Li et al. BMC Genetics 2018, 19(Suppl 1):67
https://doi.org/10.1186/s12863-018-0638-3

BMC Genetics

RESEARCH Open Access

Indirect effect inference and application to GAW20 data

Liming Li[1], Chan Wang[1], Tianyuan Lu[1], Shili Lin[2] and Yue-Qing Hu[1*]

From Genetic Analysis Workshop 20
San Diego, CA, USA. 4-8 March 2017

Abstract

Background: Association studies using a single type of omics data have been successful in identifying disease-associated genetic markers, but the underlying mechanisms are unaddressed. To provide a possible explanation of how these genetic factors affect the disease phenotype, integration of multiple omics data is needed.

Results: We propose a novel method, LIPID (likelihood inference proposal for indirect estimation), that uses both single nucleotide polymorphism (SNP) and DNA methylation data jointly to analyze the association between a trait and SNPs. The total effect of SNPs is decomposed into direct and indirect effects, where the indirect effects are the focus of our investigation. Simulation studies show that LIPID performs better in various scenarios than existing methods. Application to the GAW20 data also leads to encouraging results, as the genes identified appear to be biologically relevant to the phenotype studied.

Conclusions: The proposed LIPID method is shown to be meritorious in extensive simulations and in real-data analyses.

Keywords: Epigenetics, Differentially methylated regions, DNA methylation

Background

In complex disease studies, genome-wide association studies (GWAS) [1] and epigenome-wide association studies [2] have been successful in identifying disease-associated single-nucleotide polymorphisms (SNPs) and DNA methylation loci. However, the mechanism of how these genetic loci affect the disease status remains unknown. To provide a possible explanation of the causal mechanisms of these genetic factors, integrative analyses using both types of data are important. Even though integration of multiple types of data sets is a promising method as it is generally more powerful than ordinary association studies [3], the method of integration itself is challenging.

Most existing methods use additional information to filter out nonsignificant loci and reduce the total number of tests, which, in return, improve power [4]. On the other hand, mediation analyses usually consider only a single mediator and require multiple testing correction [5]. An example is Zhao et al. [6] who proposed an integrative test, denoted as o-eSNP that was shown to be more powerful than traditional GWAS. Motivated by the data provided by GAW20, in which SNP and DNA methylation data for integrative analysis are available, we aim to characterize the effects of SNPs into direct and indirect effects. As the direct effect of SNPs is simple and straightforward, in this contribution we focus on the indirect effect of SNPs. With the prior knowledge that DNA methylation can be modulated by SNPs [7], we assume that some SNPs exert their effects by regulating the DNA methylation level. Hence the indirect effect of SNPs on a phenotype of interest here is taken as the combined effects of SNPs on DNA methylation and DNA methylation on the phenotype.

In this paper, we propose a novel method, LIPID (likelihood inference proposal for indirect estimation), to use both SNP and DNA methylation data to test whether there is an indirect effect of SNPs on a phenotype. The indirect effect and its variance–covariance matrix are derived, and a Wald test is conducted. An

* Correspondence: yuehu@fudan.edu.cn
[1]State Key Laboratory of Genetic Engineering, Institute of Biostatistics, School of Life Sciences, Fudan University, 2005 Songhu Road, Shanghai 200438, China
Full list of author information is available at the end of the article

extensive simulation study was done to evaluate the properties and the performance of LIPID, which was also applied to analyze the GAW20 real data.

Methods

Suppose there are n independent subjects, and for each subject, its SNP, DNA methylation, covariates, and phenotype are measured. Specifically, let $Y = (Y_1,...,Y_n)^T$ be the vector of observed phenotypes; $X = (X_1,...,X_k)$ be the $n \times k$ matrix denoting the observed values of k (non-genetic) covariates, including intercept; $S = (S_1,..., S_r)$ and $M = (M_1,..., M_p)$ be the $n \times r$ and $n \times p$ matrices regarding the genotypes of r SNPs and methylation levels of p cytosine-phosphate-guanine (CpG) sites, respectively. Assuming that phenotype Y is a continuous variable, we can use a set of linear models to capture the relationship among Y, X, S, and M as follows:

$$\mathrm{Y} = \mathrm{S}\alpha_\mathrm{S} + \mathrm{M}\alpha_\mathrm{M} + \mathrm{X}\alpha_\mathrm{X} + \varepsilon_\mathrm{Y}, \quad (1)$$

$$\mathrm{M} = \mathrm{S}\beta_\mathrm{S} + \mathrm{X}\beta_\mathrm{X} + \varepsilon_\mathrm{M}, \quad (2)$$

where $\varepsilon_Y \sim N(0, \sigma_Y^2 I_n)$, $vec(\varepsilon_M) \sim N(\Sigma_M \otimes I_n)$, $vec(\cdot)$ is the vectorization operation; $\otimes$ is the Kronecker product; and ε_Y and ε_M are independent. Note that here β_S and β_X are $r \times p$ and $k \times p$ matrices respectively, and Σ_M is a $p \times p$ positive definite matrix. Model (1) characterizes the relationship between phenotype and SNPs, DNA methylation, and covariates, while model (2) depicts the relationship between DNA methylation (as a response variable) and SNPs and covariates. It is concluded from models (1) and (2) that the direct effect of SNPs on the phenotype is α_S and the indirect effect is $\gamma = \beta_\mathrm{S}\alpha_\mathrm{M}$.

Estimation and inference

For linear models (1) and (2), we have the following maximum likelihood estimates

$$\hat{\alpha} = \left(G_1^T G_1\right)^{-1} G_1^T Y, \hat{\beta} = \left(G_2^T G_2\right)^{-1} G_2^T M$$

where $G_1 = (S, M, X), G_2 = (S, X), \ \alpha = (\alpha_S^T, \alpha_M^T, \alpha_X^T)^T$, and $\beta = (\beta_S^T, \beta_X^T)^T$. The variance–covariance matrices for $\hat{\alpha}$ and vec $(\hat{\beta})$ are, respectively:

$$Cov(\hat{\alpha}) = \left(G_1^T G_1\right)^{-1} \sigma_Y^2$$

$$Cov(vec(\hat{\beta})) = \left(G_2^T G_2\right)^{-1} \otimes \Sigma_M$$

Their corresponding block matrices are the variance–covariance matrices for $\hat{\alpha}_M$ and vec $(\widehat{\beta_S})$, respectively. According to the law of total variation, the variance–covariance matrix for $\hat{\gamma}$ is

$$\begin{aligned} Cov(\hat{\gamma}) &= Cov\left(\widehat{\beta_S \alpha_M}\right) \\ &= Cov\left(E\left(\hat{\beta}_S \hat{\alpha}_M | \hat{\beta}_S\right)\right) + E\left(Cov\left(\hat{\beta}_S \hat{\alpha}_M | \hat{\beta}_S\right)\right) \\ &= \alpha_M^T \Sigma_M \alpha_M \left(G_2^T G_2\right)_{11}^{-1} + \beta_S Cov(\hat{\alpha}_M) \beta_S^T \\ &\quad + tr(\Sigma_M Cov(\hat{\alpha}_M)) \left(G_2^T G_2\right)_{11}^{-1} \end{aligned}$$

where $(\cdot)_{11}$ represents the first $r \times r$ diagonal submatrix. As α_M and β_S are unavailable, their estimates $\hat{\alpha}_M$ and $\hat{\beta}_S$ are used. After several lines of algebra we have

$$E\left(\hat{\alpha}_M^T \hat{\Sigma}_M \hat{\alpha}_M\right) = \alpha_M^T \Sigma_M \alpha_M + tr(\Sigma_M Cov(\hat{\alpha}_M))$$

$$\begin{aligned} E\left(\hat{\beta}_S Cov(\hat{\alpha}_M) \hat{\beta}_S^T\right) &= \beta_S Cov(\hat{\alpha}_M) \beta_S^T \\ &\quad + tr(\Sigma_M Cov(\hat{\alpha}_M)) \left(G_2^T G_2\right)_{11}^{-1} \end{aligned}$$

So an unbiased estimate for the variance–covariance matrix of $\hat{\gamma}$ is

$$\begin{aligned} \hat{cov}(\hat{\gamma}) &= \hat{\alpha}_M^T \hat{\Sigma}_M \hat{\alpha}_M \left(G_2^T G_2\right)_{11}^{-1} \\ &\quad + \hat{\beta}_S \hat{Cov}(\hat{\alpha}_M) \hat{\beta}_S^T - tr\left(\hat{\Sigma}_M \hat{Cov}(\hat{\alpha}_M)\right) \left(G_2^T G_2\right)_{11}^{-1}, \end{aligned}$$

where $\hat{\Sigma}_M$ *and* $\hat{Cov}(\hat{\alpha}_M)$ are corresponding estimates of Σ_M and $Cov(\hat{\alpha}_M)$. This estimate is different from o-eSNP [6] in the last component. The Wald statistic to test if indirect effect exists is

$$LIPID = \hat{\gamma}^T \hat{Cov}(\hat{\gamma})^{-1} \hat{\gamma},$$

which asymptotically follows a chi-squared distribution with r degrees of freedom.

Adaptation to correlated subjects

For subjects with correlation between each other, linear model with mixed effect is utilized. So model (1) is changed to

$$\mathrm{Y} = S\alpha_\mathrm{S} + \mathrm{M}\alpha_\mathrm{M} + \mathrm{X}\alpha_\mathrm{X} + \mathrm{b} + \varepsilon_\mathrm{Y}$$

where b is the random effect, with mean 0 and variance–covariance matrix $2\sigma_b^2\Phi$, where Φ is the kinship coefficient matrix, and $b + \varepsilon_\mathrm{Y} \sim N(0, \sigma_Y^2 I_n + 2\sigma_b^2\Phi)$. Model (2) is not changed. The estimate α_M and its variance–covariance matrix are derived similarly, and the LIPID statistic has the same form.

Simulation study

To evaluate the performance of LIPID, simulation under various scenarios are conducted. For simplicity, we assume there are no covariates, and that S, M, and Y are all univariate. In addition, the direct effect that we are not interested in does not exist in simulation. The simulated data are generated as follows. First, SNP S is generated with a minor allele frequency (MAF) under Hardy-Weinberg equilibrium. Then, DNA methylation M is generated from a normal distribution with mean $S\beta_S$ and variance σ_M^2. Finally, phenotypic value Y is generated from a normal distribution with mean $M\alpha_M$ and variance σ_Y^2. The number

of individuals is set to be 100 and the number of replications is 10,000. As Table 1 shows, there are 5 scenarios of parameters designed to gauge the Type I error rates of LIPID, and 5 scenarios for the evaluation of power. The variance of Y and variance of M are fixed to 1 for simplicity. In scenario 1, the coefficients β_S and α_M are both 0; in scenarios 2 and 3, β_S is nonzero while α_M is 0; in scenarios 4 and 5, α_M is nonzero with β_S equal to 0. In these scenarios, the indirect effect is nonexistent, so we change the coefficient of one parameter to measure the Type I error rates under different situations. Under H_a, the indirect effect is $\beta_S\ \alpha_M \neq 0$. The coefficients are chosen so that different methods have moderate powers. From scenario 1 to scenario 5, the β_S and α_M are increased. The MAF ranges from 0.1 to 0.4.

Results

Table 2 shows the Type I error rates in the 5 scenarios, from which we can see that the Type I error rates are more or less conservative in scenarios 1 to 5 for o-eSNP and LIPID, while regressing on SNPs only (denoted as SNP in Table 2) controls the Type I error rate well. For scenario 1, where both coefficients β_S and α_M are 0, the Type I error rate is conservative; for scenarios with 1 coefficient that is not 0 but relatively small, the Type I error rates are still conservative; for scenarios with 1 large coefficient, the Type I error rates are better controlled. Compared to o-eSNP [6], the Type I error rates of LIPID are favorable, as o-eSNP is more conservative in all scenarios. Figure 1 shows the powers of 3 methods in 5 scenarios. We can see that LIPID is the most powerful in all scenarios, while SNP-only method has the least power. From scenario 1 to scenario 5, as the indirect effect increases, the performance of LIPID and o-eSNP are very close to each other.

Real data analysis

The GAW20 real data package contains genomic, DNA methylation, and phenotypic data for more than 1000 individuals from 188 pedigrees. The phenotypic data include metabolic indices, lipoproteins, and triglyceride. Dense genome-wide SNP markers make up the genomic data. DNA methylation levels are also available on CpG sites genome-wide before and after individuals are treated with fenofibrate. The level of triglycerides and the methylation level before treatment are made use of. The covariates include gender, age, smoking status, high-density lipoprotein, metabolic disorder, and center. We use SOLAR (Sequential Oligogenic Linkage Analysis Routines) [8] to obtain the heritability for triglyceride level, using 1108 subjects with phenotypic data. The total number of subjects with SNP data and DNA methylation is 716. As LIPID considers a region of multiple genetic markers, SNPs and DNA methylation loci within the range of each gene are analyzed; we analyze a total of 13,968 genes. Genes that pass false discovery rate (FDR) correction are *FAT1* (p value 9.4E-7) and *DCTN6* (p value 1.3E-6), while o-eSNP fails to find any significant genes (*FAT1* p value 2.5E-5; *DCTN6* p value 3.7E-6). We further use the BIOS QTL (quantitative trait locus) browser [9] to validate our findings. We found that rs458021 on gene *FAT1* is a *cis*-meQTL (methylation quantitative trait locus) with a p value of 3.8E-07, but we did not find any meQTL on gene *DCTN6*. *FAT1* is associated with cholesterol in DAVID (Database for Annotation, Visualization, and Integrated Discovery), whereas *DCTN6* is involved in lipid metabolism [10]. Because the eQTM (expression quantitative trait methylation) database are not widely available, we cannot further validate if these CpG sites on these genes can modulate the expression, and further influence the phenotype.

Table 1 Parameter settings under H_0: $\gamma = \beta_S\alpha_M = 0$ and H_a: $\gamma = \beta_S\alpha_M \neq 0$

Hypothesis	Parameter	Scenario				
		1	2	3	4	5
H_0	β_S	0	0.4	1	0	0
	α_M	0	0	0	0.4	1
H_a	β_S	0.2	0.3	0.2	0.3	0.4
	α_M	0.4	0.4	0.6	0.6	0.6

Table 2 Type I error rates of 3 methods in scenarios 1 to 5

MAF	Method	Scenario				
		1	2	3	4	5
0.1	o-eSNP	0.000	0.003	0.028	0.025	0.048
	LIPID	0.000	0.004	0.032	0.029	0.049
	SNP	0.054	0.054	0.056	0.051	0.048
0.2	o-eSNP	0.000	0.007	0.039	0.024	0.053
	LIPID	0.000	0.009	0.042	0.028	0.054
	SNP	0.056	0.053	0.053	0.056	0.054
0.3	o-eSNP	0.000	0.009	0.041	0.027	0.050
	LIPID	0.000	0.011	0.044	0.032	0.051
	SNP	0.059	0.054	0.052	0.053	0.057
0.4	o-eSNP	0.000	0.011	0.043	0.021	0.048
	LIPID	0.000	0.013	0.044	0.025	0.049
	SNP	0.052	0.056	0.056	0.051	0.054

The MAF changes from 0.1 to 0.4

Discussion

Compared to o-eSNP [6], LIPID controls Type I error rates better in all scenarios, and the power is higher

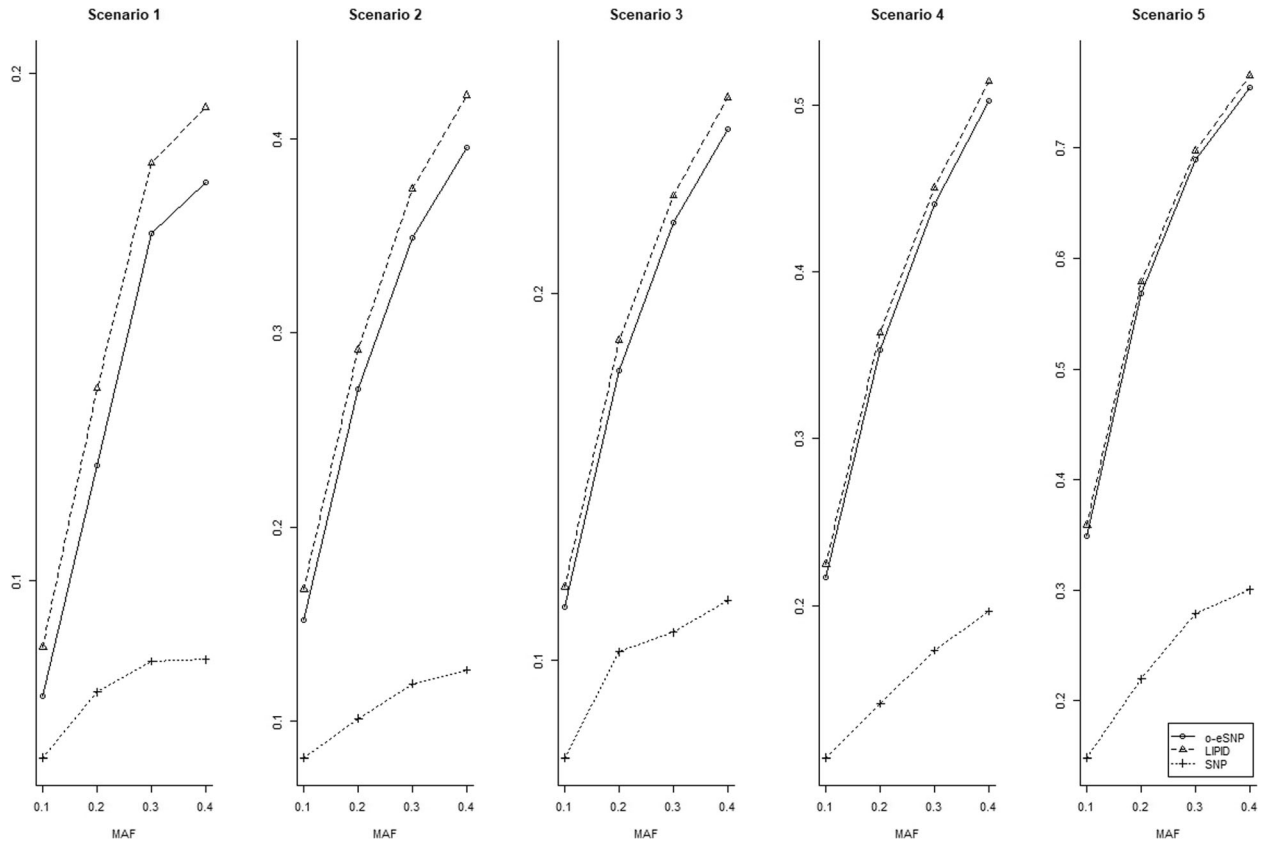

Fig. 1 Power for 3 methods under 5 scenarios. The o-eSNP, LIPID, and SNP methods are denoted with ○, Δ, and +, respectively. The MAF ranges from 0.1 to 0.4

in all scenarios. The estimate itself is the same, no matter if we regress M or $M\alpha_M$ on S, but the variance–covariance estimates are different, and LIPID has a less-biased variance–covariance estimate, which leads to the improvement of performance of LIPID. Furthermore, we also adapt o-eSNP and LIPID to correlated subjects.

Application of LIPID to the GAW20 real data indicates that LIPID is capable of detecting genes with indirect effect. The computation is efficient, and the process takes 30 min to analyze the whole GAW20 data set on a personal computer with an Intel Core i3–4150 CPU. The genes identified appear to be functionally relevant to the trait being considered, thereby substantiating the importance of these findings and leading to confidence of genes found being true discoveries.

Conclusions

For complex diseases, we propose a novel method to detect indirect effect of SNPs on a phenotype via methylation, and we demonstrate its superiority compared to 2 existing methods. LIPID is single-step and does not require multiple tests, compared to traditional mediation analysis; at the same time, multiple genetic loci can be used simultaneously to test indirect effect.

Abbreviations

DAVID: Database for annotation, visualization, and integrated discovery; eQTM: Expression quantitative trait methylation; FDR: False discovery rate; GAW: Genetic analysis workshop; LIPID: Likelihood inference proposal for indirect estimation; MAF: Minor allele frequency; meQTL: Methylation quantitative trait locus; QTL: Quantitative trait locus; SOLAR: Sequential oligogenic linkage analysis routines

Funding

Publication of this article was supported by NIH R01 GM031575. This work was supported in part by National Natural Science Foundation of China (11571082, 11171075), Key Research Project of the Ministry of Science and Technology of China (2016YFC0904400), the Scientific Research Foundation of Fudan University.

Availability of data and materials

The data that support the findings of this study are available from the Genetic Analysis Workshop (GAW) but restrictions apply to the availability of these data, which were used under links for the current study. Qualified researchers may request these data directly from GAW.

About this supplement

This article has been published as part of *BMC Genetics* Volume 19 Supplement 1, 2018: Genetic Analysis Workshop 20: envisioning the future of statistical genetics by exploring methods for epigenetic and pharmacogenomic data. The full contents of the supplement are available

online at https://bmcgenet.biomedcentral.com/articles/supplements/volume-19-supplement-1.

Authors' contributions
LL, CW, and TL conceived the study and developed the statistical methods; LL and YQH conducted the data analysis; LL and YQH wrote the paper; SL and YQH reviewed the paper. All authors have read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]State Key Laboratory of Genetic Engineering, Institute of Biostatistics, School of Life Sciences, Fudan University, 2005 Songhu Road, Shanghai 200438, China. [2]Department of Statistics, The Ohio State University, 1958 Neil Avenue, 404 Cokins Hall, Columbus, OH 43210, USA.

Published: 17 September 2018

References
1. Klein RGAWJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005; 308(5720):385–9.
2. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011;12(8):529–41.
3. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, et al. Variations in DNA elucidate molecular networks that cause disease. Nature. 2008;452(7186):429–35.
4. Ware JS, Petretto E, Cook SA. Integrative genomics in cardiovascular medicine. Cardiovasc Res. 2013;97(4):623–30.
5. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: conceptual strategic and statistical considerations. J Pers Soc Psychol. 1986;51(6):1173–82.
6. Zhao SD, Cai TT, Li H. More powerful genetic association testing via a new statistical framework for integrative genomics. Biometrics. 2014;70(4):881–90.
7. Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, Hod E, Li K, Murty VV, Schupf N, Vilain E, et al. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. Nat Genet. 2008;40(7):904–8.
8. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. 1998;62(5):1198–211.
9. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, van't Hof P, Mei H, van Dijk F, Westra H-J, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017;49(1):139–45.
10. Yang S, Chen C, Wang H, Rao X, Wang F, Duan Q, Chen F, Long G, Gong W, Zou M-H, et al. Protective effects of acyl-CoA thioesterase 1 on diabetic heart via PPARα/PGC1α signaling. PLoS One. 2012;7(11):50376.