Chapter 3

DNA tests and segregation analysis for genetic disorders

Brian Kinghorn
Twynam Chair of Animal Breeding Technologies
University of New England

Segregation analysis by example

Figure 1 shows a simple pedigree in which all females have been genotyped for a 2-allele locus (A,a). Yellow (light colored) lines emanate from females, and red (dark colored) lines from males. Can we deduce the genotypes of the males?

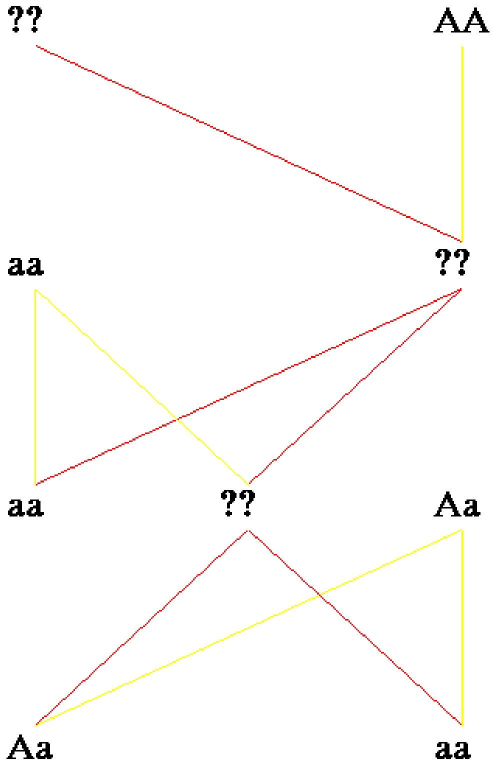


Figure 1. A simple pedigree in which females have been genotyped.

- The male in the second row is quite easy. Can you explain why he must be a heterozygote?

- The male in the first row is more difficult, and the answer depends on our prior assumption about gene frequency. This is taken as *p(A)* =.6 in Table 1.

- The male in the third row is even more difficult, but the answer in this case does not depend on gene frequency. Can you explain briefly why this is the case?

Table 1. Genotype probabilities for the males in Figure 1.

Male in row	p(aa)	p(Aa)	a(AA)
1	0.4	0.6	0
2	0	1	0
3	2/3	1/3	0

We might be able to deduce some results for simple cases like this, by relatively simple logic. But what about bigger examples in more realistic, complex pedigrees?

If we have 20 ungenotyped animals we have up to 3 to the power 20 (=3,486,784,401) 'possible' answers for the 2-allele, 3-genotype case. This makes solution by simple searching methods not feasible.

Most methods for doing this sort of thing make use of information from three sources:

1. Parents
2. Self
3. Mate(s) plus progeny,

... and use that information either recursively or iteratively over a number of cycles. Care has to be taken not to overuse information (double dipping) or mistreat loops in the pedigree (eg. as with inbreeding). At UNE, we currently use Richard Kerr's iterative method (Kerr and Kinghorn, 1996), coded in the program GENEPROB.

Example - the 'halothane gene' in pigs:

The halothane-sensitive variant of this gene causes porcine stress syndrome (PSS) and pale soft exudative (PSE) meat, as well as malignant hyperthermia on exposure to the gas halothane. It also gives higher lean percent. Breeding companies want a handle on this gene, but the DNA test is expensive.

In the example given here, the data set contains 4207 pigs in a complex pedigree structure. Money was spent genotyping 113 of these animals by DNA test at the Ryanodine receptor locus (the 'halothane' locus). Of these, 65 were normal homozygotes, 40 were heterozygotes and 8 carried two copies of the unfavourable allele.

After segregation analysis, an additional 1886 animals could be excluded form one genotype class and an additional 42 animals could be genotyped, both with 100 percent confidence. At the 90% confidence level (ie. probability of being any given genotype > 0.9) an additional 263 animals could be genotyped.

All animals had probabilities calculated for the halothane gene, and this information gave a basis for more accurate estimation of the effects of the halothane gene, using data from all pigs in the data set - not just those that were DNA tested.

The pig pedigree is shown in Figure 2. Can you deduce the genotypes of any ungenotyped animals?

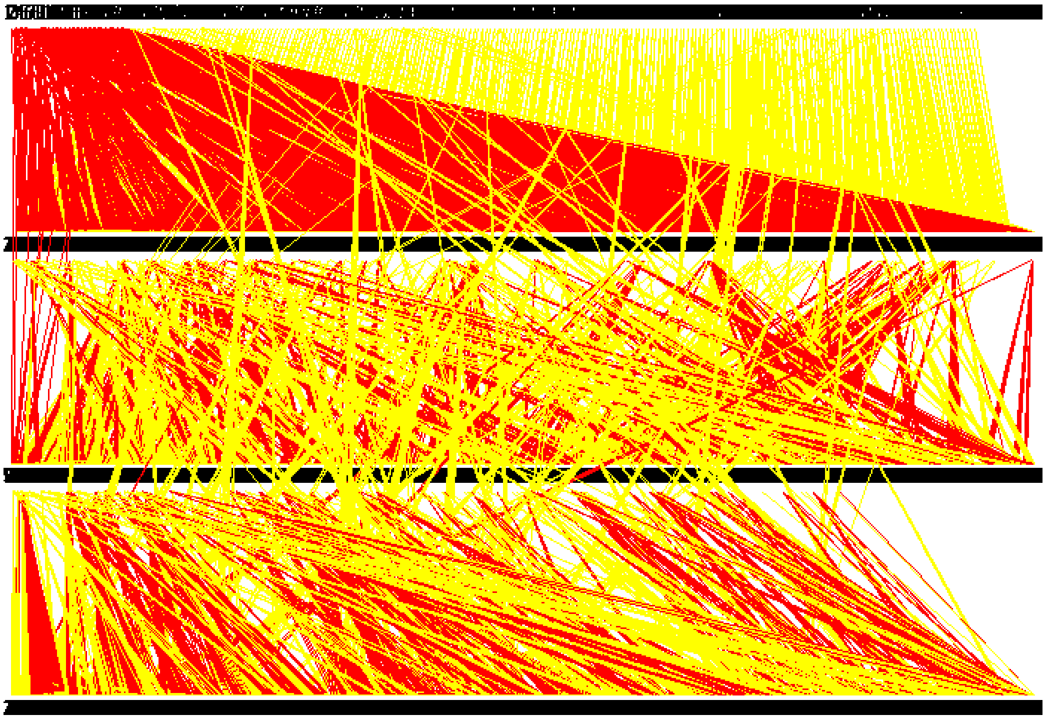


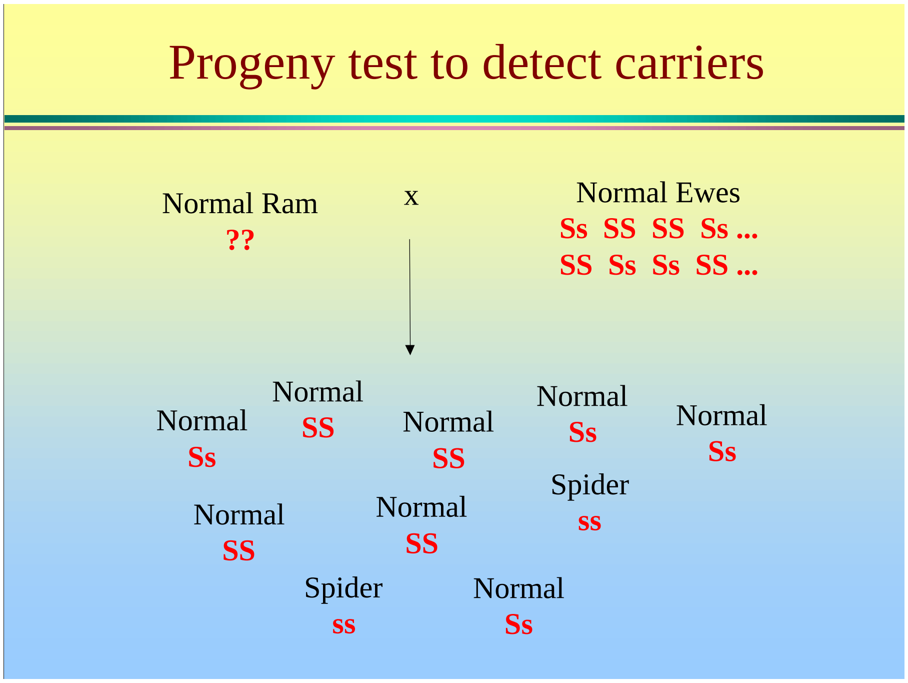
Figure 2. A pedigree diagram of the pig data set.

Of course not! You need a computer program to do the job for you. Richard Kerr and Brian Kinghorn have written 'GENEPROB' for this. It works well and fast on large data sets, but does not fully account for inbreeding loops. Output is genotype probabilities (probabilities of being *AA*, *Aa* and *aa*, summing to 1) for each individual.

Example - the spider syndrome in sheep:

The spider syndrome in Suffolk sheep is a recessive lethal condition. Development at the ends of the long bones is impaired and lambs end up on the ground with their legs splayed like a spider.

Classically, a progeny test is used to detect carriers, as illustrated in Figure 3. However, using segregation analysis, as in GENEPROB, information from all relatives can be used, such that some results are available at the time of making first breeding decisions - we do not really have to wait for the results from a progeny test.



Progeny test to detect carriers

Normal Ram
??

x

Normal Ewes
Ss SS SS Ss ...
SS Ss Ss SS ...

Normal
Ss

Normal
SS

Normal
SS

Normal
Ss

Normal
Ss

Normal
SS

Normal
SS

Spider
ss

Spider
ss

Normal
Ss

Figure 3. The classical progeny test approach to detecting carriers of deleterious recessive genes. The ram with a normal phenotype must be a carrier as he has some affected offspring.

In the example used here, a small data set containing 167 sheep included 6 spider lambs. These lambs were the progeny of just 3 rams and 3 ewes, which must be carriers. GENEPROB was run, and Figure 4 shows the resulting probabilities of being a carrier.

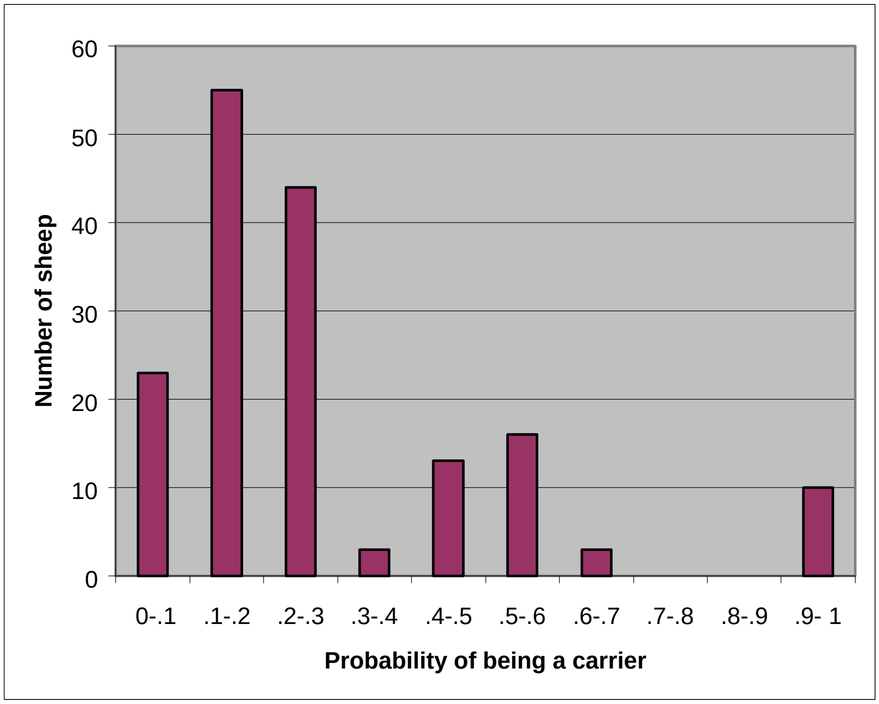


Figure 4. Frequency distribution of probability of being a spider carrier. This information comes from knowledge of the spider status of just six spider lambs.

These probabilities give breeders much more power to monitor and manipulate genes such as the spider gene. It is quite easy to include them in selection indices to provide appropriate penalty against animals probably carrying unwanted variants of such genes. TGRM (Chapter 20) can be used to help balance the speed of getting rid of such deleterious recessive genes [which is enhanced by mating to give higher incidence and thus more information in the medium term], and avoiding too much expression of the deleterious condition while doing so.

A DNA test for the spider gene has recently been developed, but this was not available for the current example. However, this example shows the power of segregation analysis in cases where such tests are not available, or too expensive.

Example – Achondroplasia in Dexter cattle:

In this example there were 8845 cattle in the dataset, of which 26 are recorded as carriers of the achonroplasia gene. This is a recessive gene, causing the lethal "bulldog" condition. A small number of known carriers gives a lot of information at the population level, as shown in figure 5. In this case, Genotype Probability Index (GPI) values are shown. These range from zero (no useful information) to 100 (full information), and are described later in this chapter.

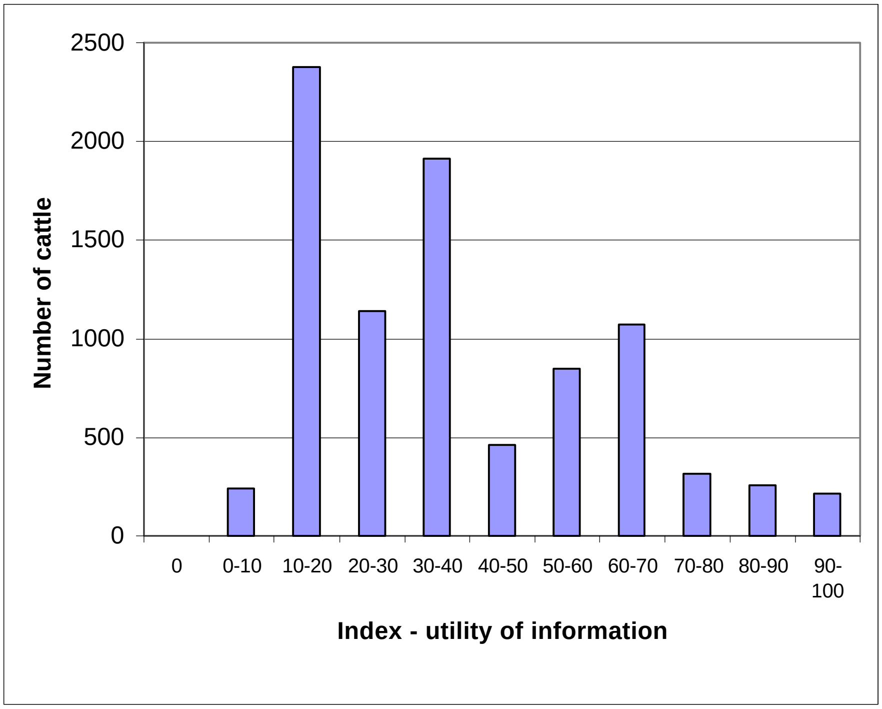


Figure 5. Frequency distribution of utility of genotype probabilities (Genotype Probability Index) for 8845 Dexter cattle. This information comes from knowledge of the carrier status of 26 bulls.

Using DNA test results

It can be noted that DNA test results can be used to drive segregation analysis, to get genotype probabilities for all untested animals. Moreover, DNA test results and incidence information can both be used together for maximum power.

The Genotype Probability Index (GPI)

As shown in figure 6, this index has a value of zero for an individual with no direct information – such that it has genotype probabilities equal to Hardy-Weinberg frequencies. Individuals that have been genotyped with full confidence have GPI values of 100 percent. This is also true for individuals that are confidently genotyped by inference from relatives – for example the known heterozygous status of progeny from parents of opposing homozygous genotypes, in the absence of mutation.

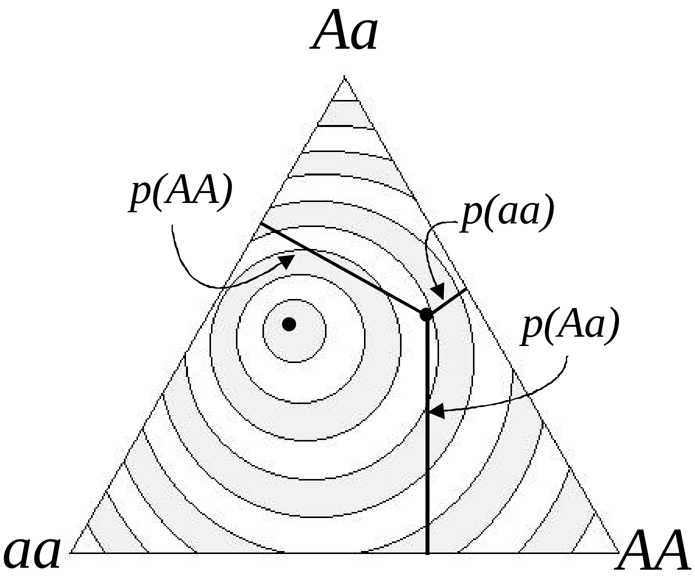


Figure 1. An illustration of the Genotype Probability Index, GPI (Kinghorn, 1997), for a biallelic locus (*A*, *a*) Probability *p(AA)* of being genotype *AA* is the perpendicular distance from the point to the edge opposite the *AA* vertex. Thus points on edges reflect zero probability for the genotype at the opposite vertex, and points at vertices represent confident genotyping for each of the three genotypes. The dot to the left represents H-W genotype frequencies for *p(A)* = 0.4. Individuals at this location have a GPI value of zero. Individuals at the vertices have GPI values of 100 percent. Individuals on an edge of the triangle generally have positive GPI values, reflecting the value of being able to exclude the possibility of one genotype. The individual plotted has genotype probabilities *p(aa)*=.1, *p(Aa)*=.5, *p(AA)*=.4 and has a GPI value of just under 40 percent, as can be seen from counting the contours.

Genotyping strategies

For conditions that have a DNA test available, there can still be some hard decisions to make because of the cost of testing. Strategies to decide on which animals to DNA test have been developed (Kinghorn, 1999). These involve an iterative approach, with batches of one or more animals DNA tested at each cycle. Segregation analysis is run between each cycle of DNA testing, with animals chosen for testing at each cycle using a criterion that gives most improvement in genotype probabilities across the whole population.

The GPI is in fact one such criterion – choosing animals with low GPI to genotype at each cycle. More useful criteria might also include one or more components, such as numerator relationship, parental status and estimated breeding value. The best criterion for the job depends on the task. For example, with genotype assisted selection there will be extra utility in gaining better information on individuals which are more likely to be selected – those with higher estimated breeding values – as genotyping them will contribute information to resulting descendants. In applications that aim to detect quantitative trait loci, the index could usefully favour genotyping individuals of extreme phenotype, as this can give more detection power.

The approach was tested on a simulated population with a single locus segregating. The objective was to gain information about a single known major gene for the live individuals in a pedigreed data set. After genotyping about 10%, 20% and 80% of the population, the resulting information contained about 50%, 60% and 100% of the utility (average genotype probability index) of full genotyping (figure 7).

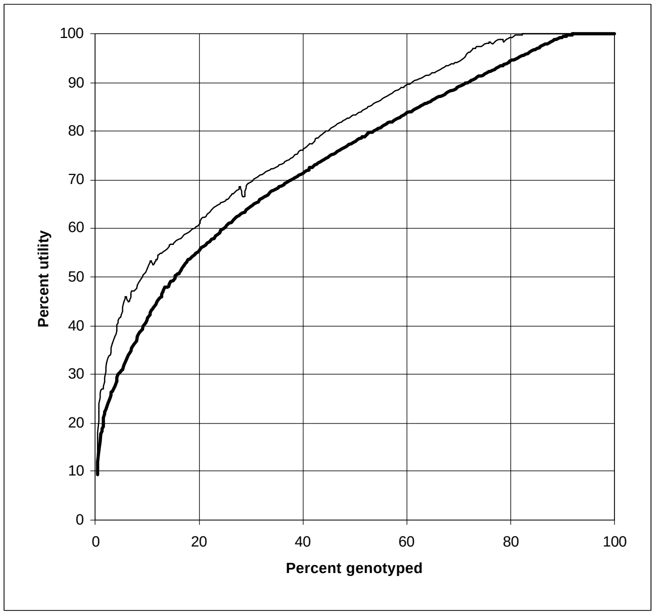
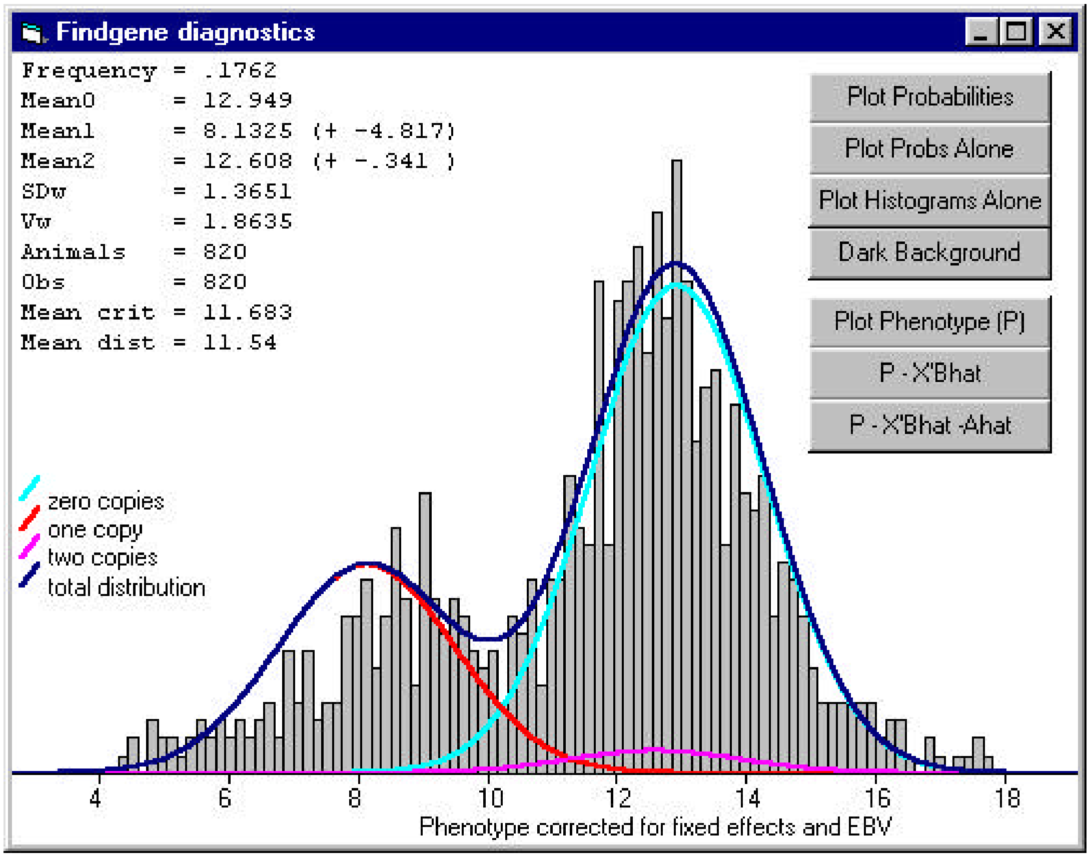


Figure 2. Plot of percent utility (population average GPI) versus percent of the population genotyped, with base population allele frequency 0.1. The upper curve is for use of the index 100*CON – GPI to rank and choose the individual to genotype at each iteration. The lower curve in bold is for random choice of individual to genotype at each iteration (average of 100 replicates, each with different random sampling).

Screening populations for evidence of segregating QTL

This is another application in which segregation analysis plays a big role.

Virtually all major genes currently used by animal breeders were first detected by noting strong familial trends during inspection of recorded data. Given the potentially large confounding influence of environment, polygenes, and segregation at the major locus, together with the large pedigreed data sets available today, such 'eyeballing' of data can be improved upon greatly by using computers to apply statistical techniques for major gene detection. The diagram below shows results from a screening test using Findgene software, described later on this page.



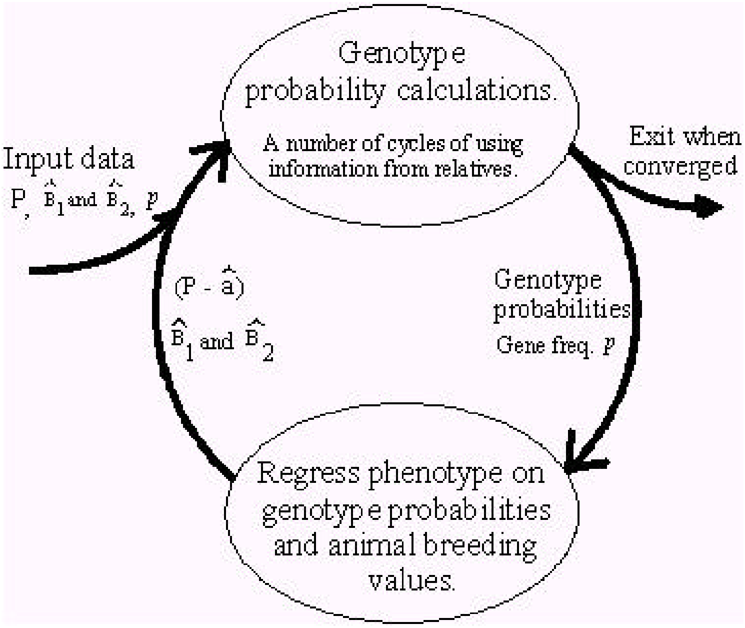
```
Frequency = .1762
Mean0     = 12.949
Mean1     = 8.1325 (+ -4.817)
Mean2     = 12.608 (+ -.341 )
SDw       = 1.3651
Vw        = 1.8635
Animals   = 820
Obs       = 820
Mean crit = 11.683
Mean dist = 11.54
```

A number of test statistics have been developed to detect major gene segregation, and these yield a single result per population, not per animal. Hill and Knott (1990) classify and discuss these. Le Roy and Elsen (1992) compare the performance of 22 statistics, and suggest their robustness may be low, especially when trait distribution is skewed. Moreover, to screen populations in order to select likely carriers, a method is needed which allows ranking of animals on the probability of carrying one or more major genes.

Genotype probability calculations can be very computationally demanding, with up to 3^n possible combinations of genotypes among n animals for a single locus with two alleles segregating (Elston and Stewart, 1971; Ott, 1979). This means that using exact maximum likelihood methods to either find the most likely combination of genotypes or to calculate genotype probabilities is only practical for small problems involving less than about 20 animals. Animal pedigrees are generally highly looped, due to the fact that sires are mated to a large number of dams, and involve inbreeding, which makes the problem considerably more complex. However, Janss *et al*. (1993) developed an iterative method, based on that of van Arendonk *et al.* (1989), and this can be used to determine genotype probabilities in large animal breeding populations spanning several generations.

For a mixed inheritance model involving both a major gene and polygenes, exact likelihood analysis is not feasible and approximations are needed. A numerical integration technique to account for polygenes (e.g. Hermite integration as applied by Knott *et al*. (1991*a*,*b*)) yields reasonable results. Numerical integration can only be applied to groups of independent sire families. This restricts the value of these methods, which are not able to exploit the extra information in data covering several generations (Janss *et al*., 1993).

Approximate maximum likelihood methods have been used by Hoeschele (1988), Knott *et al*. (1991*a*,*b*) and Hofer and Kennedy (1993) to calculate both genotype probabilities and estimate polygenic breeding values for simple pedigree structures. However, the flexible iterative approach to calculating genotype probabilities (van Arendonk *et al*., 1989; Janss *et al*., 1993) can be implemented together with a mixed-model regression step to account for the effects of polygenes under any pedigree structure (Kinghorn *et al*., 1993), implemented as 'Findgene'. Both the approximate maximum likelihood methods and this regression method can lead to estimates of genotype effects and gene frequencies for the population as well as genotype probabilities and estimated breeding values for all individuals. However, results are generally biased to a moderate degree, except, under some circumstances, in the absence of selection.


Genotype probability calculations.
A number of cycles of using information from relatives.
Exit when converged
Input data
P, $\hat{B}_1$ and $\hat{B}_2$, p
$(P - \hat{a})$
$\hat{B}_1$ and $\hat{B}_2$
Genotype probabilities
Gene freq. p
Regress phenotype on genotype probabilities and animal breeding values.

This is an illustration of the method used to arrive at converged estimates of major gene effects, b1 and b2, and calculate genotype probabilities for individuals. Genotype probabilities are calculated using segregation analysis, following the method of van Arendonk et. al. (1989), and allele frequency estimated by appropriate averaging of these probabilities. These probabilities are then fitted in a regression of phenotype on genotype probabilities and animal breeding values, using a BLUP framework. This regression yields estimates of breeding value and new estimates of b1 and b2. Phenotypes (P) are corrected for estimated breeding values ($\hat{A}$) in an attempt to reduce the influence of polygenic effects on the next calculation of genotype probabilities. The cycle illustrated is repeated sufficient times to give convergence in estimates of b1 and b2. However, to speed operation, convergence is first achieved with animal breeding values not fitted in the regressions, then animal breeding values are fitted until final convergence is reached.

Both the approximate maximum likelihood methods and this regression method can lead to estimates of genotype effects and gene frequencies for the population as well as genotype probabilities and estimated breeding values for all individuals. However, results are generally biased to a moderate degree, except, under some circumstances, in the absence of selection.

More recently, Monte Carlo integration techniques have been applied in segregation analysis (Guo and Thompson, 1992, Janss *et al*., 1993). Janss *et al*. found empirically unbiased estimates of variation due to polygenes and major gene effects in simulated data which contained multiple loops.

The methods described in this section do not make use of genetic markers, and seem unlikely to be able to detect genes reliably with an effect of less than about half a phenotypic standard deviation, even with favourable combinations of population size, population structure, and polygenic variation. However, if used routinely in parallel with genetic evaluation of pedigreed data sets, they may prove useful in identifying individuals and families which warrant closer scrutiny via test matings and use of genetic markers.

References

Elston, R.C.; Stewart, J. (1971) A general model for the genetic analysis of pedigree data. *Human Heredity* **21**, 523-542.

Fernando, R.L.; Grossman, M. (1989) Marker assisted selection using best linear unbiased prediction. Genetics, Selection,.Evolution 21, 467-477.

Goddard, M.E. (1992) A mixed model for analyses of data on multiple genetic markers. Theoretical and Applied Genetics 83, 878-886.

Guo, S.W.; Thompson, E.A. (1992) A Monte Carlo method for combined segregation and linkage analysis. *American Journal of Human Genetics* **51**, 1111-1126.

Hill, W.G.; Knott, S. (1990) Identification of genes with large effects. In: *Advances in statistical methods for genetic improvement of livestock.* Springer-Verlag.

Hoeschele, I. (1988) Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance. *Theoretical and Applied Genetics* **76**, 81-92.

Hofer, A.; Kennedy, B.W. (1993) Genetic evaluation for a quantitative trait controlled by polygenes and a major locus with genotypes not or only partly known. *Genetics Selection Evolution* **25**(6), 537-555.

Janss, L.L.G.; Van Der Werf, J.H.J.; van Arendonk, J.A.M. (1994) Identification of a major gene in several generations of data with an application to F2 and F3 crosses. *Genetics Selection Evolution*. In Press.

Kerr, R.J.; B.P. Kinghorn, 1996: An efficient algorithm for segregation analysis in large populations. J. Anim. Breed. Genet. 113: 457-469.

Kinghorn, B.P.; Kennedy, B.W.; Smith, C. (1993) A method of screening for genes of major effect. Genetics 134, 351-360.

Kinghorn, B.P. 1997. An index of information content for genotype probabilities derived from segregation analysis. Genetics. 145:479-483.

Kinghorn, B.P. and Clarke B. E. 1997. Genetic evaluation at individual QTL. Animal Biotechnology, 8:63-68.

Kinghorn, B.P. 1999. Use of segregation analysis to reduce genotyping costs. J. Anim. Breed. Genet. 116:175-180.

Knott, S.A.; Haley C.S.; Thompson, R. (1991*a*) Methods of segregation analysis for animal breeding data: a comparison of power. *Heredity* **68**, 299-311.

Knott, S.A.; Haley C.S.; Thompson, R. (1991*b*) Methods of segregation analysis for animal breeding data: parameter estimates. *Heredity* **68**, 313-320.

Knott, S.A.; Haley C.S. (1992) Maximum likelihood mapping of quantitative trait loci using full-sib families. *Genetics* **132**(4), 1211-1222.

Le Roy, P.; Elsen, J.M.; (1992) Simple test statistics for major gene detection: a numerical comparison. *Theoretical and Applied Genetics* **83**, 635-644.

Meuwissen, T.H.E. and Goddard M.E. 1997. Estimation of effects of quantitative trait loci in large complex pedigrees. Genetics 146:409-416.

Nejati-Javaremi, A., Smith, C. and Gibson, J.P. 1997. Effect of total allelic relationship on accuracy and response to selection. J. Anim. Sci. 75:1738-1745.

Ott, J. (1979) Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *American Journal of Human Genetics* **31**, 161-175.

Van Arendonk, J.A.M.; Tier, B.; Kinghorn, B.P. (1994) Use of multiple genetic markers in prediction of breeding values. Genetics 137(1), 319-329.

Van Arendonk, J.A.M.; Smith, C.; Kennedy, B.W. (1989) Method to estimate genotype probabilities at individual loci in farm livestock. *Theoretical and Applied Genetics* **78**, 735-740.

Wang, T., Fernando, R.L., Van der Beek, S, Grossman, M, and van Arendonk, J.A.M. 1995. Covariance between relatives for a marked quantitative trait locus. Genet. Sel. Evol. 27:251-274.