RESEARCH ARTICLE

An efficient method to handle the 'large *p*, small *n*' problem for genomewide association studies using Haseman–Elston regression

BUJUN MEI[1,3] * and ZHIHUA WANG[2]

[1]*Agriculture Department, Hetao College, Bayannur 015000, People's Republic of China*
[2]*Department of Civil Engineering, Hetao College, Bayannur 015000, People's Republic of China*
[3]*Department of Animal Science, Iowa State University, Iowa 50010, USA*

Abstract

The 'large *p*, small *n*' problem in genomewide association studies (GWAS) is an important subject in genetic studies. Many approaches have been proposed for this issue, but none of them successfully combine the Haseman–Elston (H–E) regression with sliding-window scan approaches in GWAS. In this article, we extended H–E regression to GWAS, and replaced original data with different measurements of phenotype of sib pairs. Meanwhile, we also applied hidden Markov model to infer identity by state. Using subsequent simulation studies, we found that it had higher statistical power than the corresponding single-marker association studies. The advantage of the H–E regression was also sufficient to capture about 48.01% of the quantitative trait locus (QTL). Meanwhile, the results show that the power decreases with the increase in the number of QTLs, and the power of H–E regression is sensitive to heritability.

[Mei B. and Wang Z. 2016 An efficient method to handle the 'large *p*, small *n*' problem for genomewide association studies using Haseman–Elston regression. *J. Genet.* **95**, 847–852]

Introduction

The analysis and comprehension of high-throughput genomic data has posed great challenges to researchers, partly due to the 'large *p* small *n*' problem (Diao and Vidyashankar 2013). So-called 'large *p* small *n*' or 'short-fat data' problem can occur if the number of available covariates is essentially larger than sample sizes in statistical inference and modelling problems. Here, *p* usually indicates the dimension of independent variables, and *n* indicates the sample size. The independent variables, the $x_{i,j}s$ is an $n \times p$ matrix X in which the ith row is $(x_{1,i}, \ldots, x_{p,i})$, which implies that X is fat and short, which means $p \gg n$. Conventionally, the design matrix of general regression contexts is $n \gg p$, so that there are no direct methods to obtain parameter estimates. In genomewide association studies (GWAS), the amount of observations, *n* is mostly more or less hundreds or thousands, whereas the amount of markers, *p* is approximately hundreds of thousands (Chen 2014). This is just known as 'large *p* small *n*' problem, one of the several problems of 'curse of dimensionality'. Further, the problem is more deteriorated when independent variables are in multiple correlations.

It requires some kind of special statistical procedures, such as penalized likelihood, variable selection, constraint or other shrinkage methods, just for the purpose of solutions existence (Shen *et al.* 2013). There have been many approaches for GWAS. Generally, such models turned the problem back into multiple testing to avoid high-dimensional models (Daetwyler *et al.* 2013). Subsequently, Bayesian methods have been developed, and ridge regressions (RR) have been proposed for GWAS (de Los Campos *et al.* 2013). These methods require computationally feasible algorithms. Simultaneously, the flexibility of statistical models is also sufficient to capture the influential genetic effects, which are also expected to be better than inflexible methods.

The aim of this study is to propose an efficient method for scanning-associated single nucleotide polymorphisms (SNPs) in GWAS based on the Haseman–Elston regression (H–E) (DeFries 2010). The proposed method uses multiple regression, which has combined advantages of H–E method, for $p \gg n$ high-dimensional problems. All programs are written in Julia language and can be run under Windows, Mac and Linus/Unix environments. And all of these are available upon request from Bujun Mei.

*For correspondence. E-mail: meibujun@iastate.edu.

Bujun Mei initiated the idea, developed the theory and derived the equations; and wrote the paper. Zhihua Wang conducted the simulation studies and obtained the analytical results. All authors approved the final version of the paper.

Keywords. Haseman–Elston regression; genomewide association studies; large *p* small *n* problem; identity by state; Julia language.

Materials and methods

Theory of the H–E regression

Haseman and Elston based on comparison of the identity by descent (IBD) sharing and the phenotype similarity (or difference) between sib pairs. It originated a linear model, $Y_{ij} = \mu + b\pi_{ij} + e_{ij}$, for scanning linkage between markers and quantitative trait locus (QTLs) in full-sib families (Barber *et al.* 2004). Here, for a particular family, Y_{ij} symbolizes the squared difference of a phenotype, $y_j = 0.5(x_{1j} - x_{2j})^2$, between a pair of relatives (Elston *et al.* 2000). Then, π_{ij} is the estimated proportion of IBD at the given marker alleles; μ is the mean of the regression; and b, the regression coefficient, is a function of $-2(1-2\theta)^2\sigma_a^2$, where $\sigma_a^2 = 2p(1-p)[\alpha + (1-2p)\delta]^2$, for a dominant trait, when $\delta = \alpha$, $\alpha_a^2 = 8p(1-p)^3\alpha^2$; while for a recessive trait, when $\delta = -\alpha$, $\sigma_a^2 = 8p^3(1-p)\alpha^2$; and e_{ij} is the residual variance (Drigalenko 1999). Using the above regression model, we can set up the null hypothesis of no linkage $H_0 : b = 0$, which is a one-sided t-test. The original H–E regression is a variance component procedure that was primarily merely recommended to linkage study and less powerful than the other same theoretical framework methods (Etzel *et al.* 2003). However, this method still donates a theoretical foundation for diversified H–E regressions that also designed for linkage disequilibrium (LD) of quantitative traits, and accordingly can be exploited to GWAS (Forrest 2001; Franke *et al.* 2005). Many extensions of the original H–E have been proposed that use various forms of H–E regression by definitions of the squared sum or squared difference as dependent variables (Garner 2002). Table 1 lists these forms (Wang and Elston 2005).

To enhance the statistical power, various weighing models have been developed for the two parameters that are estimated from measurements of dependent variables (Gerhard and Hothorn 2010); these models essentially differ in how to infer the weight values of the two regression parameters based on their estimated variances (Hadicke *et al.* 2008). Meanwhile, a two-level H–E is also developed for linkage analysis of quantitative trait and general pedigrees, which can exploit all the information stored in any general pedigree structures (Wang and Elston 2005; Yoon *et al.* 2005); at the same time, it can incorporate multilevel effects, and possibly infers different complex genetic effects (Stoesz *et al.* 1997).

H–E regression for GWAS

The H–E regression offers a reputable procedure for estimating variance components. However, this method uses IBD which seems to be slightly unfashionable in the GWAS era (Chen 2014). The reason is that H–E regression depends largely on IBD instead of identity by state (IBS). Owing to the close resemblance between IBS and conventional IBD, IBS can be employed in the GWAS within a theoretical framework of H–E regression (Xu *et al.* 2000; Sham and Purcell 2001). In this study, we reused the H–E regression to handle the 'large p, small n' problem for a GWAS.

The fundamental principle of the method is as follows. In farm animal species with half-sib families, the phenotype of ith individual is indicated as y_i, $i = 1, 2, \ldots, n$, where n is the number of animals and y follows the normal distribution $N(\mu, \sigma^2)$ (Shete *et al.* 2003). The genotype is $x_{i1}, x_{i2}, \ldots, x_{ip}$, in which p is the number of markers. Now, H–E original regression can be modified as

$$Y_{ij} = \mu + b\xi_{ij} + \varepsilon_{ij}. \quad (1)$$

Here $Y_{ij} = 0.5(y_i - y_j)^2$ denotes the function of squared difference. The symbol ξ_{ij} represents the genetic similarity between a sib pair based on IBS, and ε_{ij} is the error term (Single and Finch 1995; Stoesz *et al.* 1997). Generally speaking, given N families and n individuals per family, the number of sib pairs is $n' = \frac{n!}{2(n-2)!} > n$, so $n'/p > n/p$.

A hidden Markov model (HMM) is generally used to reconstruct the unobserved process of the IBS for the two alleles (Bercovici *et al.* 2010). The beginning allele frequencies, the conditional distribution of the phenotypes and transfer matrices of state in a Markov chain determine the joint distribution of the IBS and the hidden process in a straight way. Two fundamental algorithms, the forward and the backward algorithms, were developed for efficient computation of HMM.

Forward algorithm

We define IBS process as $J(t_m)$ at the mth marker and probability of the genotypes as $\Pr(G_m|j)$, where j is the status of the IBS. The transition probability of this HMM process might be defined in various ways. Here, we use $T_{ij} = \Pr(J(t_m) = j|J(t_{m-1}) = i)$ to denote it (Yu *et al.* 2007).

Table 1. Various forms of H–E regressions.

Method	Acronym	Measurements of dependent variables
Original	oHE	$0.5(Y_{1j} - Y_{2j})^2$
Revisited	rHE	$(Y_{1j} - \bar{Y})(Y_{2j} - \bar{Y})$
Weighed	wHE	$\frac{1}{2}\{(1-\omega)(Y_{1j} + Y_{2j} - 2\bar{Y})^2 - \omega(Y_{1j} - Y_{2j})^2\}$
Sibship sample mean	smHE	$(Y_{1j} - \bar{Y}_j)(Y_{2j} - \bar{Y}_j)$
Shrinkage mean	pmHE	$(Y_{1j} - \tilde{\mu}_j)(Y_{2j} - \tilde{\mu}_j)$

$\bar{Y}$, total mean; $\bar{Y}_j$, sibship mean; $\tilde{\mu}$, shrinkage mean; ω, weight value.

The forward algorithm computes recursively the following quantities:

$$F_m(j) = \Pr(G_1, G_2, \ldots, G_m, J(t_m) = j). \quad (2)$$

This is the joint probability distribution for the genotypes G, locus t_m and the IBS status at the locus. Using the nonafter-effect property of Markov chain, it is only on condition t_{m-1} states, thus equation (2) can be expressed as

$$\begin{aligned} F_m(j) &= \sum_i \Pr(G_1, G_2, \ldots, G_m, J(t_{m-1}) = i, J(t_m) = j), \\ &= \sum_i \{F_{m-1}(i) \times T_{ij} \times \Pr(G_m|j)\}. \end{aligned} \quad (3)$$

Backward algorithm

The backward algorithm, however, is used to compute the following quantities:

$$B_m(j) = \Pr(G_{m+1}, G_{m+2}, \ldots, G_{\tilde{m}}|J(t_m) = j). \quad (4)$$

Given the IBS status at a locus, equation (4) is the conditional probability distribution for the genotypes beyond the locus (Yu *et al.* 2007). On the given chromosome, the symbol $\tilde{m}$ denotes the indicator of the last marker. The backward algorithm starts values as $B_{\tilde{m}}(j) = 1$ and computes recursively a sum over the states of the Markov process as the following form:

$$\begin{aligned} B_m(j) &= \sum_i \Pr(G_{m+1}, \ldots, G_{\tilde{m}}, J(t_{m+1}) = i|J(t_m) = j), \\ &= \sum_i \{B_{m+1}(i) \times T_{ji} \times \Pr(G_{m+1}|i)\}. \end{aligned} \quad (5)$$

Within the given pedigree, G expresses marker on the chromosome. As

$$\Pr(G, J(t_m) = j) = \Pr(G_1, \ldots, G_{\tilde{m}}, J(t_m) = j) = F_m(j)B_m(j). \quad (6)$$

Here, equation (6) is the conditional probability, given the marker. It follows the conditional probabilities as

$$\Pr(J(t_m) = j|G) = \frac{F_m(j)B_m(j)}{\sum_i F_m(i)B_m(i)}. \quad (7)$$

It can be used to compute the conditional distribution of IBS at each locus.

Simulation study

Comparisons between GWAS using single marker regression and H–E regression were made for different scenarios. The simulated genome consisted of 10 chromosomes and each one was 1 Morgan. Each chromosome consisted of 2000 SNP markers that were almost evenly spaced and 10 randomly distributed QTLs, giving 20,000 markers and 19,900 potential QTLs in total. The gamma distribution (1.66, 0.4) was used to draw allele substitution effect (α) of the QTL (Meuwissen *et al.* 2001). All markers were biallelic with starting allele frequencies of 0.5 and a mutation rate of 2.5×10^{-8} per generation. Haldane's mapping function was used to model recombination between adjacent loci on a chromosome. And the recombination relies only on the distance between loci. Other parameters, including allele frequencies and locus positions were held constant. The simulation started with 100 unrelated individuals as a base population, followed by 1000 discrete historical generations which keep the same population size and randomly mated for creating LD between loci. In the base and following historical generation, one male mated randomly with 20 females, and each one produced two progenies. After this three additional generations were simulated. And the population size was expanded to 2000 by randomly mating one male with 50 females, and each mating produced 20 offspring. True breeding values and genotypes were simulated for all 2000 individuals, and phenotypic records of a continuous trait were assigned by $y = \mu + \alpha + e$, where e is residual variance and $e \sim N(0, \sigma_e^2)$. To study the effect of heritability and number of QTLs, two groups of scenarios were simulated. In the first group, three levels of heritability were simulated: 0.05, 0.1 and 0.5. In the second group, we simulated different numbers of QTLs: 20, 50 and 100. For two scenarios, 10 simulation replications using different random seeds were carried out to investigate the effect of the H–E regression.

Heterogeneous stock mice dataset

Solberg Woods *et al.* (2010) performed genomewide genetic association studies of complex traits in heterogeneous stock mice dataset, which consists of genotypes for 13,459 SNPs on 1,904 mice and 298 parents based on intercross mating among eight inbred strains (http://mus.well.ox.ac.uk/GSCAN/index.shtml/) (Valdar *et al.* 2006). Using the H–E regression model, we analysed a pair of immunological trait, $CD4^+/CD8^+$ ratio. After quality control, we exclude SNPs with minor allele frequency (MAF) < 0.01, Hardy-Weinberg equilibrium (HWE) $P < 10^{-4}$ or call rate < 0.99. Imputation of missing genotype data was on the basis of allelic frequencies in the dataset developed by Legarra and Misztal (2008). Then, 1,884 individuals (about 168 full-sib families) kept in the data file, each genotyped for 10,946 SNPs. Prior to the analysis, the trait was preadjusted for covariates and then SNP genotypes were ordered according to their genetic positions along the chromosome. IBD calculation for this dataset is implemented in the freely available BEAGLE software (http://www.stat.auckland.ac.nz/~bbrowning/beagle/beagle.html).

Results

Results from the simulations

Figure 1 depicts the simulated SNP effects and the QTL effects estimated by the H–E regression (figure 1a) and the single marker regression (figure 1b). The results show that the H–E

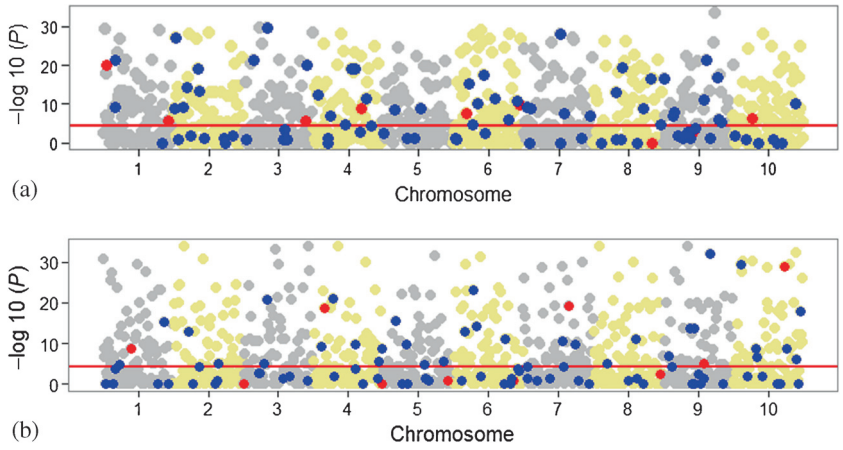

Figure 1. (a) Manhattan plots of association results for H–E regression and (b) single marker regression. Manhattan plots ordered by chromosome position. The red line indicates the genomewide significance threshold of 5.0×10^{-5}. Red points show SNP window harbouring the major QTL, and blue points illustrate SNP window containing the minor gene and its effect is one hundredth of the major QTL.

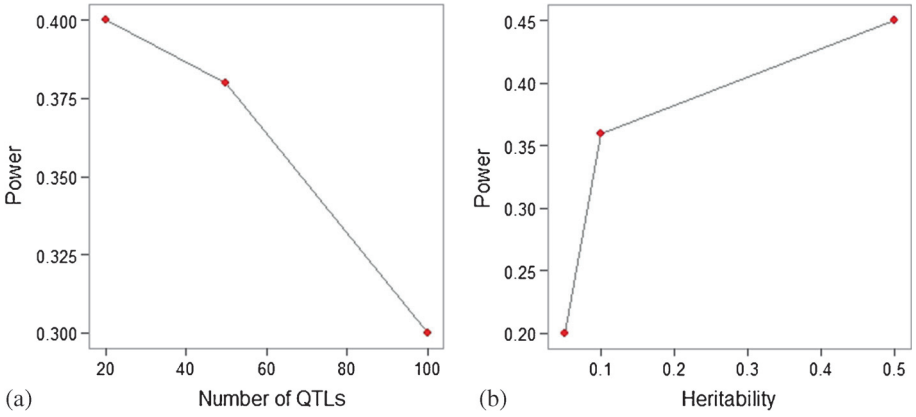

Figure 2. Power of H–E regression for different numbers of QTL and heritability in simulated datasets. The number of QTLs in figure 2a is 50, while the heritability changes from 0.05 to 0.5. The heritability in figure 2b is 0.3, while the number of QTLs increases from 10 to 100.

regression is better than single marker regression. The advantage of the H–E regression was also sufficient to capture about 48.01% of the QTL, and it was nearly 31.25% higher than the counterpart method.

Effect of number of QTLs: Figure 2a shows the power of the method under different number of QTLs. With the increase of it, the power declined consistently from 0.40 to 0.30.

Effect of heritability: As shown in figure 2b, the power of H–E regression is sensitive to heritabilities. By increasing the heritability from 0.05 to 0.5, the power of the method increases as expected from 0.20 to 0.45.

Results from the analysis of real heterogeneous stock mice dataset

Similar to simulation analyses, the H–E regression outperformed predictive abilities of the single marker regression for the $CD4^+/CD8^+$ ratio. As shown in figure 3, for the 1,884 individuals and 10,946 SNPs in the heterogeneous stock mice dataset, the effect of shrinkage for H–E regression was much stronger than single marker regression. On the basis of Atwell *et al.* (2010), the defense-related trait is essentially controlled by monogene. The result through the H–E regression should approve such a clear monogenic signal in terms of QTL identification. In table 2, two QTLs that have effect sizes less than 2%, are mapped into 50% confidence intervals (CI) of <1 Mb by bootstrapping method in Valdar *et al.* (2006).

Discussion

GWAS can be used as landmarks to locate genes that underlie complex traits. However, GWAS has several limitations and issues. One of common problems is insufficient sample size (de Los Campos *et al.* 2013). Ignoring this issue has been quoted as preventable cause for using the GWAS

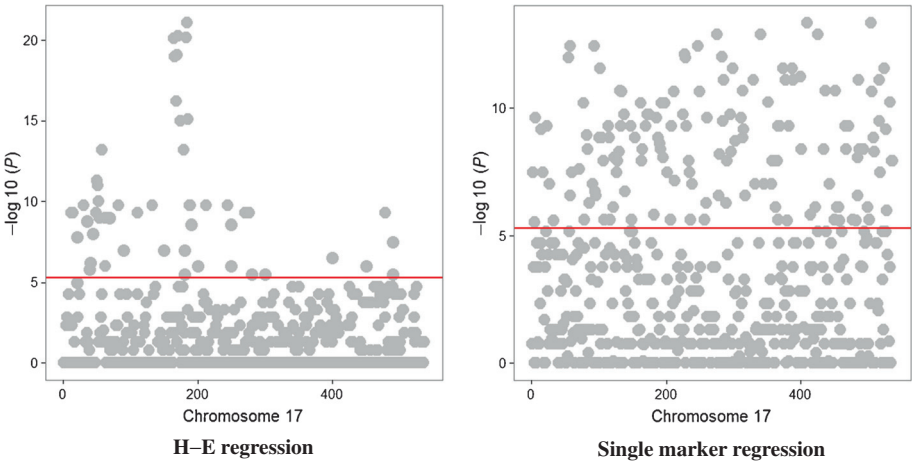

Figure 3. Ratio of CD4 to CD8 lymphocytes. The red line indicates the genomewide significance threshold of 4.6×10^{-6}.

Table 2. Highly significant QTLs of $CD4^+/CD8^+$ ratio.

Phenotype	Chr.	BPP	log *P*	Effect size	95% CI From	95% CI To
$CD4^+/CD8^+$	17	1.00	21.53	3.90	42.83	43.25
$CD4^+/CD8^+$	17	1.00	65.76	11.93	30.84	32.08

Chr., chromosome number; BPP, the bootstrap posterior probability of the QTL; log *P*, negative log 10 of the *P* value; effect size, the percentage variance explained by a given QTL; 95% CI, the bootstrap 95% CI for the QTL (in Mb).

methodology. Genomewide linkage study (GWLS) is another important tool to identify inherited phenotype, which is constantly replaced by GWAS when it became more attractive (Daetwyler *et al.* 2013). Historically, H–E regression method was the major tool for sib-pair data (Ziegler *et al.* 2001; Zhang *et al.* 2008). In this study, we modified traditional H–E regression and established a theory framework for genomewide studies. By this means, 'large *p* small *n*' problem can be partly solved with combination of sib-pair data. Further, more complicated issue–multiple correlated responses have been showed as slightly reduced by this improvement (Won *et al.* 2006). Development of the original H–E regression to GWAS can be considered as an attempt to promote the diversity of GWAS analysis methods.

Application of the H–E regression to GWAS analysis revealed that it is equivalent to the mixed model methods. And it is based on the least square framework rather than the maximum likelihood framework as mixed model methods. But this kind of equality is based on the polygenic genetic architecture of quantitative traits and genomes (Weeks and Harby 1995). And the selection in active regions resulted in covariance between loci which leads to biased estimates as a mixed model method. In addition, H–E regression is more accurate than the mixed linear model for special experiment design of QTL mapping, such as case–control studies, and is preferable when the number of causal loci is less. Suspiciously high relatedness, such as population stratification can increase spurious association. To reduce this effect, data can be easily adjusted by principal component analysis and then modelled by H–E regression. Undocumented relatedness may reduce the power of the H–E regression (Chen 2014). For instance, if a population is admixed or genetic heterogeneity, the allele frequency of the ancestor is different, and each allele has the different ancestral origin in these individuals.

Estimation of IBD in population is meaningful for population-based linkage analysis. The IBD calculations may not be accurate if they involve allele or haplotype frequencies unless we have a large group. Our method relies mostly on the accuracy of the IBD estimate, and it is computationally expensive to apply our IBD-detection method to all pairs of individuals in large samples. But in our main analysis, the computational complex of the H–E regression is asymptotically $O(2N^2)$, which shows that it is two times of square to the sample size. Given the large data sizes in GWAS, the computational intensity of H–E regression can be extraordinarily decreased.

Conclusions

In this study we resurrected traditional H–E regression in the use of GWAS. Using simulation methods and the real data, we proved that the modified approach has better statistical power than the single marker regression. Meanwhile, we evaluated the power of H–E regression for different number of QTLs and heritability. The results show that the power decreases with the increase in the number of QTLs, and the power of H–E regression is sensitive to heritability.

Acknowledgements

We thank the editor and referees for helpful comments. This work was supported by the National Natural Science Foundation of China (grant no. 31460594), China Scholarship Council (grant no. 201308155140), Hetao College teaching and research project (grant no. HTXYJZ14005).

References

Atwell S., Huang Y. S., Vilhjalmsson B. J., Willems G., Horton M., Li Y. *et al.* 2010 Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631.

Barber M. J., Cordell H. J., MacGregor A. J. and Andrew T. 2004 Gamma regression improves Haseman–Elston and variance components linkage analysis for sib-pairs. *Genet. Epidemiol.* **26**, 97–107.

Bercovici S., Meek C., Wexler Y. and Geiger D. 2010 Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping. *Bioinformatics* **26**, i175–i182.

Chen G. B. 2014 Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman–Elston regression. *Front. Genet.* **5**, 107.

Daetwyler H. D., Calus M. P., Pong-Wong R., de Los Campos G. and Hickey J. M. 2013 Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* **193**, 347–365.

de Los Campos G., Hickey J. M., Pong-Wong R., Daetwyler H. D. and Calus M. P. 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**, 327–345.

DeFries J. C. 2010 Haseman and Elston sib-pair linkage analysis: a brief historical note. *Behav. Genet.* **40**, 1–2.

Diao G. and Vidyashankar A. N. 2013 Assessing genome-wide statistical significance for large *p* small *n* problems. *Genetics* **194**, 781–783.

Drigalenko E. 1999 Matrix representation of the Haseman–Elston method. *Theor. Popul. Biol.* **55**, 157–165.

Elston R. C., Buxbaum S., Jacobs K. B. and Olson J. M. 2000 Haseman and Elston revisited. *Genet. Epidemiol.* **19**, 1–17.

Etzel C. J., Shete S., Beasley T. M., Fernandez J. R., Allison D. B. and Amos C. I. 2003 Effect of Box–Cox transformation on power of Haseman–Elston and maximum-likelihood variance components tests to detect quantitative trait loci. *Hum. Hered.* **55**, 108–116.

Forrest W. F. 2001 Weighting improves the new Haseman–Elston method. *Hum. Hered.* **52**, 47–54.

Franke D., Kleensang A., Elston R. C. and Ziegler A. 2005 Haseman–Elston weighted by marker informativity. *BMC Genet.* **6** suppl 1, S50.

Garner C. P. 2002 Nonparametric linkage analysis. I. Haseman–Elston. *Methods Mol. Biol.* **195**, 37–60.

Gerhard D. and Hothorn L. A. 2010 Rank transformation in Haseman–Elston regression using scores for location-scale alternatives. *Hum. Hered.* **69**, 143–151.

Hadicke O., Pahlke F. and Ziegler A. 2008 A general approach for sample size and power calculations based on the Haseman–Elston method. *Biom. J.* **50**, 257–269.

Legarra A. and Misztal I. 2008 Technical note: computing strategies in genome-wide selection. *J. Dairy. Sci.* **91**, 360–366.

Meuwissen T. H., Hayes B. J. and Goddard M. E. 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

Sham P. C. and Purcell S. 2001 Equivalence between Haseman–Elston and variance-components linkage analyses for sib pairs. *Am. J. Hum. Genet.* **68**, 1527–1532.

Shen X., Alam M., Fikse F. and Ronnegard L. 2013 A novel generalized ridge regression method for quantitative genetics. *Genetics* **193**, 1255–1268.

Shete S., Jacobs K. B. and Elston R. C. 2003 Adding further power to the Haseman and Elston method for detecting linkage in larger sibships: weighting sums and differences. *Hum. Hered.* **55**, 79–85.

Single R. M. and Finch S. J. 1995 Gain in efficiency from using generalized least squares in the Haseman–Elston test. *Genet. Epidemiol.* **12**, 889–894.

Solberg Woods L. C., Holl K., Tschannen M. and Valdar W. 2010 Fine-mapping a locus for glucose tolerance using heterogeneous stock rats. *Physiol. Genomics* **41**, 102–108.

Stoesz M. R., Cohen J. C., Mooser V, Marcovina S. and Guerra R. 1997 Extension of the Haseman–Elston method to multiple alleles and multiple loci: theory and practice for candidate genes. *Ann. Hum. Genet.* **61**, 263–274.

Valdar W., Solberg L. C., Gauguier D., Burnett S., Klenerman P., Cookson W. O. *et al.* 2006 Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.* **38**, 879–887.

Wang T. and Elston R. C. 2005 Two-level Haseman–Elston regression for general pedigree data analysis. *Genet. Epidemiol.* **29**, 12–22.

Weeks D. E. and Harby L. D. 1995 The affected-pedigree-member method: power to detect linkage. *Hum. Hered.* **45**, 13–24.

Won S., Elston R. C. and Park T. 2006 Extension of the Haseman–Elston regression model to longitudinal data. *Hum. Hered.* **61**, 111–119.

Xu X., Weiss S., Xu X. and Wei L. J. 2000 A unified Haseman–Elston method for testing linkage with quantitative traits. *Am. J. Hum. Genet.* **67**, 1025–1028.

Yoon S., Suh Y. J., Mendell N. R. and Ye K. Q. 2005 A Bayesian approach for applying Haseman–Elston methods. *BMC Genet.* **6** suppl 1, S39.

Yu T., Ye H., Sun W., Li K. C., Chen Z., Jacobs S. *et al.* 2007 A forward–backward fragment assembling algorithm for the identification of genomic amplification and deletion breakpoints using high-density single nucleotide polymorphism (SNP) array. *BMC Bioinformatics* **8**, 145.

Zhang Y. M., Lu H. Y. and Yao L. L. 2008 Multiple quantitative trait loci Haseman–Elston regression using all markers on the entire genome. *Theor. Appl. Genet.* **117**, 683–690.

Ziegler A., Boddeker I. R. and Geller F. 2001 A bivariate Haseman–Elston method and application to the analysis of asthma-related phenotypes on chromosome 5q. *Genet. Epidemiol.* **21** suppl 1, S216–S221.

Received 28 October 2015, in revised form 27 January 2016; accepted 17 March 2016
Unedited version published online: 21 March 2016
Final version published online: 29 November 2016

Corresponding editor: RAJIVA RAMAN