BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 19 2014, pages 2802–2807
doi:10.1093/bioinformatics/btu396

Data and text mining

Advance Access publication June 19, 2014

PGS: a tool for association study of high-dimensional microRNA expression data with repeated measures

Yinan Zheng[1,†], Zhe Fei[2,†], Wei Zhang[3], Justin B. Starren[4], Lei Liu[5], Andrea A. Baccarelli[6], Yi Li[2,*] and Lifang Hou[5,7,*]

[1]Institute for Public Health and Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, [3]Institute of Human Genetics, University of Illinois at Chicago, Chicago, IL 60612, [4]Division of Health and Biomedical Informatics, Departments of Preventive Medicine and Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, [5]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, [6]Department of Environmental Health, Harvard School of Public Health, Boston, MA 02115 and [7]The Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: MicroRNAs (miRNAs) are short single-stranded non-coding molecules that usually function as negative regulators to silence or suppress gene expression. Owning to the dynamic nature of miRNA and reduced microarray and sequencing costs, a growing number of researchers are now measuring high-dimensional miRNA expression data using repeated or multiple measures in which each individual has more than one sample collected and measured over time. However, the commonly used univariate association testing or the site-by-site (SBS) testing may underutilize the longitudinal feature of the data, leading to underpowered results and less biologically meaningful results.

Results: We propose a penalized regression model incorporating grid search method (PGS), for analyzing associations of high-dimensional miRNA expression data with repeated measures. The development of this analytical framework was motivated by a real-world miRNA dataset. Comparisons between PGS and the SBS testing revealed that PGS provided smaller phenotype prediction errors and higher enrichment of phenotype-related biological pathways than the SBS testing. Our extensive simulations showed that PGS provided more accurate estimates and higher sensitivity than the SBS testing with comparable specificities.

Availability and implementation: R source code for PGS algorithm, implementation example and simulation study are available for download at https://github.com/feizhe/PGS.

Contact: y-zheng@northwestern.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 2, 2014; revised on June 2, 2014; accepted on June 15, 2014

1 INTRODUCTION

MicroRNAs (miRNAs) are short single-stranded RNAs of nearly 20–24 nt in length that are transcribed from DNA but not translated into proteins (Singh *et al*., 2008). Most miRNAs inhibit the translation of proteins, destabilize their target messenger RNAs (mRNAs) and control many cellular mechanisms dynamically (Baek *et al*., 2008; Selbach *et al*., 2008). Even small changes in miRNA expression levels may have profound consequences for the expression levels of target genes (Reinsbach *et al*., 2012). The dynamic nature of miRNAs distinguishes it from genetics. Therefore, owing also to the reduced microarray and sequencing experiments cost, a growing number of researchers are conducting investigations that measure high-dimensional miRNA expression data using repeated or multiple measures in which each individual has more than one sample collected and measured over time (Chen *et al*., 2013; Hecker *et al*., 2013). Repeated or multiple measures data allow the researcher to exclude miRNA expression variation between individuals depending on the outcomes and pinpoint the causal role of miRNA expressions such as longitudinal study design.

The popular site-by-site (SBS) testing using generalized estimation equation (GEE) (Zeger *et al*., 1988) or linear mixed model (LMM) (Henderson *et al*., 1959) represents a feasible approach to accommodate the presence of high-dimensional miRNA expression data measured at different time points. SBS testing is a type of analysis that performs univariate tests of associations for each of the biomarker sites, followed by multiple-testing adjustments, for example, the Bonferroni's *P*-value correction or the false discovery rate (FDR) q-value (Storey and Tibshirani, 2003). However, this approach is particularly problematic in high-dimensional miRNA expression data because it ignores the underlying dependent structure between miRNAs. In addition, not like genetics, miRNA expressions are modifiable by environmental factors including diet, air pollution and other external exposures (Hamm *et al*., 2010). Hence, for better delineation of the direct effects of miRNAs, adjusting for these environment factors in models are recommended (Rakyan *et al*., 2011). But issues such as overfitting, collinearity and obscuring biomarkers with small effect sizes are usually encountered in typical regression approaches, e.g. GEE and LMM. Therefore, computationally feasible methods are required to tackle these problems.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
[*]To whom correspondence should be addressed.

We propose to apply a variable selection method with specific application in high-dimensional miRNA expression data with repeated or multiple measures. Wang *et al*. proposed a novel penalized GEE (PGEE) (Wang *et al*., 2012) method to select variables when the number of covariates is moderate in a repeated or multiple measures setting. PGEE is able to account for both within subject correlation and dependencies between different biomarkers. However, PGEE typically fails in the presence of high-dimensional biomarkers, i.e. when the number of biomarkers is larger than the sample size. To tackle this issue, we develop a prescreening-based PGEE with grid search (PGS) method. Our method consists of an iterative two-step approach. Step 1, uses the screening method to downsize the biomarkers, whereas Step 2 feeds the 'survived' biomarkers to the PGEE. We repeat these two steps in a grid search and perform K-fold cross-validation to determine tuning parameters.

We test our methods in a miRNA profiling dataset generated from our Beijing Truck Driver Air Pollution Study (BTDAS) (Baccarelli *et al*., 2011; Byun *et al*., 2013; Guo *et al*., 2014; Hou *et al*., 2012; 2013a,b). In BTDAS, we measured air pollution and health outcomes, including lung functions, twice with 1–2 weeks apart. We also collected blood samples twice for biomarker measurements, including miRNA profiling.

Based on repeated measures miRNA collected in BTDAS, we apply PGS to model the lung function levels measured by forced expiratory volume in one second (FEV1), with the goal of detecting lung function-related miRNAs and uncovering regulatory pathways. We also compare PGS with GEE/LMM SBS testing in terms of prediction error of lung function levels and enrichment of lung function-related biological pathways. In addition, a simulation study is conducted to examine the extensions of PGS.

2 METHODS

2.1 Penalized Generalized Estimating Equations

PGEE (Wang *et al*., 2012) has emerged as a useful method for analyzing repeatedly measured data with a moderate number of covariates. The algorithm is able to select non-zero effects among a number of predictors via adding penalty terms to the traditional GEE. The penalized generalized estimating functions U(β) are defined as below:

$$U(\beta) = S(\beta) - q_{\lambda}(|\beta|)\text{sign}(\beta)$$

where S(β) are the estimating functions defining a GEE; $q_{\lambda}(|\beta|)$ are the penalty functions that introduce penalties to each of non-zero β estimate, so that if certain true β_i are zero, the algorithm would force the estimates to be zeroes; sign(β) is the sign vector for β. Tuning parameter λ within the penalty functions $q_{\lambda}(|\beta|)$, is the coefficient of penalty terms, and it determines the amount of shrinkage, i.e. bigger λ leads to smaller overall size of estimated effects. To select the effect size that best fits the data, λ will be tuned among a sequence of candidate values using K-fold cross-validation.

2.2 PGEE with Grid Search

To enhance the reliability of PGEE selection result, we perform PGEE on a sequence of subsets of the biomarkers based on the ranking of significance (e.g. *P*-values) obtained by a univariate prescreening analysis adjusted for confounders. Each time a certain number of top biomarkers (denoted as P_m) enter the PGEE model, P_m becomes a tuning parameter and constitute a searching grid together with PGEE penalty parameter λ. By running PGEE throughout all parameter pairs (P_m, λ) in the grid, the best pair would be achieved in terms of the smallest prediction error calculated by 20-fold cross-validation. With the best parameter pair setting, biomarkers with absolute coefficient (β) estimates >0.001 were selected as influential biomarkers.

2.3 Model comparison between PGS and GEE/LMM

Two evaluation matrices were considered in model comparison: phenotype prediction performance and enrichment of phenotype-related biological pathway.

2.3.1 Phenotype prediction performance Smaller prediction error indicates better phenotype prediction performances, and thus higher value of disease diagnosis using the biomarkers selected from a model. To obtain comparable prediction errors between PGS and SBS testing, the 10-fold cross-validation procedure used in PGS was implemented to SBS testing.

2.3.2 Enrichment of phenotype-related biological pathway A higher enrichment of biological pathway suggests that the sites selected from a model is biologically plausible. miRNA pathway enrichment analysis was conducted using DIANA-mirPath v2.0 (Vlachos *et al*., 2012) (http://www.microrna.gr/miRPathv2), a Web-based computational tool incorporating an *in silico* miRNA target prediction tool using high prediction accuracy algorithm DIANA-microT-CDS (Paraskevopoulou *et al*., 2013). Gene union set targeted by at least one selected miRNA was identified. miRNA- and pathway-related information was obtained from miRBase 18 (Kozomara and Griffiths-Jones, 2011) and KEGG v58.1 (Kanehisa *et al*., 2012), respectively. To define a reliable miRNA-gene target prediction, microT score threshold >0.9 was used in DIANA-microT-CDS such that the miRNA-predicted genes were also predicted by miRanda (John *et al*., 2004) and/or TargetScan 5.0 (Friedman *et al*., 2009). For enrichment tests, we applied Fisher's Exact test based on jackknifing the test's probability (Hosack *et al*., 2003), which is more conservative than Fisher's Exact test so that pathways with fewer targeted genes are penalized. The Benjamini and Hochberg FDR (Benjamini and Hochberg, 1995) was calculated to adjust for multiple hypothesis testing. Signaling pathways that have been shown to be associated with lung function were identified (Zander *et al*., 2007). The enrichment score of lung function-related signaling pathway was quantified using negative log10 of the FDR.

2.4 Beijing Truck Driver Air Pollution Study

The BTDAS, conducted between June 15 and July 27, 2008, included participants with high exposure to air pollution in Beijing. All participants were examined on two separate examination days with 1–2 weeks apart. Detailed study design and data measurements were described previously (Baccarelli *et al*., 2011). Lung function was quantified by FEV1 on both examination days. Total RNA was extracted from peripheral blood collected from each participant on both examination days. For better delineation of the direct effects of miRNAs, 10 potential confounders including $PM_{2.5}$, sex, age, BMI, smoking status, usage of central heating, commuting time, working hours per day, dew point and temperature were adjusted in lung function in GEE/LMM SBS testing and PGS.

3 RESULTS

3.1 miRNA profiling data

We conducted miRNA analysis on 240 blood samples collected on two examination days separated by a 1–2 week interval from n = 120 study subjects. Detailed miRNA extraction and profiling data preprocessing can be found in the Supplementary Material.

After normalization and background correction, 568 miRNAs with complete zero expression level across all 240 blood samples were removed, leaving 166 valid miRNAs together with the 10 potential confounders in the final dataset.

3.2 Identification of lung function-related miRNAs with GEE/LMM SBS testing

GEE/LMM SBS testing with the adjustment of 10 confounders were applied to each of the 166 miRNAs. Two widely used *P*-value adjustment methods for multiple comparisons, BH-FDR (Benjamini and Hochberg, 1995) and FDR q-value (Storey, 2002), were calculated to account for multiple testing. Only one miRNA from LMM SBS testing was significant using the conventional significant threshold level of 5% FDR (Supplementary Table S1).

3.3 Identification of lung function-related miRNAs selection with PGS

Before running PGS, all 166 miRNAs were standardized with mean of zero and standard deviation of one. The ranking of miRNA significance was from a univariate prescreening model adjusted for confounders using either GEE or LMM. Starting from the top 10 miRNAs, we increased the number of top P_m miRNAs by an increment of 20, resulting in eight subsets of input miRNAs for PGS ($P_m = 10, 30, \ldots, 150$) and one additional set of whole miRNAs ($P_m = 166$). Penalty parameter λ varied from 0.06 to 0.24 by an increment of 0.02. To evaluate the reliability of the results, we repeated the above procedure for eight times. We found LMM-based PGS were more stable, as six of eight repeats yielded the same 10 selected influential miRNAs related to lung function with $P_m = 110$ and $\lambda = 0.14$, while GEE-based PGS selection results showed less consistency across the eight repeats (Supplementary Table S2). The prediction error grid of LMM-based PGS was represented by a heat map shown in Figure 1.

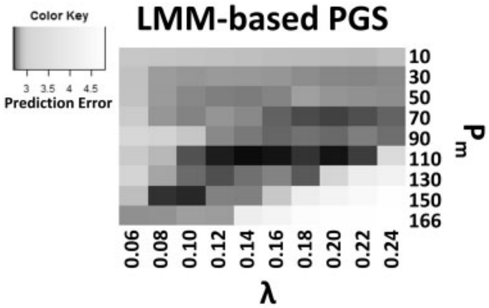

Fig. 1. Heat map of the 20-fold prediction errors from LMM-based PGS. Each gray-scale block represents a PGEE prediction error under the corresponding parameter pair (P_m, λ) in the grid. Best selection results were achieved when incorporating top 110 miRNAs ($P_m = 110$) defined by LMM prescreening model with PGEE penalty parameter $\lambda = 0.14$

3.4 Model comparison between PGS and GEE/LMM SBS testing

3.4.1 Lung function prediction performance For the purpose of model comparison, we selected a sequential number of the top ranking (top 5, 10, 15 and 20) miRNAs based on *P*-values as the identified miRNAs using SBS testing. Comparing with the lowest prediction errors of SBS testing that were achieved using the top 10 miRNAs (mean error = 5.4 for GEE and 6.1 for LMM), paired *t*-tests showed that PGS yielded a significantly lower prediction error (mean error = 5.3) (Table 1).

3.4.2 Enrichment of lung function-related biological pathways We further evaluated enrichment of phenotype-related biological pathways of both PGS and SBS testing. Seven laboratory-proven KEGG signaling pathways related to lung function were identified (Zander *et al.*, 2007). The 10 influential miRNAs selected by LMM-based PGS were used to represent the PGS approach. We also chose the top 10 miRNAs from GEE/LMM SBS testing, respectively, so that the three approaches yielded comparable amount of target genes. miRNAs identified by all three approaches were enriched in mTOR, PI3K-Akt, ErbB, Wnt and MAPK signaling pathways. In general, enrichment score of the genes targeted by miRNAs from PGS was higher than that from GEE SBS testing, and considerably higher than that from LMM SBS testing (Fig. 2), indicating miRNAs selected by PGS were more biologically plausible.

3.5 Simulation study

A simulation study was conducted to compare the performance of PGS and SBS testing in more general settings. Scenario I was comparable with the BTDAS miRNA dataset we studied with $n = 120$ subjects and number of miRNAs $p = 200$. Scenario II had a higher number of miRNAs $p = 800$ but with a smaller sample size of 60. We assumed that there were 5% of miRNAs with true non-zero effects (i.e. 10 for Scenario I and 40 for Scenario II). Corresponding 10 times 10 tuning parameter grids were set, $P_m = (20, 40, \ldots, 200)$ and $\lambda = (0.010, 0.025, \ldots, 0.145)$ in Scenario I and $P_m = (10, 20, \ldots, 100)$ and $\lambda = (0.010, 0.025, \ldots, 0.145)$ in Scenario II. $P_m > 100$ in Scenario II would not yield reliable selection results because of the relatively small sample size. Performance of GEE/LMM SBS testing as well as GEE-based and LMM-based PGS under Scenario I was shown in Table 2. Scatterplot of receiver operating characteristic can be found in Supplementary Figure S1. PGS had >90% chance to identify effect with size as small as 0.3, whereas SBS testing was good till effect size was ~0.6. Moreover, PGS provided more accurate estimates even for small effects. On average, PGS estimates only had 0.01 and 0.03 deviation from the true effects in GEE-based and LMM-based PGS, respectively. PGS gave a much higher sensitivity (>0.80) than SBS testing with comparable specificity (>0.95). For Scenario II, owing to the difficulties in detecting 40 non-zero miRNAs with sample size of 80, both SBS testing and PGS methods did not perform ideally. However, there was still significant gain from PGS over SBS testing (Supplementary Table S3 and Supplementary Fig. S1).

Table 1. Comparison of phenotype prediction performance between GEE/LMM SBS testing and PGS

Method	Mean prediction errors	
	GEE	LMM
SBS testing		
Top 5 miRNAs	5.5	5.6
Top 10 miRNAs	5.4	6.1
Top 15 miRNAs	5.7	6.2
Top 20 miRNAs	5.9	6.1
PGS (LMM-based)	5.3	
P-value	< 0.001	< 0.001

Note: Mean prediction errors were computed by 10-fold cross-validation errors from 50 repeats. The 10 influential miRNAs selected by LMM-based PGS were used to calculate the prediction errors. *P*-values were calculated by paired *t*-tests between the lowest prediction errors from GEE/LMM using top 10 miRNAs and the prediction errors from PGS using the selected 10 influential miRNAs.

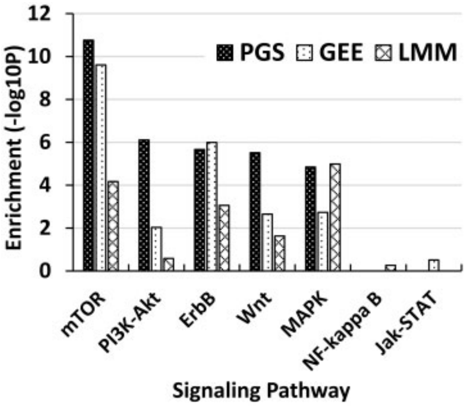

Fig. 2. Enrichment score of seven laboratory-proven KEGG signaling pathways related to lung function. PGS selection results were from LMM-based PGS

Table 2. Performances of PGS and SBS testing in simulation study with 200 miRNAs, $n = 120$ subjects and 10 non-zero true effects (Scenario I)

True effects	SBS number of selections	SBS mean estimates	PGS number of selections	PGS mean estimates
GEE SBS and GEE-based PGS				
0.1	1	0.73	4	0.10
–0.2	12	–0.42	53	–0.18
0.3	27	0.48	93	0.27
–0.4	57	–0.46	98	–0.38
0.5	76	0.55	100	0.50
–0.6	95	–0.62	100	–0.59
0.7	99	0.70	100	0.69
–0.8	99	–0.82	100	–0.80
0.9	100	0.89	100	0.89
–1.0	100	–1.00	100	–0.99
Sensitivity	0.67		0.85	
Specificity	0.95		0.95	
RMSE		0.22		0.01
LMM SBS and LMM-based PGS				
0.1	3	0.37	7	0.16
–0.2	5	–0.47	70	–0.15
0.3	29	0.46	94	0.27
–0.4	51	–0.50	100	–0.39
0.5	81	0.52	100	0.49
–0.6	95	–0.59	100	–0.60
0.7	100	0.71	100	0.69
–0.8	100	–0.81	100	–0.80
0.9	100	0.91	100	0.89
–1.0	100	–0.98	100	–0.98
Sensitivity	0.66		0.87	
Specificity	0.96		0.95	
RMSE		0.14		0.03

Note: Mean estimates were calculated from the 100 simulation runs, respectively. Number of selections were the number of times each true effect selected by models out of the 100 runs. Sensitivity and specificity for each method were the average for all 100 runs. RMSE is root mean square error between mean estimates and true effects, which describes mean accuracy of the model estimates.

4 DISCUSSION

In this study, we proposed and applied PGS method to handle high-dimensional miRNA expression data with repeated measures. We compared performances of the PGS and the traditional SBS testing. PGS performed consistently better than GEE/LMM SBS testing in terms of higher phenotype prediction performance, higher enrichment of phenotype-related biological pathway, more accurate estimates, and higher sensitivity. One of the regularity conditions for PGEE algorithm is that $p = O(n)$, which requires the number of predictors (p) in model, should be comparable with the sample size (n). To ensure the tuning penalty parameter (λ) yields an overall effect size estimation that best fits the data and avoid exceeded number of input biomarkers (P_m) for PGEE algorithm, we incorporated the grid search method. A 20-fold cross-validation was implemented to determine the best PGEE selection results in terms of the smallest prediction error among the parameter searching grid (P_m, λ). Though PGS has a relatively high computational intensity, the most common and systematic way, as we have done in current study, is through the cross-validation, which is well documented in other established penalized models LASSO and elastic net (Zou and Hastie, 2005). PGS is a data-driven and self-training analytical framework that can achieve maximum data utilization while constraining model complexity simultaneously.

One key merit of PGS is its capacity to handle a multitude of biomarkers at the same time and select the influential ones. Using SBS testing, we only found few ‘significant’ results after having corrected for a high number of multiple testing. While using PGS, we identified a set of influential and meaningful miRNAs without encountering multiple testing issues, which offers a novel perspective in analyzing high-dimensional data with repeated measures. Another distinct advantage of PGS is that it considers all informative biomarkers as a whole instead of treating them individually. Using the SBS testing, although some miRNAs successfully pass the multiple testing, it is possible that meaningful biological events might not even be detected because of the correlations and interactions among miRNAs. However, the effects of these biomarkers could be significant when modeled together. PGS is able to capture these complex features across all input biomarkers and detect influential ones that SBS testing would potentially ignore.

It is worth noting that PGS does not provide exact *P*-values, as the estimates do not follow normal distribution under the null hypotheses. Therefore, the criterion for determining influential biomarker in PGS is not based on *P*-value, but on the threshold for coefficient (β) estimates of biomarkers. Influential biomarkers are selected when the estimates are greater than the threshold. As suggested in Wang's paper (Wang *et al.*, 2012), the default threshold is 0.001, which is also proved to be a practical and robust threshold in our simulation study in terms of sensitivity and specificity. To ensure that the threshold of 0.001 works uniformly for any dataset, it is required to standardize each of the biomarkers with mean of zero and standard deviation of one as a typical procedure before running penalize model.

Robust selection of biomarkers by PGS relies on the setting of grid boundary and grid resolution. Too large λ would shrink all β estimates to zero, whereas too small λ would not shrink the estimates at all. Our numerical experience in similar settings suggests that a range of λ from 0.01 to 0.30 should cover the optimal λ for most of the cases. Within this range, an increment of 0.02 in λ provides a proper grid resolution to capture subtle effect changes of λ on biomarker selection results without much extra computational burden. On the other hand, a too large P_m might defy the application of PGS, whereas a too small P_m might underexploit the data. An optimal P_m could be found around the sample size, as is the case in our miRNA study example, where the optimal $P_m = 110$. Besides, an increment of 10 or 20 in P_m provides a sufficient resolution for capturing the effects of increasing P_m on biomarker selection. For all practical usage, the initiation of λ vector can be varied from 0.01 to 0.30, and the initiation of P_m can be a vector with a few numbers varied around the sample size minus the number of confounders (in our case, it is $120 - 10 = 110$). It is not necessary to initiate a full vector of P_m, as it adds redundant parameter pairs to the grid. Extension of λ boundary and/or P_m boundary will be considered only when the optimal λ and/or P_m hit the initial boundaries. In this article, we used different elaborate λ ranges (but all were within the 0.01–0.30 range) for better results representations, and higher grid resolution (i.e. smaller increment of λ) for more reliable method evaluations in simulation studies.

Insensitive to mis-specification of the covariance structure is a feature that distinguishes GEE from LMM. Inheriting this feature from GEE, PGS is able to estimate the correlation matrix regardless of whether the structure is specified. Based on the estimation, one could either use LMM-based PGS with a solid guess of the covariance structure or use GEE-based PGS if there is no good choice of the structure.

Prescreening step prioritizes potentially influential biomarkers, which facilitates PGS to handle the situation in which the number of biomarkers is considerably larger than the sample size. GEE and LMM are two handy approaches for prescreening. However, the potential limitation of prescreening is that miRNAs with small but true non-zero effects may be excluded during the prescreening step. Nevertheless, with a given sample size, the grid search method in PGS ensures that PGEE can include as many biomarkers as possible.

5 CONCLUSION

The performance of PGS is comparable with the approaches being benchmarked, i.e. SBS GEE/LMM. However, PGS is more suitable for high-dimensional miRNA expression data with repeated measures in that, by exploiting underlying dependent structures, it relies on variable selection in the context of a multiple regression model, which circumvents multiple testing issues. PGS is also applicable to other longitudinally collected high-dimensional quantitative data, such as epigenomics, mRNA transcriptomics, proteomics, metabolomics, etc. The growing number of studies conducting high-dimensional profiling dataset using different platforms requires a more comprehensive evaluation of PGS in various study settings.

ACKNOWLEDGEMENTS

Many thanks to Dr. Lan Wang and Jianhui Zhou for kindly providing the R source code of PGEE and for their many thoughtful suggestions in developing PGS.

Funding: This work was supported by R21 ES020010 and R21 ES020984-01 from the National Institute of Environmental Health Sciences (NIEHS).

Conflicts of Interest: none declared.

REFERENCES

Baccarelli,A. *et al.* (2011) Effects of particulate air pollution on blood pressure in a highly exposed population in Beijing, China: a repeated-measure study. *Environ. Health*, **10**, 108.

Baek,D. *et al.* (2008) The impact of microRNAs on protein output. *Nature*, **455**, 64–71.

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

Byun,H.M. *et al.* (2013) Effects of airborne pollutants on mitochondrial DNA methylation. *Part. Fibre Toxicol.*, **10**, 18.

Chen,C.L. *et al.* (2013) Changes in microRNA expression profile in hippocampus during the acquisition and extinction of cocaine-induced conditioned place preference in rats. *J. Biomed. Sci.*, **20**, 96.

Friedman,R.C. *et al.* (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.

Guo,L. *et al.* (2014) Effects of short-term exposure to inhalable particulate matter on DNA methylation of tandem repeats. *Environ. Mol. Mutagen*, **55**, 322–335.

Hamm,C.A. *et al.* (2010) Microenvironment alters epigenetic and gene expression profiles in Swarm rat chondrosarcoma tumors. *BMC Cancer*, **10**, 471.

Hecker,M. *et al.* (2013) MicroRNA expression changes during interferon-beta treatment in the peripheral blood of multiple sclerosis patients. *Int. J. Mol Sci.*, **14**, 16087–16110.

Henderson,C.R. *et al.* (1959) The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **15**, 192–218.

Hosack,D.A. *et al.* (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.

Hou,L. *et al.* (2012) Air pollution exposure and telomere length in highly exposed subjects in Beijing, China: a repeated-measure study. *Environ. Int.*, **48**, 71–77.

Hou,L. *et al.* (2013a) Inhalable particulate matter and mitochondrial DNA copy number in highly exposed individuals in Beijing, China: a repeated-measure study. *Part. Fibre Toxicol.*, **10**, 17.

Hou,L. *et al.* (2013b) Altered methylation in tandem repeat element and elemental component levels in inhalable air particles. *Environ. Mol. Mutagen*, **55**, 256–265.

John,B. *et al.* (2004) Human MicroRNA targets. *PLoS Biol.*, **2**, e363.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.

Paraskevopoulou,M.D. *et al*. (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res*., **41**, W169–W173.

Rakyan,V.K. *et al*. (2011) Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet*., **12**, 529–541.

Reinsbach,S. *et al*. (2012) Dynamic regulation of microRNA expression following interferon-gamma-induced gene transcription. *RNA Biol*., **9**, 978–989.

Selbach,M. *et al*. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.

Singh,S.K. *et al*. (2008) MicroRNAs–micro in size but macro in function. *FEBS J*., **275**, 4929–4944.

Storey,J. (2002) A direct approach to false discovery rates. *J. Royal Stat. Soc. Series B Stat. Methodol*., **64**, 479–498.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Vlachos,I.S. *et al*. (2012) DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways. *Nucleic Acids Res*., **40**, W498–W504.

Wang,L. *et al*. (2012) Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, **68**, 353–360.

Zander,D.S. *et al*. (2007) *Molecular Pathology of Lung Diseases*. Springer, New York, NY, U.S.A.

Zeger,S.L. *et al*. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. Royal Stat. Soc.: Series B Statist. Methodol*., **67**, 301–320.