Theory and Practice in Quantitative Genetics

Daniëlle Posthuma[1], A. Leo Beem[1], Eco J. C. de Geus[1], G. Caroline M. van Baal[1], Jacob B. von Hjelmborg[2], Ivan Iachine[3], and Dorret I. Boomsma[1]

[1] *Department of Biological Psychology, Vrije Universiteit Amsterdam, The Netherlands*
[2] *Institute of Public Health, Epidemiology, University of Southern Denmark, Denmark*
[3] *Department of Statistics, University of Southern Denmark, Denmark*

With the rapid advances in molecular biology, the near completion of the human genome, the development of appropriate statistical genetic methods and the availability of the necessary computing power, the identification of quantitative trait loci has now become a realistic prospect for quantitative geneticists. We briefly describe the theoretical biometrical foundations underlying quantitative genetics. These theoretical underpinnings are translated into mathematical equations that allow the assessment of the contribution of observed (using DNA samples) and unobserved (using known genetic relationships) genetic variation to population variance in quantitative traits. Several statistical models for quantitative genetic analyses are described, such as models for the classical twin design, multivariate and longitudinal genetic analyses, extended twin analyses, and linkage and association analyses. For each, we show how the theoretical biometrical model can be translated into algebraic equations that may be used to generate scripts for statistical genetic software packages, such as Mx, Lisrel, SOLAR, or MERLIN. For using the former program a web-library (available from http://www.psy.vu.nl/mxbib) has been developed of freely available scripts that can be used to conduct all genetic analyses described in this paper.

"Genetic factors explain *x*% of the population variance in trait Y" is an oft heard outcome of quantitative genetic studies. Usually this statement derives from (twin) family research that exploits known genetic relationships to estimate the contribution of unknown genes to the observed variance in the trait. It does not imply that any *specific* genes that influence the trait have been identified. Given the rapid advances made in molecular biology (*Nature* Genome Issue, February 15, 2001; *Science* Genome Issue, February 16, 2001), the near completion of the human genome and the development of sophisticated statistical genetic methods (e.g., Dolan et al., 1999a, 1999b; Fulker et al., 1999; Goring, 2000; Terwilliger & Zhao, 2000), the identification of specific genes, even for complex traits, has now become a realistic prospect for quantitative geneticists. To identify genes, family studies, specifically twin family studies, again appear to have great value, for they allow simultaneous modelling of observed and unobserved genetic variation. As a "proof of principle", genomEUtwin will perform genome-wide genotyping in twins to target genes for the complex traits of stature, body mass index (BMI), coronary artery disease and migraine. To increase power, epidemiological and phenotypic data from eight participating twin registries will be simultaneously analysed.

In this paper the main theoretical foundations underlying quantitative genetic analyses that are used within the genomEUtwin project will be described. In addition, an algebraic translation from theoretical foundation to advanced structural equation models will be made that can be used in generating scripts for statistical genetic software packages.

Observed, Genetic, and Environmental Variation

The starting point for gene finding is the observation of population variation in a certain trait. This "observed", or phenotypic, variation may be attributed to genetic and environmental causes. Genetic and environmental effects interact when the same variant of a gene differentially affects the phenotype in different environments.

About 1% of the total genome sequence is estimated to code for protein and an additional but still unknown percentage of the genome is involved in regulation of gene expression. Human individuals differ from one another by about one base pair per thousand. If these differences occur within coding or regulatory regions, phenotypic variation in a trait may result. The different effects of variants ("alleles") of the same gene is the basis of the model that underlies quantitative genetic analysis.

Quantifying Genetic and Environmental Influences: The Quick and Dirty Approach

In human quantitative genetic studies, genetic and environmental sources of variance are separated using a design that includes subjects of different degrees of genetic and environmental relationship (Fisher, 1918; Mather & Jinks, 1982). A widely used design compares phenotypic resemblance of monozygotic (MZ) and dizygotic (DZ) twins. Since MZ twins reared together share part of their environment and 100% of their genes (but see Martin et al., 1997), any resemblance between them is attributed to these

Address for correspondence: Daniëlle Posthuma, Vrije Universiteit, Department of Biological Psychology, van der Boechorststraat 1, 1081 BT, Amsterdam, The Netherlands. Email: danielle@psy.vu.nl

two sources of resemblance. The extent to which MZ twins do not resemble each other is ascribed to unique, non-shared environmental factors, which also include measurement error. Resemblance between DZ twins reared together is also ascribed to the sharing of the environment, and to the sharing of genes. DZ twins share on average 50% of their segregating genes, so any resemblance between them due to genetic influences will be lower than for MZ pairs. The extent to which DZ twins do not resemble each other is due to non-shared environmental factors and to non-shared genetic influences.

Genetic effects at a single locus can be partitioned into additive (i.e., the effect of one allele is added to the effect of another allele) or dominant (the deviation from purely additive effects) effects, or a combination. The total amount of genetic influence on a trait is the sum of the additive and dominance effects of alleles at multiple loci, plus variance due to the interaction of alleles at different loci (*epistasis*; Bateson, 1909). The expectation for the phenotypic resemblance between DZ twins due to genetic influences depends on the underlying (and usually unknown) mode of gene action. If all contributing alleles act additively and there is no interaction between them within or between loci, the correlation of genetic effects in DZ twins will be on average 0.50. However, if some alleles act in a dominant way the correlation of genetic dominance effects will be 0.25. The presence of dominant gene action thus reduces the expected phenotypic resemblance in DZ twins relative to MZ twins. Epistasis reduces this similarity even further, the extent depending on the number of loci involved and their relative effect on the phenotype (Mather & Jinks, 1982). Depending on the nature of the types of familial relationships within a dataset, additive genetic, dominant genetic, and shared and non-shared environmental influences on a trait can be estimated. For example, employing a design including MZ and DZ twins reared together allows decomposition of the phenotypic variance into components of additive genetic variance, non-shared environmental variance, and either dominant genetic variance or shared environmental variance. Additive and dominant genetic and shared environmental influences are confounded in the classical twin design and cannot be estimated simultaneously. Disentangling the contributions of shared environment and genetic dominance effects requires additional data from, for example, twins reared apart, half-sibs, or non-biological relatives reared together.

Similarity between two (biologically or otherwise related) individuals is usually quantified by covariances or correlations. Twice the difference between the MZ and DZ correlations provides a quick estimate of the proportional contribution of additive genetic influences (a^2) to the phenotypic variation in a trait ($a^2 = 2[r_{MZ} - r_{DZ}]$). The proportional contribution of the dominant genetic influences (d^2) is obtained by subtracting four times the DZ correlation from twice the MZ correlation ($d^2 = 2r_{MZ} - 4r_{DZ}$). An estimate of the proportional contribution of the shared environmental influences (c^2) to the phenotypic variation is given by subtracting the MZ correlation from twice the DZ correlation ($c^2 = 2r_{DZ} - r_{MZ}$). The proportional contribution of the non-shared environmental influences (e^2) can be obtained by subtracting the MZ correlation from unit correlation ($e^2 = 1 - r_{MZ}$).

These intuitively simple rules are described in textbooks on quantitative genetics and can be understood without knowledge of the relative effects and location of the actual genes that influence a trait, or the genotypic effects on phenotypic means. These point estimates, however, depend on the accuracy of the MZ and DZ correlation estimates and the true causes of variation of a trait in the population. For small sample sizes (i.e., most of the time) they may be grossly misleading. Knowledge of the underlying biometrical model becomes crucial when one wants to move beyond these twin-based heritability estimates, for instance, to add information of multiple additional family members or simultaneously estimate from a number of different relationships the magnitude of genetic variance in the population.

The Classical Biometrical Model: From Single Locus Effects on a Trait Mean to the Decomposition of Observed Variation in a Complex Trait

Although within a population many different alleles may exist for a gene (e.g., Lackner et al., 1991), for simplicity we describe the biometrical model assuming one gene with two possible alleles, allele *A1* and allele *A2*. By convention, allele *A1* has a frequency *p,* while allele A2 has frequency *q*, and $p + q = 1$. With two alleles there are three possible genotypes: *A1A1*, *A1A2*, and *A2A2* with genotypic frequencies p^2, $2pq$, and q^2, respectively, under random mating. The genotypic effect on the phenotypic trait (i.e., the genotypic value) of genotype *A1A1*, is called "*a*" and the effect of genotype *A2A2* "*-a*". The midpoint of the phenotypes of the homozygotes *A1A1* and *A2A2* is by convention 0, so *a* is called the increaser effect, and *–a* the decreaser effect. The effect of genotype *A1A2* is called "*d*". If the mean genotypic value of the heterozygote equals the midpoint of the phenotypes of the two homozygotes ($d = 0$), there is no dominance. If allele *A1* is completely dominant over allele *A2*, effect *d* equals effect *a*. If $d \neq 0$ and the two alleles produce three discernable phenotypes of the trait, *d* is unequal to *a*. This is also known as the classical biometrical model (Falconer & Mackay, 1996; Mather & Jinks, 1982) (see Figure 1).

Statistical derivations for the contributions of single and multiple genetic loci to the population mean of a trait are given in several of the standard statistical genetic textbooks (e.g., Falconer & Mackay, 1996; Lynch & Walsh, 1998; Mather & Jinks, 1982), and some of these statistics are summarized in Table 1.

The genotypic contribution of a locus to the population mean of a trait is the sum of the products of the frequencies and the genotypic values of the different genotypes. Complex traits such as height or weight, are assumed to be influenced by the effects of multiple genes. Assuming only additive and independent effects of all of these loci, the expectation for the population mean (μ) is the sum of the contributions of the separate loci, and is formally expressed as $\mu = \sum a(p - q) + 2 \sum dpq$

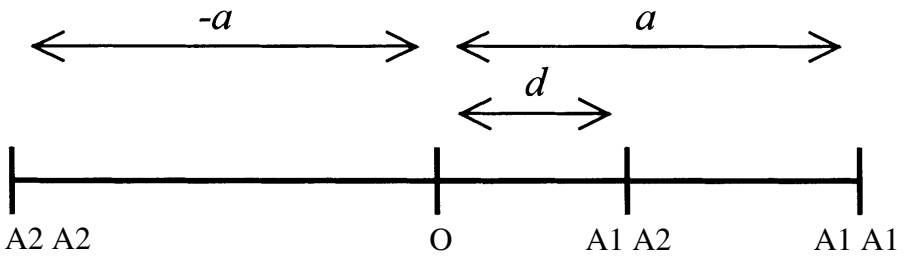

Figure 1

Graphical illustration of the genotypic values for a diallelic locus.

Table 1

Summary of Genotypic Values, Frequencies, and Dominance Deviation for Three Genotypes A1A1, A1A2, and A2A2

Genotype	A1A1	A1A2	A2A2
Genotypic value	*a*	*d*	*–a*
Frequency	p^2	*2pq*	q^2
Frequency x value	$a\,p^2$	*2dpq*	$-a\,q^2$
Deviation from the population mean	$2q(a - dp)$	$a(q - p) + d(1 - 2pq)$	$-2p(a + dq)$
Dominance deviation	$-2q^2d$	*2dpq*	$-2p^2d$

Decomposition of Phenotypic Variance

Although Figure 1 and Table 1 lack environmental effects, quantitative geneticists assume that the individual phenotype (P) is a function of both genetic (G) and environmental effects (E): P = G + E, where E refers to the environmental deviations, which have an expected average value of zero. This equation does not include the term GxE, and thereby assumes no interaction between the genetic effects and the environmental effects.

The *variance* of the phenotype, which itself is defined by G + E, is given by $V_P = V_G + V_E + 2cov_{GE}$ where V_P represents the variance of the phenotypic values, V_G represents the variance of the genotypic values, V_E represents the variance of the environmental deviations, and cov_{GE} represents the covariance between G and E. GE-covariance or GE-correlation can be modelled in a twin design that includes the parents of twins (e.g., Boomsma & Molenaar, 1987a ; Fulker, 1988) or in a design that includes actual measurements of the relevant genetic and environmental factors. For simplicity we assume that $V_P = V_G + V_E$. Statistically the total genetic variance (V_G) can be obtained by the standard formula for the variance: $\sigma^2 = \sum f_i(x_i - \mu)^2$, where f_i denotes the frequency of genotype i, x_i denotes the corresponding mean of that genotype (as given in Table 1) and μ denotes the population mean. Thus, $V_G = p^2[2q(a - dp)]^2 + 2pq[a(q - p) + d(1 - 2pq)]^2 + q^2[-2p(a + dq)]^2$. Which can be simplified to $V_G = 2pq[a + d(q - p)]^2 + (2pqd)^2 = V_A + V_D$ (see e.g., Falconer & Mackay, 1996).

If the phenotypic value of the heterozygous genotype lies midway between A1A1 and A2A2 (i.e., the effect of *d* in Figure 1 equals zero), the total genetic variance simplifies to $2pqa^2$. If *d* is not equal to zero, the "additive" genetic variance component contains the effect of *d*. Even if $a = 0$, V_A is usually greater than zero (except when $p = q$). Thus, although V_A represents the variance due to the additive influences, it is not only a function of *p*, *q*, and *a*, but also of *d*. The consequences are that, except in the rare situation where all contributing loci are diallelic with $p = q$, V_A is usually greater than zero. Models that decompose the phenotypic variance into components of V_D and V_E only, are therefore biologically implausible. When more than one locus is involved and it is assumed that the effects of these loci are uncorrelated and there is no interaction (i.e., no epistasis), the V_Gs of each individual locus may be summed to obtain the total genetic variances of all loci that influence a trait (Fisher, 1918; Mather, 1949). In most human quantitative genetic models the observed V_P of a trait is not modelled directly as a function of *p*, *q*, *a*, *d* and environmental deviations (as all of these are usually unknown), but instead is modelled by comparing the observed resemblance between pairs of differential, known genetic relatedness, such as MZ and DZ pairs. Ultimately *p*, *q*, *a*, *d* and environmental deviations are the parameters quantitative geneticists hope to "quantify".

Comparing the observed resemblance in MZ twins and DZ twins allows decomposition of observed variance in a trait into components of V_A, V_D (or V_C, for shared environmental variation which is not considered at this point), and V_E. As MZ twins share 100% of their genome, the expectation for their covariance is $COV_{MZ} = V_A + V_D$.

The expectation for DZ twins is less straightforward: as DZ twins share on average 50 per cent of their genome as stated earlier, they share half of the genetic variance that is transmitted from the parents (i.e., $^1/_2\ V_A$). As V_D is not

transmitted from parents to offspring it is less obvious where the coefficient of sharing for the dominance deviations (i.e., the 0.25 mentioned earlier) derives from. If two members of a DZ twin pair share both of their alleles at a single locus they will have the same coefficient for *d*. If they share no alleles or just one parental allele they will have no similarity for the effect of *d*. The probability that two members of a DZ pair have received two identical alleles from both parents is the coefficient of similarity for *d* between them. The probability is $^1/_2$ that two siblings (or DZ twins) receive the same allele from their father, and the probability is $^1/_2$ that they have received the same allele from their mother. Thus, the probability that they have received the same two ancestral alleles is $^1/_2 \times {}^1/_2 = {}^1/_4$, and the expectation for the covariance in DZ twins is $COV_{DZ} = {}^1/_2 V_A + {}^1/_4 V_D$.

Path Analysis and Structural Equation Modelling

The expectations for variances and covariances of MZ twins and DZ twins or sib pairs reared together may also be inferred from a path diagram (Wright, 1921, 1934), which is an often convenient non-algebraic representation of models such as are discussed here (see e.g., Neale & Cardon, 1992, for a brief introduction into path analysis) and which can be translated directly into structural equations. The parameters of these equations can be estimated by widely available statistical software (e.g., Mx and Lisrel). Structural Equation Modelling (SEM) has several advantages over merely comparing the MZ and DZ correlations (Eaves, 1969; Jinks & Fulker, 1970). If the model assumptions are valid, SEM produces parameter estimates with known statistical properties, while the correlational method merely allows parameter calculation. SEM thus also allows determination of confidence intervals and of standard errors of parameter estimates and quantifies how well the specified model describes the data. One can either directly derive structural equations from a theoretical model, or use path analysis as a non-algebraic intermediary to derive the structural equations.

Extended Twin Design

A convenient feature of SEM is the flexible handling of unbalanced data structures. This enables the relative easy incorporation of data from a variable number of family members. In Figure 2 a path diagram is drawn for a univariate trait measured in families consisting of a twin pair and one additional sibling. As additive, dominant and shared environmental effects are confounded in the twin design, the path diagram includes additive genetic influences (A), dominant genetic influences (D) and non-shared environmental influences (E), but not shared environmental influences (C). Note that a path diagram for shared environmental influences can be obtained by substituting 0.25 (for the DZ correlation for dominant genetic influences) for 1.00 (for the DZ correlation for shared environmental influences) and replacing the D with C for a latent shared environmental factor.

To rewrite the model depicted in Figure 2 into structural equations using matrix algebra we introduce three matrices **X**, **Y**, and **Z** of dimensions 1×1, containing the path coefficients *x*, *y*, and *z*, respectively. The matrix algebra notation for V_p is $\mathbf{XX}^T + \mathbf{YY}^T + \mathbf{ZZ}^T$, where ^{T} denotes the transpose of the matrix (and corresponds to tracing forwards through a path, see Neale & Cardon, 1992). The expectation for the MZ covariance is $\mathbf{XX}^T + \mathbf{YY}^T$ and the expectation for the DZ covariance is $0.5\mathbf{XX}^T + 0.25\mathbf{YY}^T$. Including additional siblings in this design is straightforward, and in this diagram the expectation for sib pair covariance is also $0.5\mathbf{XX}^T + 0.25\mathbf{YY}^T$ (but one should note that it is not necessary to assume the same model for sib-sib covariance as for DZ covariance).

The inclusion of non-twin siblings (if available) will not only enhance statistical power (Dolan et al., 1999b; Posthuma & Boomsma, 2000), but also provides the opportunity to test several assumptions, such as whether the covariance between DZ twins equals the covariance between non-twin siblings (which is often assumed, but can now be tested), whether the means and variances in twins are similar to the means and variances observed in siblings, or whether twin-sib covariance is different from sib-sib covariance, across males and females.

As in practice some families may for example consist of a twin pair and six additional siblings while other families consist of twins, the correlational method cannot be applied to estimate genetic and environmental contributions to the variance. Fortunately, these so-called non-rectangular (or unbalanced nested) data structures can be handled with ease using a SEM approach.

Threshold Models for Categorical Twin Data

So far we have considered quantitative traits. Several observed traits, however, are measured on a non-continuous scale, such as dichotomous traits (e.g., disease vs. no disease; smoking vs. non-smoking) or ordinal phenotypes (e.g., underweight/normal weight/overweight/obesity/severe obesity), yielding summary counts in contingency tables instead of means and variances/covariances. Reducing continuous scores like BMI to a categorical score like obese/non-obese should be avoided, as the statistical power to detect significant effects is much lower in categorical analyses (Neale et al., 1994).Contingency tables typically contain the number of (twin) pairs (for each zygosity group) for each combination (e.g., concordant non-smokers, concordant smokers, discordant on smoking). Because of the inherent polygenic background of complex traits, these data are often treated by assuming that an underlying quantitative liability exists with one or more thresholds, depicting the categorization of subjects. Although the liability itself cannot be measured, a standard-normal distribution is assumed for the liability. The thresholds (*z*-values in the standard normal distribution) are chosen in such a way that the area under the standard normal curve between two thresholds (or from minus infinity to the first threshold, and from the last threshold to infinity) reflects the prevalence of that category.

For one variable measured on single subjects, the prevalence of category *i* is given by:

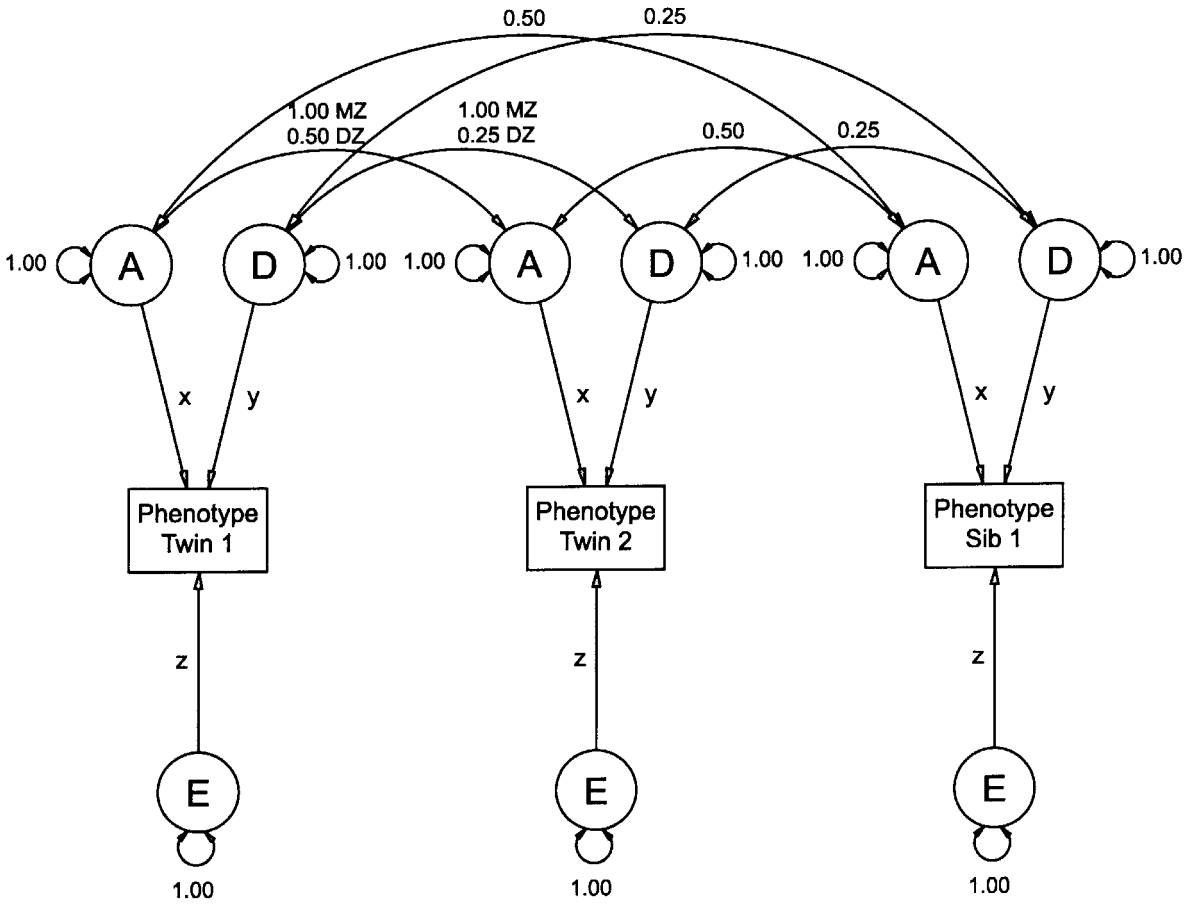

Figure 2

Path diagram representing the resemblance between MZ or DZ twins and an additional sibling, for additive genetic influences (A), dominant genetic influences (D), and non-shared environmental influences (E), for a univariate trait.

The latent factors A, D, and E have unit variance, and *x*, *y*, and *z* represent the respective path coefficients from A, D, and E to the phenotype P. The path coefficients can be regarded as standardized regression coefficients. The phenotypic variance (V_P) for the trait (the same for both members of a twin pair) equals $x^2 + y^2 + z^2$ which equals $V_A + V_D + V_E$. By applying the tracing rules of path analysis, the covariance between DZ twins (and sib pairs) is traced as $0.50x^2 + 0.25y^2$, which equals $^1/_2 V_A + ^1/_4 V_D$. The covariance between MZ twins is traced as $x^2 + y^2$ which equals $V_A + V_D$.

$$\int_{t_{i-1}}^{t_i} \phi(v)\,dv$$

were $t_{i-1} = -\infty$ for $i = 0$ and $t_i = \infty$ for $i = p$ for p categories, and $\phi(v)$ is the normal probability density function

$$\frac{e^{-0.5v^2}}{\sqrt{2\pi}},$$

where $\pi = 3.14$.

This can easily be extended to twin data, where one variable is available for both members of a twin pair (e.g., obesity yes/no in twin 1 and twin 2). In this case, the underlying bivariate normal probability density function will be characterized by two liabilities (that can be constrained to be the same) and by a correlation between them:

$$\int_{t_{i-1}}^{t_i} \int_{s_{i-1}}^{s_i} \phi(v)\,dv$$

with t_{i-1}, $s_{i-1} = -\infty$ for $i = 0$ and t_i , $s_i = \infty$ for $i = p$. In this case,

$$\phi(v) = |2\pi\Sigma|^{-n/2} \times e^{-\frac{1}{2} v_i^T \times \Sigma^{-1} \times v_i},$$

where Σ is the predicted correlation matrix.

Although the underlying bivariate distribution cannot be observed, its shape depends on the correlation between the two liability distributions. A high correlation results in a relatively low proportion of discordant twin pairs, whereas a low correlation results in a much higher proportion. Based on the contingency tables one may calculate tetrachoric (for dichotomous traits) or polychoric (for ordinal traits) twin correlations between the trait measured in twin 1 and the trait measured in twin 2 for each zygosity

group. Subsequently, genetic analyses can be conducted that use these correlations in the same way as is done with continuous data. The practical extension of this method to the analysis of more than two categorical variables may not be straightforward, especially if many variables are involved. The method then needs an asymptotic weight matrix (see Neale & Cardon, 1992) and the tetrachoric or polychoric correlations should be estimated simultaneously, not pairwise, as the latter method often produces improper correlation matrices. The computations may then become very time consuming. The preferred method for categorical multivariate designs is to conduct analyses directly on all available raw data, fitting each twin or sib pair individually using the multivariate normal probability density functions given above. Again, the underlying correlation matrix of family data can be decomposed into genetic and environmental influences in the same way as can be done for continuous data.

Multivariate Analysis of Twin Data

The univariate model can easily be extended to a multivariate model when more than one measurement per subject is available (Boomsma & Molenaar, 1986, 1987b; Eaves & Gale, 1974; Martin & Eaves, 1977) or for longitudinal data when the same subject is assessed repeatedly in time. Figure 3 is a path diagram for a bivariate design (two measurements per subject; four measures for a pair of twins or siblings). The corresponding matrix algebra expressions for the expected MZ or DZ variances and covariances are the same as for the univariate situation, except that the dimensions of matrices **X**, **Y**, and **Z** are no longer 1×1. An often convenient form for those matrices is lower triangular of dimensions $n \times n$ (where n is the number of variables assessed on a single subject; in Figure 3, $n = 2$). The subscripts of the path coefficients correspond to matrix elements (i.e., x_{ij} denotes the matrix element in the i-th row and j-th column of matrix **X**). The path coefficients subscripted by $_{21}$ reflect the variation that both measured phenotypes have in common. For example, if the path denoted by x_{21} is not equal to zero, this suggests that there are some genes that influence both phenotypes.

Thus, multivariate genetic designs allow the decomposition of an observed correlation between two variables into a genetic and an environmental part. This can be quantified by calculating the genetic and environmental correlations and the genetic and environmental contributions to the observed correlation.

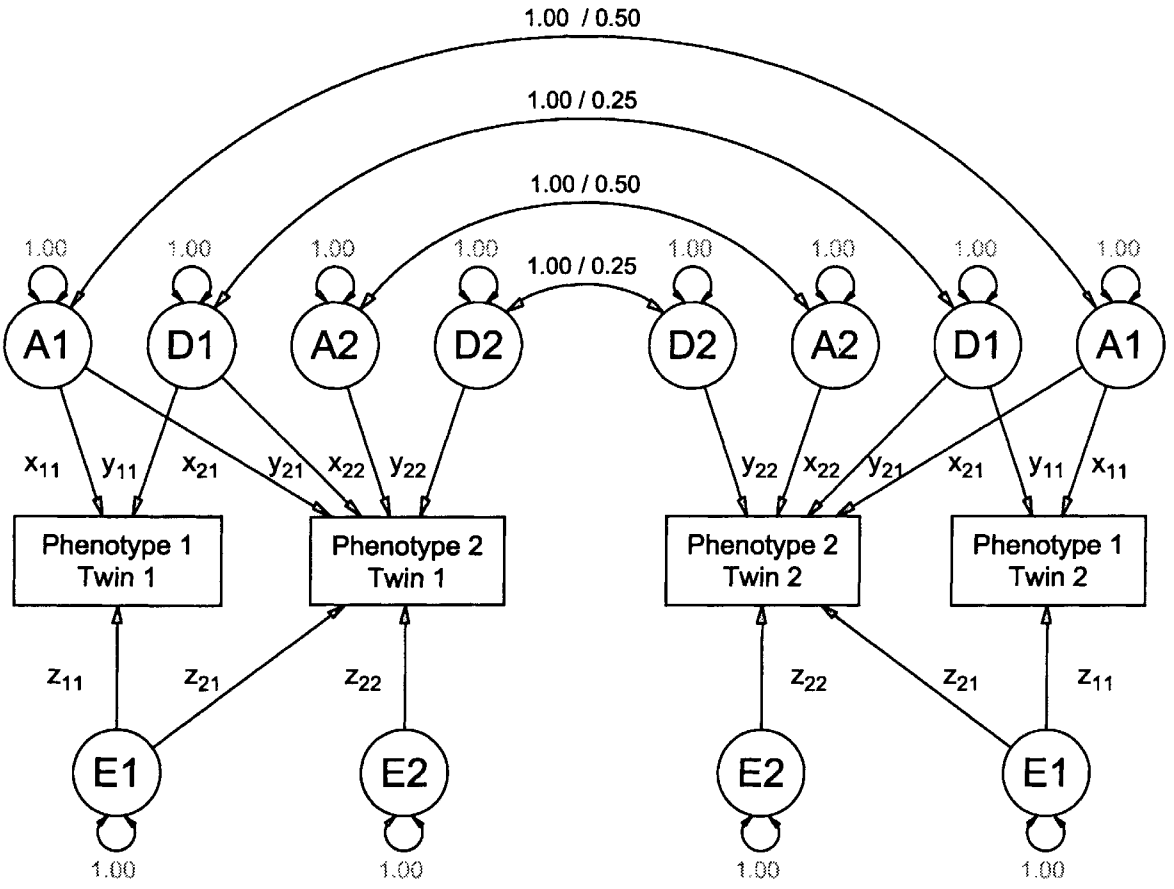

Figure 3

Path diagram representing the resemblance between MZ or DZ twins, for additive genetic influences (A), dominant genetic influences (D), and non-shared environmental influences (E), in a bivariate design.

The additive, dominance and environmental (co)variances can be represented as elements of the symmetric matrices $\mathbf{A} = \mathbf{X}\mathbf{X}^T$, $\mathbf{D} = \mathbf{Y}\mathbf{Y}^T$, and $\mathbf{E} = \mathbf{Z}\mathbf{Z}^T$. They contain the additive genetic, dominance, and non-shared environmental variances respectively on the diagonals for variables 1 to *n*. **X**, **Y** and **Z** are known as the Cholesky decomposition of the matrices **A**, **D** and **E**, which assures that these matrices are nonnegative definite. The latter is required for variance-covariance matrices.

The *genetic correlation* between variables *i* and *j* (r_{gij}) is derived as the genetic covariance between variables *i* and *j* (denoted by element *ij* of matrix **A**; a_{ij}) divided by the square root of the product of the genetic variances of variables *i* (a_{ii}) and *j* (a_{jj}):

$$r_{gij} = \frac{a_{ij}}{\sqrt{a_{ii} \times a_{jj}}}.$$

Analogously, the *environmental correlation* (r_{eij}) between variables *i* and *j* is derived as the environmental covariance between variables *i* and *j* divided by the square root of the product of the environmental variances of variables *i* and *j*:

$$r_{eij} = \frac{e_{ij}}{\sqrt{e_{ii} \times e_{jj}}}.$$

The phenotypic correlation *r* is the sum of the product of the genetic correlation and the square roots of the standardized genetic variances (i.e., the heritabilities) of the two phenotypes and the product of the environmental correlation and the square roots of the standardized environmental variances of the two phenotypes:

$$r = r_{gij} \times \sqrt{\frac{a_{ii}}{(a_{ii} + e_{ii})}} \times \sqrt{\frac{a_{jj}}{(a_{jj} + e_{jj})}}$$

$$+ r_{eij} \times \sqrt{\frac{e_{ii}}{(a_{ii} + e_{ii})}} \times \sqrt{\frac{e_{jj}}{(a_{jj} + e_{jj})}}$$

(i.e., observed correlation is the sum of the genetic contribution and the environmental contribution).

The genetic contribution to the observed correlation between two traits is a function of the two sets of genes that influence the traits and the correlation between these two sets. However, a large genetic correlation does not imply a large phenotypic correlation, as the latter is also a function of the heritabilities. If these are low, the genetic contribution to the observed correlation will also be low.

If the genetic correlation is 1, the two sets of genetic influences overlap completely. If the genetic correlation is less than 1, at least some genes are a member of only one of the sets of genes. A large genetic correlation, however, does not imply that the overlapping genes have effects of similar magnitude on each trait. The overlapping genes may even act additively for one trait and show dominance for the second trait. A genetic correlation less than 1 therefore cannot exclude that all of the genes are overlapping between the two traits (Carey, 1988). Similar reasoning applies to the environmental correlation.

Genetic correlations do not provide information on the direction of causation. In fact, genes may influence one trait that in turn influences the second trait. Or, there may be genes that act in a pleiotropic way (i.e., they influence both traits but neither trait influences the other). Genetic correlations do not distinguish between these situations, but merely provide information on the nature of the causes of covariation between two traits.

Longitudinal Analysis of Twin Data

The aim of longitudinal analysis of twin data is to consider the genetic and environmental contributions to the dynamics of twin pair responses through time. In this case the phenotype is measured at several distinct time points for each twin in a pair. To analyse such data one must take the serial correlation between the consequent measurements of the phenotype into consideration. The classical genetic analysis methods described in previous sections are aimed at the analysis of a phenotype measured at a single point in time and provide a way of estimating the time-specific heritability and variability of environmental effects. However, these methods are not able to handle serially correlated longitudinal data efficiently.

To deal with these issues the classic genetic analysis methods have been extended to investigate the effects of genes and environment on the development of traits over time (Boomsma & Molenaar, 1987b; McArdle 1986). Methods based on the Cholesky factorization of the covariance matrix of the responses treat the multiple phenotype measurements in a multivariate genetic analysis framework (as discussed under "Multivariate Analysis of Twin Data"). "Markov chain" (or "Simplex") model methods (Dolan, 1992; Dolan et al., 1991) provide an alternative account of change in covariance and mean structure of the trait over time. In this case the Markov model structure implies that future values of the phenotype depend on the current trait values alone, not on the entire past history. Methods of function-valued quantitative genetics (Pletcher & Geyer, 1999) or the genetics of infinite-dimensional characters (Kirkpatrick & Heckman, 1989) have been developed for situations where it is necessary to consider the time variable on a continuous scale. The aim of these approaches is to investigate to what extent the variation of the phenotype at different times may be explained by the same genetic and environmental factors acting at the different time points and to establish how much of the genetic and environmental variation is time-specific.

An alternative approach for the analysis of longitudinal twin data is based on random growth curve models (Neale & McArdle, 2000). The growth curve approach to genetic analysis was introduced by Vandenberg & Falkner (1965) who first fitted polynomial growth curves for each subject and then estimated heritabilities of the components. These methods focus on the rate of change of the phenotype (i.e., its slope or partial derivative) as a way to predict the level at a series of points in time. It is assumed that the individual phenotype trajectory in time may be described by a parametric growth curve (e.g., linear, exponential, logistic etc.) up to some additive measurement error. The parameters of the growth curve (e.g., intercept and slope, also called latent variables) are assumed to be random and individual-specific. However, the random intercepts and slopes may be

dependent within a pair of twins because of genetic and shared environmental influences on the random coefficients. The basic idea of the method is that the mean and covariance structure of the latent variables determines the expected mean and covariance structure of the longitudinal phenotype measurements and one may therefore estimate the characteristics of the latent variable distribution based on the longitudinal data.

The random growth curve approach shifts focus of the genetic analysis towards the new phenotypes — the parameters of the growth curve model. This framework permits to investigate new questions concerning the nature of genetic influence on the dynamic characteristics of the phenotype, such as the rate of change. If the random parameters of the growth curve would be observed, they might have been analysed directly using the classical methods of multivariate genetic analysis. However, their latent nature requires a more elaborate statistical approach. Since the growth curve model may be formulated in terms of the mean and covariance structure of the random parameters one might simply take the specification of the mean and covariance structure of a multivariate phenotype as predicted by the classical methods of multivariate genetic analysis and plug it in into the growth curve model. The resulting two-level latent variable model would then allow for multivariate genetic analysis of the random coefficients.

In the following sections we consider the bivariate linear growth curve model applied to longitudinal twin data using age as timescale. The approach may be extended to other parametric growth curves (e.g., exponential, logistic etc.) using first-order Taylor expansions and the resulting mean and covariance structure approximations (Neale & McArdle, 2000). We also describe a method for obtaining the predicted individual random growth curve parameters using the empirical Bayes estimator. These predictions may be useful for selection of most informative pairs for subsequent linkage analysis of the random intercepts and slopes.

Linear Growth Curve Model

A simple implementation of the random effects approach may be carried out using linear growth curve models. In this case each individual is characterized by a random intercept and a random slope, which are considered to be the new phenotypes. In a linear growth curve model the continuous age-dependent trait (Y_{1t}, Y_{2t}) for sib 1 and sib 2 are assumed to follow a linear age-trajectory given the random slopes and intercepts with some additive measurement error: $Y_{it} = \alpha_i + \beta_i\, t + \varepsilon_{it}$, for sibling i, where $i = 1, 2$, t denotes the timepoint ($t = 1, 2, \ldots, n$), α_i and β_i are the individual (random) intercept and slope of sib i, respectively, and ε_{it} is a zero-mean individual error residual, which is assumed to be independent from α_i and β_i. The aim of the study is then the genetic analysis of the individual intercepts (α_i) and slopes (β_i). The model may easily be extended to include covariates.

Assume the trait (Y_{it}) is measured for the two sibs at $t = 1, 2, \ldots, n$. The measurements on both twins at all time points may be written in vector form as $\mathbf{Y} = (Y_{11},\ldots, Y_{1n}, Y_{21}, \ldots, Y_{2n})^T$ (where ^{T} denotes transposition). Furthermore, if $\mathbf{L}$ denotes the vector of the random growth curve parameters, the matrix form of the linear growth curve model is: $\mathbf{Y} = \mathbf{DL} + \mathbf{E}$

where

$$\mathrm{L} = \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix},\ \boldsymbol{D} = \begin{pmatrix} F & 0 \\ 0 & F \end{pmatrix},\ \boldsymbol{F} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & n \end{pmatrix},\ \mathbf{E} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n} \end{pmatrix}$$

Note, that the linear growth curve model actually represents a Structural Equation Model with latent variables α_i and β_i (and ε_{it}'s) with loadings of the latent variables on the observed responses Y_{it} given by either 1 or t . This implies, that this model may be analysed using the general SEM techniques. In particular, parameter estimation may be carried out via the Maximum Likelihood method under multivariate normality assumptions using the fact that the moment structure ($\boldsymbol{m}_Y$, Σ_Y) of $\mathbf{Y}$ can be expressed in terms of $\boldsymbol{m}$, Σ and Var(ε_{it}), where $\boldsymbol{m} = \mathbf{E}(\mathbf{L})$, $\Sigma = \mathrm{Cov}(\mathbf{L}, \mathbf{L})$: $\boldsymbol{m}_Y = \mathbf{D}\boldsymbol{m}$, $\Sigma_Y = \mathbf{D}\,\Sigma\,\mathbf{D}^T + \Sigma_\varepsilon$, where $\Sigma_\varepsilon = \mathrm{Cov}(\mathbf{E}, \mathbf{E})$.

As described previously in the section on multivariate genetic analysis, the two-dimensional phenotype (α_i, β_i) may be analysed by modelling the covariance matrix Σ for MZ and DZ twins using the Cholesky factorisation approach. The two-level model construction leads to a parameterisation of the joint likelihood for the trait in terms of the variance components, the respective mean vectors and residual variances. This yields estimates of the two heritability values of α_i and β_i (and respective variabilities of the environmental effects) and also estimates of correlations between the genetic and environmental components of α_i and β_i, as described earlier under "Multivariate Analysis of Twin Data".

Predicting the Random Intercepts and Slopes

In the following, we briefly describe a method for obtaining the predicted individual random growth curve parameters. These predictions may be useful for selection of most informative pairs for subsequent linkage analysis of the random intercepts and slopes.

The prediction of the individual random growth curve parameters may be given by the empirical Bayes estimates, that is, the conditional expectation of the random growth curve parameters given the measurement $\mathbf{Y} = \mathbf{y}$. As noted above, $\mathbf{Y}$ and $\mathbf{L}$ are assumed multivariate normal, i.e., $\mathbf{Y} \sim \mathrm{MVN}(\boldsymbol{m}_Y; \Sigma_Y)$ and $\mathbf{L} \sim \mathrm{MVN}(\boldsymbol{m}; \Sigma)$. The joint distribution of $\mathbf{Y}$ and $\mathbf{L}$ is given by

$$\begin{bmatrix} \mathbf{L} \\ \boldsymbol{Y} \end{bmatrix} \sim \boldsymbol{MVN}\left\{ \begin{bmatrix} \boldsymbol{m} \\ \boldsymbol{m}_Y \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma\boldsymbol{D}^T \\ \boldsymbol{D}\Sigma & \Sigma_Y \end{bmatrix} \right\} .$$

By multivariate Gaussian theory, the predicted random growth curve parameters of a pair with measurement $\mathbf{Y} = \mathbf{y}$ may then be given by $\hat{\mathbf{L}}(y) = \mathbf{m} + \Sigma\mathbf{D}^T\,\Sigma_Y^{-1}(\mathbf{y} - \mathbf{m}_Y)$ for estimated parameters. The estimator is the best linear unbiased predictor of the individual random growth curve parameters (see Harville, 1976). Furthermore, an assessment of the error in estimation is provided by the variance

of the difference $\hat{\mathbf{L}} - \mathbf{L}$ given by: $\mathrm{Var}(\hat{\mathbf{L}} - \mathbf{L}|\mathbf{Y}) = \Sigma - \Sigma\mathbf{D}^T \Sigma_Y^{-1} \mathbf{D}\Sigma$.

Additional Components of Variance

In the aforementioned models the absence of effects of *genes × environment interaction*, of a *genes–environment correlation* and of *assortative mating* was assumed. Using the appropriate design these effects can relatively easy be incorporated in structural equation models.

G×E interaction occurs when the effects of the environment are conditional on an individual's genotype, such as when some genotypes are more sensitive to the environment than other genotypes. Genetic studies on crops and animal breeding experiments have shown that G×E interaction is extremely common (see summary in Lynch & Walsh, 1998. pp. 657–686). However, in general G×E interaction accounts for less than 20% of the variance of a trait in the population (Eaves, 1984; Eaves et al., 1977).

G×E interaction can be modelled according to two main methods. In the first method the presence of G×E interaction is explored by studying the same trait in two environments (or at two time points). A genetic correlation between the two measurements that is less than one indicates the presence of G×E interaction (Boomsma & Martin, 2002; Falconer, 1952). However, a genetic correlation equal to one does not need to imply the absence of G×E interaction (see Lynch & Walsh, 1998). In human quantitative genetic analyses it is often not possible to control the environmental or genetic influences, unless the specific genotype and specific environmental factors are explicitly measured (see Boomsma et al., 1999; Dick et al., 2001; Kendler & Eaves, 1986; Rose et al., 2001; but see Molenaar et al., 1990, 1999).

In a second method the presence of G×E interaction is explored by correlating the MZ intra-pair differences and MZ pair sums (Jinks & Fulker, 1970). Assuming that MZ twin similarity is purely genetic, a relation between MZ means and standard deviations suggests the presence of G×E interaction.

G×E interaction is often not included in quantitative genetic models. However, if the true world does include GxE interaction, assuming its absence may lead to biased estimates of G and E (Eaves et al., 1977). For example if G×E interaction was truly gene by non-shared environment interaction, a model without G×E interaction will result in overestimation of the effects of the non-shared environment. If, however, G×E interaction was interaction between genes and shared environmental influences, assuming its absence will result in overestimation of the effect of genes on the phenotype, as well as in overestimation of the influence of the shared environmental on the phenotype. The separate detection of these two biased effects in the presence of genes by shared environmental interaction necessitates the inclusion of twin pairs reared apart (Eaves et al., 1977; Jinks & Fulker, 1970).

GE-correlation occurs when the genotypic and environmental values are correlated. Three different forms of GE-correlation have been described (Plomin et al., 1977; Scarr & McCartney, 1983). Passive GE-correlation occurs for example when parents transmit both genes and environment (cultural transmission) relevant for a certain trait (Eaves et al., 1977). Effects of cultural transmission can be measured using a twin design that also includes the parents of twins (Boomsma & Molenaar, 1987a; Fulker, 1988). Active GE-correlation is the situation where subjects of a certain genotype actively select environments that are correlated with that genotype. Reactive GE-correlation refers to the effects of reactions from the environment evoked by an individual's genotype. The presence of a positive GE-correlation leads to an increase in the phenotypic variance. It is difficult to measure GE-correlation, however, as active and reactive GE-correlation necessitate the direct measurement of these influences (Falconer & Mackay, 1996). Falconer and Mackay (1996) state that GE-correlation is best regarded as part of the genetic variance because "… the non-random aspects of the environment are a consequence of the genotypic value …"

Assortative mating refers to a correlation between the phenotypic values of spouses (e.g., Willemsen et al., 2002). Assortative mating can be based on social homogamy (i.e., preferential mating within one's own social class) or may occur when mate selection is based on a certain phenotype (P; which in turn is a function of G and E). Phenotypic assortative mating tends to increase additive genetic variation (and therefore in the overall phenotypic variation; Lynch & Walsh, 1998) and consequently increases the resemblance between parents and offspring as well as the resemblance among siblings and DZ twins. Statistically, it may conceal the presence of non-additive genetic effects and overestimate the influence of additive genetic factors (Eaves et al., 1989; Cardon & Bell, 2000; Carey, 2002; Heath et al., 1984; Posthuma et al., in press). Assortative mating is known to exist for traits such as intelligence, exercise behavior and body height and weight (e.g., Aarnio et al., 1997; Boomsma, 1998; Vandenberg, 1972), but is to a large extent an unexplored topic in most human populations.

Gene Finding

In addition to the estimation of the effects of unmeasured genes and environmental influences on traits, SEM can also be used to test the effects of *measured* genetic and environmental factors. In the following sections we will concentrate on the detection of the actual genes that influence a trait. Two methods are currently employed: Quantitative trait loci (QTL) linkage analysis and association analysis.

QTL Linkage Analysis

Quantitative trait loci (QTL) linkage analysis establishes relationships between dissimilarity or similarity in a quantitative trait in genetically related individuals and their dissimilarity or similarity in regions of the genome. If such a relationship can be established with sufficient statistical confidence, then one or more genes in those regions are possibly involved in trait (dis)similarity among individuals.

Linkage analysis depends on the *co-segregation* (i.e., a violation of Mendel's law of independent assortment which applies only to inheritance of different chromosomes) of alleles at a marker and a trait locus (Ott, 1999). If a pair of offspring has received the same haplotype from a parent in

a certain region of the genome, the pair is said to share the parent's alleles in that region identical by descent (IBD). Since offspring receive their haplotypes from two parents, the pair can share *0*, *1* or *2* alleles IBD at a certain locus in a region. The IBD status of a pair is usually estimated for a number of markers with (approximately) known location along the genome and is then used as the measure of genetic similarity at the marker. The IBD status at a marker is informative for the IBD status at any other locus along the chromosome as long as the population recombination fraction between the marker and the locus is less than 0.5. In that case the IBD status at the marker and the locus are correlated in the population and hence similarity at the marker is informative for similarity at the locus. The locus may be a gene or may be located near a gene. If variation in the gene and variation in the trait are related (i.e., the gene is a QTL), then variation in the IBD status at the locus and thus also at the marker will be related to variation in trait similarity. Identity by descent is distinct from identity is distinct from identical by state (IBS), which denotes the number of physically identical alleles that a pair of offspring has received, and that have not necessarily been inherited from the same parent. Table 2 gives all possible sib pairs from a A1A2 by A1A2 mating and their IBD / IBS status.

A commonly made distinction is between parametric and nonparametric models for linkage analysis. Parametric models, which are discussed at length by Ott (1999), require a fairly detailed specification of population characteristics of the gene, such as allele frequencies and penetrances. Nonparametric models are not nonparametric in the usual statistical meaning of nonparametric, but these are parametric models that require fewer assumptions than nonparametric linkage models. Parametric models use a fairly simple relationship between the contribution of a QTL to the covariance of the trait values of pairs of individuals, the pairs' IBD status at a certain location on the genome and the recombination fraction between the locus and the QTL. This relationship is usually expressed as a function of the proportion π_i of alleles shared identical by descent, $\pi_i = i/2$ for $i = 0, 1, 2$. For sibling pairs, for instance, the contribution to the covariance given the proportion π_i is $f(\theta)\ \pi_i\ \sigma^2$, where σ^2 is the QTL's contribution to the population trait variance, θ is the recombination fraction between the locus and the QTL, and the monotonic function $f(\theta)$ equals *0* and *1* for recombination fractions *0.5* and *0*, respectively. Since IBD status is not always unambiguously known, it is usually estimated probabilistically from the specific allele pattern across chromosomes of two or more siblings (Abecasis et al., 2002; Kruglyak et al., 1996). The estimate of π is referred to as $\hat{\pi}$, and can be calculated as (Sham, 1998): $\hat{\pi} = {}^1/_2\ p_{IBD1} + p_{IBD2}$.

We call the correlation between the dominance values within a population of sib pairs δ, which in the population can be estimated as: $\hat{\delta} = p_{IBD2}$, where p_{IBD2} is the proportion of sib pairs that share two alleles IBD in this population. This can be incorporated in a path diagram (Figure 4).

The path coefficients v and w (Figure 4) and the relative contribution of the factors *Am* and *Dm* to the phenotypic variation are a function of the recombination fraction between the marker and the trait locus and the magnitude of the genetic effects of the trait locus. Relatively small effects of the factors *Am* and *Dm* can thus either reflect a situation with small effects at the trait locus and a small recombination fraction (close to zero) or may reflect large effects at the trait locus in combination with a large (close to 0.5) recombination fraction between the marker and the trait locus.

The expectation for the variance is in algebraic terms $x^2 + y^2 + v^2 + w^2 + z^2$, the expectation for the covariance among MZ twins is $x^2 + y^2 + v^2 + w^2$, and for the covariance among DZ twins is ${}^1/_2\ x^2 + {}^1/_4\ y^2 + \hat{\pi}v^2 + \hat{\delta}w^2$. Translating this into matrix algebraic terms we introduce matrix **Q** (i.e., the product of matrices **V** and $\mathbf{V}^T$, representing the additive genetic influences on the phenotype at the marker site) and matrix **R** (i.e., the product of matrices **W** and $\mathbf{W}^T$, representing the dominant genetic influences on the phenotype at the marker site). Written in matrix algebra the expectation for the variance equals $\mathbf{XX}^T + \mathbf{YY}^T + \mathbf{ZZ}^T + \mathbf{VV}^T + \mathbf{WW}^T$ (or $\mathbf{A} + \mathbf{D} + \mathbf{E} + \mathbf{Q} + \mathbf{R}$), the expectation for the covariance of MZ twins equals $\mathbf{XX}^T + \mathbf{YY}^T + \mathbf{VV}^T + \mathbf{WW}^T$ (or $\mathbf{A} + \mathbf{D} + \mathbf{Q} + \mathbf{R}$), and the expectation for the covariance of DZ twins equals $0.5\mathbf{XX}^T + 0.25\mathbf{YY}^T + \hat{\pi}\mathbf{VV}^T + \hat{\delta}\,\mathbf{WW}^T$ (or $0.5\mathbf{A} + 0.25\mathbf{D} + \hat{\pi}\mathbf{Q} + \hat{\delta}\,\mathbf{R}$).

Testing whether the elements of matrices V and W are statistically different from zero provides a test for linkage at a particular marker position. In a genome screen this test is conducted for each marker along the genome. Those marker positions for which the χ^2 difference exceeds a certain critical value are believed to be linked to a QTL. Apart from calculating $\hat{\pi}$'s and $\hat{\delta}$'s to model the linkage component, one may also apply a mixture model. In this model for each sib pair three models (for IBD = 0, IBD =

Table 2

IBD / IBS Status from all Possible Sib Pairings from Parental Mating Type A1A2 (father) × A1A2(mother)

	Sib 1			
	A1A1	**A1**A2	**A2**A1	**A2**A2
Sib 2				
A1A1	2 / 2	1 / 1	1 / 1	0 / 0
A1A2	1 / 1	2 / 2	0 / 2	1 / 1
A2A1	1 / 1	0 / 2	2 / 2	1 / 1
A2A2	0 / 0	1 / 1	1 / 1	2 / 2

Note: **A1A2** (Father in bold) × A1A2 (Mother in normal text).

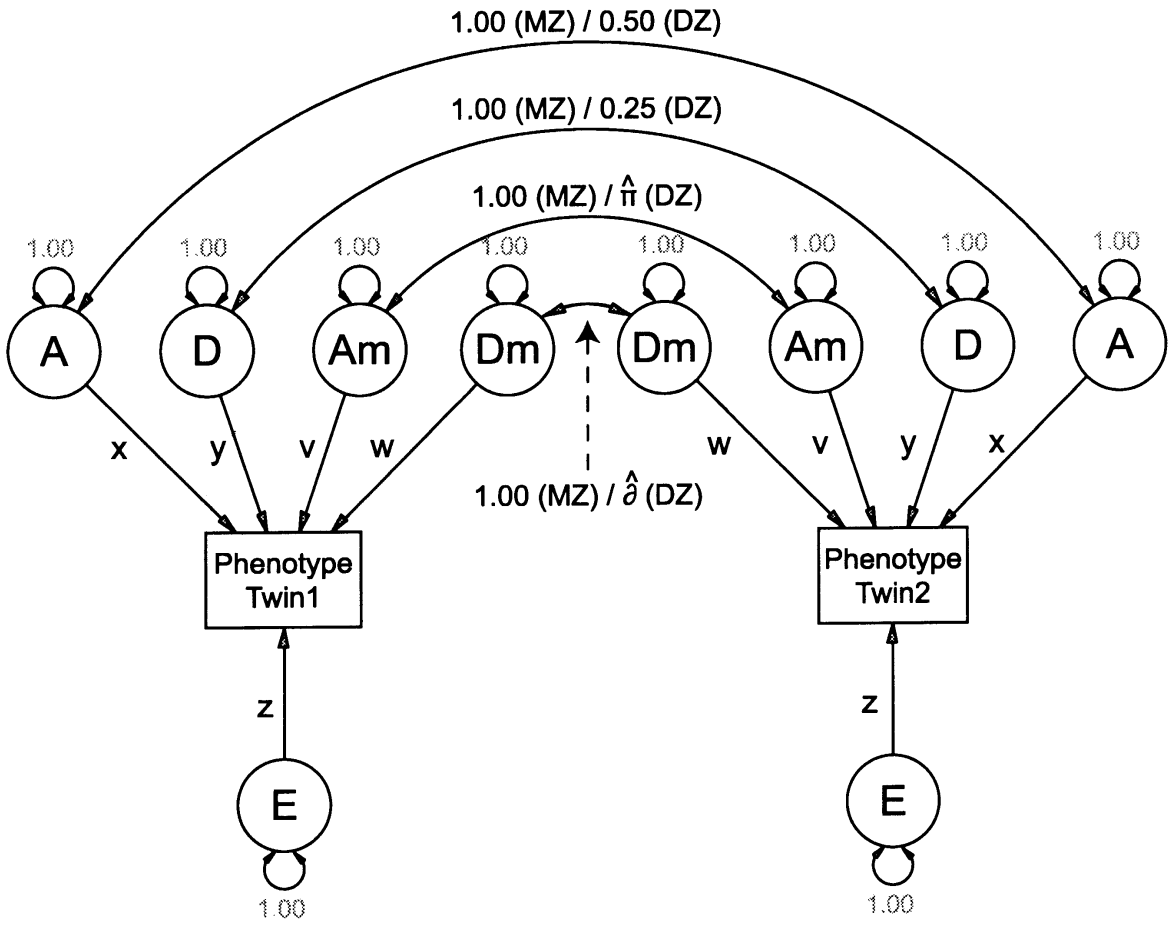

Figure 4

Path diagram representing the resemblance between MZ or DZ twins, for background additive genetic influences (A), background dominant genetic influences (D), additive genetic influences due to the marker site (Am), dominant genetic influences due to the marker site (Dm), and non-shared environmental influences (E), in a univariate design.

1, IBD = 2) are fitted to the data that are weighted by their relative probabilities.

Apart from these variance components methods for linkage analyses, other statistical methods for conducting a QTL linkage analysis have been proposed. Haseman and Elston (1972) developed a now classical model for QTL linkage analysis. In the HE model the squared difference of the trait values y_1 and y_2 of siblings *1* and *2* is regressed on the value of π_i at a particular location on the genome. The average of the squared difference in a given population equals $var\ (y_1) + var\ (y_2) - 2\ cov\ (y_1, y_2)$. As the assignment of individuals' trait values to y_1 or y_2 is usually arbitrary this becomes $2var\ (y) - 2\ cov\ (y_1, y_2)$. With a few additional assumptions only $cov\ (y_1, y_2)$ varies as a function of the IBD status at the locus, although it does also depend on the variances of other QTL's. Then the regression equation for the squared difference is $\mu - f(\theta)2\ \sigma^2\pi_i$, where π_i is the regressor and the constant μ contains variances associated with the total environmental and genetic effects, including the QTL. A statistically significant negative estimated regression weight is suggestive of a QTL at or near the locus.

By a simple rewriting of the likelihood of the joint distribution under normality, Wright (1997) demonstrated that the difference and sum together carry all the information in the variance and covariance of the joint distribution. He suggested regression approaches using both the difference and the sum. Subsequently several authors (Drigalenko, 1998; Forrest, 2001; Sham & Purcell, 2001; Sham et al., 2002; Visscher & Hopper, 2001; Xu et al., 2000) proposed regression methods that use the information in both the squared sum and the squared difference for inference about a QTL effect. These methods were shown to be nearly as powerful as the likelihood-based variance component methods. The methods generally have the advantage that the computations are easy and fast, which is of some importance if the models are fitted for many locations along the genome. In contrast, the variance components methods can be quite time consuming and may, moreover, yield estimates that do not maximize the likelihood, which is required for the validity of the distribution theory on which inferences about the QTL are based.

Association Analysis

Ultimately we aim to quantify the specific effects of genes in terms of *a*, *d*, and *–a* (see Figure 1), using association analysis. In association analysis the effects of measured alleles on a trait are tested, by incorporating fixed effects on the means using genetically related or unrelated subjects (Moxley et al., 2002; Neale et al., 1999; Neale, 2000; Zhu et al., 1999). The measured alleles can be functional alleles at a candidate gene locus, or non-functional alleles at a marker locus that is in linkage disequilibrium (LD) with the trait locus. Traditional association studies such as case control studies may provide spurious associations as a result of the effects of the use of stratified samples (Hamer & Sirota, 2000). Any trait that has a different distribution across substrata (e.g., due to cultural differences between strata or assortative mating within strata) will show a statistical association with any allele that has a different frequency across those substrata (e.g., as a result of different ancestors or genetic drift). To control for these confounding effects, statistical methods based on genomic control may be employed (Lesch, 2003). In these methods the statistical significance level for testing genetic association is adjusted using information from assessed single nucleotide polymorphisms or unlinked microsatellite markers (Bacanu et al., 2000; Pritchard et al., 2000). Alternatively, to control for the effects of population stratification, one may employ family based methods in which locus-trait associations are compared across genetically related individuals, which, by definition, stem from the same stratum. Most family based association tests have been developed in the context of discrete traits, but recently new statistical developments have provided methods for association analyses of quantitative traits (for a review see Vink & Boomsma, 2002). A very powerful method for quantitative traits was proposed by Fulker et al., 1999, who developed a method that not only allows the simultaneous analysis of linkage and association, but also simultaneously controls for possible confounding effects of population stratification. By partitioning the association effects into a between family component and a within family component, spurious associations can be separated from genuine associations. The between family effects reflect both the genuine and the possible spurious association between locus alleles and a trait (or allelic association between locus alleles and trait locus alleles). The within family effects reflect only the genuine association.

Fulker et al. (1999) thus decomposed the additive value *a* (see Figure 1) into a "between" component (a_b: the genotypic mean of a sib pair) and a "within" component (a_w: the deviation of a sib from the genotypic mean of the sib pair). Analogously, the dominance deviation *d* can be decomposed into a between part (d_b) and a within part (d_w). For example, the genotypic mean of a sib pair with genotypes A1A1, A1A2 is $(a + d)/2$ or $^1/_2\, a_b + {}^1/_2\, d_b$. The *a* and *d* have the subscript *b* to denote the between effect, the two $^1/_2$'s are the *coefficients* with which the between effects of *a* and *d* are multiplied for this particular sib pair. The deviation of the genotypic mean for the sib with genotype A1A1 is $a - ((a + d)/2)$ or $^1/_2\, a_w - {}^1/_2\, d_w$. The subscripted *w* refers to the within effect, $^1/_2$ is the coefficient for the additive within effect for this particular sib (with genotype A1A1) from this particular pair, and $-\,^1/_2$ is the coefficient for the dominance within effect for this particular sib from this pair. For the sib with genotype A1A2 the within effect is derived as $d - ((a + d)/2)$ or $-\,^1/_2\, a_w + {}^1/_2\, d_w$, where $-\,^1/_2$ is the coefficient for the additive within effects for the sib with genotype A1A2, and $^1/_2$ is the coefficient for the dominance within effects for the sib with genotype A1A2 in this particular pair.

This derivation of coefficients for additive between and within components from sib genotypic means and deviations from the genotypic sib mean for all possible sib pairs and for a diallelic locus, can be found in Fulker et al. (1999), an extension to dominance effects, multi-allele loci, variable sibshipsizes, and the use of parental genotypic information where available can be found in Posthuma et al. (in press).

As opposed to the effects of the sharing of genomic regions between sibpairs, which are described in the model for the (co-) variance of a trait, *association effects* (i.e., allelic effects on trait means) are described in the model for the trait mean(s) for each individual, next to the effects of other covariates such as age or sex.

Formally, the model for an observed score in sib *j* from the *i*-th family (y_{ij}) is represented as: $y_{ij} = \mu + \beta_1\, age_{ij} + \beta_2\, sex_{ij} + c_{abi}\, a_b + c_{awij}\, a_w + c_{dbi}\, d_b + c_{dwij}\, d_w + \varepsilon_{ij}$, where y_{ij} is the observed score for sib *j* in the *i*-th family, μ denotes the overall grand mean or intercept (assumed to be equal for all individuals), β_1 denotes the regression coefficient for the first covariate (age in this example), β_2 denotes the effect of the second covariate (a deviation of females in this example), age_{ij} and sex_{ij} denote the observed age and sex (male = 0; female=1) respectively of sib *j* in the *i*-th family, c_{abi} is the coefficient derived from the sib genotypic mean (e.g., $^1/_2$, or $-\,^1/_2$., 1, etc) for the additive between genetic effect for the *i*-th family, c_{awij} denotes the coefficient for the additive within genetic effects for sib *j* from the *i*-th family as can be derived from the deviation of sib *j* from the sib genotypic mean, c_{dbi} is the derived coefficient for the dominant between genetic effect for the *i*-th family, c_{dwij} denotes the derived coefficient for the dominant within genetic effects for sib *j* from the *i*-th family. a_b and a_w are the estimated additive between and within effects, d_b and d_w are the estimated dominance between and within effects, and ε_{ij} denotes that part of the grand mean that is not explained by the covariates (age, sex) and genotypic effects.

Equating the between genetic effects and the within genetic effects serves as a test of the presence (and direction) of spurious associations between a locus and a trait in the dataset: when the two effects are unequal a spurious association is said to exist. This test can be conducted on DNA markers as well as candidate genes.

Combined Linkage and Association Analysis

The model formulated by Fulker et al. (1999) can also be used as a simultaneous test of linkage (using identity-by-descent information at positions across the genome) and association (using the alleles from candidate genes/markers lying within the region that shows linkage), which allows quantification of the amount of linkage that can be ascribed to the locus used in the association (Abecasis et al.,

2000, 2001; Cardon & Abecasis, 2000; Fulker et al., 1999; McKenzie et al., 2001).

Testing the significance of linkage in the presence of association against the significance of linkage when association is not part of the model, provides information on how well the locus explains the linkage result. If the locus used in the association is the actual quantitative trait locus (QTL), the evidence for linkage at that locus is expected to disappear when modeled simultaneously with association. Incomplete reduction of evidence for linkage in the presence of a significant genuine association effect of a locus within the linkage region, implies that the linkage derives from some other gene within that genomic region, or that not all relevant alleles of the locus have been genotyped, or that (part of) the observed linkage may have been artefactual (i.e., due to marker genotype errors) (Abecasis et al., 2000, 2001; Cardon & Abecasis, 2000; McKenzie et al., 2001).

Conclusion

Classical biometrical genetics provides the fundamental theory for the quantification of gene effects. Proper implementation of this theory into genetic software packages for mixed models, path analysis or structural equation modelling (e.g., Mx, Lisrel, MERLIN, or SOLAR (Neale, 1997; Jöreskog & Sörbom, 1986; Abecasis et al., 2002; Almasy & Blangero, 1998 respectively), using path analysis or structural equation modelling, provides a flexible practical tool for gene finding.

Since the identification of the first gene for a common disease in humans in 1991 (Goate et al., 1991), the successes in mapping human genes for common disease have slowly but surely accelerated (Korstanje & Paigen, 2002), impressively illustrated by the reports for Crohn's disease (Hugot et al., 2001), type-II diabetes (Horikawa et al., 2000), schizophrenia (Stefansson et al., 2002; Straub et al., 2002) and asthma (Hakonarson et al., 2002; Van Eerdewegh et al., 2002). This acceleration of successes seems primarily associated with the availability of large sample sizes (Altmüller et al., 2001).

In the genomEUtwin project, funded by the European Union, data from over 0.8 million twins and their relatives are available from twin registries in eight countries. The combined datasets on traits such as body height and weight, body mass index, cardiovascular disease and migraine, will provide the necessary statistical power for unravelling not only the genetic but also the environmental architecture of these traits. Knowledge of the genetic architecture of a trait is, for example, crucial for parametric linkage analyses such as variants of the Haseman-Elston regression, where several parameters are expected to be known a priori (see e.g., Sham & Purcell, 2001).

In addition, such a large combined dataset offers the unique opportunity for modelling effects that are otherwise difficult (if not impossible) to detect, such as effects of gene–gene interaction (epistasis), gene–environment interaction as well as differences in genetic or environmental influences across countries. For example, across countries there may be different allele frequencies (Cavalli-Sforza et al., 1994), different genotypic values, different environmental contributions to the variance, or different loci that influence the same trait in different countries. Also, large datasets especially offer the opportunity to select the most informative families for linkage (Carey & Williamson, 1991; Cardon & Fulker, 1994; Dolan & Boomsma, 1998; Eaves & Meyer, 1994; Gu et al., 1996; Risch & Zhang, 1995, 1996).

Many of the targeted traits in genomEUtwin have been measured longitudinally. Longitudinal data are extremely valuable for understanding the development of disease and elucidating molecular processes such as the on/off switching of genes with age. For example, the availability of longitudinal data enables testing whether stability in a trait is due to genetic (e.g., "*do the same genes influence body mass index at different ages?*") or environmental influences. Genes for traits such as migraine, which usually develops between the ages of 25 and 55 (Breslau & Rasmussen, 2001), will go undetected when the trait is measured prior to genetic expression. Longitudinal association analyses are therefore crucial in detecting such age-related genetic penetrance.

For the purpose of analyzing the genomEUtwin datasets on body height and weight, body mass index, and migraine a library of Mx scripts has been developed[1] containing scripts for the analysis of continuous or categorical traits, saturated or variance decomposition models, twins only or extended twins design, univariate, multivariate, or longitudinal models, or association and linkage models. The aim of providing such a scripts library is to allow a uniform method of analysis, starting from saturated models, via variance decomposition models to linkage and association models. The application of sound uniform statistical modelling to the data from the largest genetic epidemiological study cohort in the world should help us meet the ultimate goal of the genomEUtwin project: the identification of polymorphisms that influence population variation in body height, weight, body mass index, cardiovascular disease, or migraine.

Endnote

1 Available from http://www.psy.vu.nl/mxbib.

Acknowledgments

This paper originates from the genomEUtwin project which is supported by the European Union Contract No. QLG2-CT-2002-01254

References

Aarnio, M., Winter, T., Kujala, U. M., & Kaprio, J. (1997). Familial aggregation of leisure-time physical activity — A three generation study. *International Journal of Sports Medicine*, *7*, 549–556.

Abecasis, G. R., Cardon, L. R., & Cookson, W. O. (2000). A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics*, *66*, 279–292.

Abecasis, G. R., Cherny, S. S., & Cardon, L. R. (2001). The impact of genotyping error on family–based analysis of quantitative traits. *European Journal of Human Genetics*, *9*, 130–134.

Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics, 30*, 97–101.

Almasy, L., & Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics, 62*, 1198–1211.

Altmuller, J., Palmer, L. J., Fischer, G., Scherb, H., & Wjst, M. (2001). Genomewide scans of complex human diseases: True linkage is hard to find. *American Journal of Human Genetics, 69*, 936–950.

Amos, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics*, *54*, 535–543.

Bacanu, S. A., Devlin, B., & Roeder, K. (2000). The power of genomic control. *American Journal of Human Genetics, 66*, 1933–1944.

Bateson, W. (1909). *Mendel's principles of heredity*. Cambridge: Cambridge University Press.

Boomsma, D. I. (1998). Genetic analysis of cognitive failures (CFQ); A study of Dutch adolescent twins and their parents. *European Journal of Personality, 12*, 321–330.

Boomsma, D. I., Busjahn, A., & Peltonen, L. (2002). The classical twin study and beyond. *Nature Reviews Genetics*, *3,* 872–882.

Boomsma, D. I., & Dolan, C. V. (2000). Multivariate QTL analysis using structural equation modeling: A look at power under simple conditions. In T. Spector, H. Snieder, & A. MacGregor (Eds.), *Advances in Twin & Sib-Pair Analysis* (pp. 203–218). London: Greenwich Medical Media Ltd.

Boomsma, D. I., Geus, E. J. C. de, Baal, G. C. M. van, & Koopmans, J. R. (1999). Religious upbringing reduces the influence of genetic factors on disinhibition: Evidence for interaction between genotype and environment. *Twin Research, 2*, 115–125.

Boomsma, D. I., & Martin, N. G. (2002). Gene–environment interactions. In H. D'haenen, J. A. den Boer, & P. Wilner (Eds.) *Biological psychiatry* (pp. 181–187). London: John Wiley & Sons Ltd.

Boomsma, D. I., & Molenaar, P. C. M. (1986). Using LISREL to analyze genetic and environmental covariance structure. *Behavior Genetics, 16*, 237–250.

Boomsma, D. I., & Molenaar, P. C. M. (1987a). Constrained maximum likelihood analysis of familial resemblance of twins and their parents. *Acta Geneticae Medicae et Gemellologiae, 36*, 29–39.

Boomsma, D. I., & Molenaar, P. C. M. (1987b). The genetic analysis of repeated measures I: Simplex models. *Behavior Genetics, 17*, 111–123.

Breslau, N., & Rasmussen, B. K. (2001). The impact of migraine: Epidemiology, risk factors, and co-morbidities. *Neurology, 56*, S4–12.

Cardon, L. R., & Abecasis, G. R. (2000). Some properties of a variance components model for fine-mapping quantitative trait loci. *Behavior Genetics, 30*, 235–243.

Cardon, L. R., & Bell, J. I. (2001). Association study designs for complex diseases. *Nature Reviews Genetics*, *2*(2), 91–99.

Cardon, L. R., & Fulker, D. W. (1994). The power of interval mapping of quantitative trait loci, using selected sib pairs. *American Journal of Human Genetics, 55*, 825–833.

Carey, G. (1988). Inference about genetic correlations. *Behavior Genetics*, *18*, 329–338.

Carey, G. (2002). *Human genetics for the social sciences*. Thousand Oaks, CA: Sage Publications.

Carey, G., & Williamson, J. (1991). Linkage analysis of quantitative traits: Increased power by using selected samples. *American Journal of Human Genetics, 49*, 786–796.

Cavalli-Sforza, L. L., Menozzi , P., & Piazza, A. (1994). *The history and geography of human genes*. Princeton, New Jersey: Princeton University Press.

Dick, D. M., Rose, R. J., Viken, R. J., Kaprio, J., & Koskenvuo, M. (2001). Exploring gene–environment interactions: Socioregional moderation of alcohol use. *Journal of Abnormal Psychology*, *110*, 625–632.

Dolan, C. V. (1992). *Biometric decomposition of phenotypic means in human samples*. Unpublished doctoral dissertation, University of Amsterdam, The Netherlands.

Dolan, C. V., & Boomsma, D. I. (1998). Optimal selection of sib pairs from random samples for linkage analysis of a QTL using the EDAC test. *Behavior Genetics*, *28*, 197–206.

Dolan, C. V., Boomsma, D. I., & Neale, M. C. (1999a). A simulation study of the effects of assigning prior IBD probabilities to unselected sib-pairs in covariance structure modeling of a QTL test. *American Journal of Human Genetics, 64,* 268–280.

Dolan, C. V., Boomsma, D. I., & Neale, M. C. (1999b). A note on the power provided by sibships of sizes 2, 3, and 4 in genetic covariance modeling of a codominant QTL. *Behavior Genetics*, *29*, 163–170.

Dolan, C. V., Molenaar, P. C. M., & Boomsma, D. I. (1991). Simultaneous genetic analysis of longitudinal means and covariance structure in the simplex model using twin data. *Behavior Genetics, 21*, 49–65.

Drigalenko, E. (1998). How sib pairs reveal linkage. *American Journal of Human Genetics*, *63*, 1242–1245.

Eaves, L. J. (1969). The genetic analysis of continuous variation: A comparison of experimental designs applicable to human data. *The British Journal of Mathematical and Statistical Psychology*, *22*, 131–147.

Eaves, L. J. (1984). The resolution of genotype x environment interaction in segregation analysis of nuclear families. *Genetic Epidemiology*, *1*, 215–228.

Eaves, L. J., & Gale J.S. (1974). A method for analyzing the genetic basis of covariation. *Behavior Genetics*, *4*, 253–267.

Eaves, L. J., Fulker, D. W., & Heath, A. C. (1989). The effects of social homogamy and cultural inheritance on the covariances of twins and their parents: A LISREL model. *Behavior Genetics, 19*, 113–122.

Eaves, L. J., Heath, A. C., & Martin, N. G. (1984). A note on the generalized effects of assortative mating. *Behavior Genetics, 14*, 371–376.

Eaves, L. J., Last, K., Martin, N. G., & Jinks, J. L. (1977). A progressive approach to non-additivity and genotype–environmental covariance in the analysis of human differences. *The British Journal of Mathematical and Statistical Psychology, 30*, 1–42.

Eaves, L. J., & Meyer, J. (1994). Locating human quantitative trait loci: Guidelines for the selection of sibling pairs for genotyping. *Behavior Genetics, 24*, 443–455.

Eerdewegh, P., van, Little, R. D., Dupuis, J., Del Mastro, R. G., Falls, K., Simon, J., et al. (2002). Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature, 418*, 426–430.

Falconer, D. S. (1952). The problem of environment and selection. *The American Naturalist*, *86,* 293–298.

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics* (4th ed.). London: Longan Group Ltd.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52,* 399–433.

Forrest, W. F. (2001). Weighting improves the 'new Haseman-Elston' method. *Human Heredity*, *52*, 47–54.

Fulker, D. W. (1988). *Genetic and cultural transmission in human behavior*. In B. S. Weir, E. J. Eisen, M. M. Goodman, & G. Namkoong (Eds.), Proceedings of the second international conference on quantitative genetics. Sinauer Associates Inc. Sunderland, USA.

Fulker, D. W., Cherny, S. S., Sham, P. C., & Hewitt, J. K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics*, *64*, 259–267.

Goate, A., Chartier-Harlin, M. C., Mullan, M., Brown, J., Crawford, F., Fidani, L., et al. (1991). Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature, 349*, 704–706.

Gu, C., Todorov, A., & Rao, D. C. (1996). Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci. *Genetic Epidemiology, 13*, 513–533.

Hakonarson, H., Bjornsdottir, U. S., Halapi, E., Palsson, S., Adalsteinsdottir, E., & Gislason, D. (2002). Gene for asthma maps to chromosome 14q24. *American Journal of Human Genetics, 7,* 483–491.

Hamer, D., & Sirota, L. (2000). Beware the chopsticks gene. *Molecular Psychiatry, 5*, 11–13.

Harville, D. (1976). Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects. *The Annals of Statistics, 4*, 384–395.

Haseman, J. K., & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, *2*, 3–19.

Heath, A. C., Eaves, L. J., Nance, W. E., & Corey, L. A. (1987). Social inequality and assortative mating: Cause or consequence? *Behavior Genetics, 17*, 9–17.

Horikawa, Y., Oda, N., Cox, N. J., Li, X., Orho-Melander, M., Hara, M., et al. (2000). Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genetics, 26,* 163–175.

Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J. P., Belaiche, J., et al. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature, 411*, 599–603

Jinks, J. L., & Fulker, D. W. (1970). A comparison of the biometrical-genetical, MAVA and classical approaches to the analysis of human behavior. *Psychological Bulletin*, *73*, 311–349.

Jöreskog, K., & Sörbom, D. (1986). *LISREL: Analysis of linear structural relationships by the method of maximum likelihood*. Chicago: National Education Resources.

Kendler, K. S., & Eaves, L. J. (1986). Models for the joint effects of genotype and environment on liability to psychiatric illness. *American Journal of Psychiatry*, *143*, 279–289.

Korstanje, R., & Paigen, B. (2002). From QTL to gene: The harvest begins. *Nature Genetics, 31*, 235-236.

Kirkpatrick, M., & Heckman, N. (1989). A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of Mathematical Biology,, 27*, 429–450.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., & Lander, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics, 58*, 1347–1363.

Lackner, C., Boerwinkle, E., Leffert, C. C., Rahmig, T., & Hobbs, H. H. (1991). Molecular basis of apolipoprotein(a) isoform size heterogeneity as revealed by pulsed-field gel electrophoresis. *Journal of Clinical Investigation*, 2153–2161.

Lesch, K. P. (2003). Neuroticism and serotonin: A developmental genetic perspective. In R. Plomin, J. C. DeFries, I. W. Craig, & P. McGuffin (Eds.), *Behavioral genetics in the postgenomic era* (pp. 389–423). Washington, DC: American Psychological Association.

Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sunderland: Sinauer Associates.

Martin, N. G., Boomsma, D. I., & Neale, M. C. (1989). Genetic analysis of twin and family data: Structural modeling using LISREL. *Behavior Genetics, 19*, 5–7.

Martin, N. G., Boomsma, D. I., & Machin, G. (1997). A twin pronged attack on complex traits. *Nature Genetics*, *17*, 387–392.

Martin, N. G., & Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity*, *38*, 79–95.

Mather, K. (1949). *Biometrical genetics*. London: Methuen.

Mather, K., & Jinks, J. L. (1982). *Biometrical Genetics*. New York, NY: Chapman & Hall.

McArdle, J. J. (1986) Latent variable growth within behavior genetic models. *Behavior Genetics, 16*, 163–200.

McKenzie, C. A., Abecasis, G. R., Keavney, B., Forrester, T., Ratcliffe, P. J., Julier, C., Connell, J. M. C., Bennett, F., McFarlane-Anderson, N., Lathrop, G. M., & Cardon, L. R. (2001). Trans-ethnic fine mapping of a quantitative trait locus for circulating angiotensin I-converting enzyme (ACE). *Human Molecular Genetics, 10*, 1077–1084.

Molenaar, P. C. M., & Boomsma, D. I. (1987). Application of nonlinear factor analysis to genotype-environment interaction. *Behavior Genetics, 17*, 71–80.

Molenaar, P. C. M., Boomsma, D. I., & Dolan, C. V. (1999). The detection of genotype-environment interaction in longitudinal genetic models. In M. LaBuda, & E. Grigorenko (Eds.), *On the way to individuality: Current methodological issues in behavior genetics* (pp. 53–70). London: Nova Science Publishers Inc,

Molenaar, P. C. M., Boomsma, D. I., Neeleman, D., & Dolan, C. V. (1990). Using factor scores to detect GxE origin of "pure" genetic or environmental factors obtained in genetic covariance structure analysis. *Genetic Epidemiology, 7*, 93–100.

Moxley, G., Posthuma, D., Carlson, P., Estrada, E., Han, J., Benson, L. L., & Neale, M. C. (2002). Sexual dimorphism in innate immunity. *Arthritis & Rheumatism*, *46*, 250–258.

Neale, M. C. (1997). *Mx: Statistical modeling* (3rd ed.). Box 980126 MCV, Richmond VA 23298.

Neale, M. C. (2000). The use of Mx for association and linkage analysis. *GeneScreen*, *1*, 107–111.

Neale, M. C, & Cardon, L. R. (1992). *Methodology for genetic studies of twins and families*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Neale, M. C., Cherny, S. S., Sham, P. C., Whitfield, J. B., Heath, A. C., Birley, A. J., et al. (1999). Distinguishing population stratification from genuine allelic effects with Mx: Association of ADH2 with alcohol consumption. *Behavior Genetics*, *29*, 233–243.

Neale, M. C., Eaves, L. J., & Kendler, K. S. (1994). The power of the classical twin study to resolve variation in threshold traits. *Behavior Genetics*, *24*, 239–258.

Neale, M. C., & McArdle, J. J. (2000). Structured latent growth curves for twin data. *Twin Research, 3*,165–177.

Ott, J. (1999). *Analysis of human genetic linkage* (3rd ed.). Baltimore and London: The Johns Hopkins University Press.

Pletcher, S. D., & Geyer, C.J. (1999). The genetic analysis of age-dependent traits: Modelling the character process. *Genetics, 153*, 825–835.

Plomin, R., DeFries, J. C., & Loehlin, J. C. (1977). Genotype-environment interaction and correlation in the analysis of human behavior. *Psychological Bulletin, 84*, 309–322.

Posthuma, D., & Boomsma, D. I. (2000). A note on the statistical power in extended twin designs. *Behavior Genetics*, *30*, 147–158.

Posthuma, D., Geus, E. J. C. de, Boomsma, D. I., & Neale, M. C. (2004). Combined linkage and association tests in Mx. *Behavior Genetics, 34* (in press).

Pritchard, J. K., Stephens, M., Rosenberg, N. A., & Donnelly, P. (2000). Association mapping in structured populations. *American Journal of Human Genetics, 67*, 170-181.

Risch, N., & Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science, 268*, 1584–1589.

Risch, N. J., & Zhang, H. (1996). Mapping quantitative trait loci with extreme discordant sib pairs: Sampling considerations. *American Journal of Human Genetics, 58*, 836–843.

Rose, R. J., Dick, D. M., Viken, R. J., & Kaprio, J. (2001). Gene–environment interaction in patterns of adolescent drinking: Regional residency moderates longitudinal influences on alcohol use. *Alcoholism, Clinical and Experimental Research*, *25*, 637–643.

Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype-environment effects. *Child Development*, *54*(2), 424–435.

Sham, P. C. (1998). *Statistics in human genetics*. London: Arnold Publishers.

Sham, P. C., & Purcell, S. (2001). Equivalence between Haseman-Elston and variance–components linkage analyses for sib pairs. *American Journal of Human Genetics, 68*, 1527–1532.

Sham, P. C., Purcell, S., Cherny, S. S., & Abecasis, G. R. (2002). Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *American Journal of Human Genetics, 71*, 238–253.

Stefansson, H., Sigurdsson, E., Steinthorsdottir, V., Bjornsdottir, S., Sigmundsson, T., Ghosh, S., et al. (2002). Neuregulin 1 and susceptibility to schizophrenia. *American Journal of Human Genetics, 71*, 877–892.

Straub, R. E., Jiang, Y., MacLean, C. J., Ma, Y., Webb, B. T., Myakishev, M. V., et al. (2002). Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *American Journal of Human Genetics*, *71*, 337–348.

Terwilliger, J. D., & Göring, H. H. H. (2000). Gene mapping in the 20th and 21st centuries: Statistical methods, data analysis, and experimental design. *Human Biology*, *72*, 63–132.

Vandenberg, S. G. (1972). Assortative mating, or who marries whom? *Behavior Genetics*, *2,* 127.

Vandenberg, S. G., & Falkner, F. (1965) Hereditary factors in human growth. *Human Biology, 37*, 357–365.

Vink, J. M., & Boomsma, D. I. (2002). Gene finding strategies. *Biological Psychology, 61,* 53–71.

Visscher, P. M., & Hopper, J. L. (2001). Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data. *Annals of Human Genetics*, *65*, 583–601.

Willemsen, G., Vink, J. M., & Boomsma, D. I. (2003). Assortative mating may explain spouses' risk of same disease. (Letter). *British Medical Journal, 326*, 396.

Wright, S. (1921). Correlation and causation. Part 1:Method of path coefficients. *Journal of Agricultural Research*, *20*, 557–585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, *5*, 161–215

Wright, F. A. (1997) The phenotypic difference discards sib-pair QTL linkage information. *American Journal of Human Genetics, 60*(3), 740–742.

Xu, X., Weiss, S., Xu, X., & Wei, L. J. (2000). A unified Haseman-Elston method for testing linkage with quantitative traits. *American Journal of Human Genetics, 67*, 1025–1028.

Zhao, H. (2000). Family based association studies. *Statistical Methods in Medical Research*, *9*, 563–587.

Zhu, G., Duffy, D. L., Eldridge, A., Grace, M., Mayne, C., O'Gorman, L., et al. (1999). A major quantitative-trait locus for mole density is linked to the familial melanoma gene CDKN2A: A maximum-likelihood combined linkage and association analysis in twins and their sibs. *American Journal of Human Genetics*, *65*, 483–492