Biometrical genetics

David Duffy

Queensland Institute of Medical Research
Brisbane, Australia

Biometrical Genetics

Biometrical genetics refers to a set of mathematical models used to describe the inheritance of quantitative traits.

A quantitative trait is a characteristic of an organism that can be measured, giving rise to a numerical value. It can be:

Continuous: eg arterial blood pressure, stature
Meristic: *a count* eg moles (nevi), bristles, digits, worm burden
Ordinal: *a ranking* eg Fitzgerald tanning index, Norwood baldness score
Categorical: eg eye colour, type of cancer

Genotype-phenotype relationship for quantitative traits

We will represent the relationship between genotype and phenotype as a linear model:

$$Y = G + E$$

Y is the trait value for an individual,

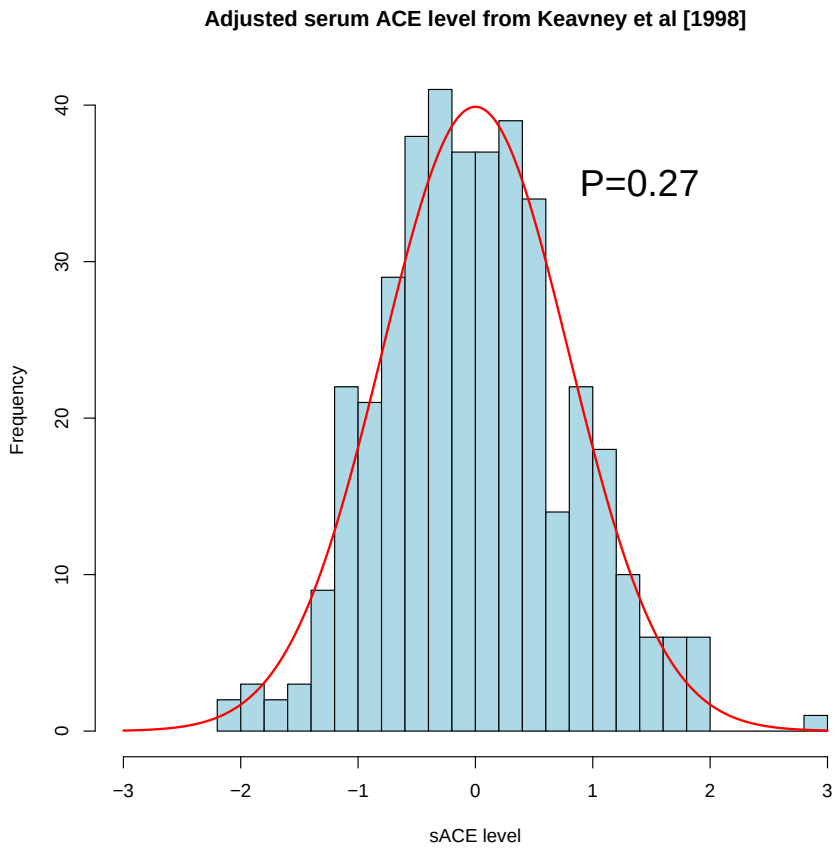
G is the effect of the individual's genotype at the **quantitative trait locus** (**QTL**), which can be one of g different values, where there are g possible genotypes,

E is the combined effect of all nongenetic factors acting on the phenotype in that individual.

Environmental Effect (E)

E is the usual "error" that appears in statistical models, and is a **random variable**, which we will treat as coming from a standard statistical distribution such as the **Gaussian** (**Normal**) distribution.

The ***E*** for the *i*th individual in a family is modelled as being a random sample from such a distribution.



QIMR

Genotype Effect (G)

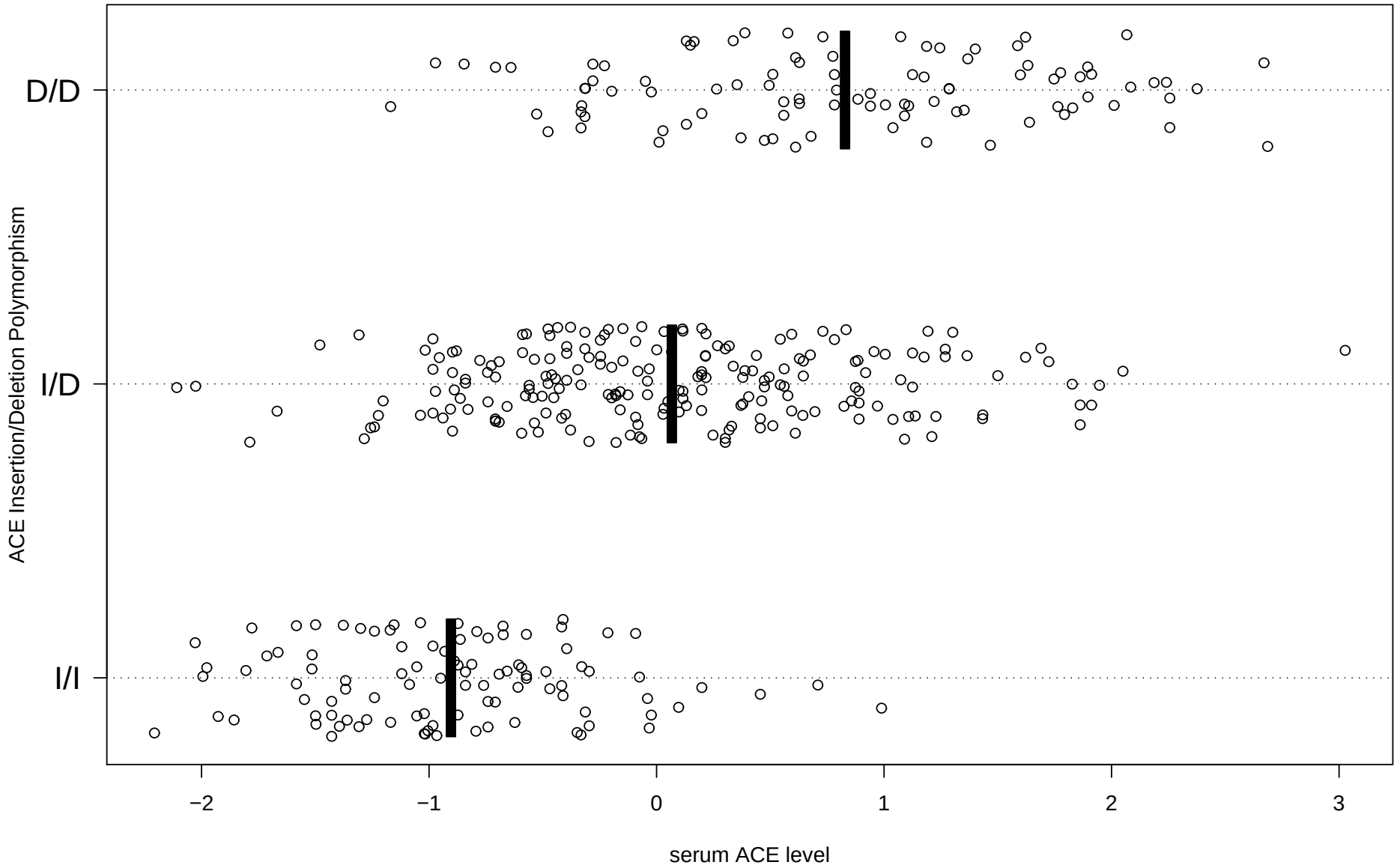
The genotypic effect is **fixed**, that is every person carrying the same genotype has the same value of ***G***.

For a diallelic autosomal gene, for example, there will be 3 **genotypic means**, which we will usually denote μ_0, μ_1 and μ_2 for the *A/A*, *A/B*, *B/B* genotypes respectively.

If we know or have estimated the value of ***G***, then we can calculate the value of ***E*** for the *i*th person, who carries genotype *j* as:

$$E_i = \mu_j - Y_i$$

ACE Indel genotype v. sACE level [Keavney et al 1998]



QIMR

Population genetics of a quantitative trait locus

The results to date apply to individuals. Unless the QTL is monomorphic, a natural population will be a mixture of genotypes, usually in Hardy-Weinberg proportions.

A/A	*A/B*	*B/B*
p^2	$2pq$	q^2
μ_0	μ_1	μ_2

The distribution of the trait values will be determined by genotype frequencies and means. It is straightforward to calculate the mean and variance of the population distribution due to the QTL.

Mean and variances of a quantitative trait

The overall **population mean** will be a weighted average of the genotypic means:

$$\mu = p^2\mu_0 + 2pq\mu_1 + q^2\mu_2$$

where p is the frequency of the A allele (q=1-p).

The total **phenotypic variance** (which I will write σ^2_T or V_T) is calculated as:

$$\sigma^2_T = \Sigma(Y_i - \mu)^2$$

The **genetic variance** (σ^2_G or V_G) is the amount of variation in the population around this global mean that is due to differences between individuals in genotype:

$$\sigma^2_G = p^2(\mu_0 - \mu)^2 + 2pq(\mu_1 - \mu)^2 + q^2(\mu_2 - \mu)^2$$

Variance Components

We started with a model for each individual:

$$Y_i = G_i + E_i$$

And can now write an equivalent equation for the phenotype variance

$$V_T = V_G + V_E$$

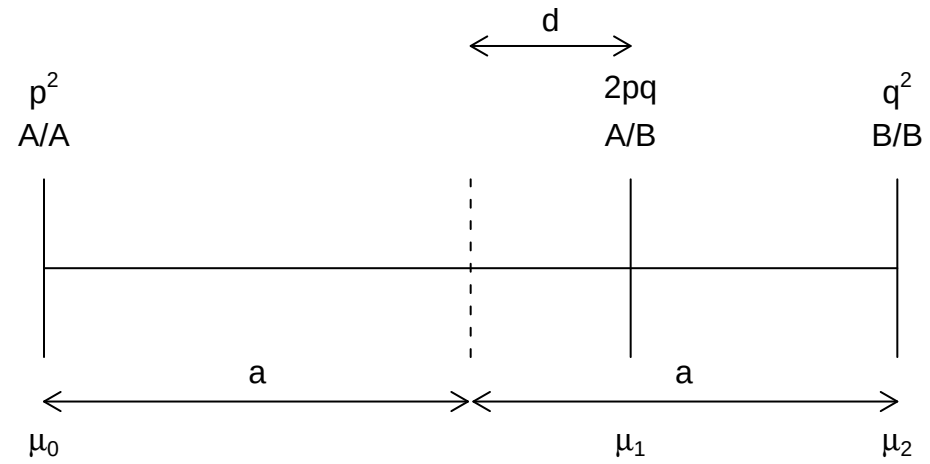
where $\boldsymbol{V_E}$ is the **environmental variance** (or environmental noise).

The **broad sense heritability** is a measure of the relative importance of the QTL:

$$h_B^2 = \frac{V_G}{V_T}$$

Allelic Effects

Because each parent only transmits one allele to offspring, it is useful to further **decompose** the genotypic means into **allelic** and **dominance** effects:



If d=0, then there is a simple linear relationship between number of the B alleles in the genotype (the **gene content**) and phenotype.

Additive and Dominance Variances

The decomposition of the genetic variance into **additive** and **dominance** variances is slightly more complex, because the **average effect** of an allele selected at random from the population is averaged over the other possible alleles of the genotype (weighted by the allele frequencies).

$$V_A = 2pq[(p-q)d + a]^2$$
$$= 2pq[p(\mu_0 - \mu_1) + q(\mu_1 - \mu_2)]^2$$

$$V_D = 4p^2q^2d^2$$
$$= p^2q^2[\mu_2 - 2\mu_1 + \mu_0]^2$$

Covariance between relatives

These results so far assume a sample of unrelated individuals.

Resemblances between particular classes of relatives on continuous traits are usually expressed as covariances between the measured values of the trait, and by various extensions of this such as **interclass** and **intraclass** correlation coefficients.

Intraclass and interclass correlations arise naturally from analysis of variance, and are very appropriate for genetic usage when there are no reasons to differentiate *within* a group of relatives eg a sibship.

Intraclass and interclass correlations

These correlations can be defined for a population containing **p** classes (eg sibships and sets of parents), with containing $\mathbf{k_p}$ members in each class on which $\mathbf{Y_{ij}}$ is the trait value for the **j**th member of the **i**th class.

$$
\begin{aligned}
E(Y_{ij}) &= \mu \\
Var(Y_{ij}) &= V_T \\
Cov_I(Y_{ij}, Y_{i'j'}) &= r_I V_T \qquad i = i', j \neq j' \\
&= 0 \qquad\quad i \neq i' \\
Cov_B(Y_{ij}, Y_{i'j'}) &= r_{ii'} V_T \qquad i = i', j \neq j' \\
&= 0 \qquad\quad i = i'
\end{aligned}
$$

$\boldsymbol{r_I}$ is the intraclass correlation and

$\boldsymbol{r_{ii'}}$ denotes the interclass correlation between the **i**th and **i′**th group.

Genetic covariance between unilineal relatives

Parents and offspring, grandparents and grandchildren etc share at most one allele in common (in the absence of inbreeding), and so are **unilineal** relatives.

Therefore, the correlation between trait values in such pairs of relatives (or the corresponding interclass correlation) represents the average effect of transmission or nontransmission of one QTL allele across all the pairs.

We do not specify the particular QTL allele is being shared – to predict the correlation, we merely need the *transmission* probability. This probability is a **kinship coefficient**.

For example, one of the two parental alleles has a 50% probability of being transmitted to a child.

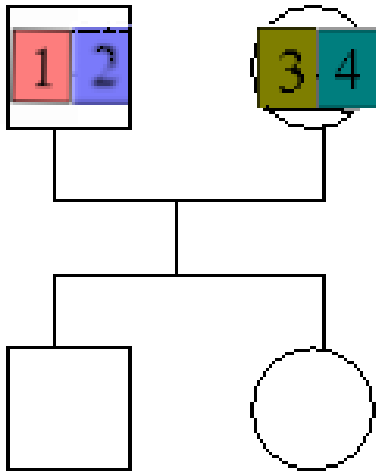
Expected genetic covariance between unilineal relatives

Relationship	Intervening meioses	Covariance	Correlation
Parent-offspring	1	$\frac{1}{2}V_A$	$\frac{1}{2}\frac{V_A}{V_T}$
Half-siblings	1	$\frac{1}{2}V_A$	$\frac{1}{2}\frac{V_A}{V_T}$
Grandparent-grandchild	2	$\frac{1}{4}V_A$	$\frac{1}{4}\frac{V_A}{V_T}$
Avuncular	2	$\frac{1}{4}V_A$	$\frac{1}{4}\frac{V_A}{V_T}$
Cousins	3	$\frac{1}{8}V_A$	$\frac{1}{8}\frac{V_A}{V_T}$

Genetic covariance between siblings

Since siblings share two parents, they are **bilineally** related, and can carry zero, one or two QTL alleles in common. This this means that the dominance variance will contribute to similarity of sibling trait values in a proportion of the population of families.

	1-3	1-4	2-3	2-4
1-3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
1-4	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
2-3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
2-4	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$



50% of sib pairs share 1 QTL allele in common and 25% share 2 QTL alleles.

Expected genetic covariance for siblings

Relationship	Covariance	Correlation
Full sibs	$\frac{1}{2}V_A + \frac{1}{4}V_D$	$\frac{1}{2}\frac{V_A}{V_T} + \frac{1}{4}\frac{V_D}{V_T}$
MZ Twins	$V_A + V_D$	$\frac{V_A}{V_T} + \frac{V_D}{V_T}$
Any	$RV_A + KV_D$	$R\frac{V_A}{V_T} + K\frac{V_D}{V_T}$

where ***R*** and ***K*** are **kinship coefficients**:

R is the **coefficient of relationship** (probability two individuals share an allele inherited from the same ancestor.

K is the **coefficient of fraternity** (probability two individuals share two alleles inherited from the same ancestors.


QIMR

Multiple QTLs

So far, we have dealt with the familial correlations arising from a single QTL.

These models can be extended to include multiple QTLs acting on the same trait. Just as the dominance variance arises from the interaction of the two alleles within a genotype at one QTL, **epistatic variance** arises from the interaction of alleles at different QTLs.

$$\begin{aligned} V_G &= V_A + V_D + V_{AA} + V_{AD} + V_{DD} \cdots \\ &= \sum_{r=1}^{n} \sum_{s}^{r+s>0} V_{r*As*D} \end{aligned}$$

and the covariance between pairs of relatives is,

$$\begin{aligned} Cov(Y_1, Y_2) &= RV_A + KV_D + R^2V_{AA} + RKV_{AD} + K^2V_{DD} \cdots \\ &= \sum_{r=1}^{n} \sum_{s}^{r+s>0} R^r K^s V_{r*As*D} \end{aligned}$$


QIMR

The polygenic model

If the individual contribution of any one QTL is small, and many QTLs are acting, then it is plausible to assume that the epistatic variance is also small.

In the infinitesimal **polygenic** model, the individual additive genetic effects of all the QTLs sum together to give the total genetic variance of the trait. This gives a justification for applying all the theoretical results we have reviewed regardless of the number of segregating QTLs.

In the absence of genotype data, it is usually not possible to determine whether a trait is under the control of one or many QTLs.

Estimating variance components

We can use observed familial correlations, therefore, to estimate the values of the different variance components whether due to a single QTL, or under certain assumptions, multiple QTLs.

Optimally, this is done by maximum likelihood, combining data from all the available different relationships, but simple algebraic estimates are useful and not too inaccurate. For example:

$$\hat{V}_A = 2r_{po}V_T$$
$$\hat{V}_D = 4(r_{sib} - r_{po})V_T$$

with r_{po} the parent offspring correlation, and
with r_{sib} the sibling correlation.

Variance components linkage analysis

To model familial correlations in the absence of information about the actual QTL genotypes, we combine data from (ideally) a large number of different types of relative pair. We use averages (expectations), including *expected* kinship coefficients.

If we have marker information, we can estimate **empirical kinship coefficients** for particular regions of the genome. This is often referred to as **identity by descent** information (*ibd*), since it allows us to infer if marker alleles in two related individuals are in fact identical copies of an allele descended from a recent common ancestor.

If a QTL affecting our trait of interest is within a region we have marker-derived *ibd* information, we can estimate the genetic variance **specific** to that QTL.

Utilizing ibd information for linkage analysis

Identity by descent	Equivalent Relationship	Covariance	Correlation
Two alleles shared IBD	MZ Twins	$V_A + V_D$	$\frac{V_A}{V_T} + \frac{V_D}{V_T}$
One allele shared IBD	Parent-offspring	$\frac{1}{2}V_A$	$\frac{1}{2}.\frac{V_A}{V_T}$
Zero alleles shared IBD	Unrelated	0	0

Maximum likelihood VC linkage analysis

To efficiently combine information from different types of relative pair, we fit an extended version of the usual biometrical model:

$$Cov(Y_i, Y_j) = \frac{(ibd)}{2} V_Q + I(ibd = 2) V_{QD} + R_{ij} V_A + K_{ij} V_D$$

where $(ibd) = 0, 1, 2$ gives the empirical kinship coefficients, and R_{ij} and K_{ij} are the expected kinship coefficients for the ijth relative pair.

Usually we further simplify this model by assuming $V_{QD} = 0$. The test for linkage (the Likelihood Ratio Test Statistic) is constructed by comparing the model likelihood when V_Q is estimated to that when V_Q is fixed to zero. This gives a *lod score* just as other types of maximum likelihood linkage analysis do.

Types of relative pair useful for VC linkage analysis

There are two types of relative pair where the empirical kinship coefficient always equals the theoretical expected kinship coefficient:

- Monozygotic twins
- Parent-offspring pairs

This type of pair therefore does not contribute any linkage information. If measuring a trait is expensive, then it is reasonable to not phenotype parents.