Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.116863/DC1

Genome-Wide Association Studies and the Problem of Relatedness Among Advanced Intercross Lines and Other Highly Recombinant Populations

Riyan Cheng, Jackie E. Lim, Kaitlin E. Samocha, Greta Sokoloff, Mark Abney, Andrew D. Skol and Abraham A. Palmer

Copyright © 2010 by the Genetics Society of America
DOI: 10.1534/genetics.110.116863
**Calculation of Relationship Matrices**

We calculated identity coefficients to obtain relationship matrices used in equation (2) by using pedigree information (“pedigreeF34.csv” in our raw data files). The pedigree data consists of 34 generations with more than 5,600 individuals. The number of individuals in the non-founder (F₀ and F₁) generations ranges from 96 to 693. The algorithm of Karigl (1981) can be applied to calculate kinship coefficients and identity coefficients. This algorithm requires generalized kinship coefficients \( \Phi_{ij} \) of two individuals \( i \) and \( j \), \( \Phi_{ijk} \) of three individuals \( i, j \) and \( k \), \( \Phi_{ijkl} \) of four individuals \( i, j, k \) and \( l \), and \( \Phi_{ij,kl} \) of two pairs, \( i \) and \( j \), and \( k \) and \( l \). One may adopt a top-down strategy that starts from the founder generation and goes through the pedigree generation-by-generation to the last generation or a bottom-up strategy that goes in an inverse direction. The calculation increases approximately exponentially with the number of generations for a bottom-up method and linearly increases for the top-down method. Therefore, a bottom-up method is not practical for our pedigree because it has too many generations. The top-down method encounters memory storage problems when the sample size in a generation is large, which is not be a problem for a bottom-up method, since one needs to store \( \Phi \)'s for intermediate generations. For instance, there are 344,291,325 unique \( \Phi_{ijkl} \)'s for 300 individuals and these numbers would require 2,626.73Mb of storage. The number of unique \( \Phi_{ij,kl} \)'s is approximately three times as large as that of unique \( \Phi_{ijkl} \)'s. To address these problems we combined the top-down and bottom-up methods, which allowed us to obtain the generalized kinship coefficients. For smaller pedigrees the R package “identity” or C source code ldCoefs [http://home.uchicago.edu/~abney/Software.html] should provide an easier solution.

Kinship coefficients were variable across individuals in our study population; the table below shows examples of the kinship confidences for twelve individuals. The figure on the following page shows the frequency of all kinship coefficients in the study.

**Table**

Kinship coefficients between twelve individuals (ID 1, 10 and 100-109) from the F₃₄ generation.
FIGURE.—Histogram of kinship coefficients between all individuals in the F_{34} generation.
**Allele Frequencies in F₂**

Figures show the frequency of the LG allele in F₂ mice as a function of position, and as a summary histogram. The excess of LG alleles at the X chromosome (red circle) reflects the unequal contribution of X chromosome markers due to the use of LG females to produce the F₁ population.

**FIGURE.**—LG allele frequencies at all typed SNP loci in the F₂ (x-axis shows genetic position; red circle shows expected deviation on the X-chromosome).

**FIGURE.**—Histogram of LG allele frequencies at all typed SNP loci in the F₂ (red circle shows the expected deviation due to the X-chromosome).
Allele Frequencies in F_{34}

In an AIL, genetic drift or unintended selection for fitness or fecundity can lead to skewed allele frequencies and even fixation of one of the two alleles. Because it was difficult to call genotypes when allele frequencies became skewed, we obtained missing genotypes in regions of low allele frequencies. The figures show the frequency of the LG allele in F_{34} mice as a function of position (upper), and as a summary histogram (lower).

To investigate whether these results were in line with the expected values due to genetic drift, we did 10,000 simulations for a single SNP given our pedigree. This simulation was very similar to the gene dropping approach described in this paper. These simulations assumed no selective advantage for any of the alleles, an assumption that may be incorrect. The results of these simulations were very similar to those observed in our real data. Due to limitations of the genotype calling algorithm, when the minor allele frequency (MAF) fell below approximately 0.2 we had difficulty calling genotypes. Thus, our real data likely contained a minority of SNPs at low MAF that we were unable to call; our study was not designed to be sensitive to such situations and so we did not vigorously pursue this issue.
**Maximum likelihood estimation**

There is generally no closed form for maximum likelihood estimates of equation (3) or (4). One needs numerical procedures to iteratively estimate the parameters. This involves manipulation of $n \times n$ matrix $V$, which can be computationally expensive when $n$ is moderately large. Our data has 688 individuals and requires a few hours to estimate the parameters in model (4) if at most two covariates are considered and Nelder-Mead algorithm (Nelder and Mead, 1965) is implemented, and it will take more time to test for putative QTL at a locus. Therefore, it is not practical to perform a scan of 3105 SNPs. Note that the random effect $g$ in model (2) is only used to control background genetic variation. A reasonable approximation will be good enough.

Assume $\frac{\sigma^2_1}{c_1} = \frac{\sigma^2_2}{c_2} = \frac{\sigma^2_3}{c_3} = \frac{\text{Cov}(a,d)}{c_4} = \frac{\mu^2}{c_5} = \sigma^2_g$. Then

$$\sigma_{ij} = [2\Phi_{ij}c_1 + \Delta_{ij,1}c_2 + \Delta_{ij,1}c_3 + (4\Delta_{ij,1} + \Delta_{ij,3} + \Delta_{ij,5})c_4 + (\Delta_{ij,1} + \Delta_{ij,2} - f_i f_j)c_5] \sigma^2_g$$

and

$$\Sigma = \tilde{\Sigma} \sigma^2_g$$

$\tilde{\Sigma}$ is a known matrix if $c$'s are known. Suppose $U$ is an orthogonal matrix such that

$$U\tilde{\Sigma}U = \begin{pmatrix} d_1^2 & 0 & \cdots & 0 \\ 0 & d_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n^2 \end{pmatrix}$$

where $d_i^2$'s are known if $c$'s are known. Then equation (3) becomes

$$l(\theta; y) = \frac{1}{(2\pi)^{\frac{n}{2}} \prod_{i=1}^n (d_i^2 \sigma_g^2 + \sigma^2)^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \left[ U(y - X\beta - x^* a - z^* d) \right]^T \left[ U(y - X\beta - x^* a - z^* d) \right] \right\}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}} \prod_{i=1}^n (d_i^2 \sigma_g^2 + \sigma^2)^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \frac{(\bar{y}_i - \bar{x}'_i \beta - \bar{x}'_i a - \bar{z}'_i d)^2}{d_i^2 \sigma_g^2 + \sigma^2} \right\}$$

(6)

where $(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n)' = U y$, $(\bar{x}'_1, \bar{x}'_2, \ldots, \bar{x}'_n)' = UX$, $(\bar{x}'_1, \bar{x}'_2, \ldots, \bar{x}'_n)' = U x^*$ and $(\bar{z}'_1, \bar{z}'_2, \ldots, \bar{z}'_n)' = U z^*$. If $c$'s are known, maximum likelihood estimates of the parameters in (6) can easily be obtained numerically. A similar idea was proposed by Kang et al. (2008), who only considered the additive polygenic variance components with $\sigma_{ij} = 2\Phi_{ij}$ in equation (2).

However, $c$'s are unknown in practice. We first estimate $c$'s from equation (4), assuming no major QTL. Then the test for QTL at a locus as well as the likelihood ratio test statistic (5) is conditional on estimated $c$'s. This is an idea of conditional tests (Pinheiro and Bates, 2000, Ch. 2).
FIGURE S1.—Differences between results for QTL mapping for response to methamphetamine using simple regression versus a mixed model; these data were generated by subtracting the values shown in Figure 3; positive values indicate a higher LOD score for regression, negative values indicate a higher LOD score for the mixed model.
FIGURE S2.—QTLs identified on each chromosome using Haley-Knott interval mapping for F₂ (red), AIL (green) and the integration of F₂ and AIL (blue). Relatedness was accounted for using a mixed model as described in the text. The genome-wide significance thresholds (p < 0.05) are indicated by horizontal lines and were determined using gene dropping. These are results for the analysis of behavior on the first test day, when saline was administered.
**FIGURE S3.**—QTLs identified on each chromosome using Haley-Knott interval mapping for F$_2$ (red), AIL (green) and the integration of F$_2$ and AIL (blue). Relatedness was accounted for using a mixed model as described in the text. The genome-wide significance thresholds ($p < 0.05$) are indicated by horizontal lines and were determined using gene dropping. These are results for the analysis of behavior on the second test day, when saline was administered.
FIGURE S4.—QTLs identified on each chromosome using Haley-Knott interval mapping for F₂ (red), AIL (green) and the integration of F₂ and AIL (blue). Relatedness was accounted for using a mixed model as described in the text. The genome-wide significance thresholds (p < 0.05) are indicated by horizontal lines and were determined using gene dropping. These are results for the analysis of behavior on the third test day, when methamphetamine was administered.
Figure S5. —Plots showing detailed views of a QTL on chromosome 8 for response to saline injection on day 2 using $F_2$ (red), AIL (green), and integrated (blue) Haley-Knott regression. Tick marks along the x-axis indicate the location of physical markers that were successfully genotyped. The genome-wide significance thresholds ($p < 0.05$) are indicated by horizontal lines and were determined using gene dropping.
FILE S2

SUPPORTING DATA

File S2 is available online as a compressed folder (.zip) at http://www.genetics.org/cgi/content/full/genetics.110.116863/DC1.