The Genetic Architecture of Quantitative Traits Cannot Be Inferred From Variance Component Analysis

Wen Huang* and Trudy F.C. Mackay*

Program in Genetics, W. M Keck Center for Behavioral Biology,

Initiative for Biological Complexity and Department of Biological Sciences,

North Carolina State University, Raleigh, NC 27695-7614 USA

* Corresponding Authors

Tel: 919-515-5810

Fax: 919-515-3355

Email: wen.huang@ncsu.edu (W.H.), trudy_mackay@ncsu.edu (T.F.C.M.)

1

Abstract

Classical quantitative genetic analyses estimate additive and non-additive genetic and environmental components of variance from phenotypes of related individuals. The genetic variance components are defined in terms of genotypic values reflecting underlying genetic architecture (additive, dominance and epistatic genotypic effects) and allele frequencies. However, the dependency of the definition of genetic variance components on the underlying genetic models is not often appreciated. Here, we show how the partitioning of additive and non-additive genetic variation is affected by the genetic models and parameterization of allelic effects. We show that arbitrarily defined variance components often capture a substantial fraction of total genetic variation regardless of the underlying genetic architecture in simulated and real data. Therefore, variance component analysis cannot be used to infer genetic architecture of quantitative traits. The genetic basis of quantitative trait variation in a natural population can only be defined empirically using high resolution mapping methods followed by detailed characterization of QTL effects.

29   **Introduction**

30

31   Classical quantitative genetic analysis begins with the premise that the amount of

32   phenotypic variation of quantitative traits in a natural population is determined by the

33   genetic variation of the trait and environmental variation that is random in nature (Fisher

34   1918; Falconer and Mackay 1996). Genetic variation of quantitative traits can be further

35   partitioned into the additive variance (V_A), dominance variance (V_D), and inter-locus

36   interaction (epistatic) variance (V_I) (Falconer and Mackay 1996). All of the genetic variance

37   components can be defined in terms of an underlying genetic model with additive,

38   dominance and epistatic genotypic effects and allele frequencies (Table 1) (Fisher 1918).

39   These quantitative genetic variance components are key parameters in theoretical

40   treatments of quantitative trait variation. The additive genetic variance V_A is of particular

41   importance because it defines the level of narrow sense heritability (h^2), which determines

42   the rate of response to natural or artificial selection (Lush 1943).

43   The relative importance of V_A, V_D, and V_I and especially the additive, dominant, and

44   epistatic gene actions they imply is controversial. Although genetic mapping studies

45   frequently detect non-additive intra- and inter-locus interactions, variance component

46   estimates nearly unanimously find that V_A is the majority of the genetic variance, and

47   dwarfs the contribution of V_D and V_I (Hill *et al.* 2008; Bloom *et al.* 2013; Mäki-Tanila and

48   Hill 2014; Zhu *et al.* 2015). These studies are the basis for the argument that V_A is the

49   overwhelming determinant of genetic variation. Therefore, it is often stated that non-

50   additive variance – and implicitly non-additive gene action – plays only a minor role in the

51   genetic architecture of quantitative traits because non-additive gene action mostly

52 contributes to V_A; and because the magnitude of V_D and V_I from dominance and epistasis,

53 respectively is small (Hill *et al.* 2008; Mäki-Tanila and Hill 2014).

54   However, there are many and potentially an infinite number of possible ways other

55 than V_A, V_D, and V_I to partition genetic variance, all of which are dependent on allele

56 frequencies in natural populations (Zeng *et al.* 2005). The implicit equivalence of just one of

57 the ways, V_A, V_D, and V_I with additive, dominant, and epistatic gene action respectively is

58 problematic. Because V_A seeks to maximize the variance it explains, unless genotype

59 frequencies meet specific conditions determined by the genetic models, all types of gene

60 actions can contribute to V_A (Hill *et al.* 2008), making it impossible to use variance

61 components to quantify the contributions of them. However, with a few exceptions, for

62 example, by distinguishing statistical and physiological epistasis (Cheverud and Routman

63 1995; Alvarez-Castro and Carlborg 2007), there has been a general lack of effort to

64 distinguish variance components from the underlying gene actions that contribute to them,

65 which has caused great confusion. Here, we show that the partitioning of genetic variance

66 components is dependent on the parameterization of allelic effects, and that arbitrarily

67 defined variance components have similar variance explaining ability as V_A, regardless of

68 genetic models. We argue that the classical definition of V_A is often inappropriate in the

69 context of explaining within-generation variance, and that partitioning of genetic variance

70 in general provides no information regarding the underlying genetic architecture of

71 quantitative traits.

72

73 **Results**

74

4

Additive genetic variance V_A is a major determinant of total genetic variance

Following conventional notation, we arbitrarily assign the genotypic value of the three possible genotypes aa, Aa, and AA at a single bi-allelic locus as $-a$, d, and $+a$ respectively (Falconer and Mackay 1996). Additive and dominant gene actions (or "genetic model") have a clear meaning with this parameterization. An "additive" genetic model refers to the situation in which $d = 0$, and hence there is a completely linear relationship between the genotypic value and the number of copies of A alleles. A "dominant" genetic model is when $d = \pm a$, or when the genotypic value is solely determined by the presence of the dominant allele. V_A (see Table 1 for this and other notations and definitions used throughout this study) accounts for the entirety of genetic variation when the true genetic model is an additive model (Figure 1a). V_A also explains the majority of genetic variation under the dominant genetic model unless the dominant allele is at high frequency (Figure 1b). Extending this single-locus model to two unlinked loci, it can be shown that V_A captures the majority of overall genetic variance unless both loci are frequent under a two-locus "additive by additive" genetic model (Figure 1c,d). These simple results have been previously shown by many authors (Falconer and Mackay 1996; Hill *et al.* 2008; Mackay 2014) but are reproduced here to set the stage for the following results.

Alternative parameterizations also capture the majority of genetic variance

The discrepancy between the well-defined genetic models and the literal implication of the term additive genetic variation (V_A), *i.e.*, apparently non-additive genetic models (dominant or additive by additive) produce significant additive variance, is confusing and counterintuitive. Realizing that V_A is a population property and relies on a specific

5

98   parameterization of genotypes, we derive alternative parameterizations and quantify the

99   variance explained by them (Table 1). Using a single-locus parameterization in which the

100   heterozygotes and the homozygotes for the dominant allele are coded identically, we define

101   an alternative dominance variance V_D' (Table 1, the prime symbol is used to distinguish this

102   variance from the conventional dominance variance V_D), which is found to capture the

103   entire genetic variance when the true genetic model is a dominant model (Figure 2a).

104   Importantly, similar to the phenomenon in the classical parameterization in which V_A

105   absorbs the majority of the genetic variance, under this parameterization V_D' can explain a

106   substantial amount of genetic variance even when the true genetic model is additive

107   (Figure 2b). Furthermore, an alternative two-locus parameterization (see Methods) allows

108   a newly defined V_{AA}'' variance component (Table 1) to explain the entire genetic variance

109   with an additive by additive genetic model (Figure 2c) while still capturing a majority of

110   genetic variance under most circumstances when the genetic model is purely additive

111   (Figure 2d).

112

113   **Conventional and alternative parameterizations capture the majority of polygenic**

114   **genetic variance**

115   To extend the single- and two-locus results to polygenic genetic models, we simulated

116   genotypes and phenotypes based on pre-defined genetic architecture and broad sense

117   heritability (H^2), and used mixed models to partition phenotypic variance under the

118   classical and alternative parameterizations described above. As expected, when the genetic

119   parameterizations and the resulting genetic covariance matrices match the true genetic

120   models, the estimated variances fully explain the total genetic variances (Figure 3).

6

121   Intriguingly, similar to the single- and two-locus models, all genetic parameterizations are

122   able to capture a major (almost always > 40%) fraction of total genetic variances regardless

123   of the true genetic architecture (Figure 3). Among the three parameterizations, the classical

124   definition of V_A appears to explain the most genetic variance when the genetic model does

125   not match its parameterization. This is likely because the genotypic coding under the

126   conventional additive parameterization is insensitive to the sign of the allelic effects; while

127   the dominance parameterization requires prior knowledge of the dominant allele, and the

128   additive by additive parameterization requires prior knowledge of the interacting pairs.

129   Nonetheless, it is remarkable that even with random assignment of the dominant allele or

130   random pairing of loci and obvious mischaracterization of the genetic model, V_D' and V_{AA}''

131   are able to explain the majority of genetic variance when the genetic architecture is

132   additive within and between loci (Figure 3b,c).

133

134   **V_A, V_D', and V_{AA}'' explain a large fraction of phenotypic variance for human height**

135   It has been previously shown that V_A accounts for a large fraction of phenotypic variance in

136   human height using a genetic covariance matrix computed from genome-wide SNP data

137   under the conventional parameterization (Yang *et al.* 2010). Based on our above results, we

138   necessarily expect this result regardless of genetic architecture and indeed recapitulated it

139   using genotype and height data for individuals from the GENEVA project (Figure 4). We

140   then asked if V_D' and V_{AA}'' under alternative parameterizations can do the same as in

141   simulated data. Remarkably, and under the naive assumptions that minor alleles are

142   recessive and randomly pairing interacting SNPs, both V_D' and V_{AA}'' can explain a substantial

143   fraction of phenotypic variance, with even larger point estimates than V_A (Figure 4).

7

144   A recent study reported a major contribution of V_A and a minor contribution of V_D

145   for a number of quantitative traits, using the classical parameterization for the additive

146   genetic variance to estimate V_A, and a frequency-dependent parameterization (Table 1)

147   orthogonal to the additive genetic value to estimate V_D (Zhu *et al.* 2015). This observation

148   led to the conclusion that dominance variation contributes little to quantitative trait

149   variation. Using the same method, we also observed similar relative contributions of V_A and

150   V_D for human height in the GENEVA data (Figure 4). However, this is only one of the many

151   possible ways of partitioning variance. Using our alternatively defined parameterizations

152   and a similar frequency-dependent parameterization orthogonal to the dominance genetic

153   value (Table 1), we find a much more substantial contribution of dominance variance, *i.e.*,

154   V_D' (Figure 4), further suggesting that partitioning of genetic variance is dependent on

155   parameterization.

156

Discussion

158

159   In animal breeding, V_A and h^2 estimated from pedigree data inform us about the proportion

160   of phenotypic variation that is "breedable", which is precisely how these genetic

161   parameters were initially defined and used. On the other hand, while broad sense

162   heritability (H^2) is the true measure of the contribution of all sources of genetic variation, it

163   is impossible to estimate in most natural populations, including humans, because no

164   replicated measures can be made on genetically identical individuals, and therefore h^2

165   provides a lower bound of H^2. Our results and those of many others clearly indicate that it

166   is a reasonable lower bound because in most cases h^2 is close to H^2.

167   While V_A is certainly useful to give an indication of the magnitude of H^2, our results

168   show that it is not the only one. Alternative parameterizations and the variance they

169   explain, V_D' and V_{AA}'', do similarly well in most cases in simulated as well as real data. The

170   emphasis on V_A in animal breeding is sensible because V_A but not V_D' or V_{AA}'' determines the

171   potential of genetic progress (Lush 1943). However, there is little reason to take this V_A

172   centric perspective when explaining within-generation genetic variation. In natural

173   populations, whether one of these variance components is more useful than another

174   depends on the true underlying genetic models. Though it is impossible to determine a

175   clear winner, it is obvious that V_A fully explains genetic variation under an additive model,

176   V_D' under a dominant model, and V_{AA}'' under an additive by additive epistatic model.

177   Therefore it seems only appropriate to define V_A as the additive variance when the genetic

178   model is additive, but to define V_D' rather than V_D as the dominance variance when the

179   genetic model is dominant, and to define V_{AA}'' as the epistatic variance when the genetic

180   model is entirely additive by additive. However, these definitions are only valid under very

181   specific and strict circumstances, limiting their use.

182   Because of such strong dependency of genetic variance components on genetic

183   models, the literal meaning implied by conclusions such as "the majority of genetic

184   variance is due to additive variance" is meaningless; the same could be said for V_D' and V_{AA}''.

185   The definition of V_A as "additive" variance is thus an inappropriate one because it reflects

186   the presumed genetic model and genetic parameterization rather than the true genetic

187   model. The relative importance of estimated genetic variance components is also

188   meaningless because it changes with the parameterizations. Therefore, genetic variance

9

189   components do not inform us about the underlying genetic architecture. This is true for V_A,

190   V'_D, and V''_{AA} and independent of the parameterization used.

191   The ability of arbitrarily defined parameterizations to capture the majority of

192   genetic variance shares analogy with the ability of the type I sum of squares to explain

193   variance that is not always attributable to the experimental factor when the experimental

194   design is not orthogonal. In genetic studies, an orthogonal design is not always achievable

195   and impossible in natural populations. While orthogonal parameterizations and

196   partitioning of variance are possible subsequently (Zeng *et al.* 2005) (e.g. the portioning of

197   V_A and V_D are based on orthogonal parameterizations, so are V'_D and V'_A), they are

198   fundamentally different from a hypothetical designed orthogonal experiment, in which

199   different gene actions can be applied independently as treatments. Importantly, the

200   additive and dominant gene actions as commonly defined (Figure 1a,b), are two

201   intrinsically inseparable terms and not independent, with a strictly additive model

202   sometimes referred to as co-dominance. This is the root of the confusing convolution of

203   different variance components, especially when not clearly defined.

204   In the modern era with genome-wide genetic polymorphism data, the segregation of

205   quantitative trait loci (QTLs) can be tracked by DNA markers in linkage or linkage

206   disequilibrium with the true QTLs; and, in rare cases, the true QTLs themselves. Therefore

207   it is possible to compare the observed genetic models with the presumed models and to

208   define genetic architecture once the participating players have been identified. Genome

209   wide association studies (GWAS) have become the most common implementation of

210   marker based approaches. However, QTLs identified by GWAS fail to explain the majority of

211   genetic variation additively, a phenomenon referred to as "missing heritability" (Manolio *et*

10

212   *al.* 2009). Simultaneously using all markers additively (Yang *et al.* 2010), as well as using

213   our newly defined parameterizations assuming alternative genetic models (Figure 3,4), can

214   explain a large fraction of genetic variation. Although this holistic method cannot identify

215   QTLs and offers no information on the underlying genetic models, it suggests that the

216   heritability is just missing but does not vanish. Strategies to find the heritability include

217   increasing sample size to find QTLs of smaller size, using sequencing to find QTLs of rarer

218   frequencies, and testing for non-additive effects. Unfortunately, the effort to model non-

219   additive effects is limited because of its requirement for much larger sample size and the

220   predominant assumption that non-additive effects are unimportant. Our results suggest

221   that this assumption is unfounded. Undoubtedly, it is experimentally and statistically

222   challenging to find all QTLs, even more so to search for cryptic genetic models other than

223   additivity and dominance within a single locus, including two way or higher order

224   interactions, and other cryptic genetic models. Like the single locus models, these more

225   complex genetic models may not ultimately explain the missing heritability. Nonetheless, it

226   is clear that the current standard single-locus approach is not adequate precisely because

227   of the missing heritability phenomenon itself.

228

Acknowledgement

232

Methods

234

11

Least squares regression interpretation of V_A

235

236   Consider a single biallelic locus in a diploid genome with alleles A and a, each with

237   frequency p and q ($p + q = 1$); and assign genotypic values $y = -a, d$, and $+a$ to genotypes

238   aa, Aa, and AA respectively. The average effects of A and a are then $q\alpha$ and $-p\alpha$

239   respectively, where $\alpha = a + d(q - p)$ is the allele substitution effect and measures the

240   change in phenotype in an individual if an allele a is substituted with A(Falconer and

241   Mackay 1996). The breeding value, defined as the expected genotypic value of the progeny

242   an individual produces, is the sum of average allelic effects each diploid individual carries,

243   and is $-2p\alpha, q\alpha - p\alpha$, and $2q\alpha$ for aa, Aa, and AA respectively. With only one locus, the

244   total genetic variation in a randomly mating (thus in Hardy-Weinberg equilibrium)

245   population can be partitioned into two orthogonal components, the additive genetic

246   variance V_A, which is defined as the variance due to breeding values, $2pq\alpha^2$, and the

247   dominance genetic variance $V_D = (2pqd)^2$ (Table 1)(Falconer and Mackay 1996).

248   Alternatively, we can define a random variable x_A as:

$$x_A = \begin{cases} 0, & genotype = aa, \\ 1, & genotype = Aa, \\ 2, & genotype = AA. \end{cases}$$

249   This parameterization has the convenient interpretation that x_A is equal to the number of A

250   alleles. It is easy to show that the allele substitution effect α as defined above is the slope of

251   the least squares regression of genotypic value y on x_A in an idealized population with

252   random   mating   (Figure   S1).   The   additive   genetic   variance   is   then

253    $V_A = Var(\hat{y}) = Var(\alpha x_A) = \alpha^2 Var(x_A) = 2pq\alpha^2$ and the dominance genetic variance V_D is

254   the residual variance. It is easy to see that the least squares solution for this regression

12

255   seeks to maximize V_A and minimize V_D. This least squares interpretation is not new and

256   dates back to the early days of quantitative genetics(Fisher 1918).

257   By extension of this least squares regression interpretation of genetic variation, if

258   we arbitrarily define any one random variable x or more than one of them and fit a linear

259   model of form $y = \beta x + \epsilon$, we can partition genetic variance due to the assumed genetic

260   model $Var(\hat{y}) = Var(\beta x)$ and residual variance $Var(\epsilon)$.

261

Derivation of dominance variance V_D' using least squares regression

263   Now we illustrate the idea of using least squares regression to partition genetic variance

264   due to dominant gene action and the remaining genetic variance. We define the random

265   variable x_D' as:

$$x_D' = \begin{cases} 0, & genotype = aa; \\ 2, & genotype = Aa \text{ or } AA. \end{cases}$$

266   The least square solution for the linear model $y = \delta' x_D' + \epsilon$ can be easily found to be

267    $\delta' = \frac{q}{1+q} d + \frac{1}{1+q} a$ (Figure S2). Therefore the variance due to x_D' is $V_D' = \frac{4pq^2}{1+q}(a + dq)^2$. The

268   residuals from this regression are $0, \frac{1-q}{1+q}(d-a)$, and $\frac{-2q}{1+q}(d-a)$, for genotypes aa, AA, and

269   AA respectively. Similar to V_A and V_D, we define the residual variance as an "additive

270   deviation" variance V_A', which can be found to be $V_A' = \frac{2p^2q}{1+q}(a-d)^2$.

271

Finding V_{AA}'' numerically

273   Extending the least squares regression interpretation of genetic variance to any arbitrary

274   random variable x and finding the solution is not always easy. However, it is

275   computationally trivial to find. For example, to numerically estimate the additive by

13

276   additive variance V_{AA}'', we define x_{AA}'' as follows for two independently segregating loci with

277   alleles A/a, and B/b respectively:

278
$$x_1 = \begin{cases} -1, & genotype = aa; \\ 0, & genotype = Aa; \\ 1, & genotype = AA. \end{cases}, x_2 = \begin{cases} -1, & genotype = bb; \\ 0, & genotype = Bb; \\ 1, & genotype = BB. \end{cases}$$

279   Then, $x_{AA}'' = x_1 x_2$. We randomly draw 100,000 individuals with the specific genotypes

280   according to pre-defined allele frequencies and assign genotypic values with pre-defined

281   genetic models. The slopes $\hat{\beta}$ can be easily found by numerically regressing y onto x. The

282   proportion of genetic variation explained by this parameterization is then just the R^2 of the

283   regression.

284

285   **Mixed model analysis of simulated and real data**

286   To extend the single- and two-locus models to polygenic models, we used mixed model

287   analysis to partition phenotypic variation in simulated and real data. To simulate

288   phenotypic data with pre-defined genetic models, we first drew from the U-shaped

289   distribution(Hill *et al.* 2008) $f(p) \propto \frac{1}{pq}$ 1,000 realizations, which took possible values of

290   0.01, 0.02, ..., 0.99. Genotypes for these $p = 1,000$ loci were randomly assigned according to

291   their Hardy-Weinberg frequencies to $n = 5,000$ individuals. Genetic values were then

292   assigned to the 5,000 individuals using this general formula $\mathbf{g} = \mathbf{X}\boldsymbol{\beta}$. Each of the columns of

293   the $n \times p$ matrix $\mathbf{X}$ was coded by the additive parameterization x_A as defined above for the

294   additive genetic model, x_D' for the dominance genetic model. Similarly for the additive by

295   additive genetic model, $\frac{p}{2} = 500$ pairs of loci were parameterized as defined above using

296   x_{AA}''. The vector $\boldsymbol{\beta}$ was drawn from standard normal distribution. The phenotypic value y

297   for each individual was then simulated by adding random noise such that $\mathbf{Y} = \mathbf{g} + \boldsymbol{\epsilon}$, where

14

298   ϵ was normally distributed with zero mean and variance equal to $Var(\boldsymbol{g}) \frac{1-H^2}{H^2}$. H^2 was the

299   broad sense heritability and was always set to 0.5.

300   We standardized columns of $\mathbf{X}$ and computed the covariance matrix as $\mathbf{X}\mathbf{X}^{\mathbf{T}}$, which

301   was further scaled by the mean of its diagonal values. A linear mixed model $\mathbf{Y} = \mu\mathbf{1} + \mathbf{Z}\mathbf{u} +$

302   ϵ was fitted to the data, where μ was the population mean, $\mathbf{Z}$ was the incidence matrix and

303   in all cases in this study the identity matrix, $\boldsymbol{u}$ was a random effect with variance

304   covariance matrix $\mathbf{G}\sigma^2$, where $\mathbf{G}$ was simply the scaled $\mathbf{X}\mathbf{X}^{\mathbf{T}}$ above and σ^2 was the part of

305   genetic variance due to the specific parameterization. We fitted this model using the GCTA

306   software(Yang *et al.* 2011) with REML and performed simulations 100 times. We defined

307   the heritability explained by σ^2 as h^2/H^2, where $h^2 = \frac{\sigma^2}{\sigma^2 + \sigma_\epsilon^2}$, and H^2 was the simulated

308   broad sense heritability.

309   To analyze real data where the genetic architecture cannot be known *a priori*, we

310   downloaded genotype and phenotype data from dbGaP for the GENEVA Genes and

311   Environment Initiatives in Type 2 Diabetes study (phs000091.v2.p1). We pruned the data

312   set to contain 5,497 unrelated (nominal genetic relationship as calculated by GCTA < 0.05)

313   individuals with European ancestry based on both self-reported ethnicity and principal

314   component analysis. We then computed genetic covariance matrices as defined above using

315   autosomal SNPs and partitioned phenotypic variance using GCTA where sex was fitted as a

316   fixed effect in the model. We used the parameterization (Table 1) as defined in a recent

317   study(Zhu *et al.* 2015) to partition phenotypic variance into V_A, V_D, and V_e. We also

318   partitioned phenotypic variance into V_A', V_D', and V_e, where V_D' was defined as above and V_A'

319   was estimated by defining a new variable x_A' (Table 1), where

15

$$x'_A = \begin{cases} 0, & genotype = aa, \\ \dfrac{1-q}{1+q}, & genotype = Aa, \\ \dfrac{-2q}{1+q}, & genotype = AA. \end{cases}$$

320

References

322 Alvarez-Castro J. M., Carlborg O., 2007 A unified model for functional and statistical

323   epistasis and its application in quantitative trait loci analysis. Genetics **176**: 1151–

324   1167.

325 Bloom J. S., Ehrenreich I. M., Loo W. T., Lite T.-L. V., Kruglyak L., 2013 Finding the sources of

326   missing heritability in a yeast cross. Nature **494**: 234–7.

327 Cheverud J. M., Routman E. J., 1995 Epistasis and its contribution to genetic variance

328   components. Genetics **139**: 1455–1461.

329 Falconer D. S., Mackay T. F. C., 1996 *Introduction to quantitative genetics*. Pearson

330   Education Limited, Essex, England.

331 Fisher R. A., 1918 The correlation between relatives on the supposition of Mendelian

332   inheritance. Trans R Soc Edinburgh **52**: 399–433.

333 Hill W. G., Goddard M. E., Visscher P. M., 2008 Data and theory point to mainly additive

334   genetic variance for complex traits. PLoS Genet **4**: e1000008.

335 Lush J. L., 1943 *Animal breeding plans*. Iowa State College Press, Amex, IA.

336 Mackay T. F. C., 2014 Epistasis and quantitative traits: using model organisms to study

337   gene-gene interactions. Nat Rev Genet **15**: 22–33.

338 Mäki-Tanila A., Hill W. G., 2014 Influence of gene interaction on complex trait variation

339   with multi-locus models. Genetics **198**: 1–27.

16

340   Manolio T. A., Collins F. S., Cox N. J., Goldstein D. B., Hindorff L. A., Hunter D. J., McCarthy M.

341   I., Ramos E. M., Cardon L. R., Chakravarti A., Cho J. H., Guttmacher A. E., Kong A.,

342   Kruglyak L., Mardis E., Rotimi C. N., Slatkin M., Valle D., Whittemore A. S., Boehnke M.,

343   Clark A. G., Eichler E. E., Gibson G., Haines J. L., Mackay T. F. C., McCarroll S. A., Visscher

344   P. M., 2009 Finding the missing heritability of complex diseases. Nature **461**: 747–753.

345   Yang J., Benyamin B., McEvoy B. P., Gordon S., Henders A. K., others, 2010 Common SNPs

346   explain a large proportion of the heritability for human height. Nat Gen **42**: 565–569.

347   Yang J., Lee S. H., Goddard M. E., Visscher P. M., 2011 GCTA: A tool for genome-wide

348   complex trait analysis. Am J Hum Genet **88**: 76–82.

349   Zeng Z. B., Wang T., Zou W., 2005 Modeling quantitative trait loci and interpretation of

350   models. Genetics **169**: 1711–1725.

351   Zhu Z., Bakshi A., Vinkhuyzen A. A. E., Hemani G., Lee S. H., Nolte I. M., Vliet-Ostaptchouk J. V

352   van, Snieder H., Esko T., Milani L., Mägi R., Metspalu A., Hill W. G., Weir B. S., Goddard

353   M. E., Visscher P. M., Yang J., 2015 Dominance Genetic Variation Contributes Little to

354   the Missing Heritability for Human Complex Traits. Am J Hum Genet **96**: 377–385.

355

17

356 **Table 1 | Notations and definitions of variance components in this study.**

Notation	Variance component	Genotype coding
V_A	$2pq[a + d(p - q)]^2$	$x_A \in \{0, 1, 2\}$
V_D	$(2pqd)^2$	$x_D \in \{0, 2p, 2(p - q)\}$
V_D'	$\dfrac{4pq^2}{1 + q}(a + dq)^2$	$x_D' \in \{0, 2, 2\}$
V_A'	$\dfrac{2p^2q}{1 + q}(a - d)^2$	$x_A' \in \{0, \dfrac{1 - q}{1 + q}, \dfrac{-2q}{1 + q}\}$
V_{AA}''	computed numerically	$x_{AA}'' \in (x_{A,1} - 1)(x_{A,2} - 1)$

357

358

359   **Figure Legends**

360

361   **Figure 1 | Additive genetic variance V_A is a major determinant of total genetic**

362   **variance.** Under additive (**a**), dominant (**b**), or additive by additive (**c, d**) models, the

363   proportion of total genetic variance explained by the additive genetic variance V_A and

364   dominance genetic variance V_D are estimated either analytically (**a, b**) or numerically by

365   simulation (**d**).

366

367   **Figure 2 | Alternative parameterizations capture the majority of genetic variance.**

368   Using an alternative parameterization that emphasizes dominant gene action, a newly

369   defined dominance variance V_D' and additive deviation variance V_A' are estimated

370   analytically under dominant (**a**) and additive (**b**) models. Using an alternative

371   parameterization that emphasizes additive by additive gene action, a newly defined

372   interaction variance V_{AA}'' is estimated numerically under additive by additive (**c**) and

373   additive (**d**) models.

374

375   **Figure 3 | Conventional and alternative parameterizations capture the majority of**

376   **polygenic genetic variance.** Simulation is used to generate data sets with the additive (A),

377   dominant (D), and additive by additive (AxA) genetic models and V_A, V_D', and V_{AA}'' are

378   estimated using linear mixed models. The results are presented as the proportion of

379   heritability explained by the genetic variance component; h_a^2 corresponds to V_A, $h_{d'}^2$ to V_D',

380   and $h_{aa''}^2$ to V_{AA}''.

381

19

382   **Figure 4 | Variance component analyses of human height data.** Phenotypic variation of

383   height (in cm) observed in the GENEVA study is partitioned into genetic variance

384   components as indicated (color-coded bars) and environmental variance (V_e, grey bar). The

385   colors of bars correspond to the colors of the text indicating the variance components.

386   Error bars indicate standard errors of the variance component estimates provided by

387   GCTA. Proportions of the components are also indicated.

388

389   **Figure S1 | Least squares regression interpretation of V_A.** This representation is

390   adapted from Fig. 7.2 of Reference 2. Grey circles indicate the genotypic value of each

391   genotype, which is coded as 0, 1, 2 for aa, Aa, and AA respectively. A regression line (red

392   line) is fitted to the data, on which the fitted values are indicated by white circles. The fitted

393   line must pass through the center of the data, as indicated by the cross. The fitted values

394   are equivalent to breeding values. The arrows between the breeding values and the

395   genotypic values are the dominance deviations, which are the same as residuals of the

396   regression. Note that the data points are weighted by their frequencies in the population. A

397   dominance model is used so that the dominance deviation can be illustrated.

398

399   **Figure S2 | Least squares regression interpretation of V_D'.** Grey circles indicate the

400   genotypic value of each genotype, which is coded as 0, 2, 2 for aa, Aa, and AA respectively. A

401   regression line (red line) is fitted to the data, on which the fitted values are indicated by

402   white circles. The fitted line must pass through the center of the data, as indicated by the

403   cross. The fitted line must also pass through the circle (half grey and half white to indicate

404   the overlap of the genotypic and fitted values) denoting genotype aa. The fitted values are

20

405   equivalent to dominance values as defined in this parameterization. The arrows between

406   the dominance values and the genotypic values are the residuals of the regression, which

407   we define as "additive deviation", therefore the residual variance is V_A'. Note that the data

408   points are weighted by their frequencies in the population. An additive model is used so

409   that the additive deviation can be illustrated.

410

a Additive

b Dominance

c Additive by additive

d

a

$-a$ 0 d $+a$
aa   Aa AA

Dominance

Proportion of genetic variance

V'_D

Allele frequency p(A)
0.01 0.2 0.4 0.6 0.8 0.99

b

$-a$ 0 d $+a$
aa   Aa   AA

Additive

Proportion of genetic variance

V'_D

V'_A

Allele frequency p(A)
0.01 0.2 0.4 0.6 0.8 0.99

c

Additive by additive

Proportion of V''_{AA}

p(B)
p(A)

d

Additive

Proportion of V''_{AA}

p(B)
p(A)

a

h_a^2 / H^2

b

h_d^2 / H^2

c

h_{aa}^2 / H^2

Figure showing variance components (cm^2):

- V_A: 46.17%
- V_D': 48.04%
- V_{AA}'': 56.96%
- $V_A + V_D$: 45.93%, 7.61%
- $V_A' + V_D'$: 41.84%, 26.60%

x-axis: Variance component (cm^2), 0 to 50

Genotypic value

$y - \bar{y} = \alpha(x - \bar{x})$

$+a$

0

$(\bar{x}, \bar{y})$

$-a$

genotypic value
dominance deviation
breeding value

| 0 | 1 | 2 |
| aa | Aa | AA |

$-a$ 0 d $+a$
aa   Aa   AA