From Linkage Analysis to Gene Detection

Ben Hayes and Mike Goddard

From Linkage Analysis to Gene Detection

Ben Hayes and Mike Goddard

PART 1. LINKAGE ANALYSIS.

1.1 OVERVIEW: THE *DGAT1* STORY

A recent article in the journal *Genome Research* (Grisart *et al*. 2002) reported the identification of a single base pair mutation in the DGAT1 gene, with major effects on milk yield and composition in cattle. The first step in identifying the mutation was a genome wide linkage analysis (genome scan), which found a region of chromosome 14 contained a QTL with a large effect on fat percentage (Georges et al. 1995). The confidence region surrounding this QTL was very large (about 20-40cM), and contained so many possible genes that could possibly be carrying the underlying mutation (candidate genes) that it was impossible to select any of the genes with confidence. The confidence region surrounding the QTL was narrowed to about 3cM using linkage disequilibrium (LD) mapping (Riquet et al. 1999) and combined linkage disequilibrium linkage analysis (LDLA) (Farnir et al. 2002). The DGAT1 gene was identified as a strong candidate in this interval, and subsequent sequencing detected a base pair mutation in this gene. The mutation caused a substitution from lysine to alanine in the DGAT1 gene (ie. a functional mutation). Further investigations showed this mutation to be associated with major effects on milk yield and composition.

This publication is the first to demonstrate that the genetic mutations underlying the variation in quantitative traits can be identified. The process was long (7 years), and required a large amount of resources. In addition, the mutation in the *DGAT1* gene has a very large effect on milk composition traits, explaining up to 50% of the variance for fat percentage. Most QTL will have smaller effects than this, making the identification of the underlying mutations even more difficult. However the rapid expansion of availability of genomic resources (eg. microsatellite markers, human genome project) should mean that many more mutations underlying variation in quantitative traits in livestock species will be identified in the next 5-10 years.

The aim of this course is to provide you with a set of criteria for designing and analysing experiments with some chance of successfully detecting QTL, and mapping these QTL with some degree of precision. Considerable attention is devoted to strategies to decrease the cost of QTL mapping experiments, as this can be a crucial factor in acquiring funding for such experiments. The course also aims to familiarise you with a small portion of the vast genomic resources that are available on the Internet, which are invaluable in selecting candidate genes.

1.2 OPTIMISING THE DESIGN OF LINKAGE MAPPING EXPERIMENTS.

The key parameters which determine the power of QTL mapping experiments to detect QTL are the distribution of the effects of genes which affect the quantitative trait of interest, the size and structure of the population used for mapping, and the statistical significance thresholds used.

1.2.1 How many big genes (detectable QTL) are there?

The power of a QTL mapping experiment ultimately depends on how many genes of large effect and how many genes of small effect are segregating in the population. Hayes and Goddard (2001) attempted to derive the distribution of gene effects from the results of QTL mapping experiments in pigs and dairy cattle. Their results indicated many genes of small effect, and few genes of large effect, Figure 1.1.

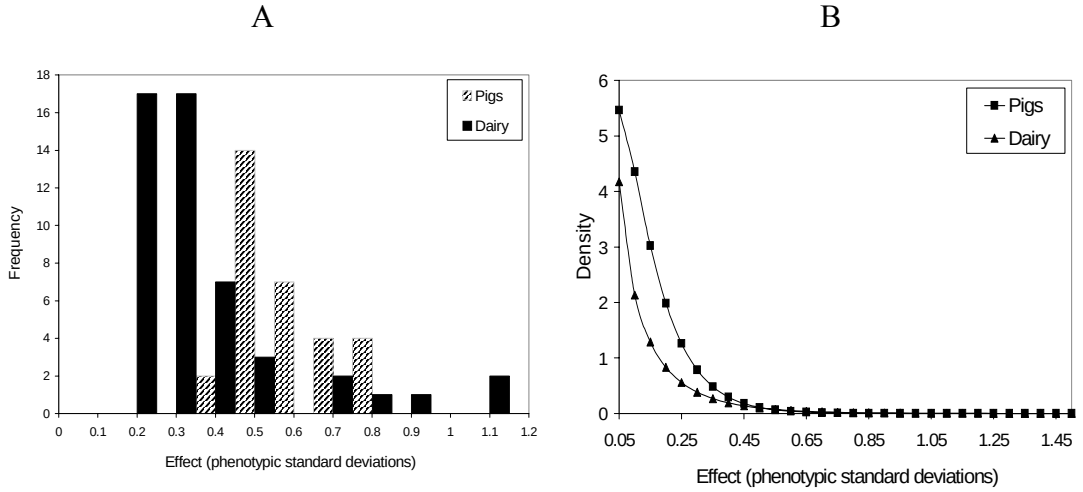

Figure 1.1 A. Distribution of additive (QTL) effects from pig experiments, scaled by the standard deviation of the relevant trait, and distribution of gene substitution (QTL) effects from dairy experiments scaled by the standard deviation of the relevant trait. B. Gamma Distribution of QTL effect from pig and dairy experiments, fitted with maximum likelihood.

Although there are few genes of large effect, these few genes contribute the majority of the genetic variance, Figure 1. 2.

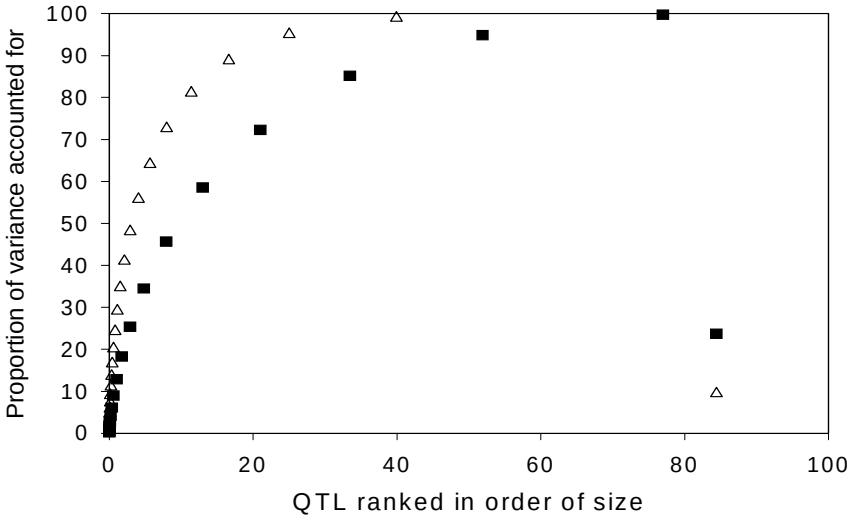

Figure 1.2 Proportion of genetic variance explained by QTL ranked in order of size of effect (■=pigs, and Δ=dairy cattle).

As a result, QTL mapping experiments need to detect the 10-20 genes with the largest effects to explain the majority of the total genetic variance. The proportion of the genetic variance explained by the detected QTL is the parameter which determines the increase in accuracy of marker assisted selection (MAS) compared with non-MAS (Meuwissen and Goddard 1996, Spelman et al. 2001).

1.2.2 Optimising the number and size of families in half sib designs

There are two key criteria which must be met for a mapping experiment to detect a QTL. If the population is outbred, and a half sib design is used, then at least one of the sires used in the mapping experiment must be heterozygous at the QTL (eg. carry a mutation at the gene locus responsible for the QTL effect on one of his chromosomes). The sire families must also be large enough to ensure that the difference between the effect of the two QTL alleles on the quantitative trait can be distinguished from environmental and other genetic effects. Typically the total number of progeny which can be genotyped will be limited by cost. There is then a choice between a large number of sires, each with a small number of progeny, and a small number of sires, each with a large number of progeny. Which is a better design for QTL mapping? This question has been investigated both using deterministic

predictions (eg. Weller et al. 1990), and simulation (eg. Hayes et al. 2001). Both approaches come to the same conclusion, which is to detect QTL of medium – large size (eg. effect of approximately 0.2 phenotypic standard deviations), very large half sib families are needed, Figure 1.3 and Table 1.1.

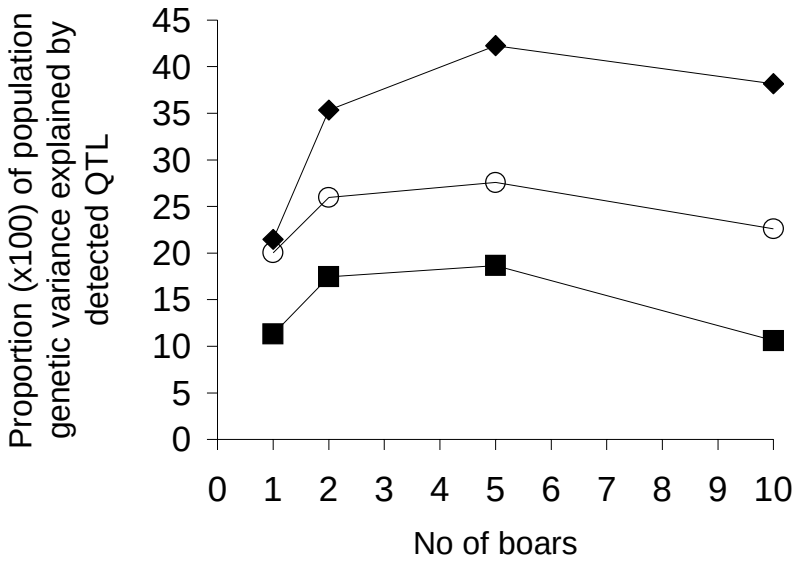

Figure 1.3. Proportion of genetic variance explained by detected QTL in genome scans with 1, 2, 5 or 10 boars and 500 (■), 1000 (O) or 2000 (♦) total progeny allocated to the mapping experiment.

In Figure 1.3, the criteria for measuring the success of the QTL mapping experiment is the total proportion of genetic variance explained by the QTL 'detected', that is, those QTL which have effects exceeding a significance threshold (in this case $P<0.05$ at the chromosome wide level). Increasing the numbers of boars increased the chance one or more boars was heterozygous for a QTL segregating in the population. However, for a given total number of progeny for the mapping experiment, increasing the number of boars decreased the progeny per boar, decreasing the chance that the QTL effect is statistically significant. Using five boars balanced these two phenomena to maximise proportion of variance explained by detected QTL, regardless of total number of progeny in the experiment.

In Table 1.1, from Weller *et al* (1990), the power of the experiment to detect QTL is expressed differently, as the probability that a QTL having an effect of a certain size

is statistically significant in the mapping experiment. Again, increasing the number of sires increases the chance of at least one sire being heterozygous at the QTL. Very large family sizes are needed to detect small QTL with any certainty.

Table 1.1 Power of half-sib design to detect a segregating QTL

Number of			Size of QTL effects[1]		
Sires	Progeny per sire	Total number of progeny	0.1	0.2	0.3
5	200	1000	0.03	0.18	0.50
	400	2000	0.07	0.44	0.80
	600	3000	0.12	0.64	0.90
	800	4000	0.18	0.76	0.94
	1000	5000	0.25	0.83	0.96
	2000	10000	0.55	0.95	0.97
10	200	2000	0.05	0.31	0.76
	400	4000	0.11	0.70	0.96
	600	6000	0.21	0.88	0.99
	800	8000	0.32	0.95	0.99
	1000	10000	0.43	0.97	0.99
	2000	20000	0.81	0.99	0.99
20	200	4000	0.07	0.56	0.95
	400	8000	0.20	0.93	0.96
	600	12000	0.38	0.99	0.99
	800	16000	0.56	0.99	0.99
	1000	20000	0.70	0.99	0.99
	2000	40000	0.97	0.99	0.99

Weller et al. (1990); [1] The size of QTL effects = a/SD, where a is half the difference between the mean trait values for the two homozygotes, and SD is the residual standard deviation.

1.2.3 Effect of statistical significance thresholds on QTL detection and accuracy of subsequent MAS

If one conducts a genome scan, with the aim of detecting QTL to use in marker assisted selection, an important question is how many QTL to take from the genome scan and use in the MAS program. The number of QTL detected in a genome scan is controlled by the number of segregating genes which affect the trait, the power of the experiment, and the level of stringency of the statistical test used to set the size of the significance threshold, above which a QTL is 'detected'. The less stringent the threshold, the greater the number of QTL detected, and the higher the proportion of genetic variance exploited by MAS using these detected QTL. The 'cost' of using less stringent thresholds is the higher number of false positives detected. False

positives reduce the accuracy of MAS, as the variance explained by marked QTL is overestimated.

A major issue in setting significance thresholds is the multiple testing problem. In most QTL mapping experiments, many positions along the genome or a chromosome are analysed for the presence of a QTL. As a result, when these multiple tests are performed the "nominal" significance levels of single test don't correspond to the actual significance levels in the whole experiment, eg. when considered across a chromosome or across the whole genome. For example, if we set a pointwise significance threshold of 5%, we expect 5% of results to be false positives. If we analyse 100 points along the chromosome (assuming for the moment these points are independent), we would expect 5 (100*0.05) false positive results! Obviously more stringent thresholds need to be set. The problem in QTL mapping is even more complex because 'tests' on the same chromosome are not independent, as the markers are linked.

Churchill and Doerge (1994) proposed the technique of permutation testing to overcome the problem of multiple testing in QTL mapping experiments. Permutation testing is a method to set appropriate significance thresholds with multiple testing (eg testing many locations along the chromosome for the presence of the QTL). Permutation testing is performed by analysing simulated data sets that have been generated from the real one by randomly shuffling the phenotypes across individuals in the mapping population. This removes any relationship between genotype and phenotype, and generates a series of data sets corresponding to the null hypothesis. Chromosome or genome scans can then be performed on these simulated data-sets. For each simulated data the position in the genome yielding the highest value for the test statistic is identified and stored. The values obtained over a large number of such simulated pedigrees are ranked yielding an empirical distribution of the test statistic under the null hypothesis of no QTL. The position of the test statistic obtained with the real data in this empirical distribution immediately measure the significance of the real dataset. Significance thresholds can then be set corresponding to 5% false positives for the entire experiment, 5% false positives for a single chromosome, and so on.

We performed an experiment to investigate the effect of the significance threshold on the subsequent accuracy of MAS. The experiment had two stages:

1. *QTL mapping*. From a simulated population of pigs, a sire was selected from the population and 200 progeny were bred from this sire for a genome scan. For each of the four marker brackets on 18 chromosomes, the sire's progeny were separated into those that inherited the sire's paternal bracket and those that inherited the sire's maternal bracket. Recombinants were ignored. If the difference between the average of the phenotypes of the two groups exceeded the significance threshold, a QTL was detected. The location of a detected QTL was considered to be at the centre of the bracket with the largest estimated effect on the quantitative trait. Five significance thresholds of decreasing stringency were set by permutation testing. The probabilities of a false positive for the five thresholds when testing an individual marker bracket were 0.0008 (corresponding to less than 5% false positives for the whole experiment), 0.014 (less than 5% false positives for each chromosome tested), 0.05, 0.10 and 0.25.
2. *Accuracy of MAS*. For each marker bracket with a significant effect, the effect of the four possible sire haplotypes (paternal, maternal, paternal-maternal recombinant and maternal-paternal recombinant) were estimated from the 200 progeny by solving the equation, $[\mathbf{Z}'\mathbf{Z} + \mathbf{I}\lambda]\hat{\mathbf{u}} = [\mathbf{Z}'\mathbf{y}]$, where $\mathbf{Z}$ is a design matrix allocating records to haplotypes, $\mathbf{I}$ is used to approximate $\mathbf{G}$, the matrix of haplotype co(variances), $\lambda = \frac{\sigma_e^2}{\sigma_g^2}$ where σ_e^2 is the error variance and σ_g^2 is the within sire variance for the QTL, $\hat{\mathbf{u}}$ is a vector of the estimates of the haplotype effects, and $\mathbf{y}$ is a vector of phenotypic records. A further 500 progeny were bred from the sire used in the genome scan. These progeny were genotyped for the marker haplotypes surrounding the detected QTL. The breeding values of a progeny were estimated as the sum over the marked QTL of the estimates for the marker haplotypes which the progeny carried. The correlation of estimated breeding values with true breeding values for these 500 progeny was the accuracy of MAS.

The results are shown in Figure 1.4. When breeding values are predicted in marker assisted selection, the QTL variance is required. Breeding values were predicted with both the true QTL variance at a significant QTL (TRUE), and the QTL variance estimated from a least squares analysis (DIRECT).

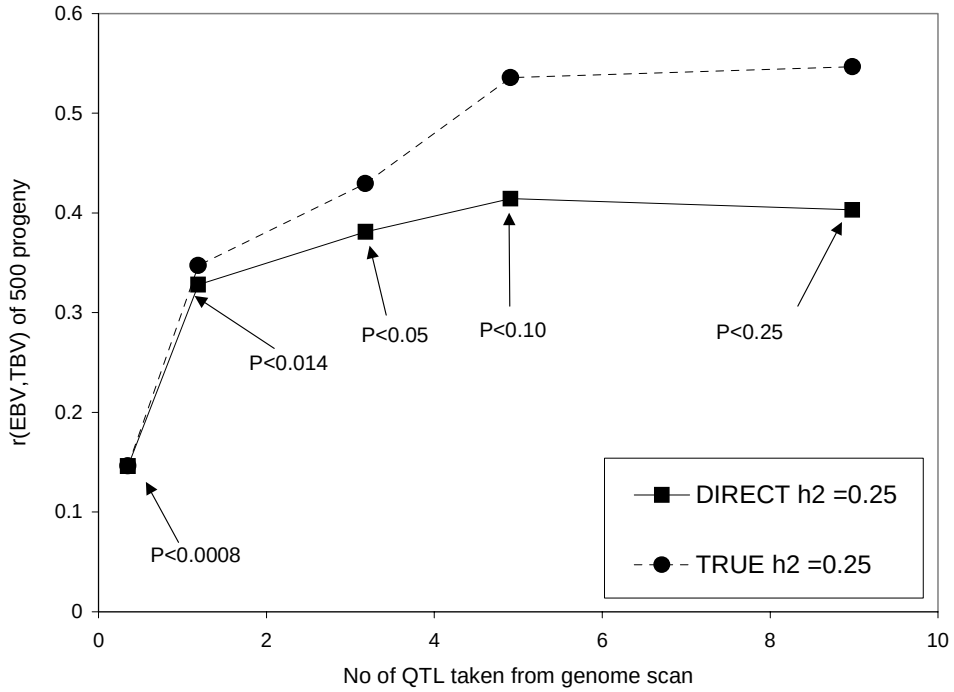

Figure 1.4. Number of QTL detected in the genome scan and accuracy of MAS.

The accuracy of MAS rose rapidly as the significance threshold for QTL detection was lowered from $P<0.0008$ to $P<0.014$. This is a result of the greater number of QTL detected explaining a greater proportion of the within sire variance. When DIRECT was used to estimate the variance from each QTL, the greatest accuracy of MAS was obtained when $P<0.10$. Lowering the stringency from $P<0.10$ to $P<0.25$ greatly increased the FDR, Table 1. This indicates that the majority of additional QTL detected by lowering the significance from $P<0.10$ to $P<0.25$ are false positives. These additional QTL therefore explain very little of the additive variance (also indicated by the plateau of accuracy using TRUE to estimate the variance), and in fact reduce the accuracy of MAS as the proportion of variance accounted for by detected QTL is overestimated.

A useful statistic is the false discovery rate (FDR). FDR is the expected proportion of detected QTL that are in fact false positives (Weller 1998). FDR can be calculated for

a QTL mapping experiment as mP_{max}/n, where P_{max} is the largest P value of QTL which exceed the significance threshold, n is the number of QTL which exceed the significance threshold and m is the number of chromosomes tested.

The FDR was calculated for each of the significance thresholds above, Table 1.2. The proportion of detected QTL which are in fact false positives rises rapidly as the stringency of the significance threshold is reduced, until when P<0.25, more than half the detected QTL are in fact false positives.

Table 1.2 False discovery rates for different significance thresholds (h^2=0.25)

P value	QTL detected	False discovery rate (FDR)
0.0008	0.35	0.04
0.014	1.3	0.20
0.05	3.2	0.24
0.1	4.9	0.34
0.25	9	0.58

The accuracy of MAS was greatest when *P*<0.10 was the criteria for taking QTL from a genome scan, if the QTL variance was estimated by DIRECT. While the *P*<0.10 threshold gave the greatest accuracy, it may not be the most profitable criteria for taking QTL from a genome scan to MAS. For example, using criteria *P*<0.05 would reduce the number of markers to be typed from 10 to 6 while only reducing accuracy by 8%. In addition, the number of progeny per sire in the mapping experiment was large (200), which meant both the QTL position and QTL variance was accurately estimated. For QTL mapping experiments with smaller families, the optimum stringency threshold for smaller QTL mapping experiments (where the intention is to use detected QTL in MAS) is likely to be more stringent (eg. chromosome or genome wide).

1.2.4 Precision of QTL mapping

The precision of positioning a putative QTL along a chromosome is usually expressed as an interval (in centiMorgan, cM) that contains the QTL with a level of statistical certainty, e.g. a 95% confidence interval. One method to assign confidence intervals to QTL locations is based on the likelihood ratio test (Lander and Bostein, 1989; Zeng, 1994; Zou, 2001). The likelihood ratio test is performed at any position covered by markers across the whole genome. The location with the highest likelihood is the most likely putative QTL position. The confidence interval (CI) is calculated by moving sideward (left and right) of the estimated position to the locations corresponding to a decrease in the LOD score of one or two units. The total width corresponding to a one- or two-LOD drop-off is then considered as the 96.8 or 99.8%CI, respectively (Mangin *et al.,* 1994). In the Lander and Bostein method, estimates of QTL position and its effects are approximately unbiased if there is only one QTL segregating on a chromosome (Zeng, 1994). Haley and Knott (1992) adopted a similar approach in a regression framework.

Visscher *et al.* (1996) proposed a bootstrap method to determine approximate confidence intervals for QTL position. For data on *N* individuals, a bootstrap sample is created by sampling with replacement *N* individual observations from the dataset. Each observation has marker genotype and phenotype. In the bootstrap sample some records can appear more than once. This process is repeated *n* times to generate *n* bootstrap samples. The Haley and Knott (1992) interval mapping method was used to detect QTL from the bootstrap samples. The empirical central 90 and 95% confidence intervals (CI) of the QTL position are determined by ordering the *n* estimates and taking the bottom and top fifth and 2.5th percentile, respectively.

Darvasi and Soller (1997) proposed a formula for estimating the 95% CI for QTL location for daughter and granddaughter designs provided genetic maps with high density of markers: $CI=3000/(kN\delta^2)$, where *N* is the number of individuals genotyped, δ the substitution effect in units of the residual standard deviation, *k* the number of informative parents per individual, which is equal 1 for half-sibs and backcross designs and 2 for F_2 progeny, and 3000 is about the size of the cattle genome in centi-Morgans. For instance, given a QTL with a substitution effect of 0.5 residual standard deviations, and 1000 progeny genotyped, the 95% CI would be 12cM.

Another commonly used method to improve the mapping precision is to increase the marker density on the chromosome. In practise the effectiveness of this strategy is limited, as enormously large half sib families would be required to generate recombinants between closely spaced markers in order to refine the QTL position. Linkage disequilibrium (LD) mapping is a possible solution and will be discussed later.

1.3 STRATEGIES TO MINIMISE THE NUMBER OF GENOTYPINGS

The largest cost in any QTL mapping experiment is the cost of genotyping animals at the DNA markers. For example, the cost of genotyping alone (ignoring DNA extraction costs, etc) for a genome scan in pigs, with say 10 markers on each of 18 chromosomes, at $4/marker/animal, and in a resource population of 5 sires with 200 progeny each, would be $723600.00. More than pocket money! As a result, there has been considerable effort into identifying strategies to reduce the genotyping cost.

1.3.1 Selective genotyping

Selective genotyping is a method of QTL mapping in which the analysis of linkage between markers and QTL is carried out by genotyping only individuals from the high and low phenotypic tails of the trait distribution in the population (Darvasi and Soller, 1992). In half sib designs, the selective genotyping is usually done with each sire family, eg. for each sire there will be a high and low progeny group. Individuals most deviating from the mean are considered to be most informative for linkage, as their genotypes at the QTL can be inferred from their phenotypes more clearly than can those with average phenotypes, Figure 1.5.

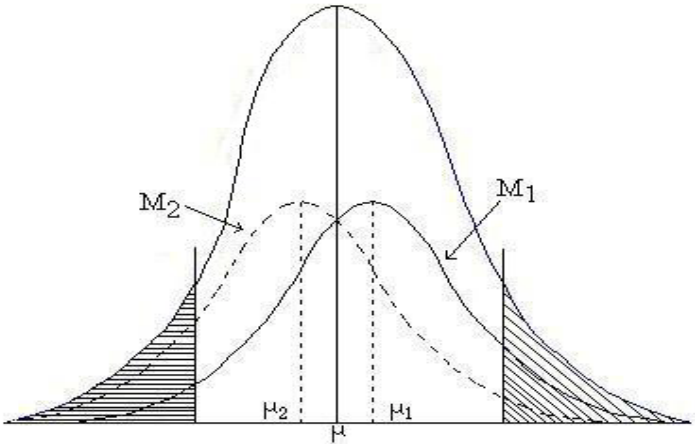

Figure 1.5 Distribution of the progeny of a M_1Q_1/M_2Q_2 sire, with high and low phenotypes for the production trait.

In fact Darvasi and Soller (1992) demonstrated that it is not necessary to genotype more than 50% of a population to get maximum power from the design.

For a constant number of genotyped progeny, selective genotyping can actually increase the power of the mapping experiment (Bovenhuis and Meuwissen 1996). Figure 1.6, from Bovenhuis and Meuwissen (1996), shows the power for different selected fractions as a function of the total number of animals with phenotypic records The design consists of five sires with a large number of daughters. Other parameters are: heritability 0.1, type I error 0.05, gene effect $0.2\sigma_p$, and recombination fraction 0. For a given number of animals genotyped, and no restrictions on the number of animals available for phenotypic trait evaluations, the power can be increased dramatically by using selective genotyping. This increase in the power results from the increased contrast between individuals carrying different marker genotypes (Bovenhuis and Meuwissen, 1996). Nevertheless it is recommended that the selection not be lower than 10% in either tail, (Bovenhuis and Meuwissen, 1996) because the data might contain outliers representing artefacts.

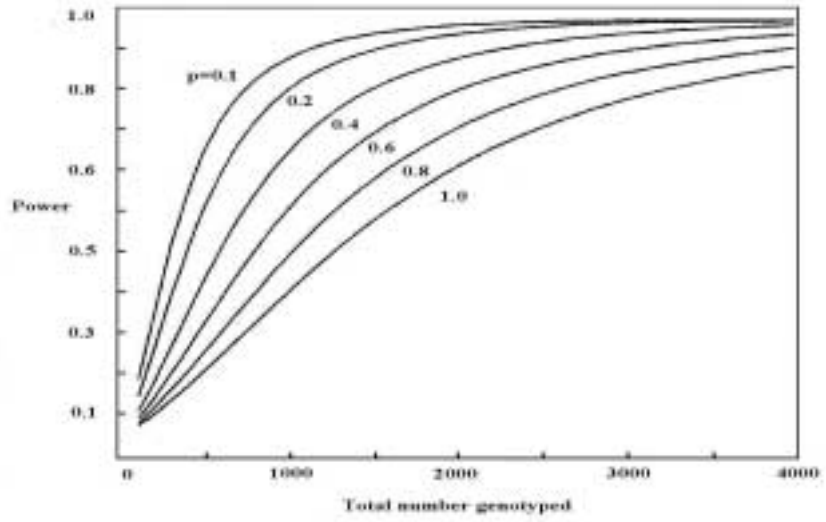

Figure 1.6. The power of a daughter design (probability of detecting a QTL with effect of 0.2σp, as a function of the number of individuals genotyped, for different selection fractions (p) (Bovenhuis and Meuwissen, 1996).

One drawback with selective genotyping is that the estimated QTL effect is severely biased (upwards) if only genotyped individuals are used to estimate the effect (Darvasi and Soller, 1992), Figure 1.7.

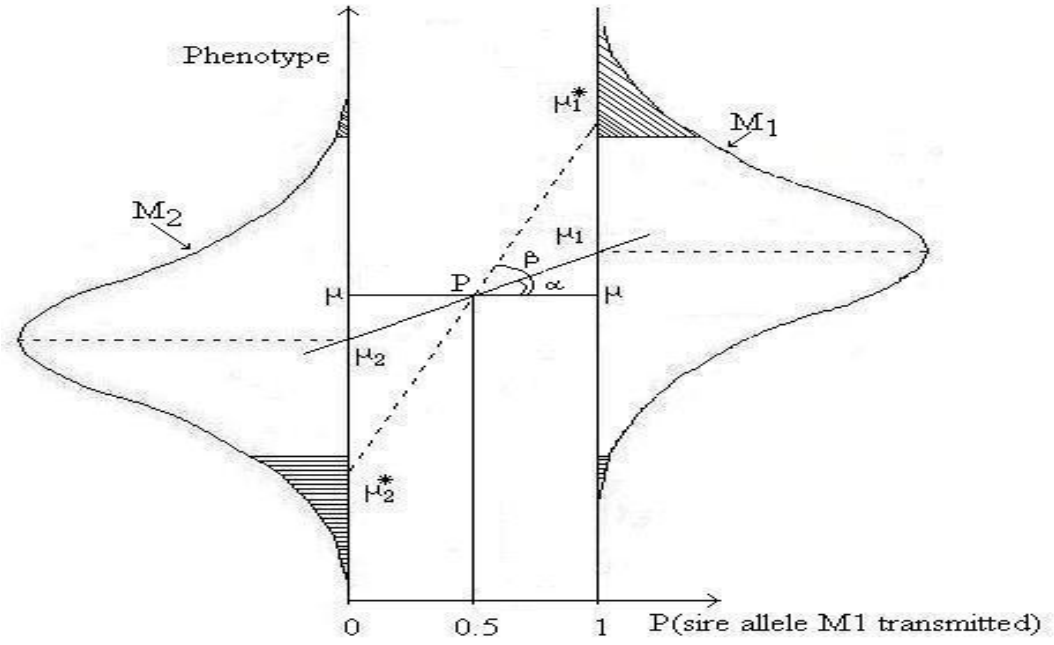

Figure 1.7. Overestimation of the QTL effect with selective genotyping. Progeny are from a single sire heterozygous at a marker locus (allele M_1 and M_2) and a closely linked QTL. The slope of the regression line between the mean of the distribution of

phenotypes of progeny inheriting the M1 allele and the mean of the distribution of phenotypes of progeny inheriting the M2 allele is the estimate of the allele substitution effect of the QTL allele. When 100% of progeny are genotyped, the slope of the regression line is line $\mu_1\mu_2$ is α. With selective genotyping, the trait means of the upper and the lower tails are μ_1^* and μ_2^*, respectively. The regression line is now $\mu_1^*\mu_2^*$, with the slope of β, which is greater than α. Thus the sire QTL effect is overestimated with selective genotyping. The higher the selection intensity, the greater the β is, and the more the QTL effect is overestimated.

This is a major problem if the QTL are to be used in MAS, as the overestimation of the QTL variance will erode the advantage of using the marker information. Darvasi and Soller (1992) suggested a method to derive the actual QTL effect as a function of the observed effect and the selection fraction (see Darvasi and Soller (1992) for more details). An alternative which can be applied in a wide range of situations is to include the pedigree and phenotype information from the ungenotyped animals in a variance component analysis. An identical by descent (**IBD**) matrix tracing the inheritance of sire alleles to progeny is constructed. Ungenotyped animals are given a probability of 0.5 of inheriting either sire QTL allele. Consider the following example, with three sires having two progeny each. The first two sires are heterozygous at the marker (which is very closely linked to the QTL), the second is homozygous. The model assumes that each sire carries two different QTL alleles, and these alleles are unique to that sire.

Table 1.3. Example of complete marker information

Sire	ID	1		2		3	
	Marker alleles	A and B		C and D		E and E	
Progeny	ID	4	5	6	7	8	9
	Sire ID	1	1	2	2	3	3
	Marker alleles inherited from the sire	A	A	C	D	E	E

The IBD matrix for this data set is:

Progeny ID	4	5	6	7	8	9
4	1	1	0	0	0	0
5	1	1	0	0	0	0
6	0	0	1	0	0	0
7	0	0	0	1	0	0
8	0	0	0	0	1	0.5
9	0	0	0	0	0.5	1

Both progeny of sire 1 inherit the A allele (and therefore the same QTL allele), so their covariance is 1. Animals 6 and 7 are progeny of sire 2. They received different marker alleles from their sire (and different QTL alleles), so their covariance at the QTL is 0. Sire 3 is homozygous at the marker, and so his progeny both received the E allele. In this case the marker cannot be used to infer the QTL allele which the progeny received, and in fact given the marker data the progeny are equally likely to have received either allele.

Now consider the case where animals 5, 6 and 7 have not been genotyped because they do not fall in the top or bottom phenotypic tail.

Table 1.4. Example of missing of marker information

Sire	ID	1		2		3	
	Marker alleles	A and B		C and D		E and E	
Progeny	ID	4	5	6	7	8	9
	Sire ID	1	1	2	2	3	3
	Marker alleles inherited from the sire	A	*	*	*	E	E

* Marker information is missing

The IBD matrix is now:

Progeny ID	4	5	6	7	8	9
4	1	0.5	0	0	0	0
5	0.5	1	0	0	0	0
3	0	0	1	0.5	0	0
4	0	0	0.5	1	0	0
5	0	0	0	0	1	0.5
6	0	0	0	0	0.5	1

As animal 5 has not been genotyped, we do not know which marker allele was inherited from the sire. In the absence of marker information, there is a 50% chance of inheriting either sire QTL allele. The covariance between animals 4 and 5 is therefore 0.5.

The variance component model used was : $Y = \mu + Zu + Zv + e$, where Y is a vector of observation, μ the overall mean, Z the design matrix relating animals to phenotypes, u the vector of additive polygenic effects, v the vector of additive paternal QTL effects, e the residual vector. The random effects u, v, and e are assumed to be distributed as follows: $u \sim (0, \sigma_u^2 \boldsymbol{A})$, $v \sim (0, \sigma_v^2 \boldsymbol{IBD})$, $e \sim (0, \boldsymbol{I}\sigma_e^2)$, where σ_u^2, σ_v^2, and σ_e^2 are the polygenic variance, the additive variance of one QTL allele, and the residual variance, respectively; $\boldsymbol{A}$ is the standard additive genetic relationship matrix, and $\boldsymbol{IBD}$ is a matrix whose ij element $\boldsymbol{IBD}_{ij}$ = Prob(QTL alleles i and j are IBD), described in the tables above, and $\boldsymbol{I}$ is an identity matrix.

This method gave unbiased estimates of the QTL variance in simulations, Table 1.5.

Table 1.5. Estimates of the variance of a QTL segregating in a half-sib design, 30 sires with 100 progeny each, and either 100% of progeny genotyped, 20% of progeny genotyped (top and bottom 10% within each sire family), or 20% of progeny genotyped but including the ungenotyped animals in the analysis.

Strategy	QTL size
True	0.32
100% genotyped	0.30±0.02
20% genotyped	0.93±0.02
20% genotyped, ungenotyped animals included in the analysis	0.31±0.02

1.3.2 Selective DNA pooling

A really clever strategy to greatly reduce the number of genotypings was proposed by an Israeli mapping group (Darvasi and Soller, 1994, Lipkin et al. 1998). In DNA pooling, the determination of linkage between a marker and a QTL is based on the distribution of parental alleles among pooled DNA samples of the extreme high and low phenotypic groups of offspring. The concept is illustrated in Figure 1.8.

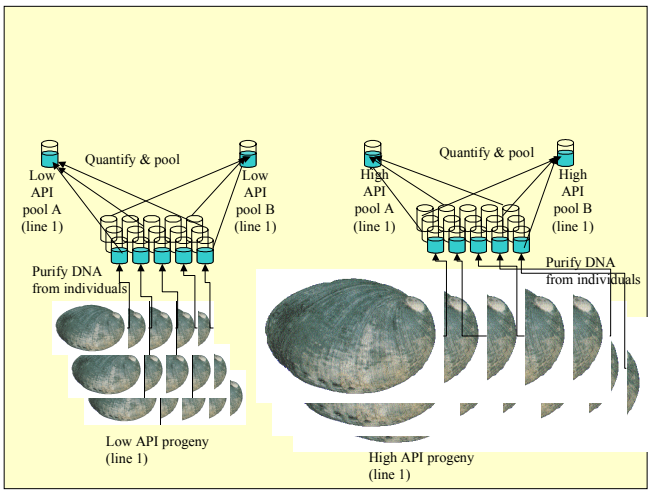

Figure 1.8. Design for selective DNA pooling of high and low abalone profit index (API) progeny within an abalone family line. Two pools of lows and two pools of highs are created for replication.

For a particular sire, if the 150 marker allele is linked to the increasing QTL allele (Q), and the 160 marker allele is linked to the decreasing QTL alleles (q), then we

would expect more of the 150 marker in the high pool than and in the low pool, and more of the 160 allele in the low pool than the high pool, Figure 1.9.

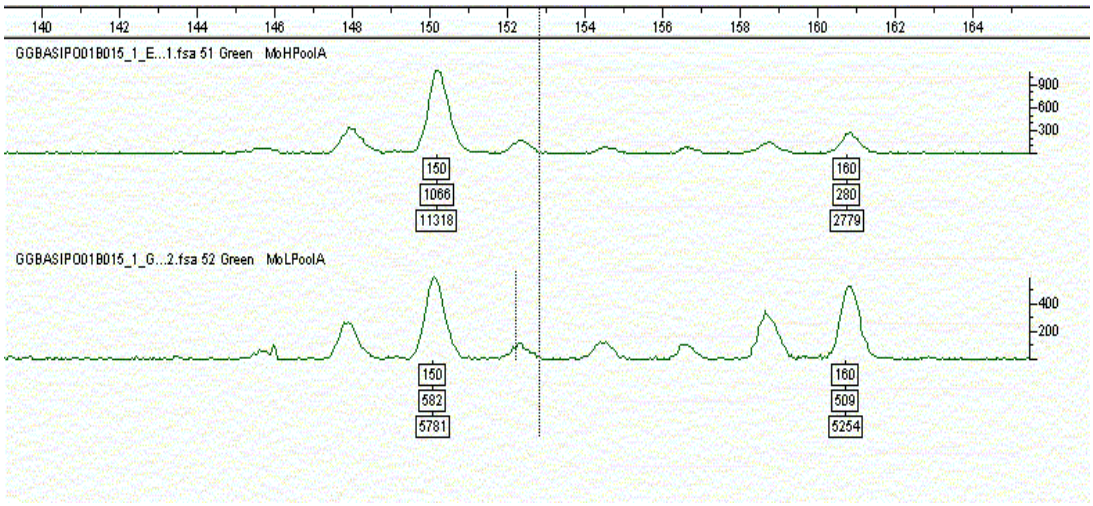

Figure 1.9. Electropherogram of marker allele abundance in high and low phenotype pools of progeny (top and bottom line respectively) from a single sire.

There are three potential difficulties with DNA pooling. One is that the amount of DNA in each pool to be genotyped must be quantified very accurately in order to estimate the frequencies of alleles in each pool with any precision, Figure 1.10.

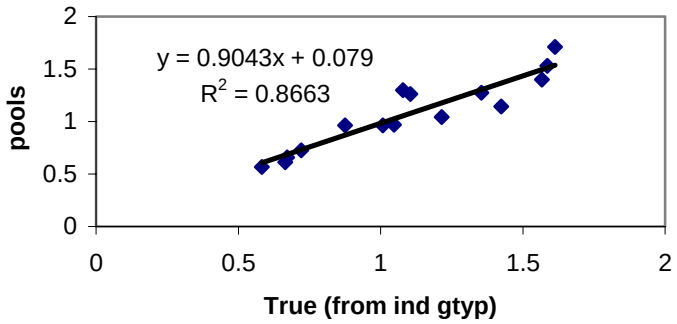

Figure 1.10. Allele frequencies in pools estimated from DNA pooling compared with pool frequecies estimated from individual genotyping.

The second is that with poly(TG) microsatellites, estimates of allele frequencies from pooled genotypes are confounded by "shadow" ("stutter") bands (eg Figure 1.9). Correction procedures have been developed on the basis of an observed linear regression between shadow band intensity and allele TG repeat number.

The third difficulty is that the selective DNA pooling experiment has the power to detect QTL affecting only the trait from which the pools of high and low phenotype progeny were made. The power to detect QTL for other traits will be extremely low, unless there is a high genetic correlation between the traits.

One parameter which should be considered in the design of selective DNA pooling experiments is the proportion of the half sib progeny from a sire which should be placed into each pool. The extra information from including a higher proportion of animals (and reduction of error due to outliers) must be balanced with minimising the number of DNA extractions to reduce time and cost. Figure 1.11, using simulated data, indicates most of the information is captured if high and low pools consist of 10% of the progeny.

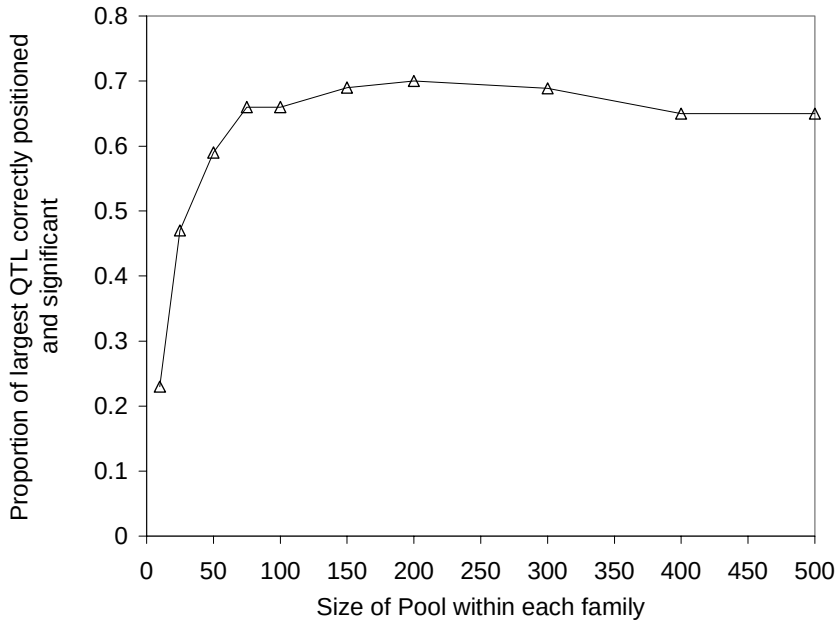

Figure 1.11. Power of QTL detection in a selective DNA pooling experiment in abalone, with three families of 1000 progeny each.

A large DNA pooling experiment has been carried out in Israeli-Holstein dairy cattle to detect QTL affecting milk protein percentage (Lipkin et al. 1998). Selective DNA pooling accessed 80.6% and 48.3% respectively of the information that would have been available through individual selective genotyping or total population genotyping. In effect, the statistical power of 45,600 individual genotypings was obtained from

328 pool genotypings. The experiment detected 5 QTL with highly significant effects on protein percentage.

To quote from Lipkin et al. (1998) “The DNA pooling methodology can make genome-wide mapping of QTL accessible to moderately sized breeding organisations." Extremely large QTL detection experiments using DNA pooling are currently underway in humans using dense single nucleotide polymorphism (SNP) markers.

PART 2. LINKAGE MAPPING IN COMPLEX PEDIGREES

In some species (eg. humans), it is difficult to create large half sib families for QTL mapping. An alternative to creating large half-sib families is to genotype animals in the existing pedigree. Potentially a larger number of recombination events (essential for positioning the QTL) can be accessed with this approach than with half-sib families. In practise, the large number of missing marker genotypes reduces the power of complex pedigrees for QTL mapping.

A two step process has been suggested for QTL mapping in complex pedigrees (George et al (2000):

1. For each QTL position on the chromosome segment, calculate the (co) variance matrix associated with the QTL. This matrix is also called the **G** or **IBD** (identical by descent matrix), and has elements *ij* = Prob(QTL alleles *i* and *j* are identical by descent or IBD).
2. For each position considered in step 1., construct a linear model to estimate QTL variance and other parameters, test for the presence of a QTL.

2.1 Calculating the IBD matrix

The **IBD** matrix has the dimensions 2*the number of animals x 2*the number of animals, eg two QTL alleles for each animal in the pedigree. The IBD matrix traces the transmission of the alleles of the founder animals (those at the top of the pedigree) through the population. If the marker information was complete, and could be used to perfectly infer the transmission of QTL alleles, this matrix would contain 1s and 0s only. At the other extreme, if there is no marker information, the **IBD** matrix will be identical to the **A** matrix (the average genetic relationship matrix), ie many elements of 0.5 signifying equal probability of inheriting either allele from a parent.

The inference of QTL genotypes from marker genotypes is considerably more complicated in complex pedigrees than simple half-sib pedigrees, as marker alleles need to be tracked over multiple generations. This can lead to a large number of missing genotypes. These genotypes have to be inferred in some way, which has proved to be very difficult in populations with many inbreeding and or marriage loops

(eg. most livestock populations). Considerable effort has been invested in creating strategies to infer genotypes with missing marker information and complex pedigrees (eg. Heath 1997, Kinghorn et al. 1993 and many others). Most strategies now use simulation based methods, predominantly Monte Carlo Markov Chain (MCMC) approaches. See George et al. (2000) for a good review of methods to calculate the IBD matrix when genotypes are missing, and Tier and Henshall (2002) for an approach that appears to work in the presence of inbreeding loops.

2.2 Variance component approaches for estimation of QTL parameters.

A further complication of using complex pedigrees for QTL mapping is that simple regression or maximum likelihood analysis can not be used easily to analyse linkage in complex pedigrees. Rather, a variance component approach is required (George et al. 2000). The model used to analyse such data is (ignoring fixed effects):

$$Y = \mu + Zu + Zv + e\ ,$$

where Y is a vector of observation, μ the overall mean, Z the design matrix relating animals to phenotypes, u the vector of additive polygenic effects, v the vector of additive QTL effects, e the residual vector. The random effects u, v, and e are assumed to be distributed as follows: $u \sim (0, \sigma_u^2 A)$, $v \sim (0, \sigma_v^2 G)$, $e \sim (0, \sigma_e^2 I)$, where σ_u^2, σ_v^2, and σ_e^2 are the polygenic variance, the additive QTL variance, and the residual variance, respectively; A is the standard additive genetic relationship matrix, and G is a matrix whose ij element G_{ij} = Prob(QTL alleles i and j are identical by descent or IBD).

2.3 Linkage mapping in complex pedigrees and MAS

One major advantage of linkage mapping in complex pedigrees is that the variance component model predicts marker assisted breeding values for all animals in the analysis. As some of these animals are likely to be the current crop of selection candidates, MAS can be implemented immediately. In the half-sib designs the situation is quite different – the sires used in the design will be one or two generations older than the current crop of selection candidates.