

Combining QTL and microarray data: the current state of play

Peter Baker^{1,2}

¹CSIRO Mathematical and Information Sciences, St Lucia, QLD

²NSW Agricultural Genomics Centre, Wagga Wagga, NSW

20 July, 2004

Abstract

In general, microarray experiments do not take into account other sources of information. For instance, molecular marker or phenotype data may also be recorded. Several articles have suggested that more insight might be gained by combining such data with microarray data. This article provides a review of current trends in combining such data as well as results of recent studies. Schadt et al (2003) coined the term 'eQTL' or expression QTL for studies where gene expression is considered to be a trait and QTL analysis is conducted. To efficiently combine such data it would seem either that simple but fast tests like single marker intersection tests or some form of dimension reduction, such as PCA or PLS, is required. In the latter approach, by reducing the space of gene expression levels, more realistic QTL models than are currently employed for eQTL studies may be adopted. If the latter approach is adopted, simulation studies indicate that the benefit of combining molecular and gene expression data increases as the number of cDNA levels affecting the trait and the size of the QTL effect decreases and when genes are co-expressed. Also, it is clear from several eQTL studies that considering a single gene at a time may be inappropriate due to epistatic interactions.

1 Background

Currently, there is a growing interest in the possible benefits of combining quantitative trait loci (QTL) linkage studies with microarray data. Many reviews have appeared recently but these are often so general that specifics are lacking (Gibson and Mackay, 2002; Boake et al., 2002; Chao, 2002; Cheung and Spielman, 2002; Eggen and Hocquette, 2003; Gracey and Cossins, 2003; Liu, 2003). In their review, Jansen and Nap (2001) suggested mapping gene expression levels via QTL methodology. This was the approach adopted, either in crosses varying for the trait of interest or in human populations, by Brem et al. (2002); Steinmetz et al. (2002); Watts et al. (2002); Schadt et al. (2003) and Gladney et al. (2004).

QTL linkage studies are conducted in order to map a region or regions of the genome which affect a continuous or quantitative trait. In agriculture, as soon as markers linked to QTL are found for economically important traits, these markers can be used for selecting individuals in breeding programmes. In human studies, the aim is often to identify markers indicating disease susceptibility. Current techniques for measuring markers are usually relatively slow and laborious. Newer DNA technology, such as SNP or single nucleotide polymorphisms (Kwok, 2001b; Patil et al., 2001; Kwok, 2001a) or single feature polymorphisms or SFP (Hazen

and Kay, 2003), provide a high throughput alternative to traditional markers but these are not widely used at present.

On the other hand, microarrays offer the promise of high throughput parallel assessment of gene expression for large numbers of genes or cDNA fragments in different tissues and organisms (Schena et al., 1995). Instead of finding molecular markers to map regions on the chromosome, actual gene expression is measured at a particular time under appropriate conditions and related to traits of interest, often by comparing two groups differing for the trait.

Darvasi (2003) states that there is an “undeclared dispute among researchers who study complex traits . . . On one side are classical geneticists . . . on the other are proponents of gene expression analysis . . .”. Darvasi goes on to outline the possible advantages of combining these techniques over and above either technique alone. In addition to better correlating genetic and phenotypic information, such analyses may help define better the trait into genetically more homogeneous groups as well as helping to better identify candidate genes that affect a given trait.

Dumas et al. (2000) provide a study, not employing microarrays, to find QTL for gene expression in twenty rats from a cross of two recombinant inbred rat strains. Following immobilisation stress, five specific heat stress genes (*hsps*) mRNA expression levels were determined and related to 475 polymorphic markers. Two loci were found to be linked to differential expression of *hsps*. The locus on chromosome 7 explained 42% of the variation of *hsp* levels in response to stress.

For microarrays, Jansen and Nap (2001) introduced the idea of carrying out genome-wide gene-expression QTL analysis and coined the term ‘genetical genomics’. Brem et al. (2002) applied these ideas to budding yeast and Schadt et al. (2003) use them to study mice, maize and man. Schadt et al. also coined the term ‘eQTL’ for expression QTL.

In a similar vein, 34 candidate genes for ovariole number in *Drosophila melanogaster* were identified by Wayne and McIntyre (2002) by combining mapping and arraying; and Eaves et al. (2002) studied diabetes in mice using microarrays and crosses of mouse congenic strains.

Due to the very large number of QTL analyses performed, simple statistical QTL linkage techniques are usually employed. However, given the perceived inaccuracies of such approaches, Doerge (2002) argued that more complex techniques for QTL analysis will be necessary for eQTL studies and other functional genomics approaches. More recently, Chesler et al. (2003) and Wang et al. (2003) provide examples where more sophisticated statistical techniques are employed to analyse eQTL.

Finally, Perez-Enciso et al. (2003) provides a simulation study for analysing a binary disease outcome related to underlying cDNA levels by partial least squares (PLS). Results of these analyses were combined with molecular marker data. They argued that instead of conducting QTL linkage analyses for each individual gene expression that employing the linear combination obtained by PLS as the trait variate results in a better analysis.

This article is organised as follows. QTL mapping methods are briefly outlined in Section 2 and short review of elementary methods for QTL linkage analysis is provided in Appendix A. In Section 3, current techniques for combining molecular marker and gene expression data are outlined. Finally, these methods are discussed and suggestions for future research are outlined in Sections 4 and 5.

2 QTL mapping methods

Quantitative trait loci (QTL) are regions on the chromosome that affect a quantitative or continuous trait. The main agricultural applications of QTL are to improve selection efficiency in breeding important plants or animals. In medical and veterinary research, finding QTL facilitate opportunities in improving medical diagnosis or disease management.

In the early 1990's, with the advent of good molecular maps for many species, several successes were achieved in finding molecular makers linked to human diseases such as cystic fibrosis, Huntington's disease and familial dysautonomia (primary references cited in Doerge et al., 1997). Since these diseases appear to be monogenic, the relationship between molecular maker and susceptibility was clear cut and, in general, there was little ambiguity over which individuals were susceptible. On the other hand, many complex traits of interest appear to be polygenic and so more complex genetic models and statistical methods are required. It should be noted that, in nearly all cases, estimation of QTL parameters is not trivial (Weller, 2001).

Classical genetic theory of quantitative continuously distributed traits, due to Fisher (1918), states that a very large number of genetic determinants each contribute a small amount to the value of the trait. Such a model poses a much greater challenge than simple Mendelian traits, especially in humans or other natural (outbred) populations (McPeck, 2000). QTL studies in model or agricultural organisms such as *Arabidopsis* or mice have a tremendous advantage, in that breeding can be controlled. In some cases, such as in Doubled-Haploid studies where purely homozygous are produced by chemical means, a further reduction in genetic variability may be achieved. In contrast to classical genetic theory, Hunt et al. (1995) found two QTL which accounted for 59% of the variation of the amount of pollen stored in honey bee colonies.

Lynch and Walsh (1998) outlines line-cross mapping and the types of designs that may be employed for animals and plants. Put simply, the idea behind these types of mapping designs is that linkage disequilibrium is created between the marker(s) and loci that differ between the lines, which in turn creates associations between marker loci and segregating QTLs. Various designs have advantages. For instance, Lynch and Walsh state that

“The F_2 design has an advantage over designs using backcross, RIL (recombinant inbred lines) or DHL (doubled-haploid lines), . . . which allows the the estimation of the degree of dominance of detected QTLs.”

but point out that various designs or linkage analysis methods have advantages and disadvantages and need to be chosen for the question at hand.

In general, there are two common statistical approaches for detecting and/or locating QTL. Both approaches involve moving along the chromosome and considering data for one or several markers at a time and relating these to the traits of interest.

The first approach is based on information from a single DNA marker, where trait data are related to different marker classes via ANOVA or simple linear regression. In this approach, a genetic map is not necessary even though the genetic distance of the QTL to the marker may be estimated for some types of line-cross design.

The second, more realistic but complicated approach, is based on two or sometimes more DNA markers and their estimated genetic map positions. Either a likelihood based or regression method incorporating the estimated recombination fraction (estimated genetic distance) between the DNA markers and putative QTL position is employed to estimate the size of the QTL effect and the QTL location. These models can be extended to incorporate more than one QTL and interactions between QTL. In general, these methods are considered to provide a

more precise means of detecting QTL and estimating their location and effects than the single marker methods.

Many practical considerations, such as missing markers, non-informative markers, linkage phase determination and segregation ratio distortion also need to be taken into account.

Appendix A provides a brief overview of genetic markers, maps and elementary methods for QTL linkage analysis. Due to the large literature available, the review is necessarily of limited scope. The emphasis is on agricultural species rather than human studies, for which there is also a large QTL literature.

3 Current techniques for combining mapping and arraying

3.1 Genetical Genomics

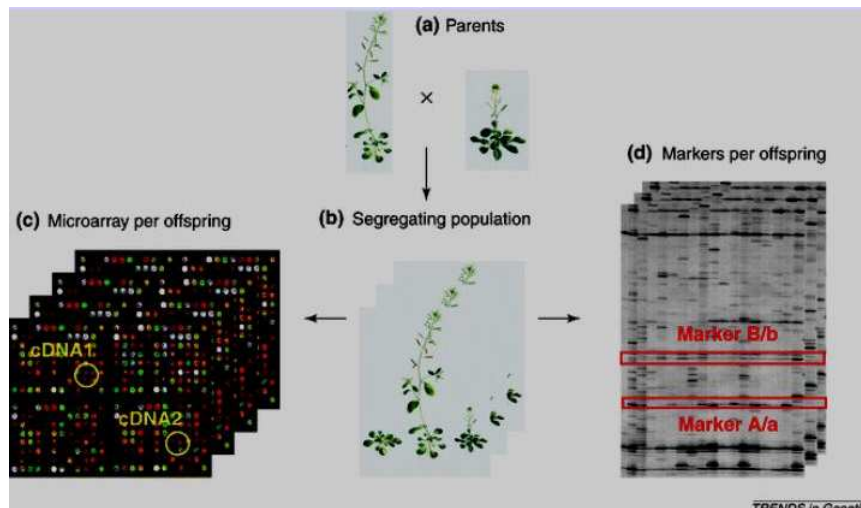


Figure 1: Expression profiling in combination with molecular marker analysis of a hypothetical segregating population of *Arabidopsis*. Parental lines (a) are crossed to form any type of segregating population, such as F₂ or RIL (b). Each individual in (b) is then employed for expression profiling (c) and molecular marker analysis (d). **Source: Jansen and Nap (2001)**

Jansen and Nap (2001) coined the term 'genetical genomics' when they outlined ideas on combining molecular marker and microarray data for segregating populations. While not actually applying their methods to experimental data, they argued that standard microarray experiments overlooked the power of genetics and that further insight may be gained by incorporating marker-based fingerprinting of each individual (see Figure 1) in a segregating population.

Any experimental design or study involving segregating lines would be suitable for such an approach. Jansen and Nap (2001) regard self-compatible plants such as *Arabidopsis* or maize as ideal, since a large pedigree of F₂, F₃, recombinant inbred lines (RIL) or near-isogenic lines (NIL) can easily be obtained.

To avoid measuring gene expression profiles on each individual, the simpler approach of bulked segregant analysis may be adopted. However, Jansen and Nap note that this type of analysis zooms in on preselected regions of the genome and hence genes with important

influence may be missed. This appears to be analogous with selective DNA pooling. Selective genotyping, where only individuals of the tails of the trait distribution are genotyped for molecular markers, is known to be efficient for mapping QTL (Darvasi and Soller, 1992; Muranty and Goffinet, 1997). However, when similar samples are pooled rather than measured individually, considerable power to detect QTL may be lost although this approach may still be more practical than conducting a study where all individuals are genotyped (Darvasi and Soller, 1994; Lipkin et al., 1998).

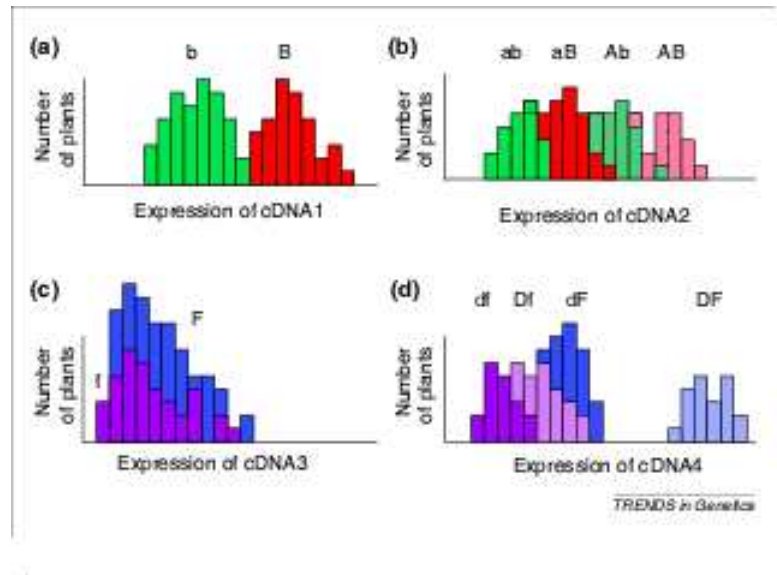


Figure 2: Combined graphical analysis of expression profiling in combination with molecular marker analysis of a hypothetical segregating population of *Arabidopsis*. 4 hypothetical cDNAs are presented for individuals with various marker types are differently coloured for 4 markers A/a, B/b, d/D and f/F. Distributions of expression levels are seen where (a) cDNA1 shows a clear qualitative distribution based on marker A/a (b) shows a quantitative distribution for cDNA2 which can be resolved by grouping the markers of A and B (c) cDNA3 cannot be resolved based on marker F and (d) cDNA4 can be resolved based markers of D and F. ANOVA may be used to ascertain differences between expression for different marker alleles. One way ANOVA would compare alleles for a single marker, two way ANOVA for two markers and so on. Note that (d) exhibits an epistatic interaction. **Source: Jansen and Nap (2001)**

In its simplest form, a combined analysis may take the form of plotting (see Figure 2). While some gene expressions for particular cDNA spots may be clearly different for some markers (Fig. 2 (a)) and hence appear to be qualitatively distributed, it is more likely that expression levels will be quantitatively distributed (Fig. 2 (b-d)). Indeed, Jansen and Nap note that than one marker may be required to determine expression groupings. Also, they suggested that multi way ANOVA may provide indications of epistatic interaction or of gene expressions influenced by more than one chromosomal region.

If cDNA spots can be mapped, then Jansen and Nap (2001) suggest that putative pathways may be deduced by the process outlined in Figure 3. Basically, the steps involve allocating cDNAs close together in the pathway if they are linked to the same markers. However, it is clear that an exact pathway is not easy to deduce by logic alone and in any case, uncertainty

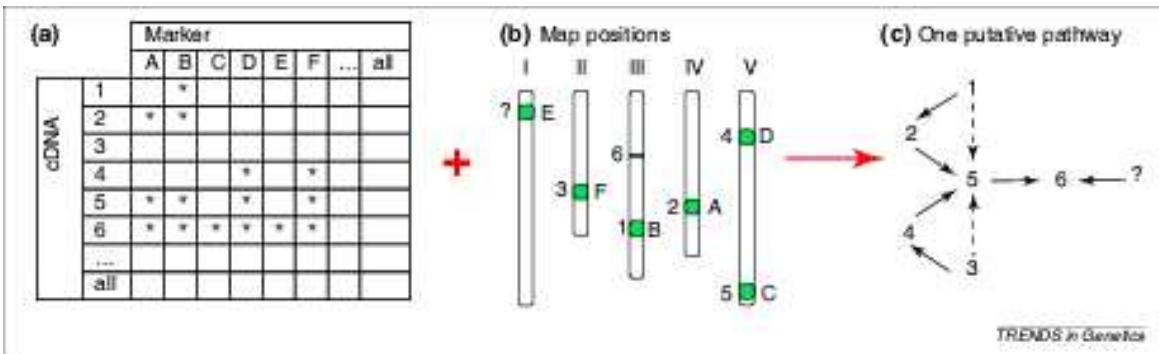


Figure 3: Hypothetical pathway reconstruction and pathway memory. (a) QTL analysis is employed to find significant linkages between molecular markers which are located as in (b). (c) A putative pathway is determined as follows: cDNA5 is mapped to markers A, B, D and F. Since cDNA6 maps to these markers and as well as C and E, and noting that cDNA5 is close to C, it is deduced that cDNA5 influences cDNA6. A similar line of reasoning can be used to deduce that cDNA2 and cDNA4 act on cDNA5; and that cDNA1 acts on cDNA2. **Source: Jansen and Nap (2001)**

in both QTL map position or gene expression levels can not be taken into account. Also, while the explanation provided by Jansen and Nap appears to be straight forward to apply to a small number of markers and cDNAs, it is not clear how well this would scale up to realistic numbers of cDNA and markers expected in such experiments.

Indeed, several authors Eaves et al. (2002); Wallace et al. (2002); Brem et al. (2002); Schadt et al. (2003) have attempted to combine microarray data with marker data by treating gene expression levels as phenotypes and conducting separate QTL analyses on each gene. However, in addition to the usual problem of assessing significance levels under multiple statistical tests, it is possible that gene-by-gene analysis may not be informative enough (Steinmetz et al., 2002; Perez-Enciso et al., 2003; Darvasi, 2003). This is partly due to the fact that these genes are often correlated and partly because many genes are expressed and regulated in concerted action. Also, in each study several hundred or more gene expression QTLs were found. The large number of eQTL can lead to difficulties in interpreting these eQTL biologically. It would also conceivably prove problematic to systematically deduce pathways by the method of Jansen and Nap (2001).

3.2 eQTL studies

Wallace et al. (2002) mapped 70 rat genes and ESTs from a set of differentially expressed genes obtained by comparing a spontaneously hypertensive rat strain and a normotensive control strain. The emphasis of this study was to physically map the differentially expressed genes and thus, using the mouse as a model, identify potential genes or previously unknown conserved chromosomal regions for insulin resistance or hypertension in humans. Radiation hybrid mapping (RHM) was employed since non-polymorphic DNA markers are not required. Instead, RHM is a physical mapping that makes use of the frequencies of X-ray induced DNA breakage to infer distances between markers (see <http://opbs.okstate.edu/~melcher/MG/MGW1/MG1228.html>). Of the 70 genes mapped, only 5 had previously been mapped in QTL studies and so many potential candidate genes were identified.

Brem et al. (2002) found 1528 differentially expressed genes and 570 eQTL when studying budding yeast. Of the 570 eQTL, approximately half or 308 genes were also differentially expressed. These loci fell largely into two categories: cis-acting modulators of single genes and trans-actors of many genes. They found eight trans-actor genes affecting the expression of 7 to 94 genes. Brem et al. concluded that most ($\approx 80\%$) expression profiles were affected by multiple loci, not just one or two. They also found that 185 of the 570 messages had levels which were linked within 10 kb of their own gene. In summary, their results indicated that even in a single cell organism, genetic variation in phenotypes was found to be highly complex and gene expression typically had a polygenic basis.

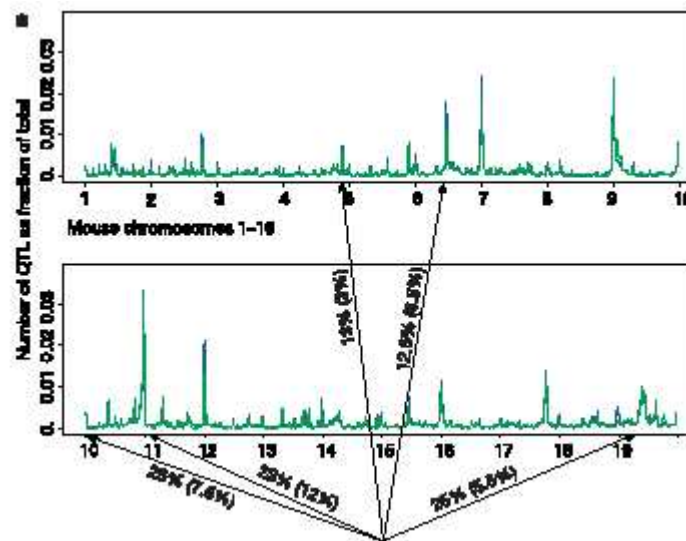


Figure 4: Percentage of eQTL as calculated by calculating LOD score thresholds across 920 evenly spaced bins, each 2cM wide, covering the mouse genome **Source: Schadt et al. (2003)**

Schadt et al. (2003) studied liver tissues in 111 F2 mice constructed from two inbred lines. Using standard interval mapping, of the 23,574 mouse genes on the microarray, 4,339 eQTL were found to have LOD scores over 4.3 while 784 had LODs over 7.0. Using the Celera Mouse Genome Database, 18,460 or 78% of the genes on the array could be mapped to a chromosome. Of the 7,681 differentially expressed genes, they found 2,123 eQTL or genes with significant QTL ($P < 0.00005$). eQTL hotspots were pinpointed by considering how many eQTLs might be expected by chance and highlighting regions containing significantly more eQTL. Hotspots were found on chromosomes 2,6,7,9,10,16 and 17 (See Figure 4). In order to avoid the problem of multiple testing, Schadt et al. chose very conservative p-values for assessing eQTL significance. However, identifying eQTL hotspots may be problematic. Darvasi (2003) and Perez-Enciso et al. (2003) point out that the significance of hot spots may be falsely inflated due to high correlations or interactions among sets of genes and because of the multiple testing employed.

It is interesting to note that 71% of mapped genes with strong evidence of linkage to eQTL ($\text{LOD} > 7$) had physical location close to the estimated eQTL position. A similar figure of 80% was found in *Zea mays* using 76 F3 crosses of maize. Only 34% of such mouse genes with $\text{LOD} > 4.3$ had a similar coincident position.

To cluster similar genes that may affect obesity, Schadt et al. then conducted two way

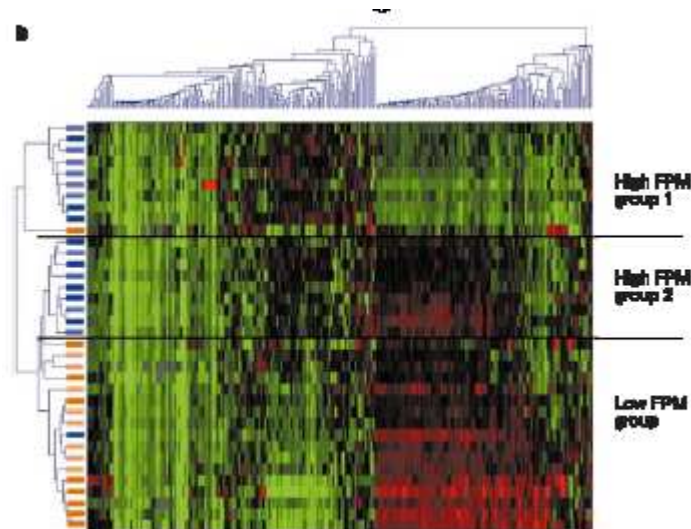


Figure 5: Two way clustering of gene expression levels by fat pad mass (FPM) in obese mice
Source: Schadt et al. (2003)

clustering on the 280 most differentially expressed genes by fat pad mass (FPM) which is a measure of obesity (Fig. 5). Some of the candidate genes on chromosome 2 produced by this clustering also appeared to be orthologues of genes in the homologous regions of chromosome 20 in humans. In general, the obesity trait FPM was found to be a complex trait in that it had 4 QTL regions associated with it and two distinct gene expression patterns associated with high FPM mice (Fig. 5). To add to this complexity, it is worth noting that 40% of genes with a LOD score over 3.0 had more than one eQTL and approximately 4% had more than three eQTL.

Both Brem et al. (2002) and Schadt et al. (2003) found that, in general, eQTL with high LOD score or high significance are observed to be cis-acting whereas moderately significant QTL are trans-acting. This is consistent with the expectation that first order effects (DNA variations affecting transcription of the gene itself) were easier to detect than second-order effects where genes act on the transcription of other genes.

Eaves et al. (2002) outline an approach of combining microarrays with mapping using congenic strains of mice. Congenic mapping may be employed to confirm significant linkage and subsequently for fine mapping (Consortium, 2003). Congenic strains are bred by introgressing the QTL interval or critical region into an inbred line for many, say 10, generations. The resultant congenic strain should then exhibit different phenotypes to the inbred strain but only be chromosomally different at the QTL region. Eaves et al. bred several congenic strains to have varying resistance to diabetes. By finding the chromosomal locations of differentially expressed genes and comparing these gene's expression levels in the relevant congenic and non-diabetic strains, they showed that it was possible to gain insight into the action of specific genes in relation to QTL regions affecting diabetes resistance. In particular, of the more than 400 gene expression differences, some of the highly or lowly expressed genes were related to particular marker alleles and hence different phenotypes. For the *ldd9.1* locus they found eight new candidate genes for further study. However, this was not the case in general and Eaves et al. indicate that more targeted experiments will be required to identify the genes and pathways for six other *ldd* loci conferring disease protection.

In a similar fashion, de Haan et al. (2002) studied recombinant inbred and congenic strains of mice and found three distinct clusters of eQTL co-localised for hematopoietic stem cell cycles. QTL analysis was carried out to map cell cycle phenotypes. Unlike several other studies, QTL analysis was not performed on gene expression levels but instead using sequence information, genes from three clusters of differentially expressed genes were physically mapped to the same approximate location as cell cycle phenotypes QTL on chromosome 11 plus two other regions. Using sequence information, de Haan et al. successfully assigned the chromosomal location of 156 out of 440 differentially expressed genes. The evidence is that these clustered genes are inherited collectively since these genes mapped to recombination “coldspots”,

Wayne and McIntyre (2002) describe an experiment to identify candidate genes for ovariole number in *Drosophila melanogaster*. To identify candidate genes, they fine mapped eQTL for ovariole number. However, as expected, this led to hundreds of potential candidate genes. To reduce this number, only significantly differentially expressed genes known to be located within the QTL regions were considered as candidate genes. The shortened list of 34 genes was obtained from the original 548.

Prows et al. (2003) studied inbred mice strains to assess genetic susceptibility to nickel induced lung injury. Human exposure is quite extensive and so the mouse was used as a model system. In addition to finding QTL for the phenotype or lung injury data, QTL were mapped for large cDNA gene expression levels. Results were combined in a qualitative way, noting that some significantly expressed genes mapped to similar regions as the nickel-induced injury QTL. Similarly, Okuda et al. (2004) studied 102 F2 mice bred from a salt-sensitive and insensitive parental lines to study genes related to hypertension. eQTL were assessed via ANOVA and simple correlations were undertaken to relate gene expression and blood pressure traits. While finding strong evidence for one gene on chromosome 10 and three other genes implicated in hypertension, the statistical analyses are very descriptive and appear to be poorly carried out. For instance, multiple testing was not taken into account in that all tests were conducted at the 0.05 level and genes were declared to be differentially expressed based on a 2.0 fold change.

3.3 eQTL studies employing complex QTL mapping

Doerge (2002) advocated the use of sophisticated statistical methods for QTL analysis, possibly on many traits simultaneously. However, methods for QTL mapping which employ more complicated models are also slower computing wise. Hence, some authors have employed dimension reduction methods (see Section 3.4) in order to employ more realistic but computationally expensive models while others have undertaken high throughput methods for obtaining markers.

An alternative method for complex QTL mapping was introduced by Steinmetz et al. (2002), namely reciprocal-hemizyosity analysis. Two strains of yeast, one with low (S288c) and other with high (YJM789) high-temperature-growth phenotypes were chosen for the study. Figure 7 shows results from the parental yeast lines and the hybrid YJM789/S288c which exhibited heterosis (hybrid vigour) in that at 41°. To identify DNA markers, total genomic DNAs from the two parental strains were hybridized separately to Affymetrix oligonucleotide arrays. A total of 3,444 biallelic markers were identified from probes that showed decreased signal strength in the YJM789 line. Genomic DNA from 19 of the 104 high lines were hybridized and recombinations identified by tracing signal levels. Simple probability calculations or t-tests, with no adjustment for multiple testing, were performed to find QTL linkages. Initial analysis revealed two regions, one on chromosome XIV and another

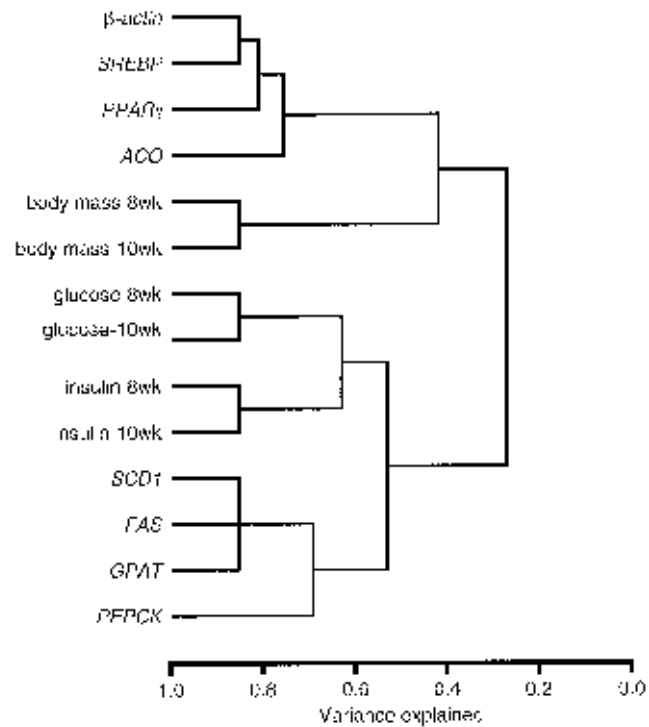


Figure 6: Hierarchical clustering of eight mRNA traits and eight phenotypic traits in an experiment to study F2 mice for type 2 diabetes **Source: Lan et al. (2003)**

on XVI.

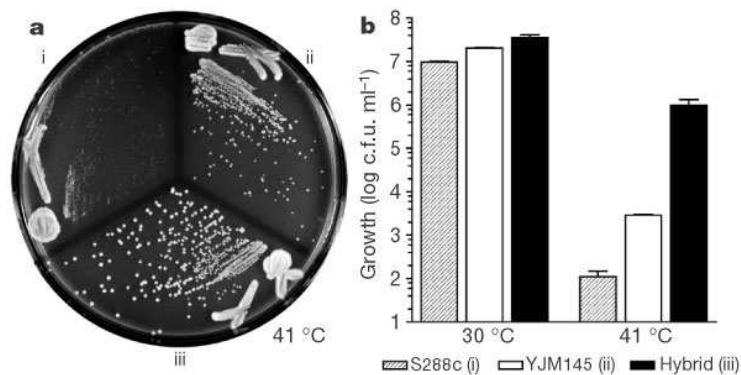


Figure 7: Average high-temperature-growth phenotype (Htg). a, Qualitative differences measured by colony size; b, quantitative differences in growth measured by competition assay after 48 h at 30 and 41 °C: (i) S1029, an Htg⁻ S288c strain, (ii) YAG040, an Htg⁺ YJM145 strain, and (iii) XHS123, a YJM145/S288c hybrid. Bars indicate s.e.m. (n = 6). **Source: Steinmetz et al. (2002)**

Multiple molecular markers were then developed for these regions. Chromosome XVI was studied in further detail via detailed sequence analysis on half a dozen Htg⁻ and Htg⁺ strains. Isogenic lines were then constructed for the 15 genes found and phenotypes subsequently compared. Three tightly linked QTL regions were found, two of them on Chromosome XVI. In essence, Steinmetz et al. conclude that these results demonstrate a deficiency in current QTL analysis approaches. However, their technique may not be very practical for other species and so this points to the need to extend current QTL linkage methods for mapping multiple loci to include epistatic interactions.

3.4 Studies reducing the dimensionality of the eQTL space

Instead of analysing each individual gene expression level to map eQTL, Lan et al. (2003) and Perez-Enciso et al. (2003) undertook the dimension reduction techniques of principal components analysis (PCA) and partial least squares (PLS), respectively.

Lan et al. (2003) describe an experiment using microarrays to map mRNA abundance as quantitative traits. They used either PCA, if no external grouping factor or trait such as disease status was available, or hierarchical clustering including the trait(s) if such trait(s) were available (see Figure 6). Once genes (mRNAs) were clustered, PCA was applied within a cluster in order to determine QTL for the principal components. They applied this technique to an F2 population derived previously to study QTL in obesity and diabetes. Once the first and second principal components were obtained, these were assessed for QTL linkage using a variety of sophisticated methods, some of which allowed for an arbitrary number of QTL. In addition to principal components analysis, multiple trait mapping was employed to map QTL for several genes simultaneously with similar results.

Specifically, Lan et al. applied multiple interval mapping (MIM: Kao et al. (1999)) and Bayesian interval mapping (BIM: Satagopan et al. (1996a)) for QTL discovery (see Figure 8). Also, no QTL were found for the principal components for genes (mRNA) clustered together with traits but which exhibited small correlation. In contrast, two QTL were found for the

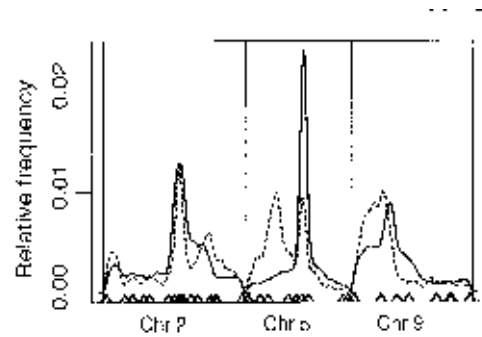


Figure 8: Bayesian interval mapping of *SCD1* mRNA abundance (solid line) and first principal component of all mRNAs (dashed line) using a model for three or more QTL. The method detects two QTL on chromosome 2 and one QTL on chrom. 9 **Source: Lan et al. (2003)**

principal components of several genes, highly correlated with insulin. These and other genes were subjected to extensive analysis which showed evidence of multiple QTL, complicated genetic architecture and possible epistatic interactions.

Perez-Enciso et al. (2003) proposed a method for analysing a binary disease outcome and relating this to both microarray and marker data. They consider the probability of disease to arise from an underlying continuous latent variable (liability). The liability is then related to cDNA levels by partial least squares (PLS). The estimated, rather than observed, liability is subsequently employed for QTL analysis. Perez-Enciso et al. concentrate on case-control studies but the method can be applied to any design. Their objective was to study the issue of whether or not cDNA can be used to refine genomic position estimates of genes that affect a complex trait, in this case disease.

Perez-Enciso et al. conducted a simulation study to explore the conditions under which combining microarray and marker data proved beneficial. Since simple assumptions, such as independent expression levels, may prove too unrealistic, They argue that simulating the underlying liability conditional on observed expression levels induces the more realistic correlations between gene expression levels. Different scenarios gave rise to the following conclusions:

1. Effectiveness of using microarray data for gene mapping increases when number of cDNA levels in the underlying liability and QTL effect decrease and when genes are co-expressed.
2. Correlation between estimated and true liability appears to be large under all conditions studied.
3. It is unlikely that cDNA identified as significant are actually those responsible.
4. The number of significant cDNA levels increases if cDNAs are co-expressed in a cluster.
5. Data reduction is necessary to smooth out the variability in expression levels when these are analysed individually. Indeed, by conducting PLS not only are the data smoothed but missing values also become less influential.

3.5 eQTL software tools

Software tools have recently become available for combining microarray, trait and marker data (Wang et al., 2003; Fischer et al., 2003). While these appear to be of limited value at this time, this may change in the near future.

Chesler et al. (2003) conducted microarray experiments using Affymetrix chips on female mice from 30 well studied recombinant inbred lines. The 12,422 gene expression levels, physical attribute data and previously assembled behavioral data from WEBQTL (Wang et al., 2003) were correlated using a Pearson correlation coefficient. Given the large number of tests of significance conducted, an unspecified false discovery rate was implemented. Traits and also gene expressions were also correlated with high density marker map available at WEBQTL to search for QTL. While interval mapping to better position QTL is also apparently possible, little detail is provided in Chesler et al. (2003) of how or whether this was actually carried out or whether particular trait QTL were already known. Genes with expression levels highly correlated to particular traits were then tested for QTL linkage at the same site as the trait or elsewhere.

In contrast to WEBQTL which provides analytic tools for QTL and microarrays, Fischer et al. (2003) provides software for visualising relationships between “linkage scores” for particular traits, which are presumably LODs, and gene expression levels. Some form of regression analysis is carried out but unfortunately, little detail is provided.

4 Discussion

Recent studies have advanced our knowledge about combining microarray, marker and other data. However, it is clear that these have several limitations.

Firstly, complex traits exhibit a complex genetic architecture (Steinmetz et al., 2002; Schadt et al., 2003; Perez-Enciso et al., 2003). Current QTL mapping methodologies usually focus on simple models such as a single QTL or multiple QTL without epistatic interaction. While this may be overcome by a somewhat complicated and detailed experimental followup in yeast, as was done by Steinmetz et al. (2002), this is unlikely to be practical in non-model organisms. Instead, it has been proposed that QTL methodology and software may need to be extended to incorporate more complex but realistic models (Doerge, 2002).

Given current computer hardware, computation for such models will generally prove to be infeasible for the large number of gene expressions under consideration. Coffman et al. (2003) carried out a simulation study to compare single marker intersection tests and interval mapping and found that single marker tests are, under some conditions, just as powerful as interval mapping but not in general. They concluded that performing both single and multiple marker analyses should be used to guide researchers in order to identify regions for further analysis. Unfortunately, this may prove infeasible for eQTL studies. Instead, Perez-Enciso et al. (2003) and Lan et al. (2003) suggest dimension reduction via PLS and PCA, respectively. Smoothing the data in such a way may prove beneficial or information may potentially be lost. Further study is required.

Secondly, an alternative attempt to reduce the number of genes under consideration by choosing a subset of genes for subsequent QTL analysis may give biased results. To cut down the number of expression traits to consider, de Haan et al. (2002) and Wayne and McIntyre (2002) only considered differentially expressed genes. However, Brem et al. (2002) found that 262 out of 570 significant eQTL for non-significantly differentially expressed genes. Interestingly, in the related subject of clustering similar genes, (Ambroise and McLachlan, 2002)

provide an example where leave one out cross-validation gives biased results for cancer classification and instead they advocate the .632 bootstrap estimate.

Thirdly, only a portion of cDNAs can be physically mapped to the chromosome. For instance, Schadt et al. (2003) found that 78% of 23,754 mouse genes could be mapped to the chromosome while de Haan et al. (2002) could only map 156 out of 440 differentially expressed genes for cell cycles in mice.

Fourthly, given the multivariate nature of the gene expression data, either multivariate techniques or multivariate QTL analysis (Jiang and Zeng, 1995; Henshall and Goddard, 1999; Kao et al., 1999), may be appropriate. Of course, these methods are essentially linear and were developed for small numbers of complex traits and so they will probably not scale up to the analysis of thousands of eQTL.

Finally, finding QTL for gene expression is only part of the story. Ideally, the aim is to find networks of genes or pathways affecting phenotypes or complex traits. While Jansen and Nap (2001) provide a simple method for deducing pathways, it is not clear how this would scale up to larger datasets. Alternative approaches are being developed. For instance, Dobra et al. (2004) used Bayesian regression analysis and large-scale, but sparse, graphical models to generate joint distributions of genes where each gene has a small number of neighbors. Thus Dobra et al. model, not causal gene networks, but associations to form observed gene networks. While this approach is in essence exploratory, it represents a step forward in modelling such networks and pathways. However, any approach will need to be augmented by biological interpretation through integration of biological and literature data,

5 Conclusion

There is a growing interest in combining microarray and other sources of data, such as marker data. The jury is still out on the benefits of combining such data, over and above simply using microarray or marker data alone.

However, little data has been produced thus far. It is likely that benefits will be gained by combining such data when it becomes available but further research is required. This will involve, not only large and expensive microarray studies of segregating or mapping populations, but also the development of new and sophisticated analytical methodologies to interpret the large amounts of complex data from such studies.

A QTL Linkage Analysis

Several excellent review articles and books are available.

Firstly, Kearsey and Pooni (1996), which is written as a self-instructional text, provides a good introduction to quantitative genetics and QTL studies. In addition to outlining elementary QTL linkage analysis, Kearsey and Pooni also contains descriptions of molecular markers, maps and mating designs. Lynch and Walsh (1998) provides a more comprehensive and mathematical account of QTL linkage analysis and quantitative genetics. Corrections and draft chapters of a second volume *Evolution and Selection of Quantitative Traits* may be found at <http://nitro.biosci.arizona.edu/zbook/book.html>. Two other books, namely Liu (1998) and Weller (2001), while factually correct and comprehensive, are somewhat terse and can not recommended as introductory textbooks. However, these may prove very useful for reference purposes.

Doerge et al. (1997); Jansen (2001) and Doerge (2002) provide good reviews of statistical aspects of methodology for QTL linkage analysis.

A.1 Genetic Markers and Maps

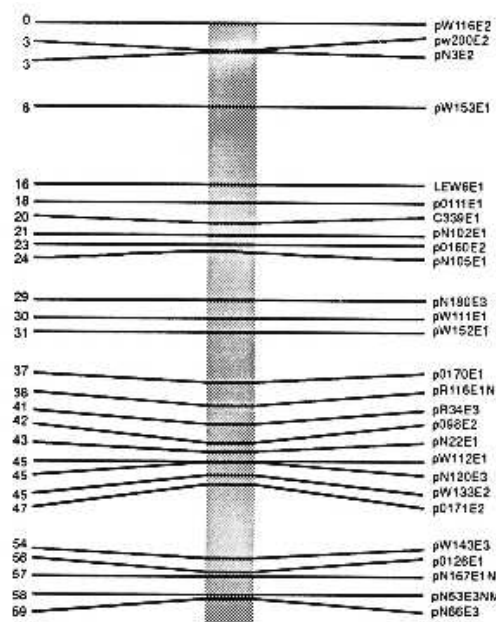


Figure 9: A typical molecular marker map. This example is for about a third of chromosome 3 of *Brassica oleracea*. Source: Kearsey and Pooni (1996).

Genetic maps consist of a series of markers or identifiable features at known, or perhaps best described as estimated, locations on the genome (see Figure 9).

For some discrete traits, simple Mendelian inheritance is followed and the phenotype has a one to one correspondence with the genes controlling it. These are so called morphological markers, which were then related to continuous or quantitative traits of interest. Examples are shape, colour, size or height in particular varieties of peas, as studied by Mendel. For another example, see Appendix A.2.

Prior to 1980, markers were predominantly morphological or physical. Following this time, a huge explosion in the number and type of markers occurred. Since 1980, advances in PCR technology enabled the invention of many types of DNA markers including RFLPs, minisatellites, RAPDs, AFLPs, microsatellites, STSs, ESTs and others. For details see Liu (1998) or Kearsey and Pooni (1996).

Genetic map distances are not measured but estimated. In agricultural species, often these maps are estimated by observing recombinations in offspring of designed experiments employing segregating populations. The markers can then be assigned to a chromosome and positioned (albeit with error) on the chromosomes. To obtain estimated distances the estimated recombination frequencies, which are not additive are converted in centiMorgan (cM) distances via a mapping function such as those of Haldane or Kosambi. For further detail see Kearsey and Pooni (1996, Chapter 6) or Lynch and Walsh (1998, Chapter 14).

A.2 Linkage

Methods for genetic mapping of QTLs are based on the idea that genetic markers which tend to be transmitted together with specific values of the trait are likely to be close together. If this is the case they are said to be linked.

Table 1: Seed weight and colour genotype in an F2 experiment of the bean *Phaseolus vulgaris*. Individuals with PP colour are heavier than those with pp with heterozygotes pP being intermediate in weight. Source: Sax (1923)

No. Plants	Average weight (g)	Colour genotype
41	0.264±0.03	pp (white)
80	0.283±0.03	Pp (intermediate)
45	0.307±0.06	PP (pigmented)

Payne (1918) provided the first account of linkage of a quantitative trait to a major gene locus. Payne found that the X chromosome from selected lines of *Drosophila* contained multiple factors influencing bristle number. Subsequently, several studies demonstrated associations between Mendelian markers and quantitative traits in line crosses. For instance, Sax (1923) studied the inheritance of seed weight and colour in an F2 cross between two lines of the bean *Phaseolus vulgaris*. While seed weight is a continuous variable, a single segregating gene P/p was found to be responsible for bean colour. Sax found the mean seed weight differed between the colour genotypes as shown in Table 1.

Throughout the subsequent 60 years, several successful experiments in *Drosophila* and wheat identified chromosomes or parts of chromosomes affecting quantitative traits. However, there were several factors which mitigated against the widespread use of single gene markers in hunting for QTL (Kearsey and Pooni, 1996). Firstly, there were very few morphological markers available. Even if they existed, they often exhibited detrimental effects on the quantitative trait. Secondly, experiments were usually too small to have the power needed to detect such linkage. While there are now many markers and a diverse range of methods available to generate new ones, the problem of small experiments still continues today.

With the introduction and subsequent explosion of molecular marker technology in the 1980's this changed rapidly and so by the early 1990s various species were studied for QTL and this also stimulated the development of new statistical approaches.

A.3 Single Marker Methods

In general, single marker methods are not well regarded and so are not used widely for experiments where several traits are measured. For instance Mott et al. (2000) considers an example where only 2 QTL for fearfulness in mice were found by single marker analyses, whereas 5 were found by a more complicated approach. On the other hand, Wright and Kong (1997) point out that single gene models, while not as efficient as interval mapping, are robust to such effects as “ghost” QTL and thus may beneficially be used as a first step in scanning the genome. In microarray experiments, where gene expression may be regarded as a trait and many such traits are recorded, these methods have been widely adopted.

For the data of Sax (1923) in Table 1, a simple One-Way ANOVA of the seed weight on the three colour genotypes PP, pP and pp can be used to detect linkage. Alternatively, a simple model with additive and dominance genetic parameters may, be fitted by multiple linear regression. This is just a reparameterisation of the One-Way ANOVA. In particular, consider the expected means

$$\begin{aligned} E(\mu_{pp}) &= \mu - a, \\ E(\mu_{Pp}) &= \mu + d, \text{ and} \\ E(\mu_{PP}) &= \mu + a, \end{aligned} \tag{1}$$

where a is the additive genetic effect of the QTL, d is the dominance effect or a measure of hybrid vigour associated with the heterozygote Pp, μ is the overall mean and μ_{pp} , μ_{Pp} and μ_{PP} are the mean seed weights of the respective colour classes. Note that applying the one-way ANOVA then the recombination between the marker and QTL can not be estimated and that a genetic marker map is not required.

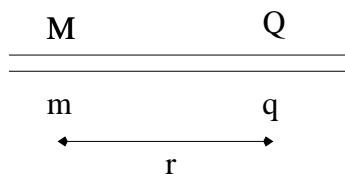


Figure 10: F1 individual formed by crossing homozygous parental lines with marker/QTL genotypes $MMQQ$ and $mmqq$. All F1 individuals have marker/QTL genotype $MmQq$. The recombination rate between the marker and QTL is r , meaning that $100r\%$ individuals are expected to have one recombination in this interval.

Consider the cross between two homozygous parental lines with the marker QTL genotypes $MMQQ$ and $mmqq$. All F1 individuals have marker/QTL genotype $MmQq$ (see Figure 10). Individuals from the parental lines, have QTL genotypes QQ and qq with expected values $\mu + a$ and $\mu - a$, respectively, where where the mean of the overall mean of the trait is μ and the additive genetic effect is a .

One of the simplest QTL study designs is the backcross. In these designs, F1 individuals are crossed to the parental lines which creates linkage disequilibrium necessary for such studies. If the recombination rate between the marker and QTL in Figure 10 is r and the F1's are crossed to the parental lines $MMQQ$ then the conditional probabilities of QTL genotypes given marker genotypes are

$$Pr(QQ|MM) = Pr(Qq|Mm) = 1 - r, \text{ and}$$

$$Pr(Qq|MM) = Pr(QQ|Mm) = r. \quad (2)$$

If r denotes the recombination rate, then for trait y_j , the likelihood l may be written as

$$l(y_j|\mu_{QQ}, \mu_{Qq}, r, \sigma^2) = \begin{cases} (1-r) \cdot \phi(y_j, \mu_{QQ}, \sigma^2) + r \cdot \phi(y_j, \mu_{Qq}, \sigma^2), & \text{for marker MM} \\ r \cdot \phi(y_j, \mu_{QQ}, \sigma^2) + (1-r) \cdot \phi(y_j, \mu_{Qq}, \sigma^2), & \text{for marker Mm} \end{cases} \quad (3)$$

given that the trait y is normally distributed with mean $\mu_{..}$, variance σ^2 and with pdf ϕ . Hence, the likelihood $\prod_j l(y_j|\mu_{QQ}, \mu_{Qq}, r, \sigma^2)$ may be maximised to obtain estimates $\hat{\mu}_{QQ}$, $\hat{\mu}_{Qq}$, \hat{r} and $\hat{\sigma}$. Haley and Knott (1992) show that the appropriate likelihood ratio test comparing this model to one with no QTL is unbiased if there are linked QTL present but biased if there are other linked QTL.

A.4 Methods employing two markers at a time

Commonly, two or more markers are considered at a time.

Simple interval mapping by means of maximum-likelihood (ML), was introduced by Lander and Botstein (1989). Under interval mapping or flanking-marker analysis, a separate analysis is performed for each pair of adjacent marker loci. The use of such two-locus marker genotypes results in $n - 1$ tests if there are n markers on a chromosome.

The likelihood functions, initially derived by Lander and Botstein (1989) and Jensen (1989) using the same idea as in (3) are more complicated than the corresponding single-marker methods. For instance, from Lynch and Walsh (1998), consider the likelihood for two markers in an F_2 cross design between two inbred lines. If the homozygous parental lines have alleles MQN and mqn respectively, then the likelihood of the trait y , given marker alleles $MMNN$ is

$$l(y|MMNN) = \frac{(1-r_1)^2(1-r_2)^2}{(1-r_{12})^2} \cdot \phi(y, \mu_{QQ}, \sigma^2) + \frac{2r_1r_2(1-r_1)(1-r_2)}{(1-r_{12})^2} \cdot \phi(y, \mu_{Qq}, \sigma^2) + \frac{r_1^2r_2^2}{(1-r_{12})^2} \cdot \phi(y, \mu_{qq}, \sigma^2) \quad (4)$$

where the three recombination parameters are: r_1 is the recombination between first marker with alleles M and m and qtl, r_2 is the recombination between the QTL and the second marker and r_{12} is the assumed known distance between the two markers. The recombination rate r_2 can be derived from the other two using mapping functions since the relationships are non-linear. The full likelihood may be obtained by addition of the other eight flanking marker combinations.

Lander and Botstein (1989) performed interval mapping by moving the position of the putative QTL along the chromosome and calculating the so called LOD (likelihood of odds) score at each point. LOD scores, which are closely related to the likelihood ratio statistic, were plotted along the chromosome (see Figure 11). It should be noted that the plots should be regarded as likelihood plots in that they display the support for a QTL at each point. The peak of the likelihood map corresponds to the ML estimate of the QTL position and its significance given by a likelihood-ratio test. Various techniques, including the bootstrap (Efron, 1979) have been employed to assess significance of the QTL.

Given the computationally intensive nature of ML, Haley and Knott (1992) and Martinez and Curnow (1992) introduced a simple linear regression method which provided an excellent approximation to ML. Using the notation in (1), denote the genotypic means as

$$\mu_{QQ} = \mu + a, \quad \mu_{Qq} = \mu + d, \quad \text{and} \quad \mu_{qq} = \mu - a, \quad (5)$$

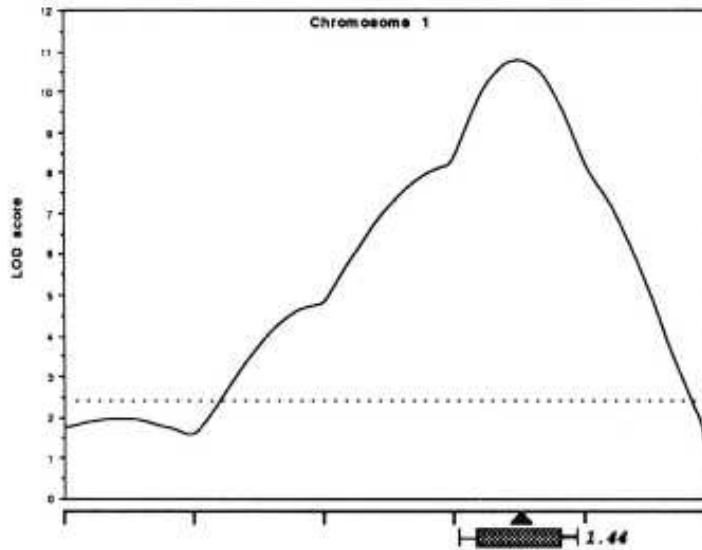


Figure 11: LOD scores for a hypothetical quantitative trait. The LOD scores are based on simulated data for 250 backcross progeny. The LOD threshold is shown as a dotted line but given subsequent advances in multiple testing and permutation testing, this threshold may need to be revised upwards. Source: Lander and Botstein (1989).

and consider the regression

$$y_j = \mu + a.x(M_j) + d.y(M_j) + e_j, \quad (6)$$

where the variables x and y , which depend on flanking-marker genotypes of the marker and the assumed map position of the putative QTL, may be approximated as follows.

Taking the expectation of (6) over all individuals with marker genotype M_j gives

$$\mu_{M_j} = \mu + a.x(M_j) + d.y(M_j) \quad (7)$$

and from the likelihood equation

$$l(y_j|M_j) = \sum_1^N \phi(y_j, \mu_Q, \sigma^2) \cdot Pr(Q_k|M_j), \quad (8)$$

then

$$\mu_{M_j} = \mu + a.[Pr(QQ|M_j) - Pr(qq|M_j)] + d.Pr(Qq|M_j). \quad (9)$$

Equating terms in (7) and (9) yields

$$x(M_j) = Pr(QQ|M_j) - Pr(qq|M_j) \quad \text{and} \quad y(M_j) = Pr(Qq|M_j). \quad (10)$$

For example, in the F_2 design described above and from (4), then for individuals with markers MMNN

$$\begin{aligned} x(MMNN) &= \frac{(1-r_1)^2(1-r_2)^2}{(1-r_{12})^2}, & \text{and} \\ y(MMNN) &= \frac{2r_1r_2(1-r_1)(1-r_2)}{(1-r_{12})^2}. \end{aligned}$$

It is interesting to note that conventional wisdom states that flanking marker or multiple marker analyses, such as outlined here, are more powerful than single marker tests. Coffman et al. (2003) considered single marker intersection tests. The intersection test was derived by conducting single marker tests for adjacent markers and forming the compound null hypothesis of no association for both markers with a simple Bonferroni adjustment for multiple testing. Coffman et al. conducted simulation studies and demonstrated that multiple marker tests were not always more powerful. Indeed, in general there appeared to be little difference except when the QTL was close to a marker, in which case the intersection method was more powerful. In an approximately equal number of times, where the QTL was in the middle of the interval this produced the opposite result of the two marker test being more powerful. However, while this simple method appears to be efficient and does not exhibit well known ghost QTL problems, the case when two or more QTL are present was not studied.

A.5 More advanced methods

Both single-marker and interval mapping are biased when multiple QTLs are linked to the marker/interval being considered (Haley and Knott, 1992; Knott and Haley, 1992; Jansen, 1993). Several methods have been proposed to circumvent this problem.

Composite interval mapping (CIM) (Zeng, 1993, 1994; Jansen, 1993), which is also known as multiple QTL mapping (MQM) (Jansen and Stam, 1994), considers a marker interval plus a few other well-chosen single markers in each analysis. Kao (2000) compare ML and regression approaches. Bayesian methods (Satagopan et al., 1996b; Heath, 1997; Sillanpää and Arjas, 1998, 1999) and multiple interval mapping (MIM) (Kao et al., 1999; Zeng et al., 1999) are regarded as being more powerful than simple interval mapping. The multipoint approaches of Wu and Li. (1994); Hyne and Kearsey (1995) and Wu and Li (1996) and the Bayesian approaches of Heath (1997); Sillanpää and Arjas (1998) and Sillanpää and Arjas (1999) model all markers of a single chromosome simultaneously and hence require one analysis per chromosome.

Another advantage of the latter methods is the potential to fit more realistic genetic models such as models that include epistasis or QTL interactions. However, given the substantial computational requirements their potential for use in eQTL studies, where many thousands of expression traits would need to be considered, is limited. Employing PCA or PLS to reduce the dimensionality of the gene expression space may be one possibility.

Mixed model approaches have also proved popular, especially in QTL studies for animals. Chapter 7 of (Weller, 2001) provides an outline of methods where QTL are assumed to be random effects. Piepho (2000) described mixed model analysis for multi-environment barley QTL analysis.

Multiple trait and multivariate methods have also been proposed but these are for the more conventional situations with a small number of traits and a somewhat larger number of individuals or samples (Jiang and Zeng, 1995; Kao et al., 1999). Knott and Haley (2000); Korol et al. (2001) considered multivariate regression QTL models for multiple traits, while Henshall and Goddard (1999) considered logistic regression for multiple traits.

Finally, when it is not possible to produce segregating populations by means of designed matings, association studies are often used (Lewis, 2002). Emahazion et al. (2001) note that association studies are well known to be difficult in finding replicable results. This may be due to the wide variability exhibited by disequilibrium (Nordborg and Tavaré (2002) or see Perez-Enciso et al. (2003, p 1604)).

A.6 New technologies

Despite the plethora of marker technologies currently available, obtaining markers is a slow laborious process.

Several high throughput alternatives have been proposed.

Grupe et al. (2001) used SNP markers to map QTL for several phenotypes in F2 mice.

As discussed in Section 3, Steinmetz et al. (2002) used genes with differences in gene expression between two lines of yeast using Affymetrix chips. They named this approach reciprocal-hemizyosity analysis. This article is somewhat lacking in details about exactly what was mapped and how.

Borevitz et al. (2003) developed high-throughput genotyping by means of single feature polymorphisms via an RNA expression Gene Chip (AtGenome1). They applied this technology both to a recombinant inbred line and bulk segregant analysis in *Arabidopsis*.

With the advent of new technologies it is likely that "saturated" marker maps with so called fine resolution will be routinely possible. Fine mapping techniques have been investigated and are outlined in chapter ten of Weller (2001) or see Meuwissen et al. (2002); Lee and van der Werf (2004).

References

- Ambrose, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS*, 99(10):6562–6566.
- Boake, C. R. B., Arnold, S. J., Breden, F., Meffert, L. M., Ritchie, M. G., Taylor, B. J., Wolf, J. B., and Moore, A. J. (2002). Genetic tools for studying adaptation and the evolution of behavior. *American Naturalist*, 160:S143–S159.
- Borevitz, J. O., Liang, D., Plouffe, D., Chang, H. S., Zhu, T., Weigel, D., Berry, C. C., Winzeler, E., and Chory, J. (2003). Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res*, 13(3):513–523.
- Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755.
- Chao, W. S. (2002). Contemporary methods to investigate seed and bud dormancy. *Weed Science*, 50:215–226.
- Chesler, E. J., Wang, J. T., Lu, L., Qu, Y. H., Manly, K. F., and Williams, R. W. (2003). Genetic correlates of gene expression in recombinant inbred strains - a relational model system to explore neurobehavioral phenotypes. *Neuroinformatics*, 1:343–357.
- Cheung, V. G. and Spielman, R. S. (2002). The genetics of variation in gene expression. *Nat Genet*, 32 Suppl:522–525.
- Coffman, C., Doerge, R., Wayne, M. L., and McIntyre, L. M. (2003). Intersection tests for single marker QTL analysis can be more powerful than two marker QTL analysis. *BMC Genet*, 4:10.
- Consortium, C. T. (2003). The nature and identification of quantitative trait loci: a community's view. *Nat Rev Genet*, 4(11):911–916.

- Darvasi, A. (2003). Genomics: Gene expression meets genetics. *Nature*, 422(6929):269–270. Comment.
- 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.
- de Haan, G., Bystrykh, L. V., Weersing, E., Dontje, B., Geiger, H., Ivanova, N., Lemischka, I. R., Vellenga, E., and Van Zant, G. (2002). A genetic and genomic analysis identifies a cluster of genes associated with hematopoietic cell turnover. *Blood*, 100(6):2056–2062.
- Dobra, A., Hans, C., Jones, B., Nevins, J. R. J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196–212.
- Doerge, R., Zeng, Z.-B., and Weir, B. S. (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science*, 12:195–213.
- Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet*, 3(1):43–52.
- Dumas, P., Sun, Y., Corbeil, G., Tremblay, S., Pausova, Z., Kren, V., Krenova, D., Pravenec, M., Hamet, P., and Tremblay, J. (2000). Mapping of quantitative trait loci (QTL) of differential stress gene expression in rat recombinant inbred strains. *J Hypertens*, 18(5):545–551.
- Eaves, I. A., Wicker, L. S., Ghandour, G., Lyons, P. A., Peterson, L. B., Todd, J. A., and Glynne, R. J. (2002). Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: The NOD model of Type 1 diabetes. *Genome Res.*, 12(2):232–243.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26.
- Eggen, A. and Hocquette, J. F. (2003). Genomic approaches to economic trait loci and tissue expression profiling: application to muscle biochemistry and beef quality. *Meat Science*, 66:1–9.
- Emahazion, T., Feuk, L., Jobs, M., Sawyer, S. L., Fredman, D., St Clair, D., Prince, J. A., and Brookes, A. J. (2001). Snp association studies in alzheimer’s disease highlight problems for complex disease analysis. *Trends in Genetics*, 17(7):407–413. association analysis; SNP; Alzheimer’s disease; genetic variation; genetic linkage; disease association; genotyping; Molecular Medicine; Genetics; Evolution; Bioinformatics.
- Fischer, G., Ibrahim, S., Brockmann, G., Pahnke, J., Bartocci, E., Thiesen, H., Serrano-Fernández, P., and Möller, S. (2003). Expressionview: visualization of quantitative trait loci and gene-expression data in Ensembl. *Genome Biol*, 4(11).
- Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edinburgh*, 52:399–433.
- Gibson, G. and Mackay, T. F. C. (2002). Enabling population and quantitative genomics. *Genetical Research*, 80:1–6.

- Gladney, C., Bertani, G., Johnson, R., and Pomp, D. (2004). Evaluation of gene expression in pigs selected for enhanced reproduction using differential display PCR and human microarrays: I. Ovarian follicles. *J Anim Sci*, 82(1):17–31.
- Gracey, A. Y. and Cossins, A. R. (2003). Application of microarray technology in environmental and comparative physiology. *Annual Review Of Physiology*, 65:231–259.
- Grupe, A., Germer, S., Usuka, J., Aud, D., Belknap, J. K., Klein, R. F., Ahluwalia, M. K., Higuchi, R., and Peltz, G. (2001). In silico mapping of complex disease-related traits in mice. *Science*, 292(5523):1915–1918.
- Haley, C. and Knott, S. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69:315–324.
- Hazen, S. P. and Kay, S. A. (2003). Gene arrays are not just for measuring gene expression. *Trends Plant Sci*, 8(9):413–416.
- Heath, S. C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet*, 61:748–760.
- Henshall, J. M. and Goddard, M. E. (1999). Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression. *Genetics*, 151(2):885–894.
- Hunt, G. J., Page-Jr., R. E., Fondrk, M. K., and Dullum, C. J. (1995). Major quantitative trait loci affecting honey bee foraging behavior. *Genetics*, 141(4):1537–1545.
- Hyne, V. and Kearsey, M. J. (1995). QTL analysis: Further uses of marker regression. *Theor. Appl. Genet.*, 91:471–476.
- Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics*, 135(1):205–211.
- Jansen, R. C. (2001). Quantitative trait loci in inbred lines. In Balding, D., editor, *Handbook of Statistical Genetics*. Wiley, New York.
- Jansen, R. C. and Nap, J.-P. (2001). Genetical genomics: the added value from segregation. *Trends in Genetics*, 17:388–391.
- Jansen, R. C. and Stam, P. (1994). High Resolution of Quantitative Traits Into Multiple Loci via Interval Mapping. *Genetics*, 136(4):1447–1455.
- Jensen, J. (1989). Estimation of recombination parameters between a quantitative trait locus (QTL) and two marker gene loci. *Theoretical and Applied Genetics*, 78:613–618.
- Jiang, C. and Zeng, Z. B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, 140(3):1111–1127.
- Kao, C.-H. (2000). On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics*, 156(2):855–865.
- Kao, C.-H., Zeng, Z.-B., and Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics*, 152(3):1203–1216.

- Kearsey, M. J. and Pooni, H. S. (1996). The genetical analysis of quantitative traits. The genetical analysis of quantitative traits / Michael J. Kearsey and Harpal S. Pooni. Includes bibliographical references and index.
- Knott, S. A. and Haley, C. S. (1992). Maximum Likelihood Mapping of Quantitative Trait Loci Using Full-Sib Families. *Genetics*, 132(4):1211–1222.
- Knott, S. A. and Haley, C. S. (2000). Multitrait least squares for quantitative trait loci detection. *Genetics*, 156(2):899–911.
- Korol, A. B., Ronin, Y. I., Itskovich, A. M., Peng, J., and Nevo, E. (2001). Enhanced efficiency of quantitative trait loci mapping analysis based on multivariate complexes of quantitative traits. *Genetics*, 157(4):1789–1803.
- Kwok, P. Y. (2001a). Genomics. Genetic association by whole-genome analysis. *Science*, 294(5547):1669–1670. Comment.
- Kwok, P. Y. (2001b). Methods for genotyping single nucleotide polymorphisms. *Annu Rev Genomics Hum Genet*, 2:235–258.
- Lan, H., Stoehr, J. P., Nadler, S. T., Schueler, K. L., Yandell, B. S., and Attie, A. D. (2003). Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics*, 164:1607–1614.
- Lander, E. S. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185–199.
- Lee, S. H. and van der Werf, J. H. (2004). The efficiency of designs for fine-mapping of quantitative trait loci using combined linkage disequilibrium and linkage. *Genet. Sel. Evol.*, 36:145–161.
- Lewis, C. M. (2002). Genetic association studies: design, analysis and interpretation. *Brief Bioinform*, 3(2):146–153.
- Lipkin, E., Mosig, M. O., Darvasi, A., Ezra, E., Shalom, A., Friedmann, 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(3):1557–1567.
- Liu, B.-H. (1998). *Statistical genomics : linkage, mapping, and QTL analysis*. CRC Press.
- Liu, Z. J. (2003). A review of catfish genomics: progress and perspectives. *Comparative And Functional Genomics*, 4:259–265.
- Lynch, M. and Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sinauer, Sunderland, Ma.
- Martinez, O. and Curnow, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor. Appl. Genet*, 85:480–488.
- McPeck, M. S. (2000). From mouse to human: Fine mapping of quantitative trait loci in a model organism. *PNAS*, 97(23):12389–12390.

- Meuwissen, T. H. E., Karlsten, 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(1):373–379.
- Mott, R., Talbot, C. J., Turri, M. G., Collins, A. C., and Flint, J. (2000). From the Cover: A method for fine mapping quantitative trait loci in outbred animal stocks. *PNAS*, 97(23):12649–12654.
- Muranty, H. and Goffinet, B. (1997). Selective genotyping for location and estimation of the effect of a quantitative trait locus. *Biometrics*, 53:629–643.
- Nordborg, M. and Tavaré, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends in Genetics*, 18(2):83–90. linkage disequilibrium; coalescent; polymorphism; haplotype; LD mapping; Genetics; Pharmaceutical Science; Evolution; Bioinformatics.
- Okuda, T., Sumiya, T., Iwai, N., and Miyata, T. (2004). Pyridoxine 5'-phosphate oxidase is a candidate gene responsible for hypertension in dahl-s rats. *Biochemical And Biophysical Research Communications*, 313:647–653.
- Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T. N., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P. A., and Cox, D. R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–1723.
- Payne, F. (1918). The effect of artificial selection on *Drosophila ampelophila* and its interpretation. *PNAS*, 4:55–58.
- Perez-Enciso, M., Toro, M. A., Tenenhaus, M., and Gianola, D. (2003). Combining gene expression and molecular marker information for mapping complex trait genes: A simulation study. *Genetics*, 164:1597–1606.
- Piepho, H.-P. (2000). A Mixed-Model Approach to Mapping Quantitative Trait Loci in Barley on the Basis of Multiple Environment Data. *Genetics*, 156(4):2043–2050.
- Prows, D. R., McDowell, S. A., Aronow, B. J., and Leikauf, G. D. (2003). Genetic susceptibility to nickel-induced acute lung injury. *Chemosphere*, 51:1139–1148.
- Satagopan, J. M., Yandell, B. S., Newton, M. A., and Osborn, T. C. (1996a). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics*, 144:805–816.
- Satagopan, J. M., Yandell, B. S., Newton, M. A., and Osborn, T. C. (1996b). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics*, 144:805–816.
- Sax, K. (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics*, 8(6):552–560.
- Schadt, E. E., Monks, S. A., Drake, T. A., Luskis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B., and Friend, S. H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422:297–302.
- Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

- Sillanpää, M. J. and Arjas, E. (1998). Bayesian mapping of multiple Quantitative Trait Loci from incomplete line cross data. *Genetics*, 148:1373–1388.
- Sillanpää, M. J. and Arjas, E. (1999). Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics*, 151:1605–1619.
- Steinmetz, L. M., Sinha, H., Richards, D. R., Spiegelman, J. I., Oefner, P. J., McCusker, J. H., and Davis, R. W. (2002). Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, 416(6878):326–330.
- Wallace, C., Ali, S., Glazier, A., Norsworthy, P., Carlos, D., Scott, J., Freeman, T., Stanton, L., Kwitek, A., and Aitman, T. (2002). Radiation hybrid mapping of 70 rat genes from a data set of differentially expressed genes. *Mamm Genome*, 13(4):194–197.
- Wang, J. T., Williams, R. W., and Manly, K. F. (2003). Webqtl - web-based complex trait analysis. *Neuroinformatics*, 1:299–308.
- Watts, J. A., Morley, M., Burdick, J. T., Fiori, J. L., Ewens, W. J., Spielman, R. S., and Cheung, V. G. (2002). Gene expression phenotype in heterozygous carriers of ataxia telangiectasia. *Am J Hum Genet*, 71(4):791–800.
- Wayne, M. and McIntyre, L. (2002). Combining mapping and arraying: An approach to candidate gene identification. *Proc Natl Acad Sci U S A*, 99(23):14903–14906.
- Weller, J. I. (2001). *Quantitative trait loci analysis in animals*. Wallingford, UK ; New York : CABI Pub.
- Wright, F. A. and Kong, A. (1997). Linkage mapping in experimental crosses: the robustness of single-gene models. *Genetics*, 146(1):417–425.
- Wu, W. and Li, W. (1994). A new approach for mapping quantitative trait loci using complete genetic marker linkage maps. *Theor. Appl. Genet.*, 89:535–539.
- Wu, W. and Li, W. (1996). Model fitting and model testing in the method of joint mapping of quantitative trait loci. *Theor Appl Genet*, 92:477–482.
- Zeng, Z.-B. (1993). Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *PNAS*, 90:10972–10976.
- Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics*, 136(4):1457–1468.
- Zeng, Z. B., Kao, C. H., and Basten, C. J. (1999). Estimating the genetic architecture of quantitative traits. *Genet Res*, 74(3):279–289.