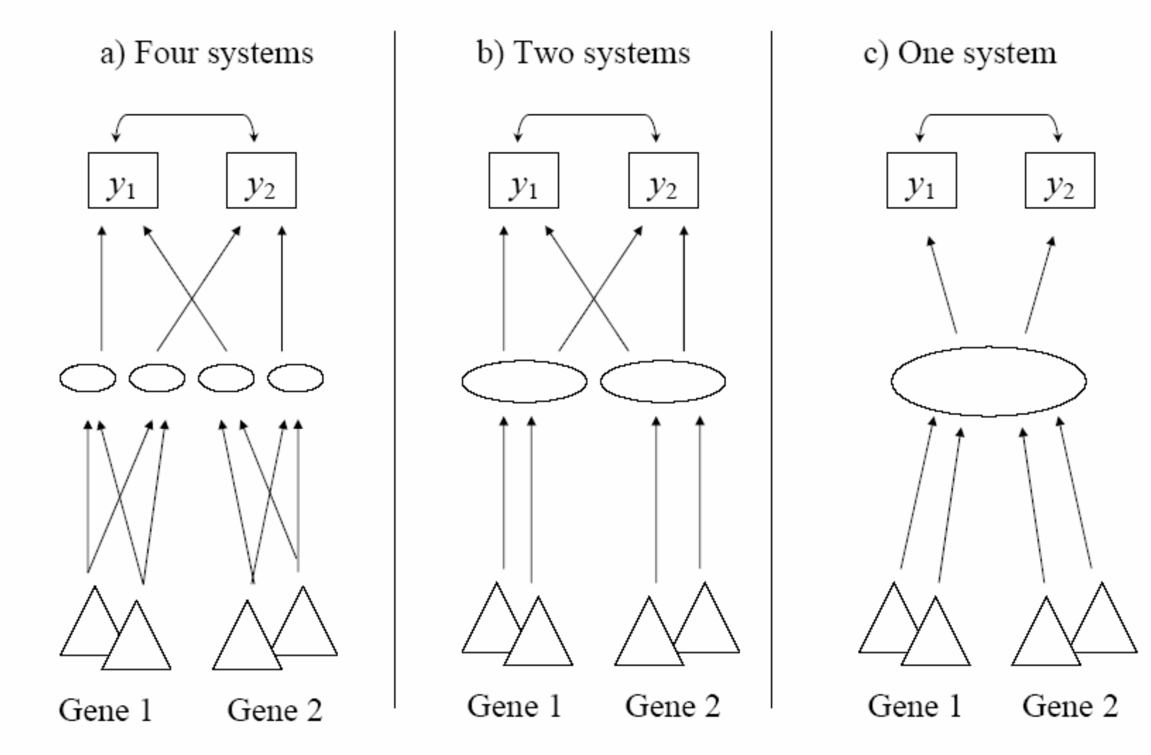
Supplementary material

This supplemental material illustrates with a simple example, how different models can be distinguished. In the figure below we assume two biallelic genes and two quantitative outcome measures y_1 and y_2. For each gene there are three genotypes (AA, Aa, and aa). The genetic contribution can therefore be decomposed into an additive and dominance effect(Falconer, 1989) that are modeled via two dummy variables (triangles). These genes affect the outcomes (squares) through unobserved biological systems (ellipses). In statistical terms these biological systems are latent variables (Bollen, 2002).

Three models could be fitted to the data to study the underlying biological systems involved. These models have a different implication for the impact of the genes on outcomes y_1 and y_2. Because the different models have different implications for means and covariances, by fitting them to the data it is possible to derive the best fitting model. For sake of illustration, we confine ourselves here to the impact on the mean. Figure a) assumes completely independent pathways for each of the genetic effects on each of the outcomes implying four biological systems. This model does not impose any constraints on the genetic effects and is technically equivalent to a traditional analysis where each combination of gene and phenotype is tested separately. Figure b assumes that each gene affects only one biological system, which in turn influences both measured variables. Each gene produces a certain amount of dominance in the biological system that it affects. Because this system in turn influences y_1 and y_2, this dominance effect should also be reflected in both outcomes that consequently would show the same amount of dominance. Figure b therefore imposes two statistical constraints on the effects of the two genes compared to Figure a. We can then expect dominance on the basis of well-established enzyme kinetic principles(Kacser & Burns, 1981) so that in practice it may often be possible to distinguish the two models. Figure c assumes one biological system. Similarly to Figure b, this implies equal amounts of dominance. In addition, if gene 1 explains more variance in the biological system, this difference will be mirrored in the two outcomes. The model therefore implies that the relative importance of gene 1 versus gene 2 is identical for y_1 and y_2 resulting in one more statistical constraint.



To show more formally that the three models have different implication for the means of measured variables y_1 and y_2 we first write the general from of the models:

$$\eta = \nu + \mathbf{Aa} + \mathbf{Dd}$$

$$\mathbf{y} = \Lambda\eta + \varepsilon$$

The components of the $n_\eta \times 1$ dimensional vector $\boldsymbol{\eta}$ are the n_η latent variables representing the unmeasured biological systems. The 2×1 vectors $\mathbf{a} = [A1,A2]^t$ and $\mathbf{d} = [D1,D2]^t$ contain the dummy variables needed to model the genetic effects. We assume that AA, Aa, and aa subjects are coded 1, 0, and -1 on the A dummy variables and 0, 1, 0 on the D dummy variables. The elements in the $n_\eta \times 2$ matrix $\mathbf{A}$ are then the additive genetic effects, and the elements in the $n_\eta \times 2$ matrix $\mathbf{D}$ the dominance effects. In our example the "latent" variables are simple linear combinations of the genes. In general this does not have to be the case and in certain situations it may be possible to estimate an additional error term that incorporates the effects of all unmeasured influences on the biological systems. The second equation indicates how the measured variables in the 2×1 vector $\mathbf{y}$ are influenced by the latent variables. The intercepts of the measured variables are in vector ν. The $2 \times n_\eta$ matrix Λ matrix contains the effects of the latent variables on the measured variables. Finally, the 2×1 dimensional vector ε contains the effect of all variables that are not measured but do affect $\mathbf{y}$. Because the biological systems may not account for all the covariance between the measured variables, these error components in $\mathbf{y}$ are allowed to correlate.

To obtain the equations for the model in Figure a we fix the effects of the four latent variables on the 2 measured variables in matrix Λ to 1. After some matrix algebra and substitution of η in the equation for $\mathbf{y}$ we get:

$$y_1 = \nu_1 + a_{11}\mathrm{A1} + d_{11}\mathrm{D1} + a_{32}\mathrm{A2} + d_{32}\mathrm{D2} + \varepsilon_1$$

$$y_2 = \nu_2 + a_{21}\mathrm{A1} + d_{21}\mathrm{D1} + a_{42}\mathrm{A2} + d_{42}\mathrm{D2} + \varepsilon_2$$

The subscripts refer to the vector and matrix elements. In the case of matrix elements the first (=row) subscript indicates the affected variable, and the second subscript (=column) indicates the causal variable that has the effect. For example, element a_{21} in matrix **A** is the effect of the first additive dummy variable A1 on the second latent variable η_2. These equations show that the model in Figure 2a simply reproduces the observed genotype means. It is essentially an unconstrained model for the effects of the two genes on the two measured variable. The model has 8 parameters because an additive (*a*) and dominance effect (*d*) is needed to model the effect of the 2 genes on each of the 2 measured variables.

Fixing the effect of the latent variables on measured variables (matrix **Λ**) to one does not have any effect on how well the model reproduces the observed genotype effects. As a matter of fact, these effects will need to be fixed to a certain value to obtain unique estimates of the parameters. The substantive interpretation is because the latent variables are not measured we do not know their mean and variance. The scale of the latent variables is arbitrary and will not affect the fit of the model. By fixing the parameters in Λ to one we are essentially giving the latent variables the same scales as the observed variables. In Figure 2b, the same latent variable affects two measured variables. If we fix for each of the latent variables one of the effects on the measured variables to one, the latent variable is scaled and it becomes possible to estimate the other effect. If we fix $\lambda_{11} = \lambda_{22} = 1$ in Figure b, we get the following equations:

$$y_1 = \nu_1 + a_{11}\mathrm{A1} + d_{11}\mathrm{D1} + \lambda_{12}(a_{22}\mathrm{A2} + d_{22}\mathrm{D2}) + \varepsilon_1$$

$$y_2 = \nu_2 + \lambda_{21}(a_{11}\mathrm{A1} + d_{11}\mathrm{D1}) + a_{22}\mathrm{A2} + d_{22}\mathrm{D2} + \varepsilon_2$$

Parameters λ_{12} and λ_{21} can be interpreted as the relative effect on a measured variable of one system versus the other system. For example, if $\lambda_{12} > 1$ the second latent variable has a larger impact on y_1 than the first latent variable. Because $\lambda_{21}d_{11}/\lambda_{21}a_{11} = d_{11}/a_{11}$ and $\lambda_{12}d_{22}/\lambda_{12}a_{22} = d_{22}/a_{22}$, this model implies that gene 1 and gene 2 show equal amounts of dominance for both measured variables. For Figure 1a this does not have to be the case and we can have: $d_{11}/a_{11} \neq d_{21}/a_{21}$ and $d_{32}/a_{32} \neq d_{42}/a_{42}$. This model estimates 6 instead of 8 parameters, and a statistical test could be performed to examine whether these two restrictions are correct.

Finally, after scaling the latent variable in Figure 2c by $\lambda_{11} = 1$, we obtain the equations:

$$y_1 = \nu_1 + a_{11}\mathrm{A}1 + d_{11}\mathrm{D}1 + a_{12}\mathrm{A}2 + d_{12}\mathrm{D}2 + \varepsilon_1$$

$$y_2 = \nu_2 + \lambda_{21}(a_{11}\mathrm{A}1 + d_{11}\mathrm{D}1 + a_{12}\mathrm{A}2 + d_{12}\mathrm{D}2) + \varepsilon_2$$

Similar to Figure b, it constrains the amount of dominance for gene 1 (d_{11}/a_{11}) and gene 2 (d_{12}/a_{12}) to be identical for both measured variables. In addition, it constrains the relative size of the effects (or explained variance) of gene 1 and gene 2 to be identical for y_1 and y_2:

$$(a_{11} + d_{11})/(a_{12} + d_{12}) = \lambda_{21}(a_{11} + d_{11}) /\lambda_{21}(a_{12} + d_{12})$$

For Figure b this does not have to be the case:

$$(a_{11} + d_{11})/ \lambda_{12}(a_{22} + d_{22}) \neq \lambda_{21}(a_{11} + d_{11})/(a_{22} + d_{22})$$

This model estimates 5 instead of 6 or 8 parameters. Tests could be performed to examine whether these constraints are consistent with the observed data.