The Heritability of a Quantitative Trait Locus

Mitchell J. Feldmann, Department of Plant Sciences, University of California, Davis
William C. Bridges, Department of Mathematical Sciences, Clemson University
Steven J. Knapp, Department of Plant Sciences, University of California, Davis

Introduction

Heritability is a fundamental concept for the study of genotypic and phenotypic variation in natural and experimental populations in biology, agriculture, and medicine. Genetic variance components are estimated through calculating the mean squares, comparing to expectations, and solving for unknown variances [1]. These particular methods can lead to hard-to-interpret, negative trait heritability [2], which can be circumvented by using biased estimators (e.g. restricted maximum likelihood (REML)) [3]. The precision of these estimates varies widely over different mating designs, sample sizes, and traits [4,5,6,7]. While issues with overestimating quantitative trait locus (QTL) effect sizes and the associated heritability have been mentioned [8,9], there has yet to be a formal discussion to address potential solutions.

Objectives

1. Demonstrate the $\hat{h}^2_{*L_i}$ estimate is an overestimation of percent variation associated with individual QTL.
2. Develop an accurate estimate of $\hat{h}^2_{*L_i}$ for balanced datasets.
3. Discover underlying pattern for higher order problems (3+ loci).

Theory: Model and Variance Components

Consider the model:

$$\mathbf{Y_{ijk} = \mu + L_{1i} + L_{2j} + L_1L_{2ij} + \epsilon_{ijk}}$$

- Y_{ijk} is the phenotypic value of a progeny with genotype i at L_1, genotype j at L_2, and measured in replication k
- μ is the overall (global) phenotypic mean
- L_{1i} is the main effect of genotype i at L_1
- L_{2j} is main the effect of genotype j at L_2
- L_1L_{2ij} is the effect of the interaction (epistasis) between L_1 and L_2 with genotype i at L_1; genotype j at L_2
- ϵ_{ijk} is random error

ANOVA will have the following form:

	df	SS	MS	E(MS)
G	df_G	SS_G	SS_G / df_G	$\sigma^2_E + r\sigma^2_G$
L_1	df_{L1}	SS_{L1}	SS_{L1} / df_{L1}	$\sigma^2_E + r\sigma^2_{L1L2} + rl_1\sigma^2_{L1}$
L_2	df_{L2}	SS_{L2}	SS_{L2} / df_{L2}	$\sigma^2_E + r\sigma^2_{L1L2} + rl_2\sigma^2_{L2}$
L_1L_2	df_{L1L2}	SS_{L1L2}	SS_{L1L2} / df_{L1L2}	$\sigma^2_E + r\sigma^2_{L1L2}$
Err	df_E	SS_E	SS_E / df_E	σ^2_E

Henderson's variance components:

$$\hat{\sigma}^2_G = \frac{SS_G/df_G - SS_E/df_E}{r}$$
$$\hat{\sigma}^2_{L_1} = \frac{SS_{L_1}/df_{L_1} - SS_{L_1L_2}/df_{L_1L_2}}{rl_2}$$
$$\hat{\sigma}^2_{L_2} = \frac{SS_{L_2}/df_{L_2} - SS_{L_1L_2}/df_{L_1L_2}}{rl_1}$$
$$\hat{\sigma}^2_{L_1L_2} = \frac{SS_{L_1L_2}/df_{L_1L_2} - SS_E/df_E}{r}$$

- r is the replicates of each genotype
- L_1 is the number of genotypes for locus 1
- L_2 is the number of genotypes for locus 2.

The definition of heritability for a QTL (L_i) used in this study is:

$$\hat{h}^2_{*L_i} = \frac{\hat{\sigma}^2_{L_i}}{\hat{\sigma}^2_G}$$

The denominator is the total genetic variance (σ^2_G), *not* total variance components for the QTL ($\sigma^2_Q = \sigma^2_{L_1} + \sigma^2_{L_2} + \sigma^2_{L_1L_2}$); the true number of QTL controlling a trait is unknown.

Theory: Overestimation and Correction

Even though $h^2_{*L_i}$ has the denominator σ^2_G instead of σ^2_Q the estimate is still an overestimation of the percent variation associated with QTL L_i. The sum of the estimates associated with all the QTL is taken as an estimator of σ^2_Q, the result is:

$$\hat{\sigma}^2_Q = \hat{\sigma}^2_{L_1} + \hat{\sigma}^2_{L_2} + \hat{\sigma}^2_{L_1L_2}$$

Substitute Henderson's variance estimators:

$$\hat{\sigma}^2_Q = \frac{SS_{L_1}/df_{L_1} - SS_{L_1L_2}/df_{L_1L_2}}{rl_2} + \frac{SS_{L_2}/df_{L_2} - SS_{L_1L_2}/df_{L_1L_2}}{rl_1} + \frac{SS_{L_1L_2}/df_{L_1L_2} - SS_E/df_E}{r}$$

After collecting terms and noting that $df_{L_1} = df_{L_2} = df_L$ and $SS_G = SS_{L_1} + SS_{L_2} + SS_{L_1L_2}$ we arrive at:

$$\hat{\sigma}^2_Q = \hat{\sigma}^2_G + (\frac{df_L}{df_G})(\hat{\sigma}^2_{L_1} + \hat{\sigma}^2_{L_2})$$

which, after solving 6 for $\hat{\sigma}^2_G$ and substituting 4 for $\hat{\sigma}^2_Q$, leads to:

$$\hat{\sigma}^2_G = \hat{\sigma}^2_Q - \left(\frac{df_L}{df_G}\right)(\hat{\sigma}^2_{L_1} + \hat{\sigma}^2_{L_2})$$
$$= \left(1 - \frac{df_L}{df_G}\right)(\hat{\sigma}^2_{L_1} + \hat{\sigma}^2_{L_2}) + \hat{\sigma}^2_{L_1L_2}$$

Since $\hat{\sigma}^2_G$ is less than $\hat{\sigma}^2_Q$ by fractional values of the variance estimate for the individual QTL a reasonable correction for the overestimation in the estimator of $\hat{h}^2_{*L_i}$ when there are 3 genotypes at each locus is:

$$\hat{h}^2_{*L_i} = \frac{\left(1 - \frac{df_L}{df_G}\right)\hat{\sigma}^2_{L_i}}{\hat{\sigma}^2_G} = \frac{\frac{3}{4}\hat{\sigma}^2_{L_i}}{\hat{\sigma}^2_G} = \frac{0.75\hat{\sigma}^2_{L_i}}{\hat{\sigma}^2_G}$$

This correction term should then be applied to all loci in the model when estimating individual QTL heritability.

Results

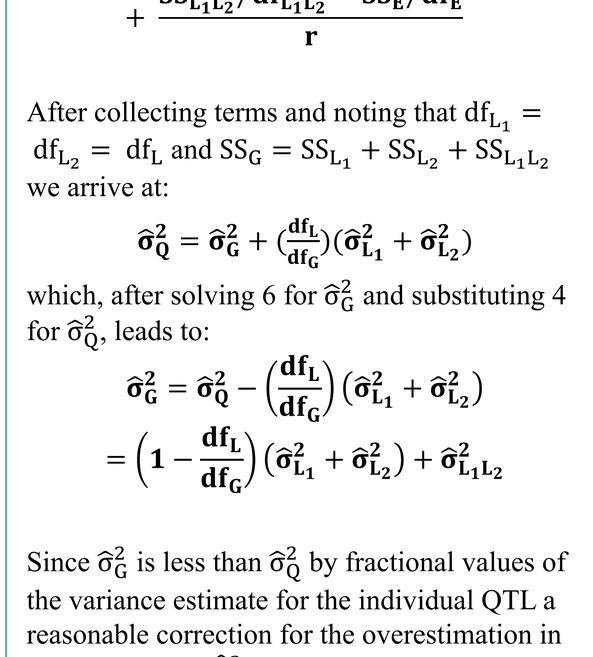

Figure 1 – Assumptions of ANOVA: Simulated data from population (upper left), data viewed by each interaction level (upper right), Residuals from ANOVA model (lower left), Q-Q plot from ANOVA model (lower right).

	ANOVA	REML	Absolute Difference
L1	226.2534326	226.2533925	4.0126824e-05
L2	62.9633623	62.9633522	1.0065570e-05
L12	11.0512475	11.0512439	3.5280943e-06
G	227.9638436	227.9638436	1.0015014e-08
Q	300.2680423	300.2679886	5.3720489e-05

Table 1 – Comparison of variance component estimates of the two-loci model from ANOVA and REML.

	ANOVA	REML	Absolute Difference
h^2 L1	0.99249700	0.99249683	1.7606636e-07
Corrected	0.74437275	0.74437262	1.3204977e-07
h^2 L2	0.27619889	0.27619885	4.4166377e-08
Corrected	0.20714917	0.20714914	3.3124782e-08
Total h^2	1.26869590	1.26869568	2.2023273e-07
Corrected	0.95152192	0.95152176	1.6517455e-07

Table 2 – Percent of heritability associated with individual QTL from the two-loci model before and after applying correction from ANOVA and REML estimates and the absolute difference of these estimates.

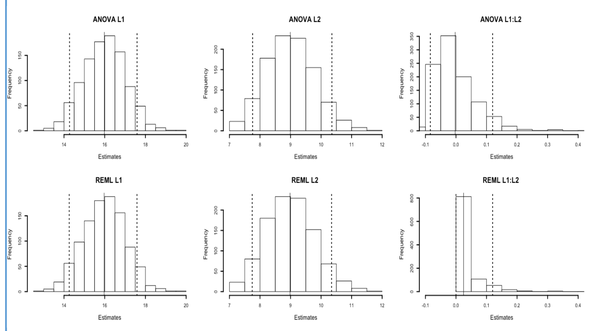

Figure 2 – Accuracy of Variance Component Estimation. ANOVA (top row) and REML (bottom row) estimators of $\beta_{L1} = 4$, $\beta_{L2} = 3$, and $\beta_{L1L2} = 0$. Mean (solid vertical line) and 95% confidence interval (dashed vertical line).

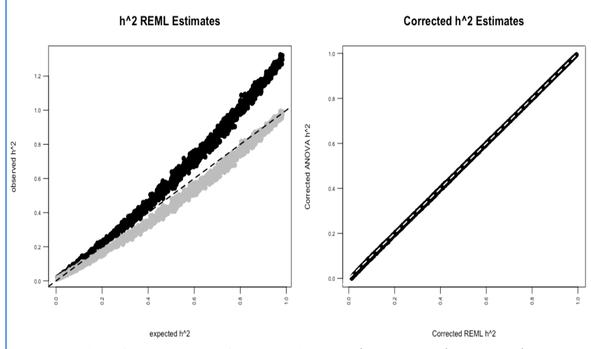

Figure 3 – Correction Comparison: The correction greatly reduces the REML overestimation of heritability calculated from two loci and the corrected h^2 estimates from ANOVA and REML coincide (right). The dashed lines $y = x$.

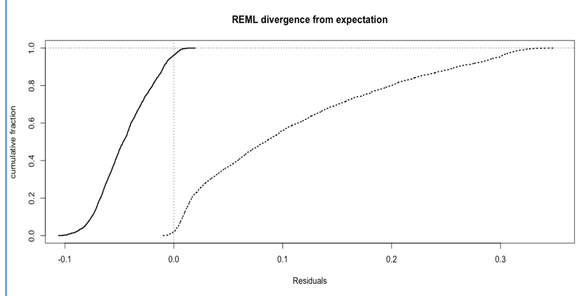

Figure 4 – Empirical cumulative distributions of the divergence of the observed REML h^2 from the expected for the corrected (solid) and uncorrected (dashed) estimates. Dashed lines at *x=0* & *y=1*.

Theory: Higher Order Patterns

Following the same logic presented, corrections for models consisting of three, four and five loci have been solved.

Three Loci

$$\hat{\sigma}^2_G = \left(1 - \frac{2df_L + df_L^2}{df_G}\right)(\hat{\sigma}^2_{L_1} + \hat{\sigma}^2_{L_2} + \hat{\sigma}^2_{L_3}) + (1 - \frac{df_L}{df_G})(\hat{\sigma}^2_{L_1L_2} + \hat{\sigma}^2_{L_1L_3} + \hat{\sigma}^2_{L_2L_3}) + \hat{\sigma}^2_{L1L2L3}$$

$$\hat{h}^2_{*L_i} = \frac{(1 - \frac{2df_L + df_L^2}{df_G})\hat{\sigma}^2_{L_i}}{\hat{\sigma}^2_G}$$

Four Loci

$$\hat{\sigma}^2_G = \left(1 - \frac{3df_{L_i} + 3df_{L_i}^2 + df_{L_i}^3}{df_G}\right)(\hat{\sigma}^2_{L1} + \hat{\sigma}^2_{L2} + \hat{\sigma}^2_{L3} + \hat{\sigma}^2_{L4}) + \left(1 - \frac{2df_L + df_L^2}{df_G}\right)(\hat{\sigma}^2_{L1L2} + \hat{\sigma}^2_{L1L3} + \hat{\sigma}^2_{L1L4} + \hat{\sigma}^2_{L2L3} + \hat{\sigma}^2_{L2L4} + \hat{\sigma}^2_{L3L4}) + \left(1 - \frac{df_L}{df_G}\right)(\hat{\sigma}^2_{L1L2L3} + \hat{\sigma}^2_{L1L2L4} + \hat{\sigma}^2_{L1L3L4} + \hat{\sigma}^2_{L2L3L4}) + \hat{\sigma}^2_{L1L2L3L4}$$

$$\hat{h}^2_{*L_i} = \frac{(1 - \frac{3df_{L_i} + 3df_{L_i}^2 + df_{L_i}^3}{df_G})\hat{\sigma}^2_{L_i}}{\hat{\sigma}^2_G}$$

Five Loci

$$\hat{\sigma}^2_G = \left(1 - \frac{4df_{L_i} + 6df_{L_i}^2 + 4df_{L_i}^3 + df_{L_i}^4}{df_G}\right)(\hat{\sigma}^2_{L1} + \hat{\sigma}^2_{L2} + \hat{\sigma}^2_{L3} + \hat{\sigma}^2_{L4} + \hat{\sigma}^2_{L5}) + \left(1 - \frac{3df_{L_i} + 3df_{L_i}^2 + df_{L_i}^3}{df_G}\right)(\hat{\sigma}^2_{L1L2} + \hat{\sigma}^2_{L1L3} + \hat{\sigma}^2_{L1L4} + \hat{\sigma}^2_{L1L5} + \hat{\sigma}^2_{L2L3} + \hat{\sigma}^2_{L2L4} + \hat{\sigma}^2_{L2L5} + \hat{\sigma}^2_{L3L4} + \hat{\sigma}^2_{L3L5} + \hat{\sigma}^2_{L4L5}) + \left(1 - \frac{2df_L + df_L^2}{df_G}\right)(\hat{\sigma}^2_{L1L2L3} + \hat{\sigma}^2_{L1L2L4} + \hat{\sigma}^2_{L1L2L5} + \hat{\sigma}^2_{L1L3L4} + \hat{\sigma}^2_{L1L3L5} + \hat{\sigma}^2_{L1L4L5} + \hat{\sigma}^2_{L2L3L4} + \hat{\sigma}^2_{L2L3L5} + \hat{\sigma}^2_{L2L4L5} + \hat{\sigma}^2_{L3L4L5}) + \left(1 - \frac{df_L}{df_G}\right)(\hat{\sigma}^2_{L1L2L3L4} + \hat{\sigma}^2_{L1L2L3L5} + \hat{\sigma}^2_{L1L2L4L5} + \hat{\sigma}^2_{L1L3L4L5} + \hat{\sigma}^2_{L2L3L4L5}) + \hat{\sigma}^2_{L1L2L3L4L5}$$

$$\hat{h}^2_{*L_i} = \frac{(1 - \frac{4df_{L_i} + 6df_{L_i}^2 + 4df_{L_i}^3 + df_{L_i}^4}{df_G})\hat{\sigma}^2_{L_i}}{\hat{\sigma}^2_G}$$

Conclusions

1. E(MS) were used to calculate variance components and demonstrated that $\hat{\sigma}^2_Q$ is fractionally larger than $\hat{\sigma}^2_G$.
2. The proposed correction properly adjusts overestimations for traits with varying heritability and locus effect sizes.
3. When models rely on 3+ loci, the expanded correction expands in a predictable pattern in a way that resembles Pascal's Triangle.

Acknowledgments

The authors would like to thank Dr. Julia Harshman for her guidance drafting the manuscript and criticisms on content.

References:

1. Henderson. 1953. Estimation of variance and Covariance components. Biometrics.
2. Gill & Jensen. 1968. Probability of obtaining negative estimates of heritability. Biometrics.
3. Bridges & Knapp. 1987. Probabilities of negative estimates of genetic variances. Theor Appl Genet.
4. Knapp et al. 1985. Exact confidence intervals for heritability on a progeny mean basis. Crop Sci.
5. Knapp. 1986. Confidence intervals for heritability for two-factor mating design single environment linear models. Theor Appl Genet.
6. Knapp & Bridges. 1987. Confidence interval estimators for heritability for several mating and experiment designs. Theor Appl Genet.
7. Knapp et al. 1987. Precision of genetic variance and heritability estimates from sorghum populations. Crop Sci.
8. Melschinger et al. 1998. Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of OTL effects. Genetics.
9. Moreau et al. 1998. Marker-assisted selection efficiency in populations of finite size. Genetics.