BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 19 2012, pages 2540–2542
doi:10.1093/bioinformatics/bts474

Genetics and population analysis

Advance Access publication July 26, 2012

Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood

S.H. Lee[1,*], J. Yang[2], M.E. Goddard[3], P.M. Visscher[1,2] and N.R. Wray[1]

[1]The University of Queensland, Queensland Brain Institute, Brisbane, QLD 4072, [2]The University of Queensland Diamantina Institute, Princess Alexandra Hospital, Brisbane, QLD 4102 and [3]Department of Agriculture and Food Systems, University of Melbourne, VIC 3010, Melbourne, Australia

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Genetic correlations are the genome-wide aggregate effects of causal variants affecting multiple traits. Traditionally, genetic correlations between complex traits are estimated from pedigree studies, but such estimates can be confounded by shared environmental factors. Moreover, for diseases, low prevalence rates imply that even if the true genetic correlation between disorders was high, co-aggregation of disorders in families might not occur or could not be distinguished from chance. We have developed and implemented statistical methods based on linear mixed models to obtain unbiased estimates of the genetic correlation between pairs of quantitative traits or pairs of binary traits of complex diseases using population-based case–control studies with genome-wide single-nucleotide polymorphism data. The method is validated in a simulation study and applied to estimate genetic correlation between various diseases from Wellcome Trust Case Control Consortium data in a series of bivariate analyses. We estimate a significant positive genetic correlation between risk of Type 2 diabetes and hypertension of ~0.31 (SE 0.14, $P = 0.024$).

Availability: Our methods, appropriate for both quantitative and binary traits, are implemented in the freely available software GCTA (http://www.complextraitgenomics.com/software/gcta/reml_bivar.html).

Contact: hong.lee@uq.edu.au

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

Received on May 20, 2012; revised on July 18, 2012; accepted on July 20, 2012

1 INTRODUCTION

Recently, we have developed new methods to estimate the proportion of variation in quantitative traits (Yang *et al.*, 2010, 2011) or in liability to disease that is associated with single-nucleotide polymorphisms (SNPs) (Lee *et al.*, 2012, 2011). The methods use very distant relationships between individuals so that estimates are unlikely to be confounded with shared family environment effects. The methodology can be extended to estimation of the genetic covariance and hence genetic correlation between different disorders that is tagged by SNPs to provide estimates of genome-wide pleiotropy. Evidence for a genetic correlation between disorders estimated directly by interrogation of the genome could have an important impact on the design of future genetic and functional studies for medical nosology and may provide new insights for novel treatments across disorders.

The aim of this study is to estimate genome-wide pleiotropy using genome-wide association studies (GWAS) case–control data for different diseases or disorders. For binary disease traits, we derive valid statistical approaches to obtain unbiased estimates of comorbidity interpretable on the scale of liability to disease. We develop computationally efficient algorithms for estimation. The method is applied to estimate the genetic correlation between hypertension (HT) and type 2 diabetes (T2D), bipolar disorder (BD) and rheumatoid arthritis (RA), BD and T2D or HT and RA from Wellcome Trust Case Control Consortium (WTCCC) GWAS data.

2 METHODS

2.1. Bivariate linear mixed model and efficient AIREML

We used a standard bivariate linear mixed model (Thompson, 1973). The models can be written as

$$\mathbf{y}_1 = \mathbf{X}_1\mathbf{b}_1 + \mathbf{Z}_1\mathbf{g}_1 + \mathbf{e}_1 \text{ for trait 1 and } \mathbf{y}_2 = \mathbf{X}_2\mathbf{b}_2 + \mathbf{Z}_2\mathbf{g}_2 + \mathbf{e}_2 \text{ for trait 2,}$$

where **y** is a vector of observations for trait, $\mathbf{b}_1$ and $\mathbf{b}_2$ are vectors of fixed effects, $\mathbf{g_1}$ and $\mathbf{g_2}$ are vectors of random polygenic effects for each individual in both trait 1 and 2 and $\mathbf{e_1}$ and $\mathbf{e_2}$ are residuals for trait 1 and 2, respectively. **X** and **Z** are incidence matrices for the effects **b** and **g**, respectively. The variance covariance matrix (**V**) is defined as

$$\mathrm{V} = \begin{bmatrix} \mathbf{Z}_1\mathbf{A}\mathbf{Z'}_1\sigma^2_{g_1} + \mathrm{I}\sigma^2_{e_1} & \mathbf{Z}_1\mathbf{A}\mathbf{Z'}_2\sigma^2_{g_1g_2} \\ \mathbf{Z}_2\mathbf{A}\mathbf{Z'}_1\sigma^2_{g_1g_2} & \mathbf{Z}_2\mathbf{A}\mathbf{Z'}_2\sigma^2_{g_2} + \mathrm{I}\sigma^2_{e_2} \end{bmatrix},$$

where **A** is the genomic similarity relationship matrix based on SNP information (Yang *et al.*, 2010) and **I** is an identity matrix, σ^2_g, σ^2_e and $\sigma^2_{g_1g_2}$, which are genetic variance, residual variance and covariance between $\mathbf{g_1}$ and $\mathbf{g_2}$. Lee and Van der Werf (2006) showed that the method of average information (**AI**) matrices derived directly from the **V** is much more efficient computationally than the original AI algorithms (Gilmour *et al.*, 1995; Johnson and Thompson, 1995). Following equation (8) in Lee and Van der Werf (2006), the **AI** matrix for the bivariate model can be derived as

*To whom correspondence should be addressed.

$$\mathrm{AI} = \frac{1}{2}\begin{bmatrix} \mathrm{y'I_1PI_1PPy} & & & & \\ \mathrm{y'I_1PI_2PPy} & \mathrm{y'I_2PI_2PPy} & & & \\ \mathrm{y'I_1PG_1PPy} & \mathrm{y'I_2PG_1PPy} & \mathrm{y'G_1PG_1PPy} & & \\ \mathrm{y'I_1PG_2PPy} & \mathrm{y'I_2PG_2PPy} & \mathrm{y'G_1PG_2PPy} & \mathrm{y'G_2PG_2PPy} & \\ \mathrm{y'I_1PCPPy} & \mathrm{y'I_2PCPPy} & \mathrm{y'G_1PCPPy} & \mathrm{y'G_2PCPPy} & \mathrm{y'CPCPPy} \end{bmatrix}$$

where $\mathrm{P} = \mathrm{V}^{-1} - \mathrm{V}^{-1}\mathrm{X}(\mathrm{X'V}^{-1}\mathrm{X})^{-1}\mathrm{X'V}^{-1}$, $\mathbf{y}' = [\mathbf{y}'_1\ \mathbf{y}'_2]$, $\mathrm{I}_1 = \partial\mathrm{V}/\partial\sigma^2_{e_1}$, $\mathrm{I}_2 = \partial\mathrm{V}/\partial\sigma^2_{e_2}$, $\mathrm{G}_1 = \partial\mathrm{V}/\partial\sigma^2_{g_1}$, $\mathrm{G}_2 = \partial\mathrm{V}/\partial\sigma^2_{g_2}$ and $\mathrm{C} = \partial\mathrm{V}/\partial\sigma^2_{g_1g_2}$.

2.2. Correlation on the scale of liability is approximately the same as that on the observed risk scale

For disease traits when the **y** phenotype vectors contain only 1 for cases and 0 for controls, a liability threshold model can be written to link unobserved continuous liability to the observed discrete scale of disease (Falconer, 1965)

$$\mathbf{l} = \mathbf{g}^* + \mathbf{e}^* \tag{1}$$

where **l** is a vector of liability phenotypes which are distributed as N(0, 1) in the population, **g*** is a vector of random additive genetic effects on the liability scale which are distributed $\mathrm{N}(0, \sigma^2_{g*})$ and **e*** is a vector of random residuals on the liability scale distributed with $\mathrm{N}(0, \sigma^2_{e*})$. The probit link function links liability to the probability of $y = 1$, and g* on the scale of liability can be approximated by a linear function of g on the observed 0–1 scale (Dempster and Lerner, 1950). Using this linear approximation, the correlation between two diseases is the same on both the observed and liability scale (Gianola, 1982; Höschele *et al.*, 1987). When samples are ascertained (typical in case–control studies), the genetic value on the observed scale can be defined with an ascertainment correcting factor as (Lee *et al.*, 2011)

$$g_{cc} \cong c + z\frac{P(1-P)}{K(1-K)}g^*, \tag{2}$$

where g_{cc} is genetic values on the observed scale in a case–control study, c is a constant, K is the disease prevalence in the population, P is the proportion of the sample that are cases and z is the height of the standard normal probability density function that truncates the proportion K. From equation (2), the covariance between genetic values on the observed scale with ascertained samples can be written as

$$\mathrm{cov}(g_{cc1}, g_{cc2}) \cong z_1\frac{P_1(1-P_1)}{K_1(1-K_1)}z_2\frac{P_2(1-P_2)}{K_2(1-K_2)}\mathrm{cov}(g_1^*, g_2^*). \tag{3}$$

From equation (3) it is clear that that even when samples are ascertained, the correlation is the same on both observed and liability scales, because an approximate linear relationship exists between the genetic values on the different scales.

2.3. Simulation study

In order to confirm the derivation that the genetic correlation is approximately the same on both observed and liability scales when samples are ascertained, we performed a simulation study. The simulation procedure was similar to that in Lee *et al.* (2011) except that two traits were simulated with a genetic correlation between them (described in the Supplementary material).

2.4. Application to genome-wide genotype data

We applied our method to estimate genetic correlation between HT and T2D, BD and RA, BD and T2D, or HT and RA using WTCCC GWAS data (WTCCC, 2007), following stringent quality control (QC) as described in the supplementary material. Since there are two control groups in the WTCCC data, i.e. 1958 cohort controls and NBS controls, we used 1958 cohort controls for the first trait, and NBS controls for the second trait. In a confirmation study, we swapped the control groups i.e. NBS for the first trait and 1958 cohort for the second trait.

We estimated a test statistic by dividing the square of the estimated genetic correlation coefficient by its approximate sampling variance and calculated a p-value from this test statistic assuming that it is distributed as a chi-square with 1 degree of freedom.

3 RESULTS

In simulations the estimated genetic correlation on the observed scale was close to the true values when using various combinations of true heritability and population prevalence (Supplementary Table S1). This confirms that the estimated genetic correlation is approximately the same on both observed and liability scales [Equation (3)]. Previously we have shown that if misdiagnosis occurs between the two disorders, then the expectation of the estimate of the genetic correlation coefficient can be non-zero even when the true genetic correlation is zero (Wray *et al.*, 2012).

The estimated genetic correlation between HT and T2D was 0.31 (SE = 0.14 and p-value = 0.023) (Supplementary Table S2), indicating that genetic factors for HT and T2D are positively correlated. However, estimates for genetic correlation between BD and RA, BD and T2D, or HT and RA were not significantly different from zero (Supplementary Table S2). In a confirmation study switching control groups between the first and second trait (Supplementary Table S2), the genetic correlation between HT and T2D was 0.32 (SE = 0.14 and $P = 0.024$). Again, none of other analyses had significant genetic correlations (Supplementary Table S2). None of the parameter estimates differed significantly between our original and confirmation analyses. We previously demonstrated that the application of our stringent QC process resulted in estimated genetic variance not significantly different from zero if we conduct a dummy case-control analysis using these two control sets but treating one set as a cases (Lee *et al.*, 2011) ($h^2 = 0.06$, SE 0.11).

ACKNOWLEDGEMENTS

We thank QBI IT team. S.H.L. acknowledges the use of the Genetic Cluster Computer for carrying out a part of simulations. The cluster is financially supported by the Netherlands Scientific Organization (NOW 480-05-003). This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the WTCCC data is available from www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113.

Funding: The Australian National Health and Medical Research Council (613672, 613601, 613608 and 1011506), the Australian Research Council (DP1093502 and FT0991360) and the US National Institute of Health (GM075091).

Conflict of Interest: none declared.

REFERENCES

Dempster,E.R. and Lerner,I.M. (1950) Heritability of threshold characters. *Genetics*, **35**, 212–236.

Falconer,D.S. (1965) The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.*, **29**, 51–71.

Gianola,D. (1982) Theory and analysis of threshold characters. *J. Anim. Sci.*, **54**, 1079–1096.

Gilmour,A.R. *et al.* (1995) Average information REML: an efficient algorithm for variance parameters estimation in linear mixed models. *Biometrics*, **51**, 1440–1450.

Höschele,I. *et al.* (1987) Estimation of variance components with quasi-continuous data using Bayesian methods. *J. Anim. Breed. Genet.*, **104**, 334–349.

Johnson,D.L. and Thompson,R. (1995) Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Sci.*, **78**, 449–456.

Lee,S.H. *et al.* (2012) Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.*, **44**, 247–250.

Lee,S.H. and Van der Werf,J.H.J. (2006) An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. *Genet. Sel. Evol.*, **38**, 25–43.

Lee,S.H. *et al.* (2011) Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.*, **88**, 294–305.

Thompson,R. (1973) The estimation of variance and covariance components with an application when records are subject to culling. *Biometrics*, **29**, 527–550.

Wray,N.R. *et al.* (2012) Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes. *Eur. J. Hum. Genet.*, **20**, 668–674.

WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Yang,J. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565–569.

Yang,J. *et al.* (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.