Jiehuan Sun[1] / Jose D. Herazo-Maya[2] / Xiu Huang[3] / Naftali Kaminski[2] / Hongyu Zhao[1]

Distance-correlation based gene set analysis in longitudinal studies

[1] Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA, E-mail: hongyu.zhao@yale.edu
[2] Internal Medicine: Pulmonary, Critical Care and Sleep Medicine, Yale School of Medcine, New Haven, CT 06519, USA
[3] Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA

Abstract:
Longitudinal gene expression profiles of subjects are collected in some clinical studies to monitor disease progression and understand disease etiology. The identification of gene sets that have coordinated changes with relevant clinical outcomes over time from these data could provide significant insights into the molecular basis of disease progression and lead to better treatments. In this article, we propose a Distance-Correlation based Gene Set Analysis (dcGSA) method for longitudinal gene expression data. dcGSA is a non-parametric approach, statistically robust, and can capture both linear and nonlinear relationships between gene sets and clinical outcomes. In addition, dcGSA is able to identify related gene sets in cases where the effects of gene sets on clinical outcomes differ across subjects due to the subject heterogeneity, remove the confounding effects of some unobserved time-invariant covariates, and allow the assessment of associations between gene sets and multiple related outcomes simultaneously. Through extensive simulation studies, we demonstrate that dcGSA is more powerful of detecting relevant genes than other commonly used gene set analysis methods. When dcGSA is applied to a real dataset on systemic lupus erythematosus, we are able to identify more disease related gene sets than other methods.

Keywords: distance correlation, gene set analysis, longitudinal gene expression study
DOI: 10.1515/sagmb-2017-0053

1 Introduction

A living cell performs its functions and responds to external stimuli through the orchestrated activities of genes, where genes of similar functions are organized into regulatory modules. Many gene sets (modules) have been inferred through decades of biomedical studies. The knowledge on these gene sets has been curated in a number of publicly available databases, such as KEGG (Kanehisa & Goto, 2000) and BIOCARTA (Nishimura, 2001), which can be utilized to conduct gene set analysis (GSA) in genomics studies to assess the associations between gene sets and clinical outcomes. Many studies have demonstrated that GSA can provide statistically more robust and biologically more interpretable results than analysis based on individual genes (Huang et al., 2003; Segal et al., 2005). In addition, GSA is generally more powerful than single gene based analysis (Efron & Tibshirani, 2007).

Longitudinal monitoring of molecular profiles can be very valuable in disease diagnosis, prognosis as well as understanding the underlying biological mechanisms (Chen et al., 2012). In particular, longitudinal molecular profiles can be extremely helpful in studying complex diseases, where subjects are highly heterogeneous (Meacham & Morrison, 2013; Jeste & Geschwind, 2014), since the heterogeneity usually can be teased apart by utilizing the longitudinal profiles of different subjects. Thanks to advances in high-throughput technologies, more and more clinical studies are now collecting longitudinal molecular profiles, especially gene expression profiles, in addition to longitudinally collected disease-related clinical variables (Xiao et al., 2011; Obermoser et al., 2013; Lévy et al., 2014; Banchereau et al., 2016). It is often of great interest and importance to identify gene sets associated with clinical outcomes over time, which might provide insights into the etiology of the disease and hence lead to better treatments.

Because of the enormous success achieved by GSA in biological studies in the past decade, many GSA methods have been developed that are adapted to different scenarios encountered in gene expression data (Curtis, Orešič & Vidal-Puig, 2005; Subramanian et al., 2005; Michaud et al., 2008; Tsai & Qu, 2008; Huang, Sherman & Lempicki, 2009; Wu & Smyth, 2012). According to Goeman and Bühlmann (2007), GSA methods can be divided into two major categories, that is competitive and self-contained methods. These two categories differ in the null hypothesis being test. The competitive GSA methods test the null hypothesis that the proportion of

Hongyu Zhao is the corresponding author.

differentially expressed genes in the given gene set is not larger than that of all other genes not in the given gene set. In contrast, the self-contained GSA methods test the null hypothesis that no genes in the given gene set are differentially expressed. The mean-rank gene set test (Michaud et al., 2008), which tests if a set of genes is highly ranked compared to all other genes given a test statistic, is one widely used competitive GSA method. Since the input to the mean-rank gene set test is the gene-wise test statistics, this method can be applied to longitudinal gene expression data, if combined with differential expression analysis methods for longitudinal gene expression data (Storey et al., 2005; Tai & Speed, 2005; Leek et al., 2006; Tai & Speed 2006; 2009).

Many self-contained GSA methods have been proposed in the past decades and we discuss some self-contained GSA methods that can be applied to longitudinal gene expression data here. One straightforward approach is to use repeated measure analysis method to model the relationship between gene sets and the outcome. For instance, a test statistic for the association between gene sets and the outcome was proposed in Tsai and Qu (2008) based on the quadratic inference function method; a more recent method proposed in Hejblum, Skinner, and Thiébaut (2015), called TcGSA, can detect both homogeneous and heterogeneous time trends of genes in one gene set. Another widely used approach, called ROAST, uses rotation (a Monte Carlo technology for multivariate regression) rather than permutation test to derive p value for each gene set and is robust to gene-wise correlation (Wu et al., 2010).

However, these model-based approaches usually require normality assumption on the data and may also need to assume the independence of genes in order to derive a theoretical null distribution, both of which, if violated, may lead to inflated type one error rate. Also, these methods do not take into account subject heterogeneity, that is the same gene set could have different effects in the clinical outcomes of different subjects due to factors such as different disease stages and genetic backgrounds (Banchereau et al., 2016), which could lead to decreased power in detecting biologically relevant gene sets. Moreover, in genomics and genetics studies, the associations between genes or SNPs and outcomes of interest can be confounded by a number of unobserved factors, such as experimental artifacts and environmental perturbations, which, if not appropriately adjusted for, could lead to misleading results (Vilhjálmsson & Nordborg, 2013; Yang et al., 2013). The confounding effect could also be a serious issue to GSA, which is usually ignored in the methods mentioned above.

Distance correlation, a novel measure of correlation, was first proposed in Székely, Rizzo, and Bakirov (2007). Compared to Pearson correlation, distance correlation has two major advantages. First, Pearson correlation measures association between only two random variables while distance correlation quantifies association between two random vectors of arbitrary dimensions. Second, Pearson correlation can only capture linear relationship whereas distance correlation can detect any form of dependence between two random vectors. These two properties make distance correlation an attractive measure for GSA, where we want to assess the dependence between clinical outcomes (univariate or multivariate) and gene sets (a group of genes) and the dependence could be in complex form. It has been shown in many genomics studies that joint analysis of multiple related clinical outcomes could boost the statistical power (Li et al., 2014) while distance correlation can naturally assess the associations between gene sets and multiple outcomes simultaneously, which could allow borrowing information across all outcomes and hence improve the power of detecting relevant gene sets.

In this paper, we extend the idea of distance correlation to longitudinal data, called LdCov, and propose a distance-correlation based gene set analysis method (dcGSA) relying on this LdCov. Our proposed method, dcGSA, is a self-contained testing method. It can detect both linear and nonlinear relationship between the gene sets and the clinical outcomes while appropriately taking into account the subject heterogeneity and adjusting for the effects of (unobserved) time-invariant confounders. Also, multiple related outcomes could be analyzed simultaneously in dcGSA to increase the statistical power. The performance of dcGSA is evaluated and compared to other commonly used GSA methods through both real data analysis and simulation studies. An R package to implement dcGSA is available in Bioconductor (http://bioconductor.org/packages/dcGSA/).

The remainder of the article is organized as follows. Section 2 details our proposed method and describes methods for comparison. Section 3 displays the performance of dcGSA in both simulation studies and real data application and comparisons to other commonly used GSA methods. We conclude the paper in Section 4.

2 Methods

2.1 Distance correlation

In the following, we first review the mathematical definition of distance correlation and its sample version, as proposed in Székely, Rizzo, and Bakirov (2007). Let $X \in R^p$ and $Y \in R^q$ be two random vectors with characteristic functions f_X and f_Y, respectively, and joint with characteristic functions $f_{X,Y}$. Since our proposed method is a self-contained GSA method, here Y represents the clinical outcome of interest and X represents the expression profile of genes in a given gene set and our goal is to test if the clinical outcome is dependent on genes in the

given gene set, which can be achieved through the distance correlation measure. The distance covariance $v^2(X, Y)$ (analogous to classical covariance) between X and Y is defined by

$$v^2(X,Y) = \int_{R^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)w(t,s)|\mathrm{d}t\mathrm{d}s, \tag{1}$$

where $w(t,s) = (c_p c_q |t|_p^{1+p} |s|_q^{1+q})^{-1}$, $|\cdot|$ is the Euclidian norm in $R^d (d = p, q)$, $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$, and $\Gamma(\cdot)$ is the gamma function.

The distance correlation $R(X, Y)$ is defined similarly to the Pearson correlation as

$$R(X,Y) = \sqrt{\frac{v^2(X,Y)}{v^2(X,X)v^2(Y,Y)}}. \tag{2}$$

A consistent estimator $\hat{v}^2(X, Y)$ for $v^2(X, Y)$, based on n independent samples (X_k, Y_k), $k = 1, ..., n$, is given by

$$\hat{v}^2(X,Y) = \hat{T}_1 + \hat{T}_2 - 2\hat{T}_3, \tag{3}$$

where $\hat{T}_1, \hat{T}_2$, and $\hat{T}_3$ are calculated based on the observed samples as

$$\hat{T}_1 = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} |X_k - X_l|_p |Y_k - Y_l|_q, \tag{4}$$

$$\hat{T}_2 = \left(\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} |X_k - X_l|_p\right)\left(\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} |Y_k - Y_l|_q\right), \tag{5}$$

$$\hat{T}_3 = \frac{1}{n^3}\sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{m=1}^{n} |X_k - X_l|_p |Y_k - Y_m|_q. \tag{6}$$

As we can see from the formulas, distance correlation (covariance) depends on the data only through pairwise distances (between all pairs of samples) among (X_k, Y_k), $k = 1, ..., n$. It was shown in Székely, Rizzo, and Bakirov (2007) that when X and Y are independent, under mild regularity conditions, the statistic $n\hat{v}^2/\hat{T}_2$ converges in distribution to a quadratic form of centered Gaussian random variables, that is

$$n\hat{v}^2/\hat{T}_2 \xrightarrow[n\to\infty]{D} \sum_{j=1}^{\infty} \lambda_j Z_j^2,$$

where the Z_j's are independent standard normal random variables and the λ_j's are nonnegative constants that depend on the distribution of (X, Y). The asymptotic distribution of $n\hat{v}^2/\hat{T}_2$ could be harnessed to test the independence between X and Y and we will call it the distance covariance statistic (dCov).

2.2 Longitudinal distance covariance statistic and dcGSA

In this section, we describe the extended distance correlation measure and the basic idea of dcGSA. In longitudinal studies, it is inappropriate to apply dCov directly to test the independence between the outcome and a gene set, due to the correlation among the repeated measures within each individual. To avoid the influence of within subject correlation, we propose to calculate dCov for each subject separately and use the average of those statistics as the test statistic (We will call it LdCov which stands for Longitudinal distance covariance statistic) for independence between the outcome and a gene set. Let $\hat{d}_i$ be the estimated dCov statistic for subject i, based on the repeated measures, then LdCov is calculated as

$$\mathrm{LdCov} = \sum_{i=1}^{N} \hat{d}_i/N, \tag{7}$$

where N is the number of subjects. Note that the subject-specific distance covariance statistic $\hat{d}_i$ can be calculated thanks to the longitudinal nature of the data, where each subject has multiple measurements.

Then, LdCov is used as the test statistic of the association between a given gene set (X) and the clinical outcome of interest (Y) in dcGSA. The null hypothesis of dcGSA is that that a gene set and the outcome are independent conditional on subjects, that is the gene set is not related to the outcome in any subject. In dcGSA, the null distribution of LdCov could be approximated by permutations, where the outcomes within each subject are permuted, and then the significance level can be inferred accordingly from the test statistics computed from the permuted samples. However, depending on the number of permutations, the p values obtained from permutations are not very precise and can be the same for all extremely significant associations, which makes the permutation p value hard to interpret and multiple testing correction methods, such as Bonferroni correction (Dunnett, 1955) and false discovery rate (FDR) (Benjamini & Hochberg, 1995), might not be applicable.

To deal with this issue, we propose to estimate the p values based on the permutations as follows. LdCov is calculated for each permutation, and we use $\hat{t}_j$ and $\hat{t}_0$ to denote the LdCov for the jth permutation and for the observed data, respectively. From the definition for LdCov, we can see that $\hat{t}_j$ is averaging over all subjects and hence is a mean type statistic, for which we know that it asymptotically follows a normal distribution based on the central limit theorem. Therefore, we can calculate a $\widehat{Z}$ score based on the $\hat{t}_j$'s as follows,

$$\widehat{Z} = \frac{\hat{t}_0 - \hat{\mu}}{\hat{s}}, \tag{8}$$

and use the standard normal distribution to infer the approximate p value based on the value of $\widehat{Z}$ score, where $\hat{\mu}$ and $\hat{s}$ are the mean and standard deviation of the $\hat{t}_j$'s across all permutations. Note that the p value for the $\widehat{Z}$ score is based on the one-side probability, since we only reject the null hypothesis when LdCov is large. The asymptotic normality of the permutation statistics and the use of the asymptotic normality to infer the approximate p values to save computation time have been discussed in Good (2005).

Since LdCov is testing the independence between the gene set (X) and the clinical outcome (Y) and the dimension of Y can be arbitrary, dcGSA can capture any form of dependence between gene sets and the outcomes, including linear and nonlinear relationship, and assess the associations between gene sets and multiple related outcomes simultaneously. Moreover, it could remove the confounding effects of (possibly unobserved) time-invariant variables, because dCov depends on the samples only through pairwise distances and the effects of time-invariant variables on the outcome within each subject will be cancelled out. Last but not least, it could also detect relevant gene sets even when the subjects are highly heterogeneous, i.e. the effects of gene sets on the clinical outcome are different across subjects, since the effects of gene sets are calculated for each subject separately and then the effects are summarized over each subject.

2.3 Methods for comparison

Since competitive and self-contained GSA methods test different null hypotheses, direct comparison of statistical power might not be meaningful. Therefore, to assess the performance of dcGSA, we compare it to two self-contained testing methods in the simulation studies: one commonly used gene set test method, ROAST (Wu et al., 2010), and one state-of-the-art method, TcGSA (Hejblum, Skinner & Thiébaut, 2015).

ROAST is a general framework and can be applied to any experimental design that can be formulated into a linear model. In addition, it is robust to gene-wise correlations and can account for the correlations among repeated measurements, as in longitudinal gene expression data. ROAST relies on rotation to calculate the p value and the precision of the p value depends on the number of rotations. TcGSA adopts the mix-effects model to identify gene sets in which the gene expressions vary significantly and display heterogeneous trends over time (hence the alternative hypothesis is wider than that of ROAST, which only identifies gene sets with gene expression profiles vary significantly over time). Both ROAST and TcGSA can detect linear as well as nonlinear relationship, which needs to be explicitly specified in the model though. In all data analyses, we use two versions for each method. For ROAST, we use ROAST.linear and ROAST.spline which specify linear term and spline terms with 2 degrees of freedom for the time variable, respectively. For TcGSA, we use TcGSA.linear and TcGSA.cubic, which specify linear term and cubic terms for the time variable, respectively.

Note that the dependent variables in both ROAST and TcGSA are the gene expression profiles, that is both methods regress the gene expression profiles on other independent variables, such as time (e.g. to see if the genes in the given gene set have some specific trends over time) and clinical groups (e.g. to see if the dynamic trends of the genes in the given gene set differ across clinical groups). In this sense, TcGSA and ROAST might not be appropriate for the clinical outcome-driven analysis. The primary goal of dcGSA is to detect gene sets that are related to the clinical outcome of interest (continuous or discrete), where the clinical outcome is the

dependent variable and the gene expression profiles are independent variables. To adapt ROAST and TcGSA to deal with continuous clinical outcomes, we treat the clinical outcome as an independent variable for both methods (we can think of the clinical outcome as the time variable).

In the real data application, we include a competitive testing method, i.e. the mean-rank gene set test (Rank test) (Michaud et al., 2008), for comparison. The Rank test investigates whether the genes in a gene set are highly ranked relative to all other genes in the entire gene list by adopting the Wilcoxon test statistic to assess the statistical significance level. The Rank test works on any type of test statistics that can be used to rank the genes.

3 Results

3.1 Simulation studies

In this section, we first compare the performance of dcGSA, TcGSA, and ROAST in terms of the type one error rate and power and then study the effect of unequal numbers of repeated measures for different subjects on the performance of dcGSA.

Although dcGSA can detect any form of dependence between the clinical outcome and gene sets, the statistical power can vary for different forms of dependence. Thus, we study the performance of the three methods under different forms of dependence as specified below. Here, we fix the dimension of Y (clinical outcome) to be 1, which is usually the case in real data applications. With the consideration of multiple related outcomes, we expect the power to improve. Let Y_{ij} denote the clinical outcome for subject i at the jth visit and X_{gij} denote the measured expression level of the gth gene in a given gene set for subject i at the jth visit, where $i = 1, \ldots, n$, $j = 1, \ldots, r_i$, $g = 1, \ldots, G$, n is the total number of subjects, r_i is the number of repeated measures for subject i, and G is the total number of genes in the given gene set. In all of our simulations, the gene expression profile X was generated from a multivariate Gaussian distribution with mean zero and an AR1 covariance structure to account for the correlations among genes, if not specified. The AR1 covariance structure is given by

$$\Sigma_\rho = \begin{pmatrix} 1 & \rho & \cdots & \rho^{G-2} & \rho^{G-1} \\ \rho & 1 & \ddots & \cdots & \rho^{G-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{G-2} & \cdots & \ddots & \ddots & \rho \\ \rho^{G-1} & \rho^{G-2} & \cdots & \rho & 1 \end{pmatrix}_{G\times G},$$

where $\rho = 0.3$ in all our simulations.

An important parameter that affects the performance of all GSA methods in GSA is the number of genes (s) in a given gene set that are truly related to Y. Given the above notations and set-up, Y was generated in different ways for different forms of dependence on X as follows. Note that we only include a random intercept for the clinical outcomes within each subject for simplicity and hence the covariance structure for the clinical outcome is compound symmetry.

- Linear dependence. In this scenario, the clinical outcome Y is linearly dependent on the genes in X, that is

$$Y_{ij} = \mu_i + \sum_{g=1}^{s} \beta_g X_{gij} + \epsilon_{ij},$$

  where the μ_i's are generated from a standard Gaussian distribution representing random effects for each subject, the ϵ_{ij}'s are generated from a standard Gaussian distribution representing the measurement errors, s is the number of genes related to the clinical outcome in the given gene set, and we set all the coefficients β_g's for these genes to be 1.0 for simplicity.

- Nonlinear dependence. In this scenario, the clinical outcome Y depends on the genes in X in a quadratic form to mimic the nonlinear relationship observed in the real data, that is

$$Y_{ij} = \mu_i + \sum_{g=1}^{s} \beta_g X_{gij}^2 + \epsilon_{ij},$$

  where μ_i's, ϵ_{ij}'s, and β_g's are generated in the same way as above.

- Random effects. In this scenario, the effects of genes in X on the clinical outcome Y are different for different subjects (i.e. subject heterogeneity), that is

$$Y_{ij} = \mu_i + \beta_i \left(\sum_{g=1}^{s} X_{gij} \right) + \epsilon_{ij},$$

where μ_i's and ϵ_{ij}'s are generated in the same way as above, and the β_i's are generated from a standard Gaussian distribution representing different effects of genes X on Y for different subjects.

- Confounding effects. In this scenario, the effects of genes in X on the clinical outcome Y are confounded by an unobserved variable Z, that is

$$\begin{aligned} Y_{ij} &= \mu_i + \beta Z_i + \epsilon_{ij}, \\ X_{gij} &= X^*_{gij} + \beta Z_i, \end{aligned}$$

where μ_i's and ϵ_{ij}'s are generated in the same way as above, the Z_i's are generated from a standard Gaussian distribution, the X^* are generated from a multivariate Gaussian distribution with mean zero and an AR1 covariance structure, and we set $\beta = 1.0$ to represent a modest confounding effect. For this scenario, since the correlations between the clinical outcome (Y) and the gene set (X) is induced by the confounding variable (e.g. Z could be age), a robust GSA method should not claim such gene sets as related to the clinical outcome, that is the lower power of the GSA method is for this scenario, more robust is the method to the confounding effects.

3.1.1 Comparison of dcGSA, TcGSA, and ROAST

In this section, we compare the performance of dcGSA, TcGSA, and ROAST in terms of power and type one error rate at a given statistical significance level. More specifically, the number of repeated measures (r) is fixed to be 5 and the number of subjects (n) is chosen from {20, 40, ..., 100}. Here, the number of repeated measures is the same for all subjects. In real data, different subjects usually have different numbers of repeated measures and this will be discussed in Section 3.1.2. Also, as mentioned above, the size of the gene set (G) and the number of significant genes (s) in the set can also affect the performance of the test. Here, we fix G to be 50 and s to be 10. Then, for each number of subjects (n) and each of the four different forms of dependence, we simulated two gene sets, where the gene expression values (X) are simulated as specified above and then the clinical outcome (Y) is simulated in the given dependence form based on the gene expression values from the first gene set (this gene set is related to the clinical outcome) while the second gene set is treated as the null gene set, which is used to estimate the type one error rate. The variance of measurement errors is tuned so that the signal to noise ratios are the same for all four forms of dependence and thus we can compare the power and type one error rate under different forms of dependence given the same sample size. Finally, the approximate p values based on the $\widehat{Z}$ scores are used as the significance levels for the two gene sets in dcGSA. The power and type one error rate are calculated at the significance level 0.05, based on 100 simulated data sets for each scenario.

From the right panel of Figure 1, we can see that the power of all methods increased as the number of subjects increases while larger number of subjects is needed for all methods in order to detect nonlinear relationship with the same power. In the simple linear dependence scenario, we find that the power of dcGSA is worse than all the other methods while the discrepancy decreases as the number of subjects increases. For the nonlinear dependence scenario, TcGSA.cubic and dcGSA outperform the other three approaches with TcGSA.cubic being the best. However, the type one error rate of TcGSA is not well controlled, especially for TcGSA.cubic, as shown in the left panel of Figure 1. And, the performance of ROAST.spline is slightly improved over ROAST.linear. For the random effects scenario, the performance of dcGSA is better than all the other approaches and hence dcGSA is robust to subject heterogeneity. Based on the performance in the confounding effects scenario, we can see that dcGSA is robust to confounding effects while all the other methods are not. The type one error rates for dcGSA and ROAST are well controlled as shown in the right panel of Figure 1.

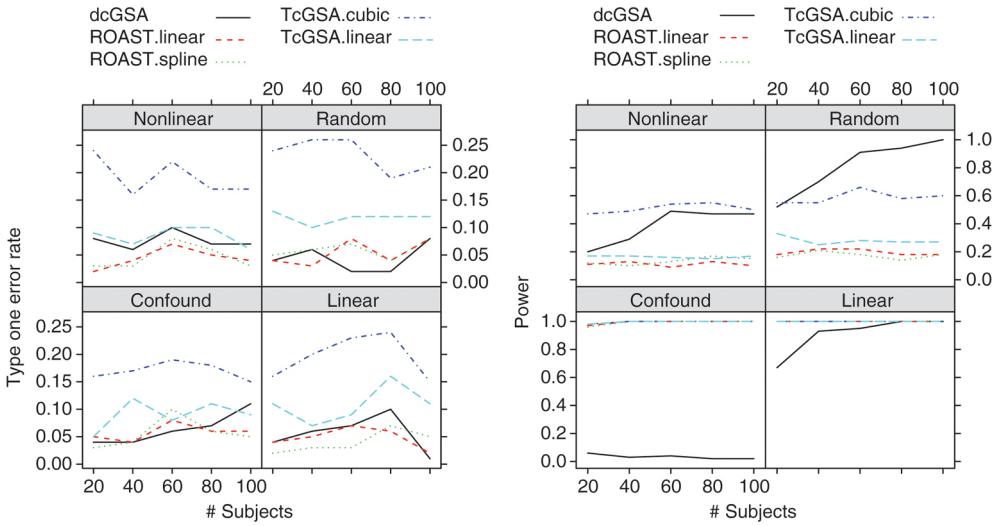

Figure 1: Comparison of dcGSA, TcGSA, and ROAST in terms of power and type one error rate given different forms of dependence and numbers of subjects.

Possible explanations for the inflated type one error rate of TcGSA are given below. First, the p values from TcGSA are calculated based on an approximate null distribution, which requires the independence of the tested effects and hence might be inflated if the genes are highly correlated in the gene sets. Second, TcGSA algorithm might not converge sometimes and hence leads to unstable results, especially when cubic function is applied and the number of repeated measurements is small. Third, TcGSA has wider alternative hypotheses compared to ROAST and dcGSA, as mentioned in Section 2.3.

3.1.2 Effect of unequal number of repeated measures (r) on dcGSA

For the simulation settings above, the number of repeated measures is assumed to be the same for all subjects. LdCov is calculated by taking the average of dCov over all subjects. However, different numbers of repeated measurements could yield dCov with different variances and hence influence the power and type one error rate of dcGSA. Here, we study the effect of unequal numbers of repeated measures (r) on the power and type one error rate of dcGSA. Specifically, the number of repeated measures (r) is drawn uniformly from {3, ..., 7} for each subject (the expected number of repeated measures is 5) and the number of subjects (n) is chosen from {20, 40, ..., 100}, and we fix the size of the gene set (G) to be 50 and the number of related genes (s) to be 10. Then, for each n and each of the four scenarios, we simulate the data and calculate the power and type one error rate in the same way as that in Section 3.1.1. We compare the power and type one error rate to the case where the number of repeated measures is fixed at 5 for all subjects.

As shown in Figure 2, compared to the case where all the subjects have the same number of repeated measures, the power and type one error rate of the test are similar in all scenarios when different subjects have different numbers of repeated measures, but with the same expected number of repeated measures, i.e. the total sample size is the same.

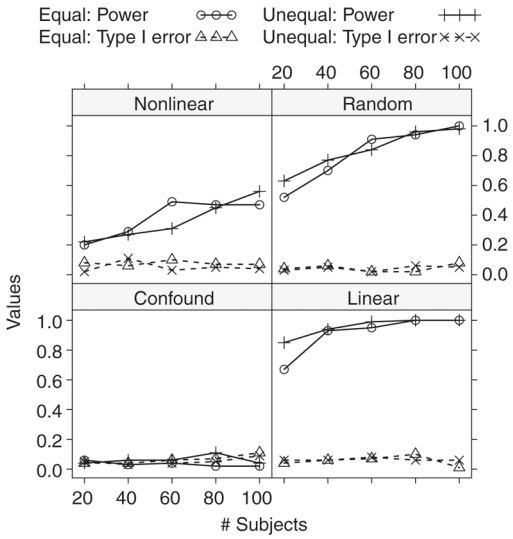

Figure 2: Effect of unequal numbers of repeated measurements on the power and type one error rate of dcGSA.

3.1.3 Computation time of dcGSA

The computational cost of dcGSA depends on the number of subjects, the number of repeated measures, the number of gene sets, the size of the gene sets, and the number of permutations. Here, we report the computation time of dcGSA for the simulation configuration of the linear dependence scenario in Section 3.1.1, where the number of repeated measures (r) is 5, the size of the gene set (G) is 50, the total number of gene sets is 200, and the number of permutations performed is 100. On a IBM NeXtScale nx360 M4 computer with 2.20 GHz Intel Xeon E5-2660 V2 CPU, the computation time of dcGSA using a single core is 0.05, 0.25, 0.75, 1.42, and 3.19 h for the number of subjects being 20, 40, 60, 80, and 100, respectively, suggesting that dcGSA is computationally efficient.

3.2 Real data analysis

In this section, we compare the performance of dcGSA with other methods on the systemic lupus erythematosus (SLE) dataset (Banchereau et al., 2016), where multiple clinical outcomes of interest are available. In addition, we show that it is important for GSA method to have the ability to detect nonlinear relationship and account for subject heterogeneity.

In the SLE dataset, the blood transcriptomes of 158 SLE subjects were longitudinally measured up to 4 years in addition to rich clinical information. SLE is a systemic autoimmune disease with unpredictable disease course and periods of remissions. SLE Disease Activity Index (SLEDAI), a summary score based on 24 clinical items and laboratory test results, is one of the validated measures for lupus activity and commonly used for diagnosis of SLE (Bombardier et al., 1992). Here, we try to identify gene sets that are associated with SLEDAI, which might lead to discovery of new biomarkers for SLE and hence better understanding of the disease progression and better diagnosis.

Our analysis considers 101 SLE patients with at least three visits, where the blood samples were taken irregularly within 4 years of follow up and the mean number of time points is 7.7. For dcGSA, SLEDAI is treated as the clinical outcome. As input to the Rank test, the t-statistic measuring association of each gene and SLEDAI (dependent variable) is calculated using linear mixed effects model with a random intercept after adjusting for potential confounders including age, gender, and race. For the two versions of TcGSA and ROAST, we also include age, gender, and race as potential confounders.

The numbers of significant gene sets after Bonferroni correction by dcGSA, TcGSA.linear, TcGSA.cubic, ROAST.linear, ROAST.spline, and Rank test are 212, 1138, 1179, 292, 213, and 49, respectively. As shown in our simulation studies, the type one error rates of TcGSA are not well controlled, which may explain why almost all gene sets are claimed to be significant by the two versions of TcGSA. The Rank test is a competitive testing method and hence it is not meaningful to compare its number of significant gene sets to the others. For dcGSA and the two versions of ROAST, the numbers of significant gene sets are similar. However, as shown in our simulation studies, ROAST is not robust to confounding effects and some of the significant gene sets might be spurious. Having said that, we find that there are 68 pathways in common among the pathways identified by

dcGSA, ROAST.linear, and ROAST.spline, which accounts for about 30% of the significant pathways discovered by dcGSA. This suggests that some pathways of potentially large effects can be identified by different methods and these gene sets are more likely to have linear or nonlinear effects on the clinical outcome, in which cases both dcGSA and ROAST are powerful. Moreover, the top gene sets identified by dcGSA are related to the immune system and cell cycle, which is consistent with the results on the same data in Banchereau et al. (2016). In addition, the most directly related pathway to SLE, "KEGG SYSTEMIC LUPUS ERYTHEMATOSUS", ranks 22nd and 41st in gene set lists given by dcGSA and the Rank test, respectively, which suggests the gene set rank list given by dcGSA is more biologically relevant (see Supplementary Material for details). For ROAST and TcGSA, it is not able to rank the top 100 gene sets, as they all have the same p values.

Besides SLEDAI, several blood test results are also correlated with disease activity, such as erythrocyte sedimentation rate (ESR) and levels of complement protein C3 (Nasiri et al., 2010). As mentioned earlier, it is natural to analyze multiple related clinical outcomes simultaneously in dcGSA, which might improve the statistical power to identity disease related pathways. Here, we apply dcGSA on the SLE dataset using three clinical outcomes including SLEDAI, ESR, and C3 to see if we can detect more gene sets. As a result, more gene sets (652) are discovered compared to using SLEDAI alone (212). Moreover, the pathway "KEGG SYSTEMIC LUPUS ERYTHEMATOSUS" ranks 4th compared to 22nd and 41st originally by using SLEDAI alone and the Rank test. Also, we discover some interesting pathways that are not identified by using SLEDAI alone. For example, one of the most improved gene sets in terms of p value (raw p value using SLEDAI alone is 0.07), "BIOCARTA BAD PATHWAY", regulates B cell apoptosis, which may contribute to the development of SLE (Lipsky, 2001). These suggest that analysis of multiple related clinical outcomes could help to improve the power of identifying relevant gene sets.

4 Discussion

In this article, we have extended the idea of distance correlation to repeated measures. Based on the extended distance correlation measures (LdCov), we propose dcGSA, a self-contained gene set analysis approach for longitudinal studies of gene expression profiles. Our simulation results suggest that, with moderate sample size, dcGSA is powerful in detecting gene sets associated with clinical outcomes in linear relationship as well as nonlinear relationship, although larger sample size is required to achieve a desired statistical power for the latter case. In addition, dcGSA is robust to confounding effects of time-invariant covariates and can detect genes sets having different effects for different subjects, making it extremely useful in cases where subjects are highly heterogeneous (i.e. random effects), as commonly observed in complex disease studies. By applying dcGSA to the SLE dataset, we find that dcGSA is powerful and can detect biologically meaningful gene sets. Moreover, dcGSA allows us to analyze multiple related outcomes simultaneously, which increases the power and discover some interesting gene sets. Although it is difficult to assess if certain confounders exist in real data application, it is safe to use dcGSA.

In our proposed method, we adopt $\widehat{Z}$ score to get an approximate p value for the association between a gene set and the clinical outcome of interest. The number of permutations has an impact on the precision of the estimates for the mean and standard deviation of the null distribution and hence the $\widehat{Z}$ score. Based on our experience, 100 permutations give a reasonably good approximation for small sample size. For moderate to large sample size, 1000 or more permutations are recommended.

In the simulation study, we fix the size of the gene set and the proportion of clinical outcome related genes in the given gene set. However, the size of gene sets could vary from tens to hundreds in public databases such as KEGG and the proportion of related genes could also vary. Regardless of the size of the gene set, the type one error rate is well controlled in dcGSA, as it is based on permutation to calculate the significance level. For gene sets of larger size, dcGSA usually requires larger signal-to-noise ratio (larger sample size and/or larger number of related genes) in order to achieve a desirable power. In addition, we take the AR1 covariance structure for the genes in the simulation studies. However, the covariance structure for the genes can be more complex in real data. In fact, we find that the performance of dcGSA is robust to different covariance structures on the genes. Additional simulation results on the influence of size of the gene set on the power of the test and the robustness of dcGSA to different covariance structures are provided in the supplement article.

There are several limitations in dcGSA. First, since distance correlation (covariance) can only be calculated for subjects with at least three repeated measures, dcGSA needs to exclude those subjects with less than three measures from the analysis, which may lead to inefficiency of the method especially when the number of subjects with two or fewer time points is large. For the same reason, dcGSA cannot be applied to time course microarray data, where the measurements at different time points are taken from different subjects, or the scenario where the primary interests are in gene sets that are differentially expressed between cases and controls.

Second, dcGSA may lose some power if only simple relationship exists between the genes and the clinical outcome in the data. For example, if all genes are in linear relationship with the clinical outcome, though this is rarely the case in real data, dcGSA could be less powerful than methods designed only for detecting linear relationship. Third, the significance level is obtained via permutation test, which might not be optimal due to the computational cost. It is desirable to derive the asymptotic distribution for LdCov, based upon which the significance level could be calculated. Lastly, although LdCov could adjust for confounding effects of some unobserved covariates by its definition, it cannot deal with time-varying confounding covariates. Adjusting for time-varying covariates could be a future direction to move forwards on dcGSA.

Funding

Jiehuan Sun, Xiu Huang, and Hongyu Zhao were supported in part by the National Institutes of Health grants R01 GM59507 and P01 CA154295 (Funder Id: 10.13039/100000002). Jose D. Herazo-Maya was supported by the Harold Amos Faculty development program of the Robert Wood Johnson Foundation and the Pulmonary Fibrosis Foundation. Naftali Kaminski was supported in part by the National Institutes of Health grants U01 HL108642 and R01 HL127349 (Funder Id: 10.13039/100000002).

References

Banchereau, R., S. Hong, B. Cantarel, N. Baldwin, J. Baisch, M. Edens, A.-M. Cepika, P. Acs, J. Turner and E. Anguiano (2016): "Personalized immunomonitoring uncovers molecular networks that stratify lupus patients," Cell, 165, 551–565.

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," J. Royal Stat. Soc. B Methodol., 57, 289–300.

Bombardier, C., D. D. Gladman, M. B. Urowitz, D. Caron, C. H. Chang, A. Austin, A. Bell, D. A. Bloch, P. N. Corey and J. L. Decker (1992): "Derivation of the SLEDAI. A disease activity index for lupus patients," Arthritis Rheum., 35, 630–640.

Chen, R., G. I. Mias, J. Li-Pook-Than, L. Jiang, H. Y. K. Lam, R. Chen, E. Miriami, K. J. Karczewski, M. Hariharan and F. E. Dewey (2012): "Personal omics profiling reveals dynamic molecular and medical phenotypes," Cell, 148, 1293–1307.

Curtis, R. K., M. Orešič and A. Vidal-Puig (2005): "Pathways to the analysis of microarray data," Trends Biotechnol., 23, 429–435.

Dunnett, C. W. (1955): "A multiple comparison procedure for comparing several treatments with a control," J. Am. Stat. Assoc., 50, 1096–1121.

Efron, B. and R. Tibshirani (2007): "On testing the significance of sets of genes," Ann. Appl. Stat., 1, 107–129.

Goeman, J. J. and P. Bühlmann (2007): "Analyzing gene expression data in terms of gene sets: methodological issues," Bioinformatics, 23, 980–987.

Good, P. I. (2005): Permutation, parametric and bootstrap tests of hypotheses, Springer New York, New York, third edition.

Hejblum, B. P., J. Skinner and R. Thiébaut (2015): "Time-course gene set analysis for longitudinal gene expression data," PLoS Comput. Biol., 11, e1004310.

Huang, E., S. Ishida, J. Pittman, H. Dressman, A. Bild, M. Kloos, M. D'Amico, R. G. Pestell, M. West and J. R. Nevins (2003): "Gene expression phenotypic models that predict the activity of oncogenic pathways," Nat. Genet., 34, 226–230.

Huang, D. W., B. T. Sherman and R. A. Lempicki (2009): "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," Nucleic Acids Res., 37, 1–13.

Jeste, S. S. and D. H. Geschwind (2014): "Disentangling the heterogeneity of autism spectrum disorder through genetic findings" Nat. Rev. Neurol., 10, 74–81.

Kanehisa, M. and S. Goto (2000): "KEGG: kyoto encyclopedia of genes and genomes," Nucleic Acids Res., 28, 27–30.

Leek, J. T., E. Monsen, A. R. Dabney and J. D. Storey (2006): "EDGE: extraction and analysis of differential gene expression," Bioinformatics, 22, 507–508.

Lévy, Y., R. Thiébaut, M. Montes, C. Lacabaratz, L. Sloan, B. King, S. Pérusat, C. Harrod, A. Cobb and L. K. Roberts (2014): "Dendritic cell-based therapeutic vaccine elicits polyfunctional HIV-specific T-cell immunity associated with control of viral load," Eur. J. Immunol., 44, 2802–2810.

Li, C., C. Yang, J. Gelernter and H. Zhao (2014): "Improving genetic risk prediction by leveraging pleiotropy," Hum. Genet., 133, 639–650.

Lipsky, P. E. (2001): "Systemic lupus erythematosus: an autoimmune disease of B cell hyperactivity," Nat. Immunol., 2, 764–766.

Meacham, C. E. and S. J. Morrison (2013): "Tumour heterogeneity and cancer cell plasticity," Nature, 501, 328–337.

Michaud, J., K. M. Simpson, R. Escher, K. Buchet-Poyau, T. Beissbarth, C. Carmichael, M. E. Ritchie, F. Schütz, P. Cannon and M. Liu (2008): "Integrative analysis of RUNX1 downstream pathways and target genes," BMC Genomics, 9, 363.

Nasiri, S., M. Karimifar, Z. S. Bonakdar and M. Salesi (2010): "Correlation of ESR, C3, C4, anti-DNA and lupus activity based on British Isles Lupus Assessment Group Index in patients of rheumatology clinic," Rheumatol. Int., 30, 1605–1609.

Nishimura, D. (2001): "BioCarta," Biotech Software & Internet Report: The Computer Software Journal for Scient, 2, 117–120.

Obermoser, G., S. Presnell, K. Domico, H. Xu, Y. Wang, E. Anguiano, L. Thompson-Snipes, R. Ranganathan, B. Zeitner and A. Bjork (2013): "Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines," Immunity, 38, 831–844.

Segal, E., N. Friedman, N. Kaminski, A. Regev and D. Koller (2005): "From signatures to models: understanding cancer using microarrays," Nat. Genet., 37, S38–S45.

Storey, J. D., W. Xiao, J. T. Leek, R. G. Tompkins and R. W. Davis (2005): "Significance analysis of time course microarray experiments," Proc. Natl. Acad. Sci. USA, 102, 12837–12842.
Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub and E. S. Lander (2005): "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," Proc. Natl. Acad. Sci. USA, 102, 15545–15550.
Székely, G. J., M. L. Rizzo and N. K. Bakirov (2007): "Measuring and testing dependence by correlation of distances," Ann. Stat., 35, 2769–2794.
Tai, Y. C. and T. P. Speed (2005): "Statistical analysis of microarray time course data," In: Nuber, U., editor, DNA Microarrays. Chapman and Hall/CRC, New York.
Tai, Y. C. and T. P. Speed (2006): "A multivariate empirical Bayes statistic for replicated microarray time course data," Ann. Stat., 34, 2387–2412.
Tai, Y. C. and T. P. Speed (2009): "On gene ranking using replicated microarray time course data," Biometrics, 65, 40–51.
Tsai, G.-F. and A. Qu (2008): "Testing the significance of cell-cycle patterns in time-course microarray data using nonparametric quadratic inference functions," Comput. Stat. Data Anal., 52, 1387–1398.
Vilhjálmsson, B. J. and M. Nordborg (2013): "The nature of confounding in genome-wide association studies," Nat. Rev. Genet., 14, 1–2.
Wu, D. and G. K. Smyth (2012): "Camera: a competitive gene set test accounting for inter-gene correlation," Nucleic Acids Res., 40, e133–e133.
Wu, D., E. Lim, F. Vaillant, M.-L. Asselin-Labat, J. E. Visvader and G. K. Smyth (2010): "ROAST: rotation gene set tests for complex microarray experiments," Bioinformatics, 26, 2176–2182.
Xiao, W., M. N. Mindrinos, J. Seok, J. Cuschieri, A. G. Cuenca, H. Gao, D. L. Hayden, L. Hennessy, E. E. Moore and J. P. Minei (2011): "A genomic storm in critically injured humans," J. Exp. Med., 208, 2581–2590.
Yang, C., L. Wang, S. Zhang and H. Zhao (2013): "Accounting for non-genetic factors by low-rank representation and sparse regression for eQTL mapping," Bioinformatics, 29, 1026–1034.

Supplemental Material: The online version of this article offers supplementary material (https://-doi.org/10.1515/sagmb-2017-0053).