Linkage disequilibrium mapping (LD) mapping and Bioinformatics Approaches to identify QTL

Ben Hayes

3 Linkage disequilibrium mapping (LD) mapping

3.1 Definitions of LD.

The classical definition of linkage disequilibrium refers to the non-random association of alleles between two loci. Consider two markers, A and B, that are on the same chromosome. A has alleles A1 and A2, and B has alleles B1 and B2. Four **haplotypes** of markers are possible A1_B1, A1_B2, A2_B1 and A2_B2. If the frequencies of alleles A1, A2, B1 and B2 in the population are all 0.5, then we would expect the frequencies of each of the four haplotypes in the population to be 0.25. Any deviation of the haplotype frequencies from 0.25 is **linkage disequilibrium** (LD), ie the genes are not in random association. As an aside, this definition serves to illustrate that the distinction between linkage and linkage disequilibrium mapping is somewhat artificial – in fact linkage disequilibrium between a marker and a QTL is required if the QTL is to be detected in either sort of analysis. The difference is:

> *linkage analysis* only considers the linkage disequilibrium that exists within families, which can extend for 10s of cM, and is broken down by recombination after only a few generations.
>
> *linkage disequilibrium* mapping requires a marker allele or markers alleles to be in LD with a QTL allele across the entire population. To be a property of the whole population, the association must have persisted for a considerable number of generations, so the marker(s) and QTL must therefore be closely linked.

One measure of LD is *D*, calculated as (Hill 1981)

$$D = freq(A1\_B1)*freq(A2\_B2)-freq(A1\_B2)*freq(A2\_B1)$$

where freq (A1_B1) is the frequency of the A1_B1 haplotype in the population, and likewise for the other haplotypes. The *D* statistic is very dependent on the frequencies of the individual alleles, and so is not particularly useful for comparing the extent of LD

among multiple pairs of loci (eg. at different points along the genome). Hill and Robertson (1968) proposed a statistic, r^2, which was less dependent on allele frequencies,

$$r^2 = \frac{D^2}{freq(A1) * freq(A2) * freq(B1) * freq(B2)}$$

Where freq(A1) is the frequency of the A1 allele in the population, and likewise for the other alleles in the population. The r^2 parameter can also be calculated between two loci with multiple (more than two) alleles, which is usually necessary for microsatellites, as

$$r^2 = \sum_i^n \sum_j^m \frac{D_{ij}^2}{p_i p_j}$$

where locus 1 has *n* alleles, locus two has *m* alleles, p_i is the frequency of allele *i* of locus one, p_j is the frequency of allele *j* of locus two, and $D_{ij} = freq_{ij} - freq_i\, freq_j$, $freq_{ij}$ is the frequency of haplotype *ij*, $freq_i$ is the frequency of allele i of locus one and $freq_j$ is the frequency of allele *j* of locus two.

These classical definitions of LD, while important and widely used, are not particularly illuminating with respect to the causes of LD, and may also not be especially useful for QTL mapping. For example, statistics such as r^2 consider only two loci at a time, whereas we may wish to calculate the extent of LD across a chromosome segment that contains multiple markers. An alternate multi-locus definition of LD is the **chromosome segment homozygosity** (**CSH**) (Hayes et al. 2003). Consider an ancestral animal many generations ago, with descendants in the current population. Each generation, the ancestor's chromosome is broken down, until only small regions of chromosome which trace back to the common ancestor remain. These chromosome regions are identical by descent (IBD). Figure 3.1 demonstrates this concept.

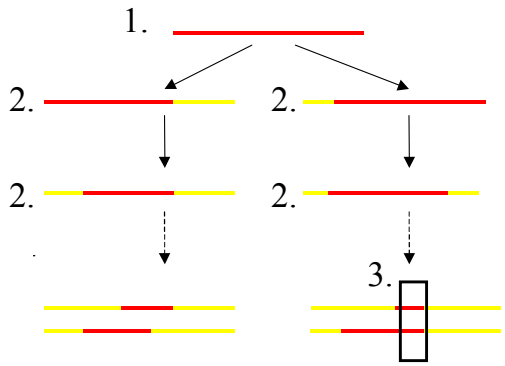


Figure 3.1 A ancestor many generations ago (1) leaves descendants (2). Each generation, the ancestors chromosome is broken down by recombination, until all that remains in the current generation are small conserved segments of the ancestor's chromosome (3). The chromosome segment homozygosity (CSH) is the probability that two chromosome segments of the same size and location drawn at random from the population are from a common ancestor.

The CSH then is the probability that two chromosome segments of the same size and location drawn at random from the population are from a common ancestor (ie IBD), without intervening recombination. CSH is defined for a specific chromosome segment, up to the full length of the chromosome. The CSH cannot be directly observed from marker data but has to be inferred from marker haplotypes for segments of the chromosome. Consider a segment of chromosome with marker locus A at the left hand end of the segment and marker locus B at the other end of the segment (as in the classical definition above). The alleles at A and B define a haplotype. Two such segments are chosen at random from the population. The probability that the two haplotypes are identical by state (IBS) is the haplotype homozygosity (HH). The two haplotypes can be IBS in two ways,

i. The two segments are descended from a common ancestor without intervening recombination, so are identical by descent (IBD), or

ii. the two haplotypes are identical by state but nor IBD

The probability of i. is CSH. The probability of ii. is a function of the marker homozygosities, given the segment is not IBD (see Hayes et al. 2003 for details). The probabilities of i. and ii. are added together to give the haplotype homozygosity (HH):

$$HH = CSH + \frac{(Hom_A - CSH)(Hom_B - CSH)}{1 - CSH}$$

This equation can be solved for CSH when the haplotype homozygosities and individual marker homozygosities are observed from the data. For more than two markers, the predicted haplotype homozygosity can be calculated in an analogous but more complex manner.

3.2 Why does LD occur?

LD can be a result of migration, mutation, selection, small finite population size or other genetic events which the population experiences (eg. Lander and Schork 1994). LD can also be created in livestock populations; in an F2 QTL mapping experiment LD is created between marker and QTL alleles by crossing two inbred lines.

In livestock populations, finite population size is generally implicated as the key cause of LD, as effective population sizes for most livestock populations are relatively small. For example, LD due to crossbreeding (migration) is large when crossing inbred lines but small when crossing breeds that do not differ as markedly in gene frequencies, and it disappears after only a limited number of generations (eg. Goddard 1991), and mutations are likely to have occurred many generations ago. While selection is probably a very important cause of LD, it's effect is likely to be localised around specific genes, and so has relatively little effect on the amount of LD 'averaged' over the genome.

3.2.1 Predicting the extent of LD with finite population size

If we accept finite population size as the key driver of LD in livestock populations, it is possible to derive a simple expectation for the amount of LD for a given size of chromosome segment. This expectation is (Sved 1971)

$$E(r^2) = 1/(4Nc + 1)$$

where N is the finite population size, and c is the length of the chromosome segment in Morgans. The CSH has the same expectation (Hayes et al. 2003). This equation predicts

rapid decline in LD as genetic distance increases, and this decrease will be larger with large effective population sizes, Fig. 3.2.

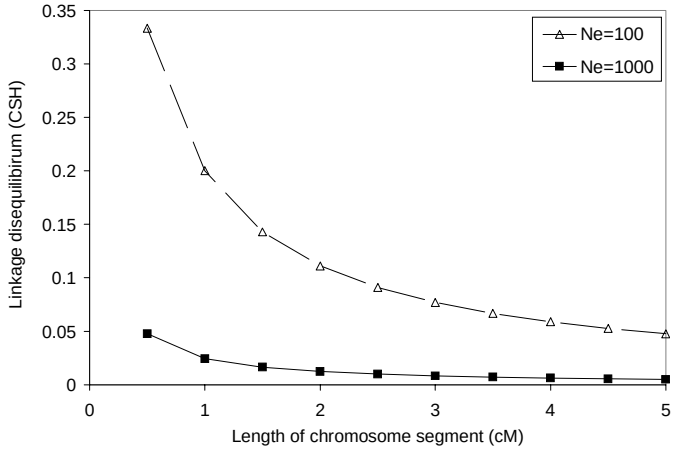


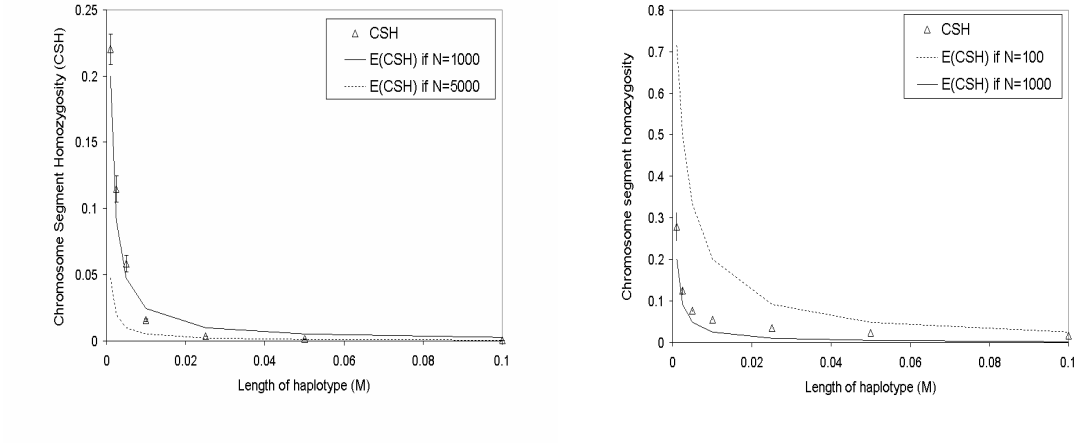
Figure 3.2. The extent of LD (as measured by chromosome segment homozygosity, CSH) for increasing chromosome segment length, for N_e=100 and N_e=1000.

3.3 The extent of LD in human and livestock populations

If LD is a predominantly result of finite population size, then the extent of LD should be many times less in humans, where the effective population size is ~ 10000 (Kruglyak 1999), than in livestock, where effective population sizes can be as low as 100 (Riquet et al. 1999). In fact, this is what is observed. Significant LD in humans typically extends less than 5kb (~0.005cM), depending on the population studied (Dunning et al. 200, Reich et al. 2001), while in cattle and sheep, considerable LD can extend up to 5-10 cM (Franir et al. 2000, McRae et al. 2002, Hayes et al. 2003).

As the extent of LD that is observed depends both on recent and historical recombinations, not only the current effective population size, but also the past effective population size are important. Effective population size for livestock species may have been much larger in the past than they are today. For example in dairy cattle the widespread use of artificial insemination and a few elite sires has greatly reduced effective population size in the recent past. In humans, the story is the opposite;

improved agricultural productivity and industrialisation have led to dramatic increases in population size. How does changing population size affect the extent of LD? To investigate this, we simulated a population which either expanded or contracted after a 6000 generation period of stability. The LD, as measured by CSH, was measured for different lengths of chromosome segment, Figure 3.3.



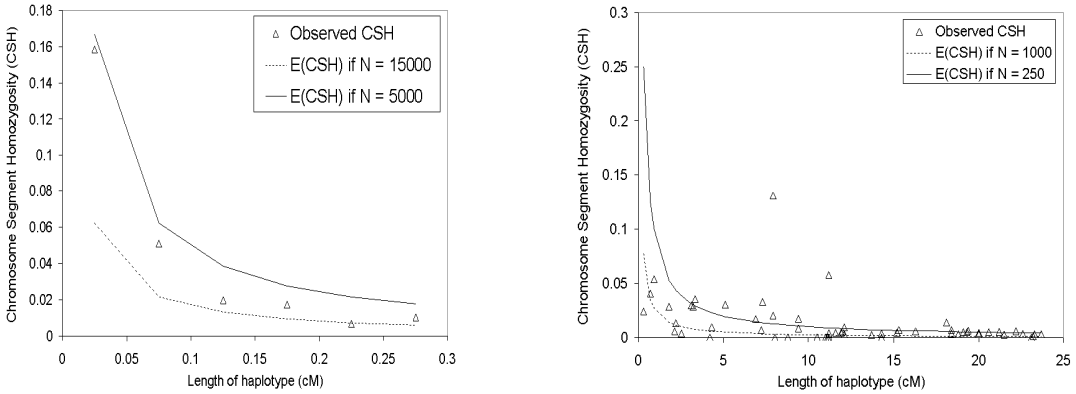
A B

Figure 3.3. Chromosomal homozygosity for different lengths of chromosome (given the recombination rate) for populations: A. Linearly increasing population size, from N=1000 to N=5000 over 100 generations, following 6000 generations at N=1000. B. Linearly decreasing population size, from N=1000 to N=100 over 100 generations, following 6000 generations at N=1000.

The general conclusion is that LD at short distances is a function of effective population size many generations ago, while LD at long distances reflects more recent population history. In fact, provided simplifying assumptions such as linear change in population size are made, it can be shown that the CSH reflects the effective population size 1/(2c) generations ago, where c is the length of the chromosome segment in Morgans.

Figure 3.4. illustrates the extent of LD, as measured by CSH, in both a human population and a Holstein Friesian population. CSH declines with distance in both species, but the levels of LD in cattle are much greater. The Figure also illustrates another feature of LD, that of extreme variability, even at the same distance. The variability in LD with a multi-

locus measure of LD like CSH is considerably lower than with two locus measures, such as r^2 (Hayes et al. 2003).



A B

Figure 3.4. A. Chromosomal homozygosity for increasing lengths of haplotype from human marker data (Moffat et al. 2000). The upper (solid) line is the expected CSH when the effective population size is 5000. The lower (dashed) line is the expected CSH when the effective population size is 15000. B. Chromosomal homozygosity from a dairy cattle data set. Also plotted are the expected values of CSH when N=1000 and N=250.

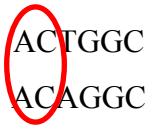
3.3.1 What type of markers are appropriate for detecting LD?

As LD extends over much shorter genetic distances than linkage within families, a denser marker map is required to detect and position QTL using LD (compared with the density required for a linkage analysis).

For linkage analysis, **microsatellite markers** are favoured, due to their high polymorphism (many alleles in the population) and ease of amplification. A microsatellite marker is usually a di-nucleotide repeat, eg, ATATAT repeated many times. Microsatellites are nearly always neutral, having no effect on gene expression. The alleles are scored according to the size, in base pairs, of the amplified DNA fragments containing the di-nuleotide repeats. For example, a heterozygous animal may have the genotype 282, 280. The microsatellite marker maps are well developed in the

major livestock species, and their position on the genetic map is known. In addition, primers can be downloaded from the web (Http://www.marc.usda.gov/genome/htmls/LinkageMap.jsp?Species=bos&Chromosome=1). In cattle, there are approximately 2141 microsatellite markers available (Ihara et al. 2002). As the cattle genome is approximately 3000cM, this gives an average marker spacing of 1.4cM. This density may be sufficient for preliminary LD mapping, however a greater density of markers would be desirable (particularly as there will be gaps in the microsatellite coverage).

In humans, where significant LD extends only very small distances, microsatellites are not suitable for LD mapping, simply because they are not sufficiently dense. Alternative markers are Single Nucleotide Polymorphisms, or SNPs. These markers are a single base pair substitution at a known site, eg



SNPs have the advantage that they occur very frequently throughout the genome, approximately 1 every 1000 bases (0.005cM). SNPs can be in either non-genomic DNA or in coding sequence. It is possible that a SNP in the LD experiment may be the actual mutation causing the QTL effect. The disadvantage of SNPs is that they are not as informative as microsatellites – they have a maximum of two alleles. About five SNPs are required to give the same amount of information as a single microsatellite. In addition, considerable laboratory effort is required for SNP detection (there are no public sites yet where primers for SNPs can be downloaded, though the United States Department of Agriculture is working on such a project), and some strategy must be used to ensure the putative SNP is not just a sequencing error.

3.4 LD mapping

The existence of LD implies there are small segments of chromosome in the current population which are descended from the same common ancestor. These IBD chromosome segments will not only carry identical marker haplotypes; if there is a QTL somewhere within the chromosome segment, the IBD chromosome segments will also carry identical QTL alleles. Therefore if two animals carry chromosomes which are likely to be IBD at a point o the chromosome carrying a QTL, then their phenotypes will be correlated. We can calculate the probability the 2 chromosomes are IBD at a particular point based on the marker haplotypes and store these probabilities in an IBD matrix (**G**). If the correlation between the animals is proportional to G there is evidence for a QTL at this position.

Before the IBD matrices can be calculated, the genotype data must be sorted into haplotypes (also called estimation of linkage phases). This can be done with Gibbs sampling, or following Mendelian rules.

3.4.1 Building the IBD matrix from marker haplotypes

We can infer the IBD status of two chromosome segments from their marker haplotypes (the CSH) as described above. For example, consider a chromosome segment which carries 10 marker loci and a single central QTL locus. Three chromosome segments were selected from the population at random, and were genotyped at the marker loci to give the marker haplotypes 11212Q11211, 22212Q11111 and 11212Q11211, where Q designates the position of the QTL. The probability of being IBD at the QTL position is higher for the first and third chromosome segments than for the first and second or second and third chromosome segments, as the first and third chromosome segments have identical marker alleles for every marker locus.

This is the basis for calculating an IBD matrix, G, for a putative QTL position from a sample of marker haplotypes (Meuwissen and Goodard 2001). Element g_{ij} of this matrix is the probability that haplotype i and haplotype j carry the same QTL allele.
The dimensions of this matrix is 2xthe number of animals x 2xthe number of animals, as each animal has two haplotypes.

Meuwissen and Goddard (2001) described a method to calculate the IBD matrix based on deterministic predictions which took into account the number of markers flanking the putative QTL position which are identical by state, the extent of LD in the population based on the expectation under finite population size, and the number of generations ago that the mutation occurred.

As with the CSH calculation, there are two ways in which marker haplotypes can be identical, either they are IBD, or the same marker haplotypes have been regenerated by recombination. The important parameters are the number of markers and the length of the haplotype (as well as the effective population size discussed above). One way to gain some insight into the effect of the number of markers in the haplotype on the IBD coefficients in the G matrix is to investigate the proportion of identical marker haplotypes which carry the same QTL allele, by calculating the proportion of QTL variance accounted for by marker haplotypes (ρ). If each unique marker haplotype is associated with a single QTL allele, this proportion will be one. For example, in a simulated population of Ne=100, and a chromosome segment of length 10cM, the proportion of the QTL variance accounted for by marker haplotypes when there were 11 markers in the haplotype was close to one, Figure 3.5. [Note that if the effective population size was larger, the proportion of genetic variance explained by a 10cM haplotype would be reduced (Goddard 1991).]

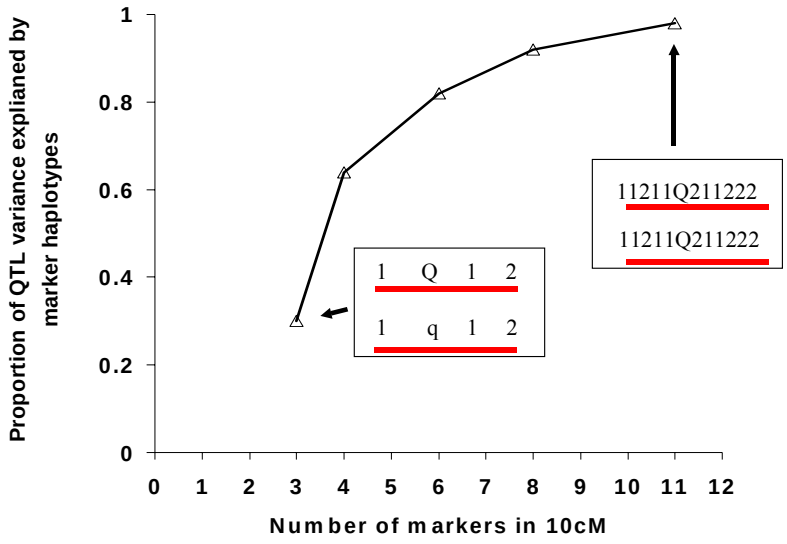


Figure 3.5. Proportion of QTL variance explained by marker haplotypes with an increasing number of markers in a 10 cM interval

Now consider a population of effective population size 100, and a chromosome segment of 10cM with eight markers. Two animals are drawn from this population. Their marker haplotypes are 12222111, 11122111 for the first animal, and 12222111 and 11122211 for the second animal. The putative QTL position is between markers 4 and 5 (ie. in the middle of the haplotype). The G or IBD matrix could look something like:

			Animal 1		Animal 2	
			Hap 1	Hap 2	Hap 1	Hap 2
			12222111	11122111	12222111	11122211
Animal 1	Hap 1	12222111	1.00			
	Hap 2	11122111	0.30	1.00		
Animal 2	Hap 1	12222111	**0.90**	0.30	1.00	
	Hap 2	11122211	0.20	0.40	0.20	1.00

The probability that the two identical haplotypes (animal 1 haplotype 1 and animal 2 haplotype 2) in the IBD matrix would be very similar to the ρ coefficient for 8 markers from the above simulation.

3.4.2 Variance component model

To estimate the additive genetic variance, we could calculate the extent of the correlation between animals with high additive genetic relationships A_{ij}. In practise, we fit a linear model which includes additive genetic value (u) with V(u) = $A\sigma_u^2$, and then estimate σ_u^2. In a similar way, to estimate the QTL variance we fit the following linear model:

$$Y = Xb + Zu + Wv + e ,$$

where Y is a vector of phenotypic observations, $\boldsymbol{X}$ is a design matrix allocating phenotypes to fixed effects, b is a vector of fixed effect, $\boldsymbol{Z}$ is a design matrix relating animals to phenotypes, u *is* a vector of additive polygenic effects, $\boldsymbol{W}$ is a design matrix relating phenotypic records to QTL alleles, v is a vector of additive QTL effects, e the residual vector. The random effects u, v, and e are assumed to be distributed as follows: $u \sim (0, A\sigma_u^2)$, $v \sim (0, G\sigma_v^2)$, $e \sim (0, \sigma_e^2 I)$, 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 $\boldsymbol{G}$ is a matrix whose element G_{ij} is the probability haplotypes i and j carry the same QTL allele (eg. are IBD at the putative QTL position).

The precision with which the QTL variance, σ_v^2, is estimated will depend on both the number of unique haplotypes sampled from the population, and the number of observations per unique haplotype. The number of unique haplotypes sampled must be large enough to be representative of the population, while the number of observations per unique haplotype determines the accuracy of estimating the haplotype effects. It is important to make a distinction here between the total number of haplotypes in the population, which will be 2 x the number of animals genotyped, and the number of unique haplotypes in the population, which is the number of different combinations of marker alleles that are present in the population. A unique haplotype can be represented many times in the population. If the marker haplotypes are to be used in MAS, the accuracy of estimating the effect of a unique haplotype will determine the amount of improvement in the accuracy of selection as a result of using the marker information.

3.5 Combined LD-LA mapping

Authors investigating the extent of LD in both cattle and sheep were somewhat surprised/alarmed to find not only was LD highly variable across any particular chromosome, but there was even significant LD between markers which were not even on the same chromosome! (Farnir et al 2002, McRae et al. 2002). These authors (and others) have suggested that LD information be combined with linkage information to filter away any spurious LD likelihood peaks. This type of QTL mapping is referred to as LDLA, for linkage disequilibrium linkage analysis.

3.5.1 IBD matrix for LDLA mapping (Meuwissen et al. 2002).

The IBD matrix for LDLA mapping will have two parts, a sub-matrix describing IBD coefficients between the haplotypes of founder animals, and a second matrix describing the transmission of QTL alleles from the founders to later generations of genotyped animals.

So for example, if we have a half sib design, we will have two haplotypes per sire, a paternal haplotype for each progeny (the one he or she inherited from dad) and a maternal haplotype from each progeny (the marker alleles the progeny did not get from dad, so must have received from mum). The sire haplotypes and the maternal haplotypes of progeny provide the LD information, and the paternal haplotypes of progeny provide the linkage information. Table 3.2, from Meuwissen et al. (2002), describes the IBD matrix for LDLA for a half-sib design.

Table 3.2. The IBD matrix

	SH	***MHP***	***PHP***
SH	***[a]***	***[a]***	***[b]***
MHP	***[a]***	***[a]***	***[b]***
PHP	***[b]***	***[b]***	***[b]***

SH: sire haplotypes; MHP: maternal haplotypes of progeny; PHP: paternal haplotypes of progeny; [a] is calculated by the method of Meuwissen and Goddard (2001); [b] is calculated by the method of Meuwissen et al. (2002).

The calculation of blocks [a] is described in Meuwissen and Goddard (2001) (and above). The calculation of blocks [b] was described in Meuwissen et al. (2002), and are very similar to the standard linkage analysis calculations (eg. Fernando and Grossman 1989). Briefly, element of blocks [b] are $P_{IBD}(X(p);Y) = r \times P_{IBD}(S(p);Y) + (1-r) \times P_{IBD}(S(m);Y)$, where

- $P_{IBD}(X(p);Y)$ is the IBD probability of the paternal QTL allele of progeny X, X(p), with any other QTL allele, Y.
- S(p) and S(m) are the paternal and maternal alleles of sire S, respectively.
- r or (1-r) is the probability that the progeny inherited the paternal or maternal QTL allele of the sire.

3.5.2 Variance component model

The variance component model for LDLA is similar to that for complex pedigrees. The model can be written as:

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

where Y is a vector of observed phenotypes, μ the overall mean, u the vector of random polygenic effects, v the vector of random haplotype effects, e the residual vector, Z a design matrix relating the phenotype records to polygenic effects, and W a design matrix relating phenotype records to QTL alleles. 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, G is the IBD matrix described in Table 3.2 above.

3.5.3 Example of the twinning QTL

The power of combining LD and LA information to filter both spurious LD and spurious LA likelihood peaks was demonstrated in a study designed to map QTL for twinning rate in Norwegian dairy cattle (Meuwissen et al 2002). Figure 3.6A is the likelihood profile from linkage only, Figure 3.6B the likelihood profile from LD analysis only, and Figure 3.6C the likelihood profile from combined LDLA.

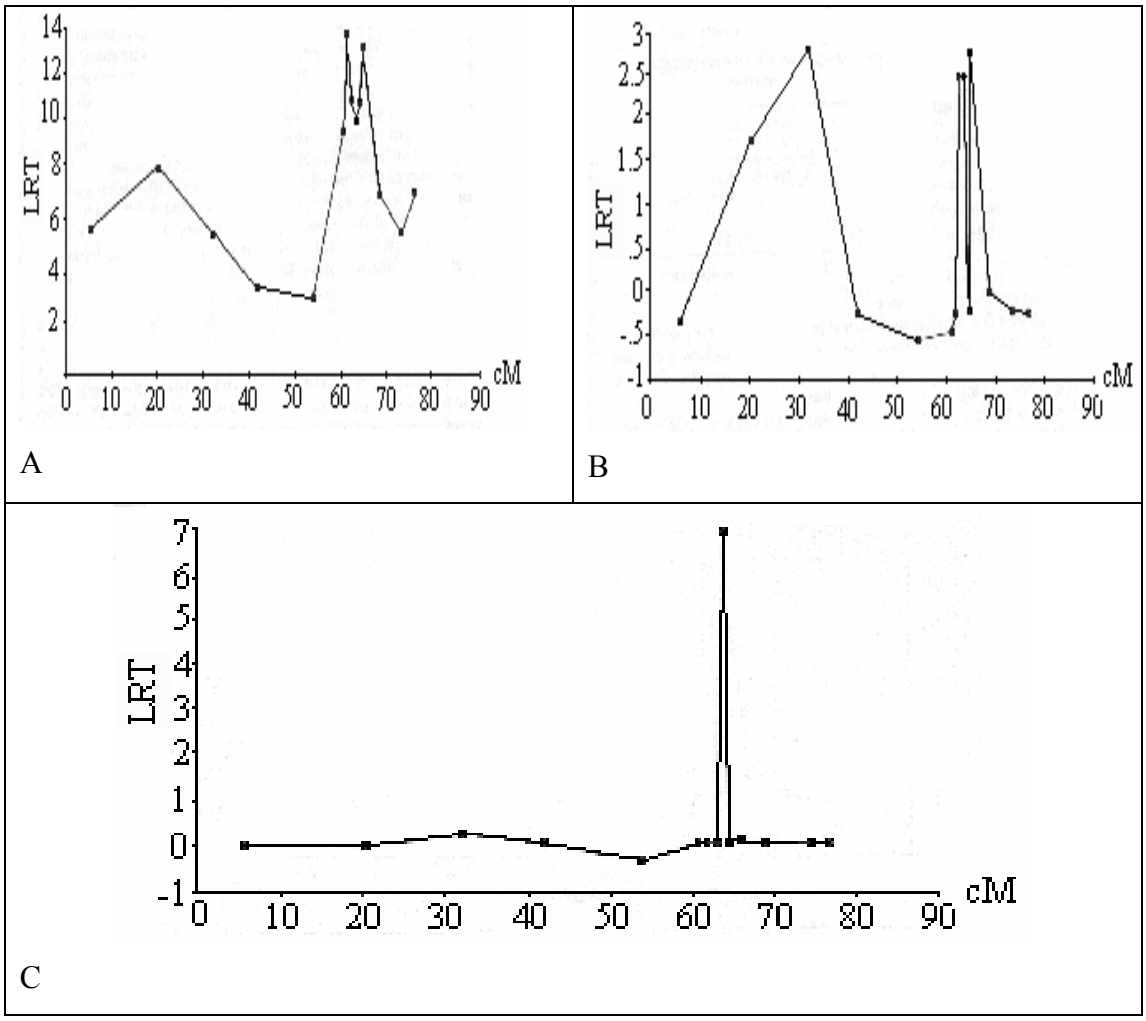


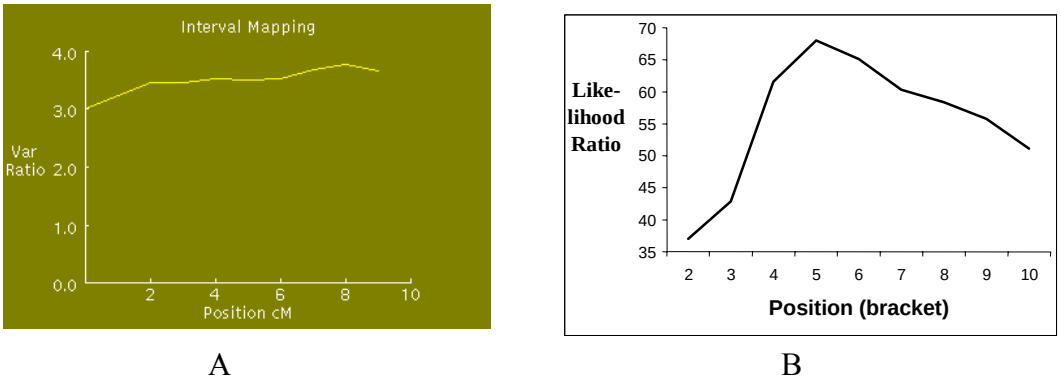
Figure 3.6. Likelihood profile from linkage analysis (A), Linkage disequilibrium analysis (B) and combined linkage disequilibrium linkage analysis (C) of marker data on chromosome 5 and twinning rate phenotypes in Norwegian dairy cattle. Meuwissen et al. (2002). Reproduced with permission from the authors.

When LDLA is performed, both linkage and linkage disequilibrium information contribute to the likelihood profile. Any peaks due to LD or linkage alone are filtered from the profile. Using LDLA, Meuwissen et al. (2002) were able to map the QTL for twinning rate to a 1cM region.

3.5.4 How much information does LD add to an LDLA analysis?

The amount of information the LD part of the LDLA analysis depends on the extent of LD. If LD extends only a fraction of a cM, as in humans, very dense markers will be required before there is any LD information in the analysis. On the other hand, if LD

extends a couple of cM, as appears to be the case in livestock, existing microsatellite maps may be dense enough to contribute some LD information. In a simulation study in pigs, with an N_e of 100 and a marker every cM, the LD information dramatically narrowed the confidence region around the QTL, Figure 3.7.



A B

Figure 3.7 A. Likelihood profile from linkage analysis only, and B., Likelihood profile from combined linkage disequilibrium linkage analysis.

In Figure 3.7A, only LA information is used to map the QTL. While the QTL is significant, the likelihood profile is very broad. When the LD information is added, the QTL is significant, but the likelihood declines rapidly as one moves away from the true QTL position (5cM).

3.5.5 Design of LD-LA experiments.

There are two design issues with LD-LA analysis. One is the density of markers required, which has already been discussed. The other is the population structure and size of experiment that is appropriate for LDLA.

An important question is 'are the large half-sib families we use in linkage analysis also suitable for LDLA analysis'? Large half sib families are of course suitable for linkage studies. LD on the other hand is a population based method (eg. the association between QTL and marker haplotype must persist across the population to be detected). To maximise the LD information, a large number of different haplotypes must be sampled, and there must be sufficient records per haplotype to estimate the effects of each

haplotype accurately. In a half sib design, the total number of founder haplotypes sampled from the population will be the number of dams (the maternal haplotype for each progeny) plus twice the number of sires (two haplotypes per sire). The number of unique haplotypes in this sample will depend on the length of the chromosome segment and the number of markers. If the markers are all in a small interval (say a few cM) the number of unique haplotypes may be small (due to LD), and there will be a considerable number of records per unique haplotype. If on the other hand the markers are widely spaced and cover the whole chromosome, there will be almost as many unique haplotypes as haplotypes sampled. In this situation only the effect of haplotypes carried by the sires are estimated with any accuracy.

Results from a simulation with Ne=100, 1 marker per cM, and varying number of half sib families, show the accuracy of LDLA (in positioning the QTL) is increased slightly by increasing the number of half sib families, Figure 3.8, but not by a great deal (Lee, S. in preparation).

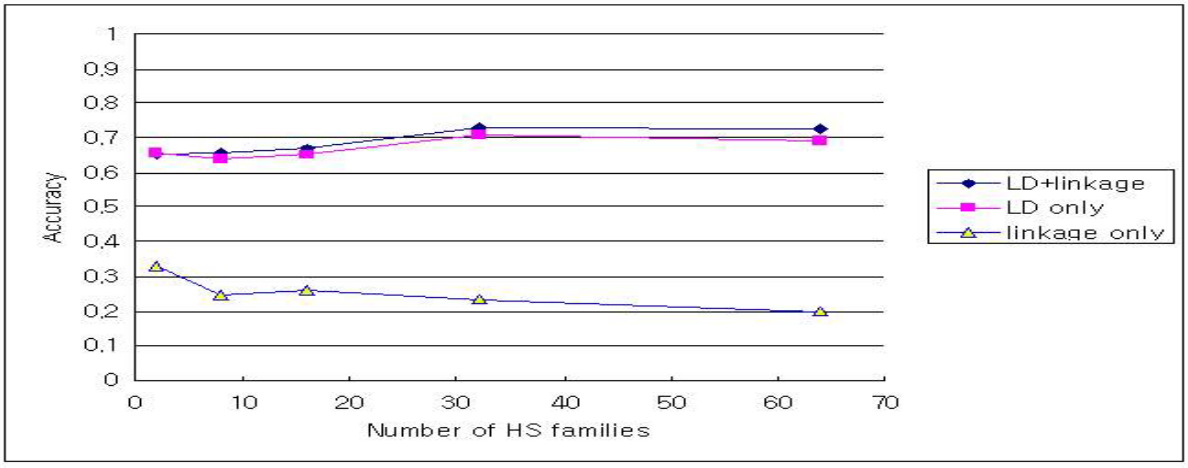


Figure 3.8 Accuracy of positioning a QTL (percentage of replicates positioning QTL in correct 1cM bracket) within a 10cM interval, with an increasing number of half sib families, 128 animals in each design. Linkage, linkage disequilibrium or combined linkage disequilibrium linkage analysis were used to position the QTL (Figure kindly provided by S. Lee).

An interpretation of this result is that the dam haplotypes are providing considerable LD information, in the designs with a small number of sires. The implications are that the designs we currently use for linkage studies should also be suitable for LDLA studies. Of course, the marker density will have to be greatly increased for the LDLA studies.

This of requires more genotyping. Another good question is can we combine the advantages of LDLA analysis with selective genotyping, to come up with a relatively cheap but powerful experiment? A simulation study was conducted, with Ne=100, 10 markers in a 10cM interval containing a QTL, and either 15 sires mated to 200 dams, 30 sires mated to 100 dams or 60 sires mated to 50 dams each, and 10 progeny per dam (so the total number of progeny in each design was 3000). Selective genotyping was conducted such that 10% of the highest phenotype and 10% of the lowest phenotype progeny were genotyped in each family (600 progeny genotyped total). The results (Table 3.3) indicate some loss of power with selective genotyping, but still a relatively high probability of correctly positioning the QTL within a 3cM bracket.

Table 3.3 Precision of QTL position estimates from LDLA. For each strategy the first number is the proportion of the progeny genotyped (100 or 20, with the progeny with the highest 10% and lowest 10% of phenotypes genotyped within each family). The second number is the number of sires used to breed the resource population (15, 30 or 60). In each design there were 3000 progeny.

	Deviation (in 1cM bracket) of estimated from correct position				
	0	1	2	3	4
100%15	44	31	9	4	5
100%30	46	32	7	3	5
100%60	40	39	8	4	2
20%15	35	36	14	7	1
20%30	32	32	15	8	6
20%60	33	37	11	8	4

Without selective genotyping, there was not a great deal of difference in the accuracy of the three designs. When selectively genotyping was implemented, only 20% of the progeny population, the 15-sire design was most accurate in estimating the QTL position. The 30-sire and 60-sire designs may have lost some linkage information during selective genotyping, resulting in less precise estimation of the QTL position.

This experiment illustrated that accurate positioning of QTL is possible with relatively few genotypings (600 progeny) by combining selective genotyping and LDLA analysis.

4. Bioinformatics approaches to selecting candidate genes

The results of the LDLA simulation studies in the last section indicate it should be feasible to narrow the QTL confidence region to approximately 3cM, given large half sib families and sufficient marker density (eg. 1 marker per cM). While this is much smaller than the confidence region from a typical linkage analysis (10-50 cM confidence region), a 3cM region contains on average approximately 90 genes (Bovenhuis and Meuwissen 1996). Relatively few genes have been mapped to bovine chromosomes, so the identity of these genes will be elusive. Additionally, searching for the mutation causing the QTL effect in each of these 90 genes is going to be extremely expensive (for example sequencing the sires heterozygous for the QTL for each of the 90 genes). Are there any approaches we can take to narrow down the list of candidates? The answer is yes, and the human genome project in particular has provided a wealth of information for this purpose. A possible approach is outlined in Figure 4.1

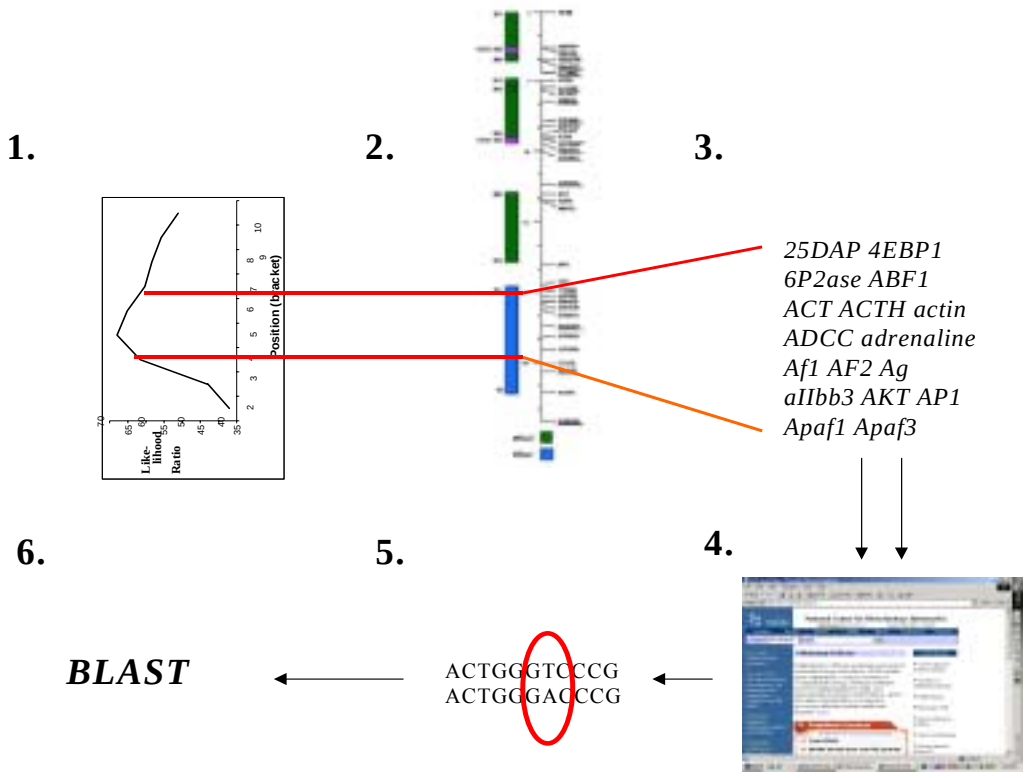


Figure 4. 1. A six step process for gene discovery. 1. The 95% confidence interval for QTL position from a combined linkage disequilibrium linkage analysis is calculated. 2. The comparative region of human chromosome corresponding to the confidence region on the bovine chromosome is located using a radiation hybrid map. 3. Using human genome project (HUGO) resources, a list of genes in this region is compiled. 4. These candidates are screened through a HUGO data base of literature based on key words (eg. lactation), and the most likely candidate(s) chosen. 5. A sire heterozygous for the QTL is sequenced for the candidate gene, and any single nucleotide polymorphisms (SNP) are assessed as the possible causative mutation in the population. 6. A BLAST search is used to assess if the SNP mutation is functional (eg. alters the protein).

4.1 Comparative mapping with humans, mice

While a few genes have been positioned on the livestock genetic maps (eg. http://www.thearkdb.org/anubis), the number is minute compared with the number of genes located on human chromosomes following the human genome project (HUGO). As the genes carried by humans and mammalian livestock species in particular are at least 98% conserved, one strategy is to use comparative mapping to find the piece of human chromosome that the bovine/livestock chromosome carrying the QTL region corresponds to, and then consider the genes in this region as candidates. There are a number of methods of comparative mapping, including radiation hybrid panels and chromosome painting. Luckily the institutes conducting such studies usually publish their results on the web, in a easily viewed format, so we don't have to do the experiments ourselves! Table 4.1 lists websites with good information from livestock species-human comparative mapping experiments.

Table 4.1. Genomic resources for the major livestock species.

Function	Species	Website
Microsatellite markers	Pigs	http://www.genome.iastate.edu/maps/marcmap.html
	Cattle*	Http://www.projects.roslin.ac.uk/cdiv/markers.html Http://www.marc.usda.gov/genome/htmls/LinkageMap.jsp?Species=bos&Chromosome=1 http://www.thearkdb.org/anubis
	Sheep*	http://www.thearkdb.org/anubis http://www.marc.usda.gov/genome/htmls/LinkageMap.jsp?Species=sheep&Chromosome=1
	Chickens	http://www.genome.iastate.edu/chickmap/
	Salmon	http://www.thearkdb.org/anubis
Genes mapped to chromosomes	Pigs	http://www.toulouse.inra.fr/lgc/pig/cyto/cyto.htm
	Cattle	http://locus.jouy.inra.fr/cgi-bin/lgbc/mapping/common/npremapping_loci.operl?BASE=cattle
	Sheep	http://www.nal.usda.gov/ttic/tektran/data/000009/52/0000095242.html http://www.nal.usda.gov/ttic/tektran/data/000011/49/0000114997.html
Radiation hybrid maps/Comparative maps	Pigs	http://abcenter.coafes.umn.edu/RHmaps/
	Cattle	Http://locus.jouy.inra.fr/cgi-bin/lgbc/mapping/common/Req_segment.pl?BASE=cattle Http://bos.cvm.tamu.edu/htmls/rhbta.html
	Sheep	Http://www.nal.usda.gov/ttic/tektran/data/000009/52/0000095242.html
Human genome project resources and mouse knockouts	Humans/ Mice	Http://www.ncbi.nlm.nih.gov/ http://tbase.jax.org/ http://www.ncbi.nlm.nih.gov/BLAST/ Http://genome.ucsc.edu/

*Note that most cattle and sheep microsatellite markers will amplify in both cattle and sheep.

4.2 Selecting candidate genes in an interval.

Once the human chromosome segment (or segments if there is a break point) has been located, the genes in this segment or segment can be listed from human genome project databases (http://www.ncbi.nlm.nih.gov/, http://genome.ucsc.edu/) (step 3 in Figure 4.1). This list is likely to be very large, and further refinement of the list of candidates will be necessary. The human genome project data bases offer facilities to search the literature for references to the genes in the list of candidates. The literature references can be screened for those containing key words, for example fat metabolism if the QTL affects fat % in milk. The list of candidates can be narrowed by only selecting those with references containing the key words.

Mouse knockout databases are also a good source of information (eg. http://tbase.jax.org/). There is a mouse knockout (where the expression of the gene is silenced) for just about every interesting gene, and if profile of effects of the knockout match the pattern of your QTL effects, this gene could be an excellent candidate.

If your happy with your choice of candidate gene, the next step is to sequence (the sires heterozygous for the QTL) for the candidate gene. Any single nucleotide polymorphisms (SNP) which alter the protein produced (ie amino acid substitutions, stop codon insertions), can be then assessed as the possible causative mutation in the population. A BLAST search on the National Centre for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov/BLAST/) can be used to determine if the SNP is a functional mutation (changes the protein) or silent (does not change the protein). If the SNP can be demonstrated to be the causative mutation in multiple populations (eg. are all animals heterozygous for the QTL also heterozygous for the SNP), there is a strong case that the SNP is the causative mutation.

4.3 Gene discovery: The example of the Inverdale gene (Galloway et al. 2001, Proceedings AAABG)

In July 2000, AgResearch of New Zealand accounted the discovery of the genetic mutation responsible for the effects on reproduction seen in Inverdale sheep. A single

copy of the gene in heterozygous ewes increases ovulation rate by about one extra egg, and litter size by about 0.6 lambs per ewe per lambing. Homozygous ewes carrying two copies of the gene have small non-functional ovaries and are infertile.

A linkage study narrowed the location of the Inverdale gene to a 10cM region on the X chromosome. At this stage several candidate genes were selected, most notably GDF9 (Growth differentiation factor 9). This gene was shown to be involved in fertility in mice, and a GDF9 knockout mouse had a similar profile of effects to the Inverdale gene in sheep.

However GDF9 was not located on the X chromosome in humans, mice or sheep (the authors mapped the gene to chromosome 5 in sheep). The authors then chose another candidate, GDF9B, with similar sequence to GDF9. GDF9B did map to the X chromosome in humans and mice. The gene was sequenced in Inverdale sheep, and a point mutation was discovered which was thought to be involved in biological activity of the protein molecule.

As the authors point out, the gene mapping studies were greatly aided by a clear and unambiguous phenotype (infertility as opposed to ovulation rate), and this is not usually the case for quantitative traits such as milk production, were the phenotype is continuos. Nevertheless, the discovery of the Inverdale gene, and particularly the discovery of the DGAT1 gene, prove that it is possible to locate the genetic mutations which cause variation in traits important for livestock production.

Acknowledgments

The authors would like to acknowledge Prof Theo Meuwissen, Duc Lu, S Lee, Maxy Mariasegaram and Helen McPartlan for their contribution. Special thanks to the gene mapping group at Victorian Institute of Animal Science.

5 References

Bovenhuis, H. and Meuwissen, T. 1996. Course *Detection and Mapping of quantitative trait loci*, 16-19 April, University of New England, Armidale, NSW, Australia.

Churchill, G. A. and Doerge, R. W. 1994. Empirical threshold values for quantitative trait mapping. *Genetics* **138**:963-971.

Darvasi, A. and Soller, M. 1992. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and Applied Genetics* **85:** 353-359.

Darvasi, A. and Soller, M. 1994. Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. Genetics 138: 1365-1373.

Darvasi, A. and Soller, M. 1997. A simple method to calculate resolving power and confidence interval of QTL map location. *Behavior Genetics* **27:** 125-132.

Dunning, A.M., Durocher, F., Healey, C.S., Teare, M.D., McBride, S.E., Carlomagno, F., Xu, C.F., Dawson, E., Rhodes, S., Ueda, S., Lai, E., Luben, R.N., Van Rensburg, E.J., Mannermaa, A., Kataja, V., Rennart, G., Dunham, I., Purvis, I., Easton, D. and Ponder, B.A.J. 2000. The extent of linkage disequilibrium in four populations with distinct demographic histories. *American Journal of Human Genetics* **67:** 1544-1554.

Farnir, F., Coppieters, W., Arranz, J.J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D. and Georges, M. 2000. Extensive genome-wide linkage disequilibrium in cattle. *Genome Research* **10:** 220-227.

Farnir, F., Grisart, B., Coppieters, W., Riquet, J., Berzi, P., Cambisano, N., Karim, L., Mni, M., Moisio, S., Simon, P., Wagenaar, D., Vilkki, J. and Georges, M. 2002. Simultaneously mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: Revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics* **161:** 275-287.

Fernando, R. and Grossman, M. 1989. Marker assisted selection using best linear unbiased prediction. *Genetics Selection Evolution* **21:** 467-477.

Galloway, S. M., McNatty, K. M., Ritvos, O. and Davis, G. H. 2002. Inverdale: a case study in gene discovery. *Proc. Assoc. Anim. Breed. Genet.* 14:7-10.

George, A.W., Visscher, P.M. and Haley, C.S. 2000. Mapping quantitative trait loci in complex pedigrees: A two-step variance component approach. *Genetics* **156:** 2081-2092.

Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A.T., Sargent, L.S., Sorensen, A., Steele, M.R., Zhao, X., Womack, J.E. and Hoeschele,

I. 1995. Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139:** 907-920.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S.J. and Thompson, R. 2002. *ASReml user guide release 1.0*. VSN International Ltd, Hemel Hempstead, HP11ES, UK.

Goddard, M.E. 1991. Mapping genes for quantitative traits using linkage disequilibrium. *Genetics Selection Evolution* **23:** 131s-134s.

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P., Spelman, R., Georges, M. and Snell, R. 2001. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine *DGAT1* gene with major effect on milk yield and composition. *Genome Research* **12:** 222-231.

Haley, C.S. and Knott, S.A. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69:** 315-324.

Haley, C.S., Knott, S.A. and Elsen, J.M. 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136:** 1195-1207.

Hayes, B., Bowman, P.J. and Goddard, M.E. 2001. Optimum design of genome scans to detect quantitative trait loci in commercial pig populations. In *Proceedings of the 8th Biennial Conference of the Australasian Pig Science Association* (ed. Cranwell, P.D.), p. 143. Australasian Pig Science Association, Victoria, Australia.

Hayes, B. J. and Goddard, M.E. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution* **33:** 209-229.

Hayes, B. J. Visscher, P. E., McPartlan, H. and Goddard, M. E*. 2002. A novel multi-locus measure of linkage disequilibrium and it use to estimate past effective population size.* Genome Research *In press.*

Heath, S*. 1997. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models.* Am. J. Hum. Genet. *61:748-760.*

Hill, W. G. 1981. Estimation of effective population size from data on linkage disequilibrium. *Genetical Research* **38:** 209--216.

Hill, W. G. and Robertson, A. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38:226-231.

Ihara, N., Takasuga, A., Mizoshita K. et al. 2002. Mapping of over 1100 bovine polymorphic micro-satellite markers to the USDA-MARC cattle linkage map. *Proc. Int. Soc. Animal Genet*, D-160.

Kinghorn, B. P. Kennedy, B. W. and Smith C. 1993. A Method of Screening for Genes of Major Effect. *Genetics* **134**: 351-360

Kruglyak, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22:** 139-144.

Lander, E.S. and Botstein, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121:** 185-199.

Lander, E.S. and Schork, N.J. 1994. Genetic dissection of complex traits. *Science* **265:** 2037-2048.

Lipkin, E., Mosig, M. O., Darvasi, A., Ezra, E., Shalom, A., Friedman, A. and Soller, M. 1998. Quantitative trait locus mapping in dairy cattle by means of selective milk DNA pooling using dinucleotide microsatellite markers: analysis of milk protein percentage. *Genetics* 149:1557-1567.

Mangin, B., Goffinet, B. and Rebai, A. 1994. Constructing confidence intervals for QTL location. *Genetics* **138:** 1301-1308.

McRae, A.F., McEwan, J.C., Dodds, K.G., Wilson, T., Crawford, A.M. and Slate, J. 2001. Linkage disequilibrium in domestic sheep. *Genetics* **160:** 1113-1122.

Meuwissen, T.H.E. and Goddard, M.E. 1996. The use of marker haplotypes in animal breeding schemes. *Genetics Selection Evolution* **28:** 161-176.

Meuwissen, T.H.E. and Goddard, M.E. 2000. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155:** 421-430.

Meuwissen, T.H.E. and Goddard, M.E. 2001. Prrediction of identity by descent probabilities from marker-haplotypes. *Genetics Selection Evolution* **33:** 605-634.

Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157:** 1819-1829.

Meuwissen, T.H.E., Karlsen, A., Lien, S., Olsaker, I. and Goddard. M.E. 2002. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* **161:** 373-379.

Moffat, M. F., Traherne, J. A., Abcasis, G. R. and Cookson, W. O. C. M. 2000. Single nucleotide polymorphism and linkage disequilibrium within the TCR α/δ locus. *Human Molecular Genetics* **9:** 1011-1019.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjlan, R., Farhadian, S.F., Ward, R. and Lander, E.S. 2001. Linkage disequilibrium in the human genome. *Nature* **411:** 199-204.

Riquet, J., Coppieters, W., Cambisano, N., Arranz, J.J., Berzi, P., Davis, S.K., Grisart, B., Farnir, F., Karim, L., Mni, M., Simon, P., Taylor, J.F., Vanmanshoven, P., Wagenaar, D., Womack, J.E. and Georges, M. 1999. Fine-mapping of quantitative trait loci by identity by descent in outbred populations: Application to milk production in dairy cattle. *Genetics* **96:** 9252-9257.

Seaton, G., Haley, C., Knott, S., Kearsey, M. and Visscher P. 2002. *QTL express: http://qtl.cap.ed.ac.uk*

Sved, J.A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2:** 125-141.

Tier, B. and Henshall, J. 2001. A sampling algorithm for segregation analysis. *Genet. Sel. Evol.* 33:587-603.

Visscher, P.M., Haley, C.S., Heath S.C., Muir W.J. and Blackwood, D.H.R. 1999. Detecting QTLs for uni and bipolar disorder using a variance component method. *Psychiatric Genetics* **9**: 75-84.

Visscher, P.M., Thompson, R. and Haley, C.S. 1996. Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143:** 1013-1020.

Weller, J.I., Kashi, Y. and Soller, M. 1990. Power of "daughter" and "granddaughter" designs for genetic mapping of quantitative traits in diary cattle using genetic markers. *Journal of Dairy Science* **73:** 2525-2537.

Weller, J. I., Song, J. Z., Heyen, D. W., Lewin, H. A. and Ron, M. 1998. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* **150**:1699-1706.

Zeng, Z.B. 1994. Precision mapping of quantitative trait loci. *Genetics* **136:** 1457-1486.

Zou, F. 2001. Efficient and robust statistical methodologies for quantitative trait loci analysis. PhD dissertation. University of Wisconsin – Madison, USA.