of Paediatric Neurology*. 2013; 17(3):225–231. [PubMed: 23069351]
46. Schraders M, Lee K, Oostrik J, Huygen PL, Ali G, Hoefsloot LH, Veltman JA, Cremers FP, Basit S, Ansar M, et al. Homozygosity mapping reveals mutations of GRXCR1 as a cause of autosomal-recessive nonsyndromic hearing impairment. *The American Journal of Human Genetics*. 2010; 86(2):138–147. [PubMed: 20137778]
47. Husain FT, Carpenter-Thompson JR, Schmidt SA. The effect of mild-to-moderate hearing loss on auditory and emotion processing networks. *Frontiers in systems neuroscience*. 2014; 8
48. Jones RM, Cadby G, Blangero J, Abraham LJ, Whitehouse AJ, Moses EK. MACROD2 gene associated with autistic-like traits in a general population sample. *Psychiatric genetics*. 2014; 24(6):241. [PubMed: 25360606]

49. Anney R, Klei L, Pinto D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Sykes N, Pagnamenta AT, et al. A genome-wide scan for common alleles affecting risk for autism. *Human molecular genetics*. 2010; 19(20):4072–4082. [PubMed: 20663923]
50. Monk CS, Peltier SJ, Wiggins JL, Weng S-J, Carrasco M, Risi S, Lord C. Abnormalities of intrinsic functional connectivity in autism spectrum disorders. *Neuroimage*. 2009; 47(2):764–772. [PubMed: 19409498]
51. Dikow N, Maas B, Karch S, Granzow M, Janssen JW, Jauch A, Hinderhofer K, Sutter C, Schubert-Bast S, Anderlid BM, et al. 3p25. 3 microdeletion of GABA transporters SLC6A1 and SLC6A11 results in intellectual disability, epilepsy and stereotypic behavior. *American Journal of Medical Genetics Part A*. 2014; 164(12):3061–3068.
52. Bonini F, McGonigal A, Trébuchon A, Gavaret M, Bartolomei F, Giusiano B, Chauvel P. Frontal lobe seizures: from clinical semiology to localization. *Epilepsia*. 2014; 55(2):264–277. [PubMed: 24372328]
53. Hernandez-Castillo CR, Galvez V, Diaz R, Fernandez-Ruiz J. Specific cerebellar and cortical degeneration correlates with ataxia severity in spinocerebellar ataxia type 7. *Brain imaging and behavior*. 2016; 10(1):252–257. [PubMed: 25917872]
54. Ginestroni A, Diciotti S, Cecchi P, Pesaresi I, Tessa C, Giannelli M, Nave RD, Salvatore E, Salvi F, Dotti MT, et al. Neurodegeneration in friedreich's ataxia is associated with a mixed activation pattern of the brain. A fMRI study. *Human brain mapping*. 2012; 33(8):1780–1791. [PubMed: 21674694]
55. Li J-M, Lu C-L, Cheng M-C, Luu S-U, Hsu S-H, Chen C-H. Genetic analysis of the DLGAP1 gene as a candidate gene for schizophrenia. *Psychiatry research*. 2013; 205(1):13–17. [PubMed: 22940546]
56. Kawashima N, Takamiya K, Sun J, Kitabatake A, Sobue K. Differential expression of isoforms of PSD-95 binding protein (GKAP/SAPAP1) during rat brain development. *FEBS letters*. 1997; 418(3):301–304. [PubMed: 9428732]
57. Ravikumar P, Liu H, Lafferty J, Wasserman L. *Proceedings of the 20th International Conference on Neural Information Processing Systems*. Curran Associates Inc; 2007. Spam: Sparse additive models; 1201–1208.

Fig. 1.

An illustration of the proposed method to detect group-wise imaging genetics associations. Each edge represents a statistically significant conditional dependence derived by projected distance correlation between a group of SNPs (gene) and a group of voxels (ROI).

Fig. 2.

A comparison of the null distribution when applying the G-DPC to real imaging genetics data. To build the null distribution, we collected fMRI data and SNP profiles from non-overlapping subjects. a) QQ plot on the original data. b) QQ plot after decorrelation.

Fig. 3.

A comparison of the null distribution when applying the G-DPC to multivariate normal vectors. The two vectors are drawn independently from a p -dimensional multivariate normal distribution with identical covariance matrix. a) $p = 5$. b) $p = 30$.

Fig. 4.

A comparison of graphs constructed by different dependence measures under linear correlation. a) TPR vs FPR when $n = 200$, b) ROC TPR vs FPR when $n = 500$.

Fig. 5.

A comparison of graphs constructed by different dependence measures under nonlinear correlation. a) TPR vs FPR when $n = 200$, b) ROC TPR vs FPR when $n = 500$.

Fig. 6.
A comparison of the computational time between OGA and LASSO in calculating the projected distance correlation.

Fig. 7.

A comparison of the FDR control with targeted FDR=0.05(left) and 0.01(right). top) linear case, bottom) nonlinear case.

Fig. 8.

A comparison of the power with targeted FDR=0.05(left) and 0.01(right). top) linear case, bottom) nonlinear case.

Fig. 9.
The distribution of the degrees of ROI. The size is proportional to the degree.

Fig. 10.
The network with significantly correlated ROI and hub genes.

TABLE I

Methods compared in the simulations and real data.

Method	Multivariate	Nonlinear
Distance correlation	Yes	Yes
Partial distance correlation	Yes	Yes
Projected distance correlation with LASSO	Yes	Yes
Projected distance correlation with OGA	Yes	Yes
Pearson's correlation	No	No
Partial correlation	No	No
Canonical correlation analysis	Yes	No

TABLE II

Comparison of the number of discoveries on PNC data.

Method	# discoveries	# Overlap with PDC
Projected distance correlation	455	–
Distance correlation	74533	288(p=1.5E-128)
Partial distance correlation	2078	25(p=3.3E-20)
Pearson's correlation	23	3(p=8.9E-09)
Partial correlation	0	0
Canonical correlation analysis	24592	112(p=1.5E-45)

TABLE III

The top 20 significantly correlated gene and ROI pairs with q-values.

Gene Symbol	ROI index	ROI name	G-PDC	DC	pDC	CCA
PDIA5	26	MFG(R)	3.85E-09	4.51E-05	4.84E-01	8.13E-02
DCDC2	26	MFG(R)	9.32E-09	1.83E-14	6.15E-05	3.18E-03
SLC6A11	60	PrG(R)	2.12E-08	1.16E-09	9.92E-01	5.04E-02
LOC101927865	7	SFG(L)	8.27E-08	3.45E-03	1.00E+00	7.33E-02
PDIA5	83	MTG(L)	1.16E-07	2.25E-03	2.91E-01	8.48E-02
DCDC2	7	SFG(L)	2.86E-07	1.10E-07	2.19E-01	6.82E-01
GRXCR1	162	PoG(R)	6.46E-07	1.14E-04	3.29E-01	2.49E-01
LYRM4	7	SFG(L)	2.00E-06	2.80E-03	4.17E-01	2.20E-02
PDIA5	6	SFG(R)	3.73E-06	2.20E-04	6.56E-01	7.13E-02
DIP2C	53	PrG(L)	6.26E-06	6.81E-02	8.03E-01	6.30E-02
PDIA5	79	STG(L)	7.61E-06	6.17E-02	1.00E+00	2.08E-01
DABI	8	SFG(R)	8.37E-06	2.47E-22	1.47E-01	5.58E-09
ACSS3	196	MVOcC(R)	1.09E-05	6.17E-02	7.65E-01	6.82E-01
LOC101927865	8	SFG(R)	1.23E-05	5.13E-03	1.00E+00	6.50E-02
DCDC2	8	SFG(R)	1.64E-05	4.43E-13	6.90E-03	3.24E-01
FBLN1	25	MFG(L)	2.50E-05	9.23E-06	1.37E-01	4.15E-01
GALNT2	26	MFG(R)	2.60E-05	2.86E-03	1.00E+00	1.87E-02
EHBP1	8	SFG(R)	2.92E-05	2.74E-04	9.78E-01	5.93E-01
MACROD2	7	SFG(L)	3.47E-05	9.72E-21	3.54E-02	2.25E-29
DLGAP1	8	SFG(R)	3.42E-05	1.00E-14	4.65E-01	1.57E-09

TABLE IV

The hub ROIs in gene-ROI network. DG represents for degree.

ROI index	ROI name	DG
6	Superior frontal Gyrus(R)	13
7	Superior frontal Gyrus(L)	28
8	Superior frontal Gyrus(R)	42
25	Middle Frontal Gyrus(L)	19
26	Middle Frontal Gyrus(R)	11
60	Precentral Gyrus(R)	16