Molecular Psychiatry (2014), 1–2

www.nature.com/mp

LETTER TO THE EDITOR

MZ twin pairs or MZ singletons in population family-based GWAS? More power in pairs

Molecular Psychiatry advance online publication, 30 September 2014; doi:10.1038/mp.2014.121

Family-based genome-wide association studies (GWAS) involve testing the genetic association of (many) genetic variants with the phenotype of interest, while taking into account the relatedness among family members. Occasionally in family-based GWAS, including monozygotic (MZ) twins, the data from one MZ twin are dropped, thus reducing the MZ pairs to singletons (for example, Loukola *et al*.,[1] Lowe *et al*.,[2] Parsons *et al*.[3] and Psychosis Endophenotypes International Consortium *et al*.[4]). From a statistical power perspective, this practice is not optimal. To evaluate the issue of power, we consider the effective sample size (N_E), that is, the number of independent cases that provides the same power as N MZ twin pairs. Given the MZ intraclass correlation of ρ, the effective sample size is calculated as $N_E = (2*N)/(1+\rho)$, where N_E ranges from N ($\rho = 1$) to $2*N$ ($\rho = 0$). For instance, given $N = 1000$ pairs, discarding data from one MZ twin reduces the sample size to 1000 singletons, that is, the N_E assuming $\rho = 1$. However, given $\rho = 0.2$ (0.4, 0.7), the N_E is 1667 (1429, 1176), so that 1000 twin pairs (2000 individuals) are equivalent—in terms of power—to 1667 (1429, 1176) unrelated individuals. To illustrate the loss in power, we consider a candidate gene explaining 1% of the variance, the power to detect the association in linear regression with $N = 1000$ MZ twin pairs ($\alpha = 0.001$). MZ singletons, that is, 1000 unrelated subjects, provide a power of 0.450. Retaining data from both MZ twins (1000 pairs), the power varies with ρ as follows: 0.884 ($\rho = 0$), 0.789 ($\rho = 0.2$), 0.643 ($\rho = 0.5$) and 0.519 ($\rho = 0.8$). We refer to Supplementary Figure 1 for more details. Importantly, the gains associated with retaining MZ pairs involve no additional genotyping costs. That is, given the almost perfect concordance rate observed in MZ twins (> 99%), genotyping one twin suffices in twins of confirmed monozygosity.

An important related question is whether retaining both MZ twins affects the type I error rate, that is, does the empirical type I error rate equal the chosen α? We checked the type I error rate by means of simulations. Our results indicate that the empirical type I error rate is correct, that is, invariably equals the chosen α (for details we refer to Supplementary Table 1 and to Supplementary Figure 2). Minică *et al*.[5] evaluated the type I error in samples involving MZ twins, full sibs and parents, and also found that the empirical α closely resembled the nominal α. We conclude that the presence of MZ twins alone, or MZ twins in combination with other family members, does not affect the type I error rate.

We note that many GWAS meta-analyses rely heavily on twin registries. For example, the educational attainment GWA[6] included more than 35% data from twin registries. Twin registries also contributed 13% cases and 9% controls to migraine meta-analysis,[7] 34% of the sample to telomere length meta-analysis[8] and 31% cases and 19% controls to the meta-analysis of GWASs for major depressive disorder.[9] These registries are rich resources of phenotypic and genotypic twin data. Whereas the MZ data may be exploited fully in primary and in meta-analyses (for example, the contribution of the Queensland Institute of Medical Research (QIMR) to Rietveld *et al*.[6]), consortia protocols often stipulate dropping MZ twins. Consider, for instance, the recent meta-analysis of GWASs for major depressive disorder.[9] Although genotypic data were available in ~1890, ~786 and ~300 MZ twin pairs at the Netherlands Twin Register, the QIMR and the TwinGene cohort, respectively, only one twin of a pair was selected for the analyses. Given an MZ correlation for depression of $\rho = 0.35$ these 2976 MZ twin pairs (5952 individuals) are equivalent in terms of power to $N_E = 4409$ unrelated subjects. By dropping 1 MZ twin, the equivalent of $4409 - 2976 = 1433$ unrelated individuals was discarded from the meta-analysis. The corresponding loss in power is notable (that is, from 0.823 power MZ twins would afford, to 0.395 power afforded by MZ singletons, given $\alpha = 10^{-8}$ and a genetic variant explaining 1% of the phenotypic variance).

Full modeling of data on families including MZs can be performed by using a mixed-effects variance components approach (for example, using MERLIN and MERLIN-offline, see http://genepi.qimr.edu.au/staff/sarahMe/merlin-offline.html). If the families are highly variable in the number and composition of participating family members, retaining all data may pose a challenge as modeling the conditional (that is, conditional on the genetic variant) covariance structure can be complicated and subject to misspecification. One tractable solution is to use generalized estimating equations (GEE) with a conditional covariance matrix containing equal covariances (that is, 'exchangeable working correlation matrix' in GEE terms), in combination with a sandwich correction for the standard errors. The use of a sandwich correction is advisable as it produces correct type I error rates, regardless of misspecification. This method fares well in terms of power, in comparison to full (correct) modeling, while the computational burden is acceptable given typical GWAS requirements.[5] We note that GEE with the exchangeable option (as implemented in R[10]) can be conducted from the Plink platform (see http://cameliaminica.nl/scripts.php).

In conclusion, the presence of MZ twin pairs does not affect the type I error rate, and reducing MZ pairs to singletons results in a loss of power. If the main interest is in the association, and not in the details of the conditional covariance matrix, adequate modeling of this matrix can be handled efficiently using GEE, with sandwich corrected standard errors.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Camelia C Minică and Jacqueline M Vink are supported by the ERC starting grant 284167. Conor V Dolan is supported by the European Research Council (Genetics of Mental Illness; grant number: ERC-230374). The statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org), which is supported by the Netherlands Scientific Organization (NWO 480-05-003), the Dutch Brain Foundation and the Department of Psychology and Education of the VU University Amsterdam.

CC Minică, DI Boomsma, JM Vink and CV Dolan
Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands
E-mail: c.c.minica@vu.nl

REFERENCES

1 Loukola A, Wedenoja J, Keskitalo-Vuokko K, Broms U, Korhonen T, Ripatti S *et al*. *Mol Psychiatry* 2014; **19**: 615–624.

2 Lowe JK, Maller JB, Pe'er I, Neale BM, Salit J, Kenny EE *et al. PLoS Genet* 2009; **5**: e1000365.
3 Parsons MJ, Lester KJ, Barclay NL, Nolan PM, Eley TC, Gregory AM. *Am J Med Genet Part B: Neuropsychiatr Genet* 2013; **162**: 431–438.
4 Psychosis Endophenotypes International Consortium, Wellcome Trust Case-Control Consortium, Bramon E, Pirinen M, Strange A, Lin K *et al. Biol Psychiatry* 2014; **75**: 386–397.
5 Minica CC, Dolan CV, Kampert MM, Boomsma DI, Vink JM. *Eur J Hum Genet* 2014; doi:10.1038/ejhg.2014.94 (e-pub ahead of print).
6 Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW *et al. Science* 2013; **340**: 1467–1471.
7 Anttila V, Winsvold BS, Gormley P, Kurth T, Bettella F, McMahon G *et al. Nat Genet* 2013; **45**: 912–917.
8 Codd V, Nelson CP, Albrecht E, Mangino M, Deelen J, Buxton JL *et al. Nat Genet* 2013; **45**: 422–427.
9 Ripke S, Wray NR, Lewis CM, Hamilton SP, Weissman MM, Breen G *et al. Mol Psychiatry* 2013; **18**: 497–511.
10 Carey VJ, Lumley T, Ripley B. 2012. http://CRAN.R-project.org/package = gee.

Supplementary Information accompanies the paper on the Molecular Psychiatry website (http://www.nature.com/mp)