Shetty et al. BMC Proceedings 2011, 5(Suppl 9):S34
http://www.biomedcentral.com/1753-6561/5/S9/S34

BMC Proceedings

PROCEEDINGS Open Access

Estimating heritability using family and unrelated individuals data

Priya B Shetty, Huaizhen Qin, Junghyun Namkung, Robert C Elston, Xiaofeng Zhu*

From Genetic Analysis Workshop 17
Boston, MA, USA. 13-16 October 2010

Abstract

For the family data from Genetic Analysis Workshop 17, we obtained heritability estimates of quantitative traits Q1 and Q4 using the ASSOC program in the S.A.G.E. software package. ASSOC is a family-based method that estimates heritability through the estimation of variance components. The covariate-adjusted mean heritability was 0.650 for Q1 and 0.745 for Q4. For the unrelated individuals data, we estimated the heritability of Q1 as the proportion of total variance that can be accounted for by all single-nucleotide polymorphisms under an additive model. We examined a novel ordinary least-squares method, a naïve restricted maximum-likelihood method, and a calibrated restricted maximum-likelihood method. We applied the different methods to all 200 replicates for Q1. We observed that the ordinary least-squares method yielded many estimates outside the interval [0, 1]. The restricted maximum-likelihood estimates were more stable than the ordinary least-squares estimates. The naïve restricted maximum-likelihood method yielded an average estimate of 0.462 ± 0.1, and the calibrated restricted maximum-likelihood method yielded an average of 0.535 ± 0.121. Our results demonstrate discrepancies in heritability estimates using the family data and the unrelated individuals data.

Background

The heritability of a trait is usually calculated using family data. The identified genetic variants found through genome-wide association studies account for only a small portion of heritability for most complex traits [1] compared with the heritability estimated from family data. This discrepancy in the estimates, the missing heritability, is of great interest because the sources of this difference are still unknown [1]. Recently, Yang et al. [2], using a novel statistical method, suggested that the missing heritability can be recovered using the genome-wide associations of unrelated samples [2]. Because the Genetic Analysis Workshop 17 (GAW17) data set included family data and unrelated individuals data for the same traits [3], we estimated the "heritability" of Q1 with the unrelated individuals data and estimated the "heritability" of Q1 and Q4 with the family data.

* Correspondence: zhu1@darwin.epbi.cwru.edu
Case Western Reserve University School of Medicine, 2103 Cornell Road, Cleveland, OH 44106, USA

For the family data, the heritability is the narrow sense heritability, estimated with the polygenetic effect model; we conducted a George-Elston transformation [4] to estimate the heritability. For the unrelated data, the heritability is the proportion of the total variance in a phenotype that can be described by all single-nucleotide polymorphisms (SNPs) under an additive model; we estimated it using the ordinary least-squares (OLS) method suggested by Yang et al. [2], a naïve restricted maximum-likelihood (REML) method, and a calibrated REML method. In all our analyses, the heritability estimates were obtained after adjustments for age, sex, and smoking status.

Methods

PEDINFO and ASSOC

For the family data, we chose to use quantitative traits Q1 and Q4 of four randomly selected data set replicates (Table 1). We used the Statistical Analysis for Genetic Epidemiology (S.A.G.E.) software and the PEDINFO and ASSOC programs. The PEDINFO program calculates summary statistics about the family data set. The ASSOC program performs a family-based association test using a

Table 1 Heritability estimates for Q1 and Q4 using the family data

Replicate number	Q1		Q4	
	Heritability	Standard error	Heritability	Standard error
1	0.608	0.063	0.754	0.106
2	0.640	0.067	0.687	0.061
52	0.698	0.103	0.773	0.117
137	0.655	0.105	0.766	0.104

polygenic mixed effect model for a quantitative trait, and it estimates the heritability through the estimation of the proportion of a polygenic component to the total trait variance. In our analysis, the heritability estimates were obtained after adjustments for age, sex, and smoking status. The George-Elston transformation was applied for normality of residual distribution [4]. We did not include any genotype variables in the model.

OLS and REML estimates

For the unrelated data, we used the OLS method suggested by Yang et al. [2] and the two REML methods to estimate the heritability of Q1 with all 200 data set replicates. Here, the heritability refers to the proportion of the variance in Q1 that can be accounted for by all SNPs under an additive model [2]. We fitted the mixed effects model:

$$\gamma = X\gamma + Zu + e \tag{1}$$

where $\gamma = (\gamma_1, ..., \gamma_n)'$ consists of trait values of n unrelated individuals, $X = [(1, x_1)', ..., (1, x_n)']'$, where $x_i = (x_{i1}, ..., x_{i3})$ consists of the sex, age, and smoking status of the ith individual, respectively, $\gamma = (\gamma_0, ..., \gamma_3)'$ consists of the effect sizes of the covariates, $Z = [z_1', ..., z_n']'$ summarizes genotype data of m unknown causal variants such that $z_i = (z_{i1}, ..., z_{im})$, and $z_{ij} = -2f_j\sigma_j^{-1}$, $(1 - 2f_j)\sigma_j^{-1}$, or $2(1 - f_j)\sigma_j^{-1}$ if the genotype of the ith individual at the jth causal variant is aa, aA, or AA, respectively, f_j is the frequency of allele A and $\sigma_j^2 = 2(1 - f_j)f_j$. Here the prime indicates the transpose of a vector or matrix.

Let the effects of m causal variants be:

$$u = (u_1, ..., u_m)' \sim N(0, \sigma_u^2 I_m) \tag{2}$$

where σ_u^2 is the variance and the residuals be:

$$e = (e_1, ..., e_n)' \sim N(0, \sigma_e^2 I_n) \tag{3}$$

where σ_e^2 is the residual variance, I_n is the identity matrix of order n,

Then the variance-covariance matrix of y is:

$$\text{var}(\gamma) = \sigma_g^2 G + \sigma_e^2 I_n \tag{4}$$

where $G = (1/m)ZZ'$ is the genetic relationship matrix of causal SNPs and $\sigma_g^2 = m\sigma_u^2$. Let X have the rank r (=*4 for the GAW17 unrelated individuals data*), and let $P = [p_1, ..., p_r]$, where $p_1, ..., p_r$ are all orthogonal eigenvectors corresponding to eigenvalue 1 of idempotent matrix $I_n - X(X'X)^{-1}X'$. Let $\tilde{\gamma} = P'\gamma$, $\tilde{Z} = P'Z$, and $\tilde{e} = P'e$. It follows that:

$$\tilde{\gamma} = \tilde{Z}u + \tilde{e} \sim N(0, \tilde{V}), \tag{5}$$

where:

$$\tilde{V} = \text{var}(\tilde{\gamma}) = \sigma_g^2\tilde{G} + \sigma_e^2 I_{n-r} \tag{6}$$

and

$$\tilde{G} = \frac{1}{m}\tilde{Z}\tilde{Z}' = P'GP \tag{7}$$

Note that

$$E\left[\left(\tilde{\gamma}_i - \tilde{\gamma}_j\right)^2\right] = 2\sigma_e^2 + \sigma_g^2(p_i - p_j)'G(p_i - p_j). \tag{8}$$

Thus the slope and intercept of the regression of:

$$\Delta\tilde{\gamma}_{ij} = (\tilde{\gamma}_i - \tilde{\gamma}_j)^2 \tag{9}$$

on $(p_i - p_j)'G(p_i - p_j)$ are σ_g^2 and $2\sigma_e^2$, respectively. Because G is unknown, it is replaced with an estimate. One naïve estimate is A, the genetic relationship of genome-wide SNPs. Yang et al. [2] established an unbiased estimate A^* for G by calibrating the prediction error of genetic relationship G of unobserved causal SNPs. Replacing G with A^* in the regression, we can estimate the heritability as:

$$\widehat{h_2}(A^*) = \frac{\widehat{\sigma_g^2}(A^*)}{\widehat{\sigma_g^2}(A^*) + \widehat{\sigma_e^2}(A^*)}. \tag{10}$$

Because this estimate is based on OLS, it does not need iteration. By replacing G with A and A^* in the model given by $\tilde{\gamma} = \tilde{Z}u + \tilde{e}$, we can constructed the

naïve and calibrated REML estimates by maximizing the likelihood of (σ_g^2, σ_e^2).

Results

Heritability estimates using the family data

In the family data, 697 individuals (202 founders and 495 nonfounders) form eight pedigrees. The pedigrees all have four generations of family members and a mean size of 87.13 individuals (range, 73–128). The pedigrees include 194 sibships with a mean size of 2.55 (range, 1–9). In the four randomly selected replicates, the heritability estimates for Q1 ranged from 0.608 to 0.698 with an average of 0.650; the heritability estimates for Q4 ranged from 0.687 to 0.773 with an average of 0.745 (Table 1).

Heritability estimates using the unrelated individuals data

The unrelated individuals data consist of genotypes of 24,487 SNPs and 200 replicates of 697 individuals for Q1. The OLS estimates of the heritability were apparently unstable (Figure 1), because many of them were outside the interval [0, 1]. We computed the mean and standard deviation of all 200 heritability estimates, including those greater than 1 or less than 0. Over the 200 replicates, the average heritability estimate for Q1 was μ = 0.555 with standard deviation σ = 0.480 after correcting for age, sex, and smoking status.

We found that the REML estimates for Q1 were more stable than estimates obtained using the OLS method (Figure 2). After accounting for age, sex, and smoking status, the 200 naïve REML estimates yielded an average heritability estimate of 0.462 ± 0.999, and the calibrated REML estimates yielded an average heritability estimate of 0.5351 ± 0.1206 for Q1.

We were unable to obtain REML estimates for Q4 because the convergence rate of the REML was extremely slow. We found that the convergence of the REML failed because no SNP contributed any phenotypic variation in the simulated model [3].

Discussion and conclusions

In our analyses, we estimated heritability using both the family data for Q1 and Q4 and the unrelated individuals data for Q1. The heritability estimates for Q1 and Q4 using the family data appeared stable and reasonable. In the simulation, Q1 has a heritability of 0.575, where 0.135 is due to the 39 causal SNPs and 0.440 is due to a polygenic component, and Q4 has a heritability of 0.70 resulting from a polygenic effect. The mean heritability estimates for Q1 and Q4 with the family data were 0.650 and 0.745, respectively.

The heritability estimates using the unrelated individuals data seem less reasonable. The OLS method did not work well for the GAW17 unrelated individuals data because the method was designed for genome-wide common SNPs. In the GAW17 unrelated individuals data, most of the SNPs are rare variants and a few of

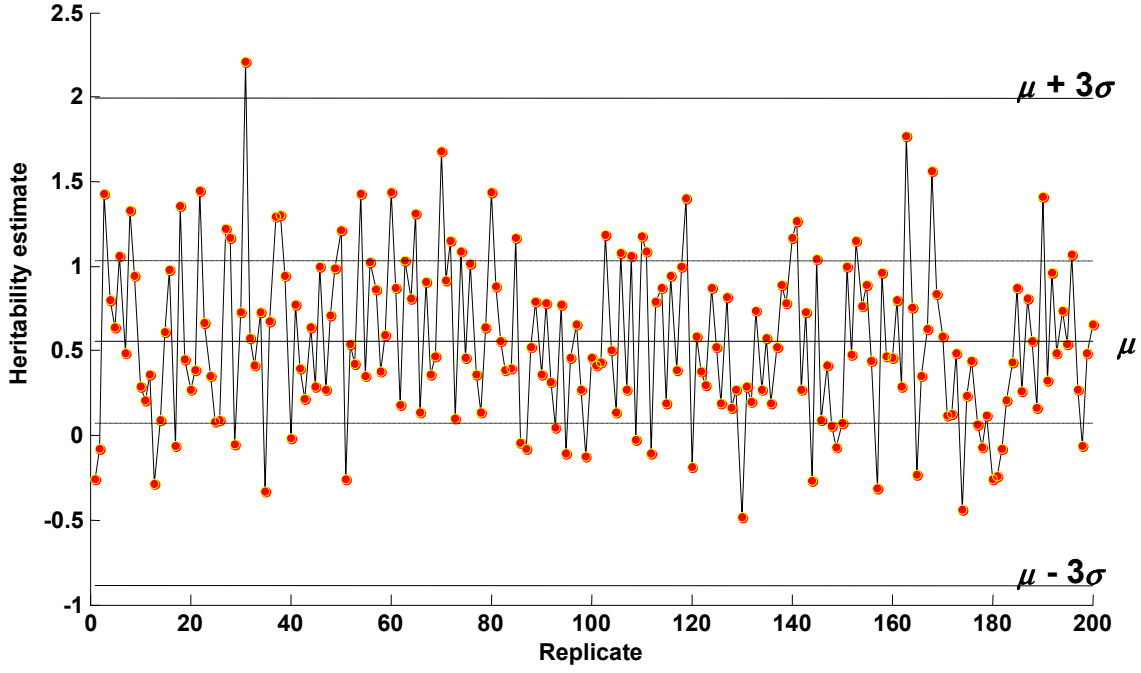

Figure 1 OLS estimates of the heritability of Q1. The estimates at many of the 200 replicates were greater than 1 or less than 0. Over the 200 estimates, the average heritability estimate for Q1 was μ = 0.5549 with standard error σ = 0.4803.

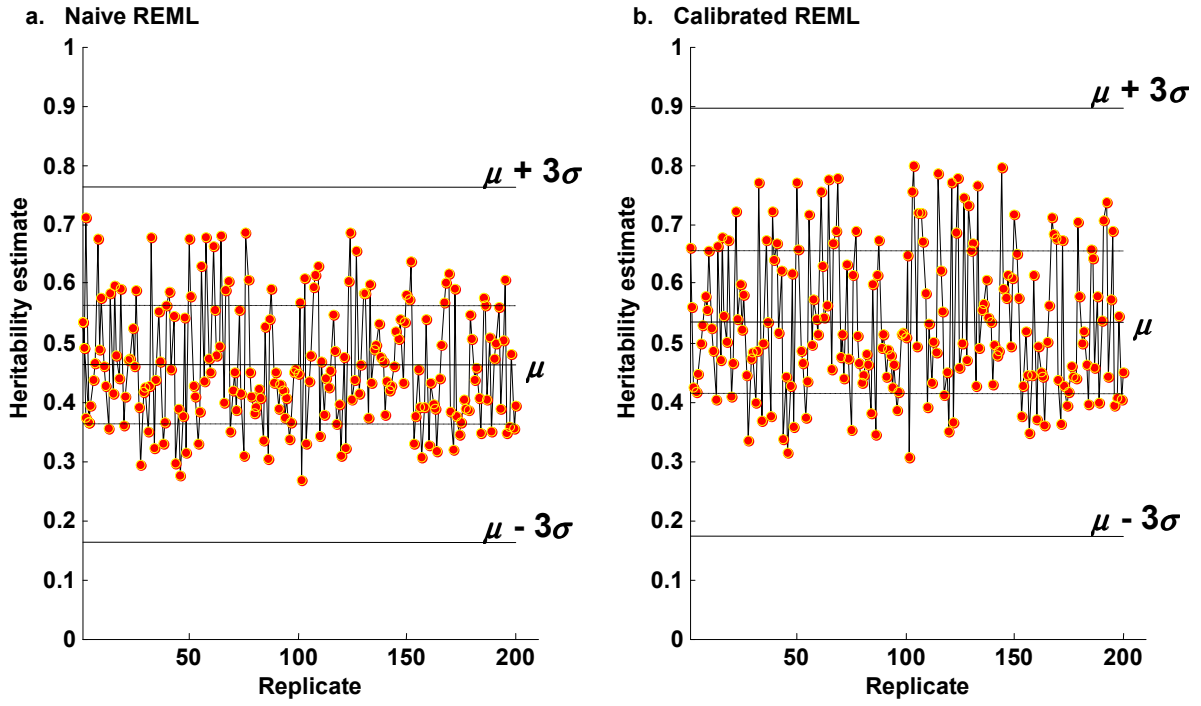

Figure 2 REML estimates of heritability of Q1. (a) The relationship A of genome-wide SNPs was used to estimate the relationship G at unobserved causal SNPs. Over the 200 replicates, the average heritability estimate was $\mu = 0.4618$ with standard error $\sigma = 0.0999$ after correcting for age, sex, and smoking status. (b) The calibrated relationship A^* was used to estimate the relationship G at unobserved causal SNPs. Over the 200 replicates, the average heritability estimate was $\mu = 0.5351$ with standard error $\sigma = 0.1206$ after correcting for age, sex, and smoking status.

them are causal variants. The genetic relationships estimated using many rare variants may be unreliable, and this results in the instability of the OLS estimates. The REML approaches appear to be more stable than the OLS method for Q1. We observed that the heritability estimates using the unrelated individuals data were less than those using the family data on average. For example, the mean of the heritability estimates for Q1 for the unrelated individuals data was 0.462 (by naïve REML), which was 0.188 less than the mean for the family data. One possible reason is that the polygenic component (0.440) in Q1 is not due to any SNPs in the GAW17 sequence data set. We should not be able to uncover the polygenic effect using unrelated samples. However, the mean naïve REML estimate (0.462) is much larger than the heritability because of the causal SNPs (0.135). The reason is that we used all 24,487 SNPs to estimate the relationships among individuals. There might be other sources contributing to the heritability estimates.

Finally, we failed to estimate the heritability for Q4 using the unrelated samples because of the convergence problem, which was the result of no genotyped exonic SNPs in the data contributing to the phenotypic variation.

Acknowledgments

The Genetic Analysis Workshop is supported by National Institutes of Health (NIH) grant R01 GM031575 from the National Institute of General Medical Sciences. This work was supported by National Cancer Institute grant P30 CAD43703 and NIH grants HL074166, HL086718, R01 HG003054 and R01 HG005854. Some of the results of this paper were obtained by using the program package S.A.G.E., which is supported by U.S. Public Health Service Resource Grant RR03655 from the National Center for Research Resources. We thank the other members of Xiaofeng Zhu's laboratory for their critiques and comments.

This article has been published as part of *BMC Proceedings* Volume 5 Supplement 9, 2011: Genetic Analysis Workshop 17. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/5?issue=S9.

Authors' contributions

PBS performed the statistical analysis of family data and HQ performed the statistical analysis of the unrelated individuals data. PBS , HQ, JN and XZ drafted and revised the manuscript. XZ conceived the project, RCE criticized and edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that there are no competing interests.

Published: 29 November 2011

References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, *et al*: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461**:747-753.

2. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, *et al*: **Common SNPs explain a large proportion of the heritability for human height.** *Nat Genet* 2010, **42**:565-569.
3. Almasy LA, Dyer TD, Peralta JM, Kent JW Jr, Charlesworth JC, Curran JE, Blangero J: **Genetic Analysis Workshop 17 mini-exome simulation.** *BMC Proc* 2011, **5(suppl 9)**:S2.
4. George V, Elston RC: **Generalized modulus power transformations.** *Commun Stat Theory Meth* 1988, **17**:2933-2952.

doi:10.1186/1753-6561-5-S9-S34
Cite this article as: Shetty *et al.*: Estimating heritability using family and unrelated individuals data. *BMC Proceedings* 2011 **5**(Suppl 9):S34.