Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray
Copyright © 2003 John Wiley & Sons, Ltd.
ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

SECTION 3

BIOINFORMATICS FOR GENETIC STUDY DESIGN

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray
Copyright © 2003 John Wiley & Sons, Ltd.
ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 8

From Linkage Peak to Culprit Gene: Following Up Linkage Analysis of Complex Phenotypes with Population-based Association Studies

IAN C. GRAY

Discovery Genetics
GlaxoSmithKline Pharmaceuticals Harlow, Essex, UK

8.1 INTRODUCTION

Linkage analysis of complex traits using family-based samples (see Chapter 11) typically results in a number of broad, ill-defined linkage peaks that represent several megabases of DNA (see for example Grettarsdottir *et al.*, 2002); beneath the expanse of each peak there may lie a gene (or genes) associated with the disease in question. Under the prior assumption that this preliminary linkage analysis has been completed, the goal of this chapter is to take the investigator through the process of characterizing and narrowing such a region using a population-based approach, with the ultimate aim of identifying candidate genes and testing them directly for association with the disease or trait in question. This is often achieved by testing markers in the linkage interval for differences in allele frequency between case and control cohorts, where the cohorts comprise unrelated individuals (although methods employing family structure, based on the difference in frequency of allele transmission in a large number of small pedigrees are also used — see below; this is covered in more detail in Chapter 11). In general, population-based methods offer large increases in both power and resolution over linkage-based approaches (McGinnis, 2000; Risch 2000; discussed in Chapter 11) and are well suited to the follow-up of preliminary (and often equivocal) linkage results. Examples of the successful application of this two-step linkage-association approach include identification of the involvement of *ApoE* in Alzheimer's disease (Strittmatter *et al.*, 1993) and the recent discovery of the role of *NOD2* in Crohn's disease (Hugot *et al.*, 2001; Ogura *et al.*, 2001). The first part of this chapter focuses on theoretical and practical considerations for good study design, whilst the second part covers a systematic approach to identification of the disease-associated gene, with emphasis on the application of methods, software tools and databases.

8.2 THEORETICAL AND PRACTICAL CONSIDERATIONS

8.2.1 Choice of Study Population

Wherever possible the study population selected for a follow-up analysis of linkage peaks should be derived from the same geographic area as the families used for the original linkage analysis. As the genetic components contributing to complex disease are likely to be varied, there is no guarantee that the predisposing genetic factors in one population will be the same in a second. If we use the term 'study population' in the broadest sense as applied to genetic association studies, a variety of study population structures may be considered. Three of the most common configurations are the case–control cohort, the discordant sib-pair cohort (i.e. one affected and one unaffected sib) and the parent–offspring triad (affected offspring with both parents) cohort. Each of these structures has advantages and disadvantages (for an evaluation of each, see Risch, 2000; Cardon and Bell, 2001).

Case–control cohorts simply consist of one group of individuals (cases) with the disease state and a second group without the disease (controls). Case–control cohorts have the advantage of being more straightforward to collect than the other two structures described above and generally provide more statistical power than similarly sized discordant sib or other nuclear family-based cohorts (McGinnis, 2000; Risch, 2000). However, case–control cohorts are prone to 'population stratification' (or substructure) effects. Population stratification occurs when the cohort under study contains a mix of individuals that can be separated on grounds other than the phenotype under study (most commonly on the basis of geographic origin). This can lead to allele frequency differences in cases and controls that are due to circumstances unrelated to the phenotypic difference under investigation, resulting in erroneous conclusions regarding association between the marker under test and the disease phenotype. Careful selection of individuals for inclusion in disease and control cohorts is necessary to ensure as homogenous a background as possible and therefore avoid stratification. If stratification is suspected it is possible to test for it using randomly selected genetic markers (Devlin and Roeder, 1999; Pritchard and Rosenberg, 1999). It is also important to match the cohorts for phenotypic or environmental variables that may otherwise confound any genetic analysis; for example, hormone replacement therapy (HRT) has a large impact on bone mineral density (BMD) and it would be necessary to account for this in a search for genetic factors influencing BMD using a cohort of post-menopausal women.

Although population homogeneity and well matched cases and controls are preferred, it may be possible to use a cohort even if stratification is present; Pritchard *et al.* (2000) have developed a method for testing for genetic association in the presence of population stratification, by using unlinked markers to make inferences about population substructure and employing this information to test for associations within the identified subpopulations. STRUCTURE and STRAT, software tools for the detection of stratification and testing for genetic association in the presence of stratification can be downloaded from http://pritch.bsd.uchicago.edu/software.html. An alternative approach to correction for population stratification, termed genomic control, measures the degree of variability and magnitude of the test statistics observed at random loci and uses this information to adjust the critical value for significance tests at candidate loci by the appropriate degree (Devlin and Roeder, 1999). However, it should be noted that correction for stratification cannot completely remove the possibility of increased false positive results under all circumstances (Cardon and Bell, 2001; Devlin *et al.*, 2001; Pritchard and Donelly, 2001) and stratification should be avoided where possible.

The main advantage of using study populations that incorporate elements of family structure (e.g. discordant sibs or trios) is that, unlike case–control cohorts, they are immune to population stratification effects. However, as mentioned above, family-based samples are typically more difficult to collect than case–control samples (particularly for late onset diseases) and generally offer less statistical power than the equivalent sized case–control cohort (McGinnis, 2000; Risch, 2000). The remainder of this chapter will focus predominantly on case–control methodology where reference to population structure is necessary; statistical methods for analysing family-based cohorts, such as the transmission disequilibrium (TDT) and sib transmission disequilibrium (S-TDT) tests, together with tools for the analysis of quantitative traits are covered in Chapter 11.

Estimation of required cohort size for a genetic study depends on a number of factors, including the size of the effect of the locus under test, the frequency of the disease-risk conferring allele and genetic nature of this ‘risk allele’, i.e. recessive, dominant, additive etc. If the causal variant is not being tested directly, the distance between the causal variant and the surrogate marker under test (see Section 8.2.3 below) is also relevant. Most of these factors are unknown prior to the start of the study and the minimum required population size is usually based on assumptions concerning these factors (see McGinnis, 2000; Risch, 2000). In reality, pragmatism typically dictates the available sample size; investigators use the largest obtainable cohort, with the caveat that the available sample may not provide sufficient statistical power to detect effects below a certain magnitude. To detect genetic factors of fairly small or moderate effect, cohorts of a several hundred to a few thousand individuals may be required (McGinnis, 2000; Risch, 2000).

8.2.2 Sequence Characterization at the Locus under Investigation

Following a whole genome linkage scan the investigator is typically faced with several genetic loci of potential involvement in the disease process, the limits of each defined by two genetic markers (usually simple tandem repeats or STRs) spanning several centiMorgans (cM). As 1 cM equates to 1 megabase (Mb) on average, and each Mb contains an estimated average of 15 genes (based on 45,000 genes in the entire 3000-Mb genome; Das *et al*., 2001), this may represent several thousand kilobases (kb) of DNA and over 100 genes per locus. The first task is to define the locus in the context of the human genome, in order to gain a comprehensive knowledge of genes and further genetic markers in the interval. Until very recently this involved the laborious laboratory process of identifying and ordering genomic clones into contigs and using those contigs as a framework for gene and marker identification. Thankfully locus characterization has become far more straightforward in the wake of the Human Genome Sequencing Project, which provides free access to assembled sequence covering 97% of the human genome at the time of writing, with the goal of complete coverage by 2003. A number of web-based tools are available for exploiting the human genome. These tools are described very briefly in Section 8.3.1 and their practical application is covered in detail in Chapters 5 and 9.

8.2.3 SNPs, Linkage Disequilibrium, Haplotypes and STRs

8.2.3.1 Introduction

In this section we provide a simple introduction to the underlying principles of the detection of genetic association using a population (i.e. non-family)-based approach. The majority of studies of this nature are undertaken using single nucleotide polymorphism (SNP) markers (see Chapter 3). Biallelic SNPs are currently the marker of choice due to their abundance in the human genome and because they are amenable to high throughput genotyping approaches. The other marker system commonly used for genetic studies is the multiallelic STR (see Chapter 3). The paragraphs below on linkage disequilibrium and haplotypes refer mainly to SNPs. The use of STRs for population-based association studies is discussed at the end of this section.

8.2.3.2 Linkage Disequilibrium

A polymorphism associated with a disease state (in the true, rather than statistical, sense) may either directly contribute to the disease process, or may be a surrogate marker which is co-inherited with an adjacent functional variant that contributes to the disease state. This co-inheritance of the surrogate marker with the disease allele can occur to varying degrees and is termed 'linkage disequilibrium' (LD). By strict definition, LD is said to be present if co-occurrence of the two polymorphisms happens with a frequency greater than would be expected by chance. A number of measures of LD are used, two of the most commonly employed being Δ and D. Both measures are based on the difference between the observed and expected (assuming independence) number of haplotypes (see below) bearing specified alleles of two markers (see Chapter 11 for a complete explanation of D and Devlin and Risch (1995) for a discussion of D, Δ and other measures of LD). Although by the strict definition given above LD can occur between unlinked variants, for example in the presence of recent population admixture, in the following paragraphs we refer specifically to LD between two linked markers.

Clearly, the greater the extent of LD between two polymorphisms, the larger the chance of detecting the phenotypic influence of one by genotyping the other in a case–control experiment. The degree of LD is dependent on the history of the two adjacent markers and is influenced by the relative times of appearance of the two polymorphisms in the population and the degree of recombination between them. An extreme example would be two polymorphisms that appeared simultaneously on the same chromosome through spontaneous mutation and between which no recombination events have occurred over 2000 generations. During this period, these two linked polymorphisms have attained a population frequency of 20% through chance (random genetic drift) and are in absolute linkage disequilibrium. Imagine an alternative scenario, where a new polymorphism arises adjacent to an ancient polymorphism which has already attained a frequency of 20% over the previous 1000 generations; over the subsequent 1000 generations, there is a high degree of recombination between the markers, eroding the LD (Figure 8.1). Clearly the former case would be more favourable for using one of the markers as a surrogate for detecting the phenotypic influence of the other.

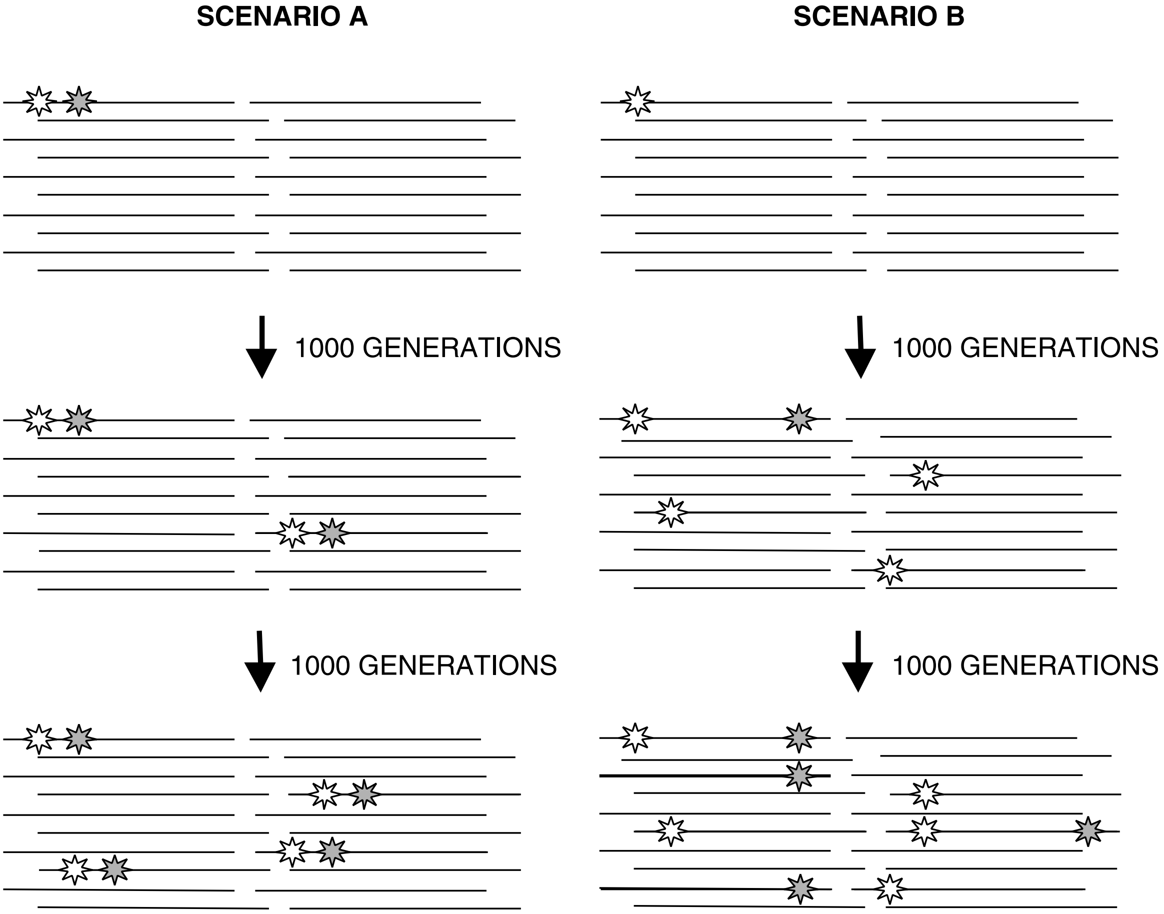


Figure 8.1 Alternative hypothetical scenarios depicting the evolution of a relationship between two SNPs. Identical stretches of DNA within a population are represented by black lines. In scenario A, two adjacent polymorphisms, represented by a white star and a grey star, arise simultaneously and by random drift achieve a population frequency of 0.1 after 1000 generations, increasing to 0.2 after 2000 generations, at which time they are still co-segregating as a tightly linked unit. In scenario B, a lone polymorphism (white star) reaches a frequency of 0.2 after 1000 generations, at which point a new polymorphism (grey star) arises spontaneously, some distance away. Note that although the grey polymorphism only occurs on a background bearing the white polymorphism, the association is less clear-cut than scenario A due to the chromosomes bearing the white polymorphism in the absence of the grey polymorphism. During the subsequent 1000 generations, association between the two polymorphisms is further clouded by recombination between the two SNPs and divergence through random drift. Unfortunately for the genetics investigator, scenario A is idealized and scenario B is more typical.

8.2.3.3 Haplotypes

A haplotype is a string of co-inherited alleles of different markers which are arranged in a successive fashion along a given stretch of DNA, hence each haplotype represents a linear section of DNA rather than the single point corresponding to a single marker. The extent of discernible haplotype length varies widely for different regions of the genome; well-defined haplotypes (characterized by moderate or high LD) are punctuated by regions of extremely low LD, suggesting that the recombination processes, selective pressures and other factors that dictate the degree of LD vary widely in an abrupt fashion across the genome (Goldstein, 2001). Although the length of preserved haplotypes shows dramatic variation from haplotype to haplotype, recent data suggest that the typical length of a discernible haplotypic block is 10–100 kb in the Caucasian population (Daly *et al.*, 2001).

In certain circumstances, statistical analysis of haplotypes is more powerful than single SNP analysis. This is because an SNP usually has only two allelic states, whereas a stretch of DNA can typically be represented by several different haplotypes; the chance that one of the many haplotypes shows strong association with a functional variant (i.e. a variant that influences the phenotype) is higher than the odds of a strong, pure correlation with one of only two possible alleles for a single SNP. In this sense, a series of haplotypes is analogous to a multi-allelic STR marker (although regarded as more stable — see below). Clearly if the functional variant itself is under test, or a polymorphism which shows perfect co-segregation with the functional variant, haplotypic analysis offers no advantage. It should also be noted that haplotype analysis is a double-edged sword and in addition to increasing statistical power has the potential to reduce it by introducing multiple testing and possibly by diluting an association signal due to undetected recombination within the haplotypes.

Haplotypes are usually constructed by comparing the genotypes of closely related individuals at two or more linked markers and identifying groups of alleles which are co-inherited as a set from one generation to the next. However, where no family members are available and the cohort under study consists of a population of unrelated individuals, it is necessary to infer haplotypes and haplotype frequencies using statistical methods. The most common method for the estimation of haplotypes is the expectation-maximization (EM) maximum likelihood estimate (MLE; Excoffier and Slatkin, 1995). The ARLEQUIN software package developed in the Genetics and Biometry Laboratory at the University of Geneva contains an EM algorithm for this purpose. ARLEQUIN can be downloaded from http://lgb.unige.ch/arlequin/. Another popular program for haplotype construction and analysis is EHPLUS (Zhao *et al.*, 2000). EHPLUS can be downloaded from http://www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBSt/software.stm. Both packages are discussed in detail in Chapter 11. Note that haplotype construction using family inheritance patterns, although more robust than population-based MLE, also typically requires a degree of inference and resulting haplotypes may be probable rather than actual (Hodge *et al.*, 1999). For absolute definition of all haplotypes, it is necessary to physically separate the two copies of each stretch of DNA under analysis, i.e. reduction from a diploid to a haploid state, to allow unmixed analysis of a single haplotype. For very short stretches of DNA (up to approximately 10 kb), this can be achieved by allele-specific PCR (Michalatos-Beloin *et al.*, 1996); for large-scale haplotype construction it is necessary to separate entire chromosomes. This strategy has been successfully employed by the California-based company Perlegen Sciences Inc., who have used a rodent–human somatic cell hybrid technique to physically separate the two copies of human chromosome 21 for haplotype elucidation (Patil *et al.*, 2001; see below). However, most investigators employ the less laborious MLE or family-based inference methods for haplotype construction and accept a certain degree of error or loss of power.

In addition to potentially providing greater power than single markers in subsequent statistical analyses, a knowledge of the haplotypes representing the locus under study is extremely valuable for maximizing efficiency in study design. For example, two markers which always co-segregate (as in Figure 8.1, scenario A) will provide the same information, regardless of which of the two is genotyped, therefore typing both markers is inefficient as the genotype of one can be inferred from the other. Consequently, a detailed knowledge of the haplotypes across the interval theoretically allows a minimum marker set to be identified that will permit the extraction of all haplotypic information (Figure 8.2; see Johnson *et al*., 2001; Patil *et al*., 2001). David Clayton of the Medical Research Council Biostatistics Unit, UK has written software (htSNP) to aid the selection of optimum marker sets based on haplotypic information which can be downloaded from http://www-gene.cimr.cam.ac.uk/clayton/software/stata.

Before this optimized marker set can be selected, it is necessary to identify all common SNPs within the area under study and construct haplotypes. The publicly available SNPs catalogued in the SNP database hosted by NCBI (dbSNP: http://www.ncbi.nlm.nih.gov/SNP/) are far too sparse for this purpose at the time of writing (Johnson *et al*., 2001). Comprehensive identification of all common SNPs in a given interval requires sequencing of a significant number of individuals from the relevant population. For example, sequencing 24 individuals would give a 95% probability of detecting all variants with a minor allele frequency of greater than 5% (Kruglyak and Nickerson, 2001); 5% is a sensible lower cut-off point, as sample size requirements for case–control studies increase dramatically when allele frequencies fall below 5% (Johnson *et al*., 2001).

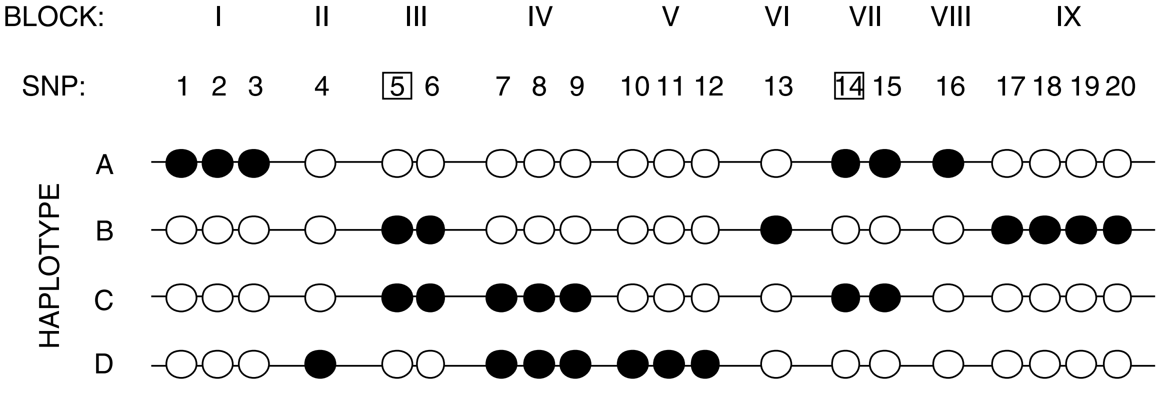


Figure 8.2 Using haplotypic information to maximize efficiency in genotyping study design. Twenty SNPs spanning four haplotypes are shown. Each SNP is represented by a circle; the circle is black or white, depending on the allelic state of the SNP. The SNPs can be grouped into nine blocks — each block contains a group of SNPs with an identical allelic pattern in the four haplotypes. Genotyping all SNPs in any given block is unnecessary, as the genotype of one SNP per block allows the genotypes of the other SNPs in the block to be inferred; for example, genotyping SNP 1 allows the genotypes of SNPs 2 and 3 to be predicted. Moreover, in this simplified example, all four haplotypes can be unambiguously identified by genotyping just two SNPs, 5 and 14 (boxed), yielding a 90% reduction in genotyping compared to a 'blind' strategy (i.e. no knowledge of haplotypic structure).

Obviously it is impractical to sequence a region covering several Mb in 24 individuals. A more realistic approach is the identification of all genes in the interval and sequencing of the coding sequence plus flanking splice sites, together with 1–2 kb of putative promoter (i.e. the region immediately upstream of the transcription start site) and any other known regulatory elements. Although not comprehensive, as unidentified regulatory elements can be intronic or several tens of kilobases away from the genes under their influence (Blackwood and Kadonaga, 1998), this approach offers a good compromise between exhaustive coverage of the locus and practicality. For SNP identification purposes, it may be preferable to use individuals derived from the disease, rather than control, population. This will give a greater chance of detecting rare functional variants (mutations) that are at a higher frequency in the disease population. For example, *NOD2* mutations predisposing to Crohn's disease were recently found to occur at a frequency of 6–12% among cases, but at <5% among controls (Hugot *et al*., 2001; Ogura *et al*., 2001).

Having identified the majority of coding and regulatory sequence SNPs with a frequency of greater than 5%, it is necessary to construct haplotypes to allow redundant SNPs to be identified and eliminated from the association study. A subset of 96 individuals from the population under study should be sufficient to detect the majority of haplotypes with a frequency of greater than 5% (B-Rao, 2001; note that studies to date indicate that common haplotypes at any given locus in Caucasian populations are restricted in number and account for the majority of all haplotypes observed; see Daly *et al*., 2001; Johnson *et al*., 2001). These haplotypes can then be used as a basis for selecting a minimal SNP set for the full association study. It should be noted, however, that SNPs which suggest a strong possibility of functional consequence (e.g. those that alter residues which are conserved between a number of species, or result in non-conservative amino acid changes; see Chapters 12–14) should not be excluded from analysis and should always be tested individually.

Clearly extensive haplotype information covering the entire human genome for a number of different populations would be an invaluable resource for all research groups undertaking association studies. Perlegen Sciences Inc. has recently released a haplotype map covering the whole of human chromosome 21. Although the number of chromosomes sampled was limited to 48, drawn from a number of different ethnic groups (Patil *et al*., 2001) this represents a good start in developing a genome-wide haplotype map. Perlegen's haplotype data have been incorporated into the Golden Path Browser (http://genome.ucsc.edu/) and can also be viewed at Perlegen's own website (http://www.perlegen.com/haplotype/).

8.2.3.4 Simple Tandem Repeat Markers (STRs)

STRs (also known as microsatellites — see Chapter 3) were the mainstay of monogenic trait linkage analysis during the 1990s, but are now frequently overlooked following the explosion of interest in SNPs for population-based studies. STRs are out of favour for two main reasons: (i) they are less amenable to cheap, high-throughput genotyping methodology than SNPs and (ii) STRs typically have a much higher mutation rate than SNPs (up to 10^{-3} per meiosis compared with an average of 10^{-9} for SNPs; Ellegren, 2000). It has been suggested that this extreme mutation rate might confound genetic association studies, as a single microsatellite allele may represent an excessive number of haplotypes, having independently arisen on the different haplotypic backgrounds through mutation events (see Moffatt *et al*., 2000). This may prevent the detection of association between the STR allele and an adjacent polymorphism associated with disease. However, comparison of the entire STR allele frequency distribution profiles for cases and controls may highlight differences which reflect a difference in the frequency of an adjacent disease-associated SNP, due to divergence of the STR profiles associated with SNP allele 1 and SNP allele 2 as a result of frequent STR mutation (Abecasis *et al*., 2001; Koch *et al*., 2000). There is also some evidence to suggest that LD can be detected over greater distances with STRs than with SNPs; possibly 10 times as far (Koch *et al*., 2000), perhaps because in some circumstances STR mutation significantly outstrips recombination at flanking sites. Given the caveat that limited empirical evidence is available at present and the extent of detectable LD is likely to be highly locus and marker specific, inclusion of STRs spaced at intervals of 50–100 kb in a preliminary case–control analysis may assist in identifying regions of potential association within the critical interval that can be prioritized for follow-up with SNPs.

8.2.4 Statistical Analysis

Methods and software for the statistical analysis of both single marker and haplotype data in both a case–control and family-based cohort scenarios are described in detail in Chapter 11. Briefly, a chi-square analysis may be used to test for departure between observed and expected allele frequencies for a biallelic marker in a case–control cohort, while multi-allelic systems (e.g. STRs) may be tested by permutation using software such as CLUMP (Sham and Curtis, 1995a; see Section 8.3.2 below). Family-based samples such as parent–offspring trios and discordant sibs can be analysed using the transmission disequilibrium test (TDT) and associated methods (Spielman *et al*., 1993; discussed in depth in Chapter 11); although the TDT was originally developed for biallelic markers, an extension of the TDT has been developed for testing multi-allelic markers and haplotypes (Sham and Curtis, 1995b). For case–control studies, haplotypes can be assessed using software such as EHPLUS (see Section 8.2.3 above) which, in addition to haplotype construction, can be used for testing for differences in haplotype frequency between cases and controls (see Chapter 11).

8.3 A PRACTICAL APPROACH TO LOCUS REFINEMENT AND CANDIDATE GENE IDENTIFICATION

Figure 8.3 gives an overview of the practical process of locus refinement, candidate gene selection and testing for phenotype–genotype association using a case–control approach. Each step is described in detail in the following sections.

8.3.1 Sequence Characterization

The most popular web tools for the purpose of human genome sequence characterization are the human genome browser hosted by the National Center for Biotechnology Information (NCBI) at http://www.ncbi.nlm.nih.gov/, the ‘Golden Path’ genome browser hosted by the University of California, Santa Cruz at http://genome.ucsc.edu/ and the Ensembl human genome browser maintained by the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute at http://www.ensembl.org/Homo_sapiens/. These browsers are described and reviewed in detail in Chapter 5 and we refer the reader to Chapter 9 for a comprehensive description of methods for defining a locus between two genetic markers at the sequence level using these three tools.

8.3.2 STR Analysis

We suggest that the first step following complete locus characterization should be an attempt to identify regions of potential association within the critical interval using STRs.

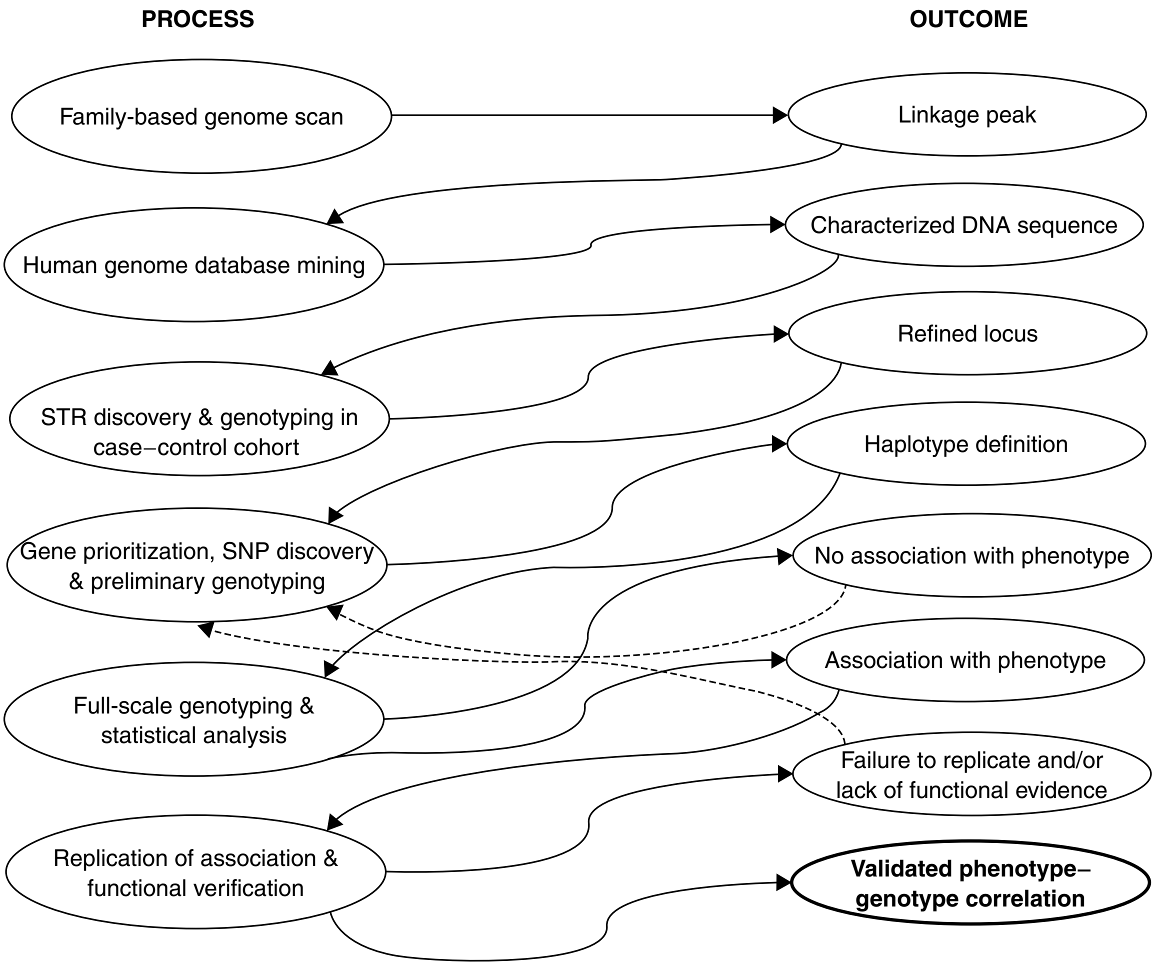

PROCESS
OUTCOME
Family-based genome scan
Linkage peak
Human genome database mining
Characterized DNA sequence
STR discovery & genotyping in case–control cohort
Refined locus
Haplotype definition
Gene prioritization, SNP discovery & preliminary genotyping
No association with phenotype
Full-scale genotyping & statistical analysis
Association with phenotype
Failure to replicate and/or lack of functional evidence
Replication of association & functional verification
Validated phenotype–genotype correlation

Figure 8.3 Flow diagram describing the logical steps in pinpointing and verifying a gene–phenotype association using a case–control follow-up to a family-based genome scan.

Any regions thus identified can then be prioritized for follow-up with SNPs. It should be noted however, that there is limited empirical data on the use of STRs to detect association in populations and lack of evidence for association with STRs should not deter the investigator from proceeding with a SNP-based association study. STRs can be identified using the tandem repeat finder software at http://c3.biomath.mssm.edu/trf.html (Benson, 1999; see Chapter 9). STR genotyping is typically performed by gel or capillary electrophoresis coupled with a fluorescence detection system (usually an adapted DNA sequencing platform); instrument and software suppliers include Molecular Dynamics, Applied Biosystems and LI-COR Biosciences.

Using a subset of 24 individuals from the control population, test the STRs for polymorphism; aim for a series of polymorphic STRs spaced at 50–100-kb intervals across the critical interval. These markers may then be typed in the entire case and control cohorts and the allele frequency distribution patterns checked for differences between cases and controls in an attempt to pinpoint areas of potential disease association. One of the most popular pieces of software for comparing STR allele frequency distributions is the CLUMP program developed by David Curtis and Pak Sham of the Institute of Psychiatry, London UK (Sham and Curtis, 1995a). CLUMP uses a Monte Carlo simulation to test for departure from expected values for a number of chi-square measures, including the intuitively appealing techniques of considering each allele in turn against the rest and grouping ('clumping') alleles to maximize the chi-squared value. CLUMP is straightforward to use and can be downloaded from http://www.mds.qmw.ac.uk/statgen/dcurtis/software.html.

8.3.3 Gene Selection, SNP Discovery and Haplotype Construction

Genes in the critical interval can be arranged in rank order for analysis, based on biological plausibility with respect to association with the disease under study or other considerations (e.g. pharmaceutical companies may wish to prioritize any tractable drug targets). We suggest sequencing coding and known or putative regulatory regions from each gene in 24 individuals selected randomly from the disease population, as discussed in Section 8.2.3 above, followed by genotyping of all SNPs thus identified in 96 random individuals derived from the control population. ARLEQUIN, EHPLUS (see Section 8.2.3 and Chapter 11) or similar software may then be used to construct haplotypes from this sub-sample, and htSNP (see Section 8.2.3 and Chapter 11) implemented to aid selection of the minimal marker set required for accurate representation of each haplotype. Note that several SNP genotyping platforms are currently available; we will not review them here and the investigator should select the most appropriate system based on cost, robustness and required throughput.

8.3.4 Genotyping and Statistical Analysis

Having selected the optimal SNP set, the whole cohort may now be genotyped. Following genotyping, an EM algorithm can be used again for haplotype construction and haplotype frequency determination (see Chapter 11). It may be beneficial to divide the cohort randomly into two case–control groups for statistical analysis, to allow the possibility of replication of any positive association using the second subset. Haplotype frequency distributions in cases and controls can be compared using CLUMP, as for STRs (see Section 8.3.3) or more specific software tools such as EHPLUS (see Section 8.2.3; discussed in detail in Chapter 11). Individual SNPs can be tested using a chi-square test (see Chapter 11). A test for Hardy–Weinberg equilibrium (HWE) is a useful prior check for ensuring that there is no (or little) population stratification and that each marker is giving the expected genotype distribution for the observed allele frequencies. Expected genotype frequencies are calculated from allele frequencies under the assumption $p^2 + q^2 + 2pq = 1$, where p and q are the allele frequencies and p^2, q^2 and $2pq$ correspond to the frequencies of the three possible genotypic states. The actual genotype frequencies are then tested for departure from the expected frequencies using a chi-square test. The calculation is simple and can be performed by hand or in a Microsoft Excel macro for biallelic markers. Alternatively the ARLEQUIN software suite includes a program for checking HWE for both biallelic and multi-allelic marker systems.

An acceptable p-value threshold for declaring association between a marker and disease is the subject of considerable debate. Clearly a nominal cut-off of $p = 0.05$ is inappropriate where multiple tests have been performed, as this value (or lower) may occur several times by chance. However, standard methods of correction for multiple testing, for example by Bonferroni correction, may be overly stringent (Cardon and Bell, 2001). The authors suggest that the investigator avoids setting thresholds that are excessively rigorous and instead follows up promising leads, adding to the weight of evidence for involvement (or lack of involvement) in the disease process by additional means (see below).

8.3.5 The Burden of Proof — is an Associated Gene Really Involved in the Disease Process?

Unfortunately detection of association between a gene and disease phenotype does not constitute definitive proof that the gene under test is involved in the disease process. Rather, it provides a single piece of evidence to suggest *possible* involvement in the disease process that requires further substantiation. Replication of the association in a second cohort considerably strengthens the argument for involvement; for example the association between the insulin gene and type 1 diabetes has been reproduced a number of times (Bennett and Todd, 1996). However, even in the event of independent replication of results, one should consider the possibility that the replication is due to chance or that the apparent disease association is due to an adjacent gene in LD with the marker under test. If the polymorphism is in protein coding sequence and causes an amino acid change, it may be possible to assess the possible impact on protein function by the nature of the change (conservative or non-conservative), the context in which it occurs (potential disruption of secondary or tertiary protein structure) and the degree of cross-species conservation and conservation within protein families. Conservation may also be used to gauge the potential impact of polymorphisms in putative regulatory elements. These areas are covered in detail in Chapters 12 and 13. However, it should also be remembered that polymorphisms that appear to be innocuous on cursory examination can have functional consequences, for example synonymous coding changes that occur in exonic splicing enhancer (ESE) regions (Liu *et al.*, 2001).

Ultimately it is likely that the investigator will wish to instigate additional laboratory-based experiments to judge the functional effect of the variant in question. These may include gene expression and cell-based reporter assays for putative promoter polymorphisms, functional enzyme or signal transduction assays for amino acid changes and *in vivo* analysis in the mouse using gene knock-out or polymorphism knock-in technology for studies in the context of the whole organism, to name but a small fraction of the available techniques.

8.4 CONCLUSION

In this chapter we have given a basic overview of the process of moving from a large genetic locus to the identification and screening of candidate genes for disease association. More detailed information on all aspects of study design and data analysis can be gleaned from the references cited in the text and further review of the literature and we strongly advise readers to broaden their knowledge beyond the limits of this chapter. Although we have highlighted a number of popular tools and techniques, several other equally valid approaches exist and we encourage investigators to actively seek out and develop further methods for comparison with those presented here. Continual development of new approaches and improvement of existing methodology is a dominant feature of this rapidly moving field; consequently there is a constant need for investigators to keep abreast of new developments to maximize the chances of success.

ACKNOWLEDGEMENTS

Many thanks to Aruna Bansal for a critical reading of the manuscript which resulted in significant improvements to this chapter.

REFERENCES

Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, *et al.* (2001). Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* **68**: 191–197.

Bennett ST, Todd JA. (1996). Human type 1 diabetes and the insulin gene: principles of mapping polygenes. *Ann Rev Genet* **30**: 343–370.

Benson G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

Blackwood EM, Kadonaga JT. (1998). Going the distance: a current view of enhancer action. *Science* **281**: 61–63.

B-Rao C. (2001). Sample size considerations in genetic polymorphism studies. *Hum Hered* **52**: 191–200.

Cardon LR, Bell JI. (2001). Association study designs for complex diseases. *Nature Rev Genet* **2**: 91–98.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. (2001). High-resolution haplotype structure in the human genome. *Nature Genet* **29**: 229–232.

Das M, Burge CB, Park E, Colinas J, Pelletier J. (2001). Assessment of the total number of human transcription units. *Genomics* **77**: 71–78.

Devlin B, Risch N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**: 311–322.

Devlin B, Roeder K. (1999). Genomic control for association studies. *Biometrics* **55**: 997–1004.

Devlin B, Roeder K, Bacanu SA. (2001). Unbiased methods for population-based association studies. *Genet Epidemiol* **21**: 273–284.

Ellegren H. (2000). Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet* **16**: 551–558.

Excoffier L, Slatkin M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12**: 921–927.

Goldstein DB. (2001). Islands of linkage disequilibrium. *Nature Genet* **29**: 109–211.

Gretarsdottir S, Sveinbjornsdottir S, Jonsson HH, Jakobsson F, Einarsdottir E, Agnarsson U, *et al.* (2002). Localization of a susceptibility gene for common forms of stroke to 5q12. *Am J Hum Genet* **70**: 593–603.

Hodge SE, Boehnke M, Spence MA. (1999). Loss of information due to ambiguous haplotyping of SNPs. *Nature Genet* **21**: 360–361.

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, *et al.* (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**: 599–603.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, *et al.* (2001). Haplotype tagging for the identification of common disease genes. *Nature Genet* **29**: 233–237.

Kruglyak L, Nickerson DA. (2001). Variation is the spice of life. *Nature Genet* **27**: 234–236.

Koch HG, McClay J, Loh EW, Higuchi S, Zhao JH, Sham P, *et al.* (2000). Allele association studies with SSR and SNP markers at known physical distances within a 1-Mb region embracing the ALDH2 locus in the Japanese, demonstrates linkage disequilibrium extending up to 400 kb. *Hum Mol Genet* **9**: 2993–2999.

Liu HX, Cartegni L, Zhang MQ, Krainer AR. (2001). A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nature Genet* **27**: 55–58.

McGinnis R. (2000). General equations for Pt, Ps, and the power of the TDT and the affected-sib-pair test. *Am J Hum Genet* **67**: 1340–1347.

Michalatos-Beloin S, Tishkoff SA, Bentley KL, Kidd KK, Ruano G. (1996). Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Res* **24**: 4841–4843.

Moffatt MF, Traherne JA, Abecasis GR, Cookson WO. (2000). Single nucleotide polymorphism and linkage disequilibrium within the TCR alpha/delta locus. *Hum Mol Genet* **9**: 1011–1019.

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, *et al.* (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**: 603–604.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, *et al.* (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.

Pritchard JK, Donnelly P. (2001). Case–control studies of association in structured or admixed populations. *Theor Popul Biol* **60**: 227–237.

Pritchard JK, Rosenberg NA. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* **65**: 220–228.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. (2000). Association mapping in structured populations. *Am J Hum Genet* **67**: 170–181.

Risch NJ. (2000). Searching for genetic determinants in the new millennium. *Nature* **405**: 847–856.

Sham PC, Curtis D. (1995a). Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Ann Hum Genet* **59**: 97–105.

Sham PC, Curtis D. (1995b). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet* **59**: 323–336.

Spielman RS, McGinnis RE, Ewens WJ. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**: 506–516.

Strittmatter WJ, Saunders AM, Schmechel D, Pericak-Vance M, Enghild J, Salvesen GS, *et al.* (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci USA* **90**: 1977–1981.

Zhao JH, Curtis D, Sham PC. (2000). Model-free analysis and permutation tests for allelic associations. *Hum Hered* **50**: 133–139.

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray
Copyright © 2003 John Wiley & Sons, Ltd.
ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 9

Genetic Studies from Genomic Sequence

MICHAEL R. BARNES

GlaxoSmithKline Pharmaceuticals
Harlow, Essex, UK

9.1 INTRODUCTION

Without risk of hyperbole, the process of definition of a locus or gene in the human genome sequence is probably the single most valuable bioinformatics process that a geneticist can carry out. This immediately places a gene or locus in a wider context. Now we can quickly find out what known genes are in the locus, what evidence exists for novel genes and what markers are available across the locus to study these genes. Digging a little deeper into the data will soon tell us where the genes are expressed and what their biological role is likely to be. We will even gain an insight into some of the common variation that exists across the locus, and also what rare mutations exist in some of the genes. Finally study of the sequence of the locus itself can tell us something about the physical nature of the region. We may even be able to draw some conclusions about the likely genetic nature of the region in terms of recombination and Linkage Disequilibrium (LD). All this is possible before setting foot in the laboratory. But we also need to be aware that there are limitations to this approach and sometimes stepping into the laboratory is the only way to resolve these limitations.

The availability of the golden path is a great advance for genetics — it is too easy to forget the pre-genome era, imprecise genetic localizations are now superseded by absolute genome locations to the nearest base pair. But the golden path must be used carefully; with proper quality checks the data can serve as an invaluable template for genetics, without these checks it can create as many problems as it solves. These caveats should not be ignored. Firstly the golden path is a *draft* dataset composed of hundreds of thousands of fragments of various sizes with many gaps. The order and orientation of the fragments is often not known from the sequencing process itself. In some cases the same part of the genome will be duplicated in more than one fragment. To address the technical challenges of whole genome assembly, the golden path is released as defined ‘builds’ on a quarterly basis (Lander *et al*., 2001; reviewed in Chapter 5). This implicitly involves some lag in availability of the most current sequence data. At the time of writing this chapter (March 2002), the December 2001 release of the golden path was in use. It is important to be aware that golden path coordinates can only be compared if both tools are using the same build version of the golden path. Finally, if complete sequence across a locus is critical to a study, additional draft sequence may be available in addition to the material in the golden path; this can be identified by searching the Human Genome BLAST database at the NCBI.

In this chapter we will take a hands-on approach to the application of genomic sequence data to genetics. We will look at the key bioinformatic steps needed to take a genetic study from an initial LOD peak to laboratory genotyping. Figure 9.1 illustrates each step of this process, with genomic sequence as a common thread through every stage.

9.2 DEFINING THE LOCUS

A geneticist may come to be interested in a gene or locus by many routes. The locus could be identified as part of a published genome scan or as part of the scientist’s own work; it could be syntenic with a mammalian disease model; it could contain a candidate gene with biological rationale. The possibilities are limitless, but how ever the locus is identified the next steps to define the region in the human genome are similar. Firstly the locus needs to be defined as accurately as possible. In some cases, this may not be easy due to insufficient or unclear data. Taking a complex disease linkage peak as an example, the linkage across the locus may be defined by a broad flat peak, or multiple

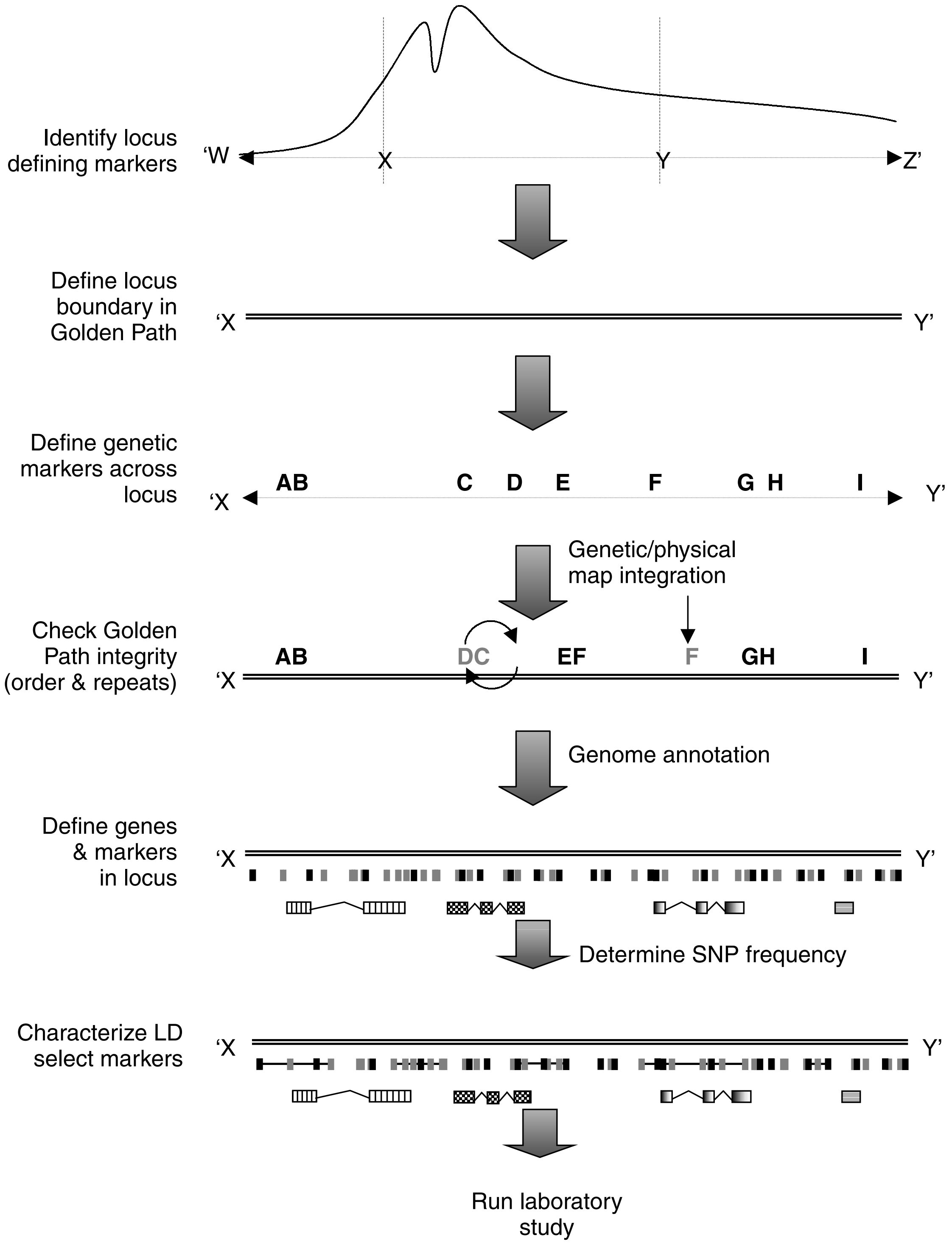

Identify locus defining markers
'W
X
Y
Z'
Define locus boundary in Golden Path
'X
Y'
Define genetic markers across locus
'X
AB
C
D
E
F
G H
I
Y'
Genetic/physical map integration
Check Golden Path integrity (order & repeats)
'X
AB
DC
EF
F
GH
I
Y'
Genome annotation
Define genes & markers in locus
'X
Y'
Determine SNP frequency
Characterize LD select markers
'X
Y'
Run laboratory study

Figure 9.1 Using the Golden Path as a template for genetics. Key bioinformatic steps to take a genetic study from an initial LOD peak to laboratory genotyping are illustrated. The reader should note the role of genomic sequence as a common thread through every stage.

peaks with no well-defined apex. In such cases, it is difficult to define a critical region; unlike monogenic diseases it is not possible to define a region by a clear recombination event between affected and unaffected family members. Instead analysis of complex traits generates an imprecise probabilistic signal based on the increased observance of an allele in affected versus unaffected individuals. Faced with such uncertainties, the best approach is to define a core region within a maximal region, based on LOD score thresholds; this gives some margin for error. Figure 9.2 shows an example of a theoretical complex disease linkage peak. In this case we define an acceptable 'core region' as any region with a LOD score of >3, with a 'maximum region' defined by markers with a LOD >2 or perhaps >1 (respectively a 10- and 100-fold drop in linkage probability). Definition of these regions is necessarily approximate. Where markers do not exactly define a locus it may be necessary to map markers on either side of the locus boundary. So for example in Figure 9.2 the core region would be defined as the region between markers E–K, while the maximum region would be encompassed between markers D and M. If linkage peaks are very flat, approximation to the nearest marker below the threshold might mean including a very large region, in such cases it may be worthwhile extrapolating between markers to identify the most probable region with an estimated LOD above the threshold.

Once the markers delineating the boundaries of the locus have been identified, they need to be mapped onto the human genome to view the full genomic context of the locus. In Chapter 5 Colin Semple reviewed the three primary tools which offer the user an opportunity to localize markers to the draft human genome assembly (the golden path). These tools are Ensembl at the EBI, the UCSC Human Genome Browser (HGB), and NCBI Map View. It is very easy to develop a preference for one or other of these tools, but each tool has its own distinct merits, so for complete characterization of a locus we recommend using all three. In practice this is easy as all three tools use the same coordinates from the same draft version of the golden path allowing reciprocal linking between each tool. Direct comparison between tools allows a second and third opinion across an identical region, which is always a good thing.

In this chapter we will describe some specific case studies which lead the reader through some of the common bioinformatics processes which can help the genetic characterization of genomic loci. In each case we will describe the use of the UCSC HGB, Ensembl or Map View to achieve a specific objective, but the reader should be aware that very similar approaches will produce similar results with each of the other tools unless otherwise mentioned. We will try to highlight the strong points of each tool, where possible with case studies, for an overview of the pros and cons of all three tools see Table 9.1.

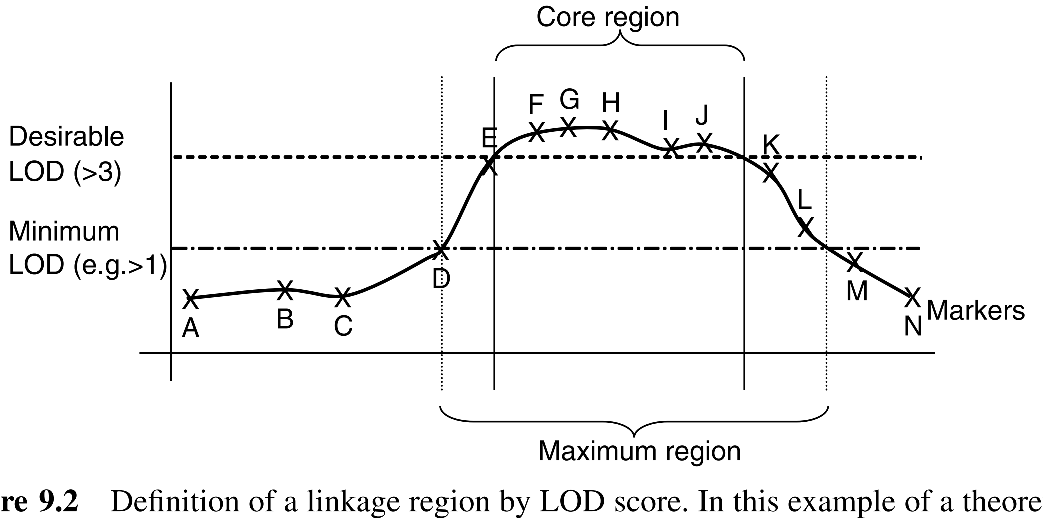


Figure 9.2 Definition of a linkage region by LOD score. In this example of a theoretical complex disease linkage peak, we define an acceptable 'core region' as any region with a LOD score of >3, with a 'maximum region' defined by markers with a LOD >2 or perhaps >1 (respectively a 10- and 100-fold drop in linkage probability).

TABLE 9.1 Pros and Cons of the Three Major Genome Viewers for Genetics

Map View	Ensembl	UCSC HGB
www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search	www.Ensembl.org/	genome.ucsc.edu/
Pros Good genetic/RH map integration Genetic/physical map focused Fully integrated with NCBI tools Comprehensive genetic markers Exclusive data: • GB4, G3 and TNG RH maps • Genethon/Marshfield maps • Mitelman morbid map • YAC contigs • FISH clones	Pros Innovative sequence annotation Sequence data focused Novel gene prediction focused Clean innovative interface Excellent data export Distributed annotation (DAS) Good integration with mouse and other genomes Free source code available Exclusive data: • Ensembl gene predictions • Detailed gene report • Eponine promoter prediction • Drosophila genome • Zebrafish genome • Mosquito genome • Many DAS tracks	Pros Innovative sequence annotation Sequence data focused Good sequence export Excellent for contig QC Archives previous golden path annotation/ converts coordinates Fast BLAT search tool Exclusive data: • Chr. 21 haplotype data • Fish genome comparison • SAGE expression data • NCI expression • GNF expression data • Identifies bridged gaps • Genome duplications • Novel tandem repeats
Cons Poor data export Limited sequence annotation Linked from Ensembl and HGB but no link back Complex and sometimes confounding interface	Cons No genetic/RH map integration No golden path version archive Accession numbers can be unstable between versions Not possible to view detailed region >1 Mb	Cons No genetic/RH map integration Data tracks appear and disappear frequently Newer golden path has less data tracks

Finally, the reader should note, that these browsers are subject to frequent change as datasets are updated and software tools improved. It is possible that the steps described here and the data returned may be superseded within months of being written! Nonetheless, the reader should not despair; we hope that the following examples may still be used as a rough overall guide for accessing the required information. Indeed, there are typically many routes to the same data and once familiar with the browser, the user will generally have no problem in retrieving the required datasets.

9.3 CASE STUDY 1: IDENTIFICATION AND EXTRACTION OF A GENOMIC SEQUENCE BETWEEN TWO MARKERS (RECOMMENDED TOOL: UCSC HUMAN GENOME BROWSER (HGB))

In Chapter 2 we reviewed methods for searching the literature for key elements of biological information. A literature search is an important preliminary stage for any study, to define the current state of knowledge in a specific research area. For the purposes of the following case studies imagine we are intending to follow up linkage results to investigate the role of a specific region in bipolar depression. Several linkage studies report a locus for bipolar depression in a region that we define between the genetic marker D21S1245 and the Radiation Hybrid marker D21S1852. To evaluate this region further — to get an idea of its physical and genetic size and number of genes — our first objective should be to locate the markers in the human genome assembly.

Marker localization to the human genome can be achieved either by searching by marker name or by searching directly with the sequence of the marker. The former approach can be problematic as no single tool contains a fully comprehensive index of genetic markers and their aliases. Map View probably contains the most comprehensive list of marker aliases but will not unambiguously localize a marker in genomic sequence. The UCSC HGB is a much more user-friendly tool for this purpose. From the home page select the most current ‘browser’ from the top left hand menu (for this exercise we used the December 2001 freeze). Type the marker names in the ‘position’ window, separated by a semicolon (e.g. D21S1245; D21S1852) and submit the request. This returns a 1.04-Mb sequence interval covering the genetic marker D21S1245 and the RH marker D21S1852. Note that in a marker name-based query HGB always returns a larger interval with 100 kb flanking either side of the markers. So in this case the markers D21S1245–D21S1852 actually encompass a 0.84-Mb interval.

The alternative strategy to searching by marker name is to search by marker sequence. This can be a useful technique to use when a marker alias is not found by the genome viewer; in such cases it may be necessary to consult other marker databases, such as GDB or dbSTS to retrieve a marker sequence (see Chapter 3). If all else fails and a marker sequence cannot be found it may be necessary to consult genetic and physical maps to find a neighbouring marker (see Chapter 8). Once a few hundred base pairs of sequence spanning each marker has been found, the sequence between the two marker locations can be identified by using the BLAT sequence search tool at HGB. Select ‘BLAT’ and enter the DNA sequence spanning the marker. Submit the search and make a note of the genomic position (take the ‘start’ position for the 5′ marker). Repeat for the second marker (take the ‘end’ position for the 3′ marker). Return to the genome browser and enter the range spanned by the two markers; this will return the exact genomic interval between the two markers. So for example, a BLAT search with D21S1245 and D21S1852 will return a 0.84-Mb locus without the flanking sequence retrieved by the marker name query. Now that this locus is defined it can be saved for future reference by simply adding a bookmark for the browser page.

9.3.1 Extracting the Genomic Sequence Across the Locus

Either of these approaches will define a genetic locus in the draft human genome sequence; once this has been achieved the sequence can be extracted to provide the ultimate physical map of the region (at 1-bp resolution!). To achieve this select the 'DNA' link in the top tool bar of the HGB, this presents the user with a number of basic options to format the DNA sequence across the selected region. If at this point you are only interested in the DNA sequence, select 'all lower case' and press the submit button. Alternatively, you can select 'lower case repeats' to highlight repeats in the sequence or you can mask them for primer design and other applications. There is also an option to reverse complement the sequence; this is particularly useful if you would like to retrieve a sequence across a gene that is in the reverse orientation in the golden path. If you would like to receive full annotation of the sequence in terms of all the features reported by the HGB then select 'extended case/colour options' and press the submit button. This will take the user to a highly sophisticated annotation interface which allows annotation of almost every available feature on the sequence, with a combination of toggled case, underlining, bold, italics and full colour lettering. This feature can be remarkably useful for preparing figures for publication etc., but bear in mind that the time to retrieve the sequence increases considerably with each added feature. In most cases toggled case annotation of repeats and exons is sufficient, also note that most sequence analysis tools will only maintain upper and lower case annotation, all other annotations (e.g. colour, underlining etc.) will be lost, unless viewed in a rich-text viewer, such as Microsoft Word.

If the sequence across the region is completely finished with no gaps (check the status of the clones in the assembly across the region) then this sequence can be used immediately for further genetic and genomic characterization (see Case study 3), however further QC is needed if the region is still in a draft form or contains gaps, as in the case of our locus between D21S1245 and D21S1852.

9.4 CASE STUDY 2: CHECKING THE INTEGRITY OF A GENOMIC SEQUENCE BETWEEN TWO MARKERS (RECOMMENDED TOOLS: UCSC HGB, NCBI MAP VIEW, NCBI EPCR)

In Figure 9.3 we show the UCSC HGB view of the locus identified between D21S1245 and D21S1852. The HGB interface can be configured to show and hide different datasets, for simplicity we only show tracks with an immediate application to the QC of the genomic sequence.

Now that the genetic locus has been identified in genomic sequence, the next key objective is to check the quality and orientation of the contig across the region. The view of this region immediately identifies two gaps, dividing the region into three contigs. The HGB differentiates between bridged and non-bridged gaps in contigs. Gaps not bridged by any other known physical clone or mRNA are indicated by a black box. If the relative order and orientation of the contigs on either side of the gap is known, then the gap is 'bridged' and indicated by a horizontal white line scored through the black box. This is a valuable feature; contigs between two non-bridged gaps should be evaluated to try to confirm the contig order and orientation. In Figure 9.3 neither gap is bridged, this makes it possible that contig NT_030187 between the two gaps could be incorrectly orientated, or even in the wrong location. This is important information to determine, as this contig constitutes one-third of the entire locus and also contains a complete gene.

9.4.1 Detecting Duplications in Genomic Assemblies

Localized duplications are a common error during genome assembly. Detection of these errors in human genome sequence is complicated by the high number of genuinely duplicated regions, which are estimated at around 3% of the total genome (Bailey *et al.*, 2001).

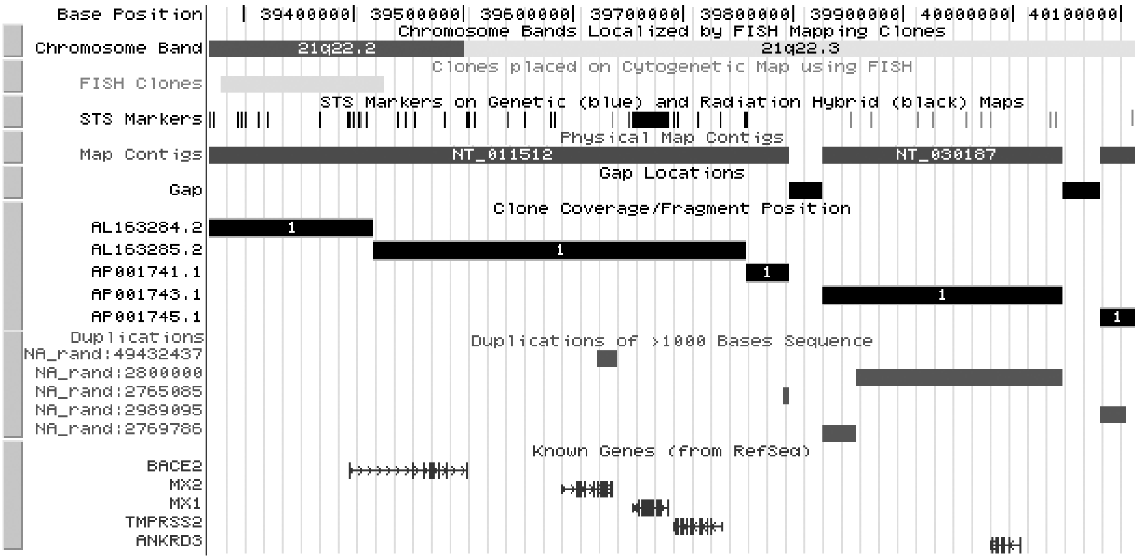
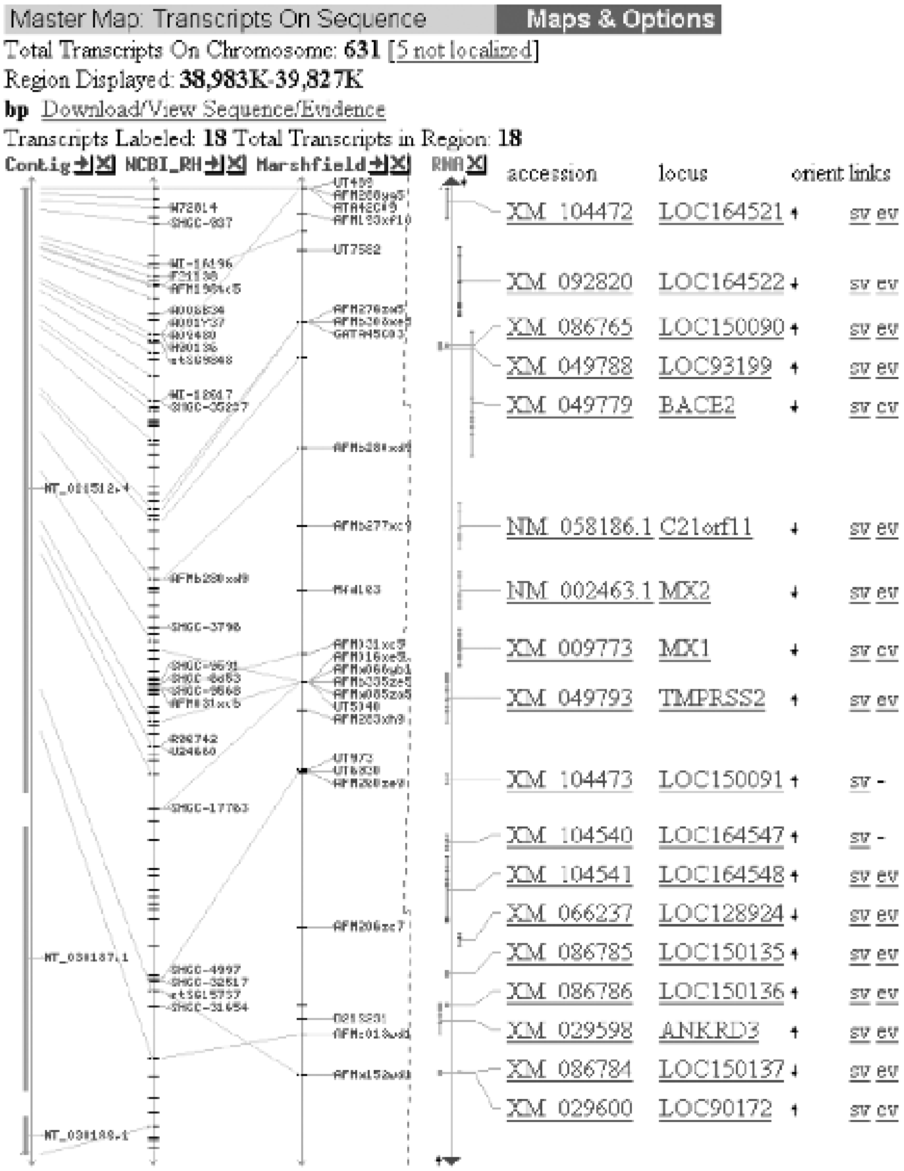


Figure 9.3 Definition of the D21S1245 and D21S1852 interval in genomic sequence using the UCSC human genome browser. The view immediately identifies five known genes, two gaps and several duplications across the region (See Colour Plates).

The HGB interface presents a very useful ‘duplication’ track. This identifies regions of >1000 bp which are duplicated in other golden path contigs. In this case five duplicated regions are identified across the region; the duplications are coloured red indicating that they have 99% or more similarity. Duplications of 98–99% and 90–98% similarity are shown in yellow and grey respectively. The left hand column reports the location of the duplicate regions. In this case the contigs are all from various ‘NA_rand’ locations — these are singleton contigs which cannot be placed in a chromosome contig. These are most probably missed overlaps between contigs, so they may not be an overt cause for concern, although this continues to suggest that this region may need careful curation.

The gaps and duplications across this region highlight a need for further QC to validate the sequence assembly before using the sequence as a basis for the construction of a laboratory study. One key *in silico* approach to validate the order and orientation of these contigs is to compare and integrate them with the RH and genetic map frameworks across the human genome. The HGB interface includes links to identical golden path regions in Ensembl and Map View. By selecting the Map View link a view of the region appears in a new window with a default view of the chromosome 21 contig map, unigene clusters and genes. To view the integrated maps, select ‘Maps & Options’. Another window will open. In this window select the following pull-down menus, the ‘NCBI RH’ map, the ‘Marshfield’ Genetic Map and ‘Transcript (RNA)’, finally select the ‘show connections’ tick box and click ‘Apply’. The main Map View window will now reload to show an integrated view of the RH maps and genetic maps across the locus. Markers shared between maps are linked by lines, which also show the location of the markers in the human genome contig. The data can also be viewed in a tabular view (with extra information) by selecting the ‘Data as Table View’ link. Figure 9.4 displays the Map View returned for this region. Examination of the NCBI Integrated RH map supports the correct ordering and orientation of the first contig (NT_011512). The genetic map is also broadly in agreement with the RH map, although there are some conflicts, however comparison with other RH maps supports the RH order for NT_011512 (not shown). Map View also supports the order of the third contig (NT_030188), however no links are drawn to support the



Master Map: Transcripts On Sequence | Maps & Options

Total Transcripts On Chromosome: **631** [5 not localized]

Region Displayed: **38,983K-39,827K**

bp Download/View Sequence/Evidence

Transcripts Labeled: **18** Total Transcripts in Region: **18**

accession	locus	orient	links
XM_104472	LOC164521	↑	sv ev
XM_092820	LOC164522	↓	sv ev
XM_086765	LOC150090	↑	sv ev
XM_049788	LOC93199	↑	sv ev
XM_049779	BACE2	↓	sv ev
NM_058186.1	C21orf11	↓	sv ev
NM_002463.1	MX2	↓	sv ev
XM_009773	MX1	↓	sv ev
XM_049793	TMPRSS2	↑	sv ev
XM_104473	LOC150091	↑	sv -
XM_104540	LOC164547	↑	sv -
XM_104541	LOC164548	↑	sv ev
XM_066237	LOC128924	↓	sv ev
XM_086785	LOC150135	↑	sv ev
XM_086786	LOC150136	↑	sv ev
XM_029598	ANKRD3	↑	sv ev
XM_086784	LOC150137	↓	sv ev
XM_029600	LOC90172	↑	sv ev

Figure 9.4 Using NCBI Map View to produce an integrated view of genetic and physical map data. Examination of the RH maps and genetic maps support the correct ordering and orientation of the first and third contigs (NT_011512 and NT_030187). No links are drawn to support the order between any of the maps and the second contig (NT_030187) spanned by the gaps, this requires further investigation.

order between the RH map and the second contig (NT_030187) spanned by the gaps. This highlights one problem with Map View: it does not comprehensively localize markers in the human genome draft sequence. In the data reviewed so far there is no firm evidence to place NT_030187 in the region under study. The only solution to this problem is to extract the genomic sequence across the region and screen it directly for matches to STS marker sequences. The tool to achieve this is Electronic PCR (ePCR) at the NCBI (http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi). This tool maps known STSs from the dbSTS and RHdb databases to a submitted sequence (Schuler, 1998).

Submission of the 0.84-Mb sequence to the NCBI ePCR server identifies 57 STS markers across the sequence. The contig we need to check, NT_030187, maps from 0.56

TABLE 9.2 RH Marker Order versus Sequence Order across Contig NT_030187

Location in Region (Base Pair)	Marker	TNG cR50000 (LOD)	GeneMap99 GB4 cR3000 (LOD)
489,493–489,804	SHGC-140546	16451 (12.3)	—
558,505	GAP1	—	—
645,681–645,988	SHGC-147821	**16519(8.7)**	—
659,283–659,421	D21S1449	**16508(14.4)**	—
689,527–689,653	stSG46899	—	**222.98(3.00)**
703,493–703,795	D21S356	—	—
711,656–711,819	stSG52786	—	**222.23(2.43)**
767,022–767,332	SHGC-148000	**16583(9.7)**	—
770,359–770,649	WI-20889	**16567(14.6)**	**225.80(3.00)**
771,401–771,520	stSG3262	**16563(17.1)**	**225.21(3.00)**
777,760	GAP2	—	—
843,780–844,059	D21S1852	16617 (11.9)	—

to 0.78 Mb, 10 STS markers span this region. In Table 9.2 we list these markers and note their map order in the other available maps. Integration of maps presents a somewhat confusing picture. Both the TNG and GeneMap99 GB4 maps show local discrepancies in marker order. These results are somewhat inconclusive, RH map resolution is unreliable below 60–100 kb so these localized discrepancies in marker order over 10–30-kb regions may be due to the lack of resolution of the maps. Alternatively it is possible that the finished BAC clone (AP001743) which constitutes most of the NT_030187 contig may contain a sequence rearrangement; further laboratory analysis would be required to confirm this. However, the overall order of the contigs in this region appears to be supported by the integrated maps across the region.

9.5 CASE STUDY 3: DEFINITION OF KNOWN AND NOVEL GENES ACROSS A GENOMIC REGION (RECOMMENDED TOOLS: ENSEMBL AND HGB)

Now that we are (relatively) sure of the order and orientation of the contig across the D21S1245–D21S1852 region, it is important to identify all the known and novel genes in the region, so that they can be evaluated as candidates or to ensure that marker maps across the region are sufficient to detect any genetic effect in genes or regulatory regions. In Chapter 4 we presented a detailed examination of the art of delineating genes in genomic sequence. The UCSC human genome browser and Ensembl are valuable assistants in this process. Both tools run the human genome sequence through sophisticated gene prediction pipelines (Hubbard *et al.*, 2002). These analyses are coupled with a detailed view of supporting evidence for genes, such as ESTs, mouse and fish genome homology, CPG islands and promoter predictions. This wealth of data probably makes further *de novo* gene prediction unnecessary in most cases; improvement on the quality of annotation provided by Ensembl and HGB would require an in-depth understanding of the intricacies of gene prediction, which we cannot hope to impart in this book (see Rogic *et al.* (2001) for an excellent review of this field). Instead we suggest that the user focuses on the available data to build gene models based on existing annotation. Genetics has one advantage over other fields of biology: detection of a genetic effect does not require a completely accurate working model of a gene (although obviously this will help). All that should be needed is an approximate gene model which can be screened by identifying proximal polymorphisms. The only exception to this might be during the analysis of functional polymorphisms where an accurate model of a gene and its regulatory regions may be critical.

For the purposes of our study, we need to identify all known and novel genes across the locus. In Figure 9.5 we show a magnified HGB view of the first contig in the region, NT_011512. Again we have configured the browser to show tracks which are directly applicable to the identification of genes in genomic sequence. Without going into an overt level of detail there are five known genes identified across the region. A number of extra tracks show pieces of evidence which support the known genes and suggest the possible existence of a further four novel genes across the locus which we indicate at the bottom of the figure. Confidence in the identification of novel genes in genomic sequence is in part dependent on the range and nature of supporting evidence. The most convincing single item of evidence is a correctly spliced mRNA transcript, either an EST or whole

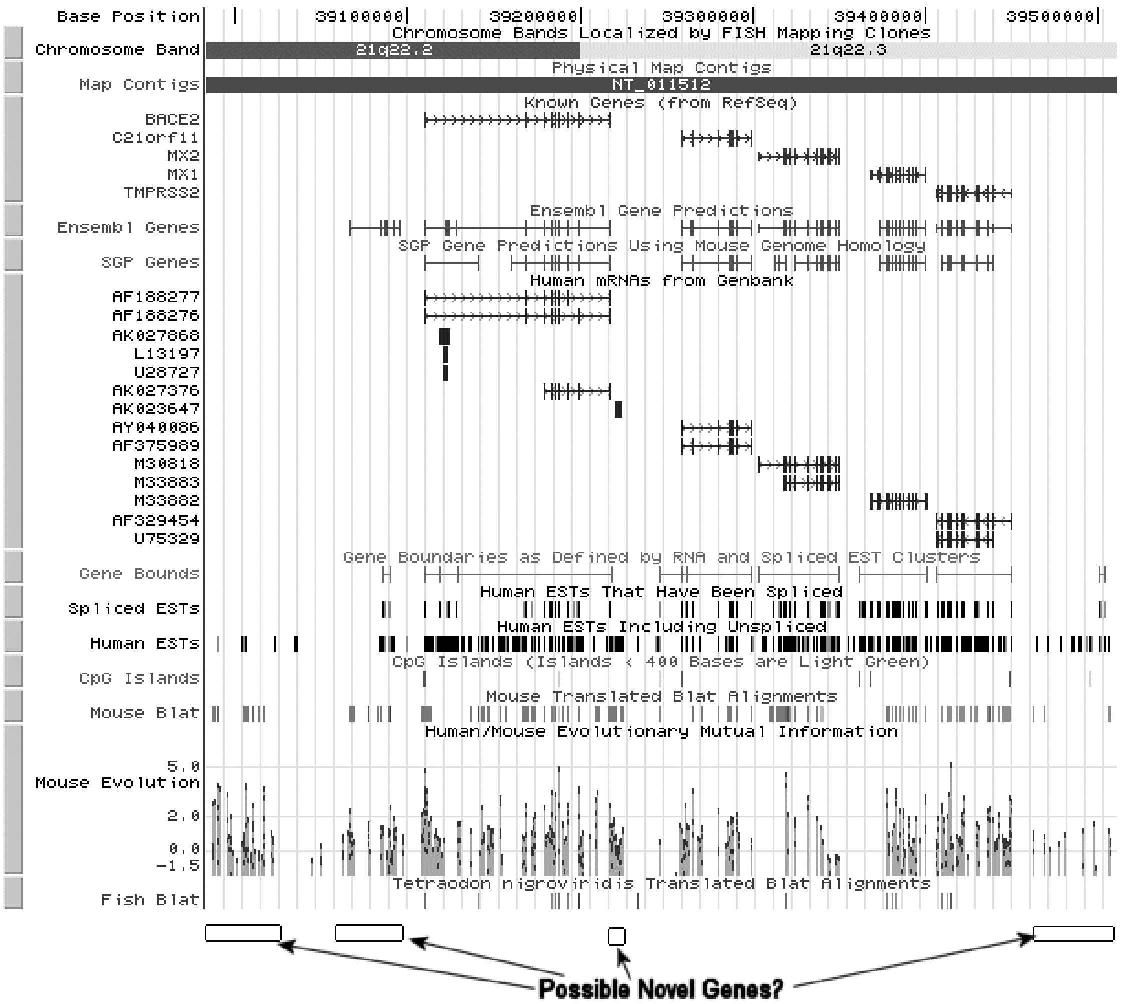


Figure 9.5 Using the UCSC human genome browser to identify known and novel genes. A range of evidence including mRNAs, ESTs and human–mouse homology supports the existence of five known genes and up to four novel genes (See Colour Plates).

transcript. In this figure several human mRNAs from GenBank are identified across the region. These include splice variants and redundant entries of the same transcript. So for example at least two splice variants are apparent for the BACE2 gene (AF188277 and AF188276). There are no novel mRNA transcripts in the GenBank track, however there are a large number of human ESTs which do not appear to map to any known gene. Human ESTs are divided into spliced and unspliced tracks. This is in recognition of the very high number of artefacts that are generated in EST libraries. Spliced ESTs, that is, ESTs which align across exons are much more reliable confirmatory evidence for genes than unspliced ESTs. A final strong source of evidence for genes are the range of tracks which show homology to non-human DNA, including non-human mRNA and comparison with mouse and fish genome sequences. Strong sequence conservation between man and other vertebrates is generally thought to be restricted to coding or regulatory regions. The four putative novel genes across this locus are supported by a range of spliced ESTs, mouse homology and gene prediction. Taken individually each of these pieces of evidence might not be sufficient to reliably support the existence of a novel gene, but taken together they are quite convincing. All that remains is to characterize these novel genes in terms of homology and putative function; we reviewed this process in Chapter 4.

9.6 CASE STUDY 4: CANDIDATE GENE SELECTION — BUILDING BIOLOGICAL RATIONALE AROUND GENES (RECOMMENDED TOOLS: HGB, ENSEMBL)

So far in our study of the D21S1245–D21S1852 region, we have identified our locus in genomic sequence and identified the known and putative novel genes in the region. Further genetic analysis of the region could now take two routes, we could perform further linkage or association studies by defining a suitable set of markers across the region (see Case study 5) or alternatively we could select specific candidate genes for follow-up studies. As we only have nine or 10 genes in our locus it would be quite viable to study each gene, but in most cases a region will contain a much larger number of genes which would make follow-up of each gene an impractical approach. An alternative in such cases would be to prioritize candidate genes based on their biological rationale in the target phenotype or trait. Criteria for biological prioritization of candidate genes are discussed throughout this book. Genes can be prioritized based on a known or putative role in the disease pathway, gene knock-out models, expression in the disease tissue, functional polymorphism and many other criteria.

In our hypothetical study we are looking for a gene with a possible role in bipolar depression, therefore to prioritize our candidate genes, we might first review the literature to search for a link between the candidate genes in the region and this disease pathway. The aetiology of bipolar disorder, like many complex diseases is poorly understood, this makes it difficult to establish a clear biological rationale for any gene in this disorder. Where biological rationale is found it could range from convincing support, such as upregulation of the gene or a related gene or pathway component in a disease model or in a similar phenotype to the most basic support, such as being expressed in a tissue affected by the disease.

Drawing together the complex strands of evidence in the literature is a skill that calls for a good background in biology and ideally a broad understanding of the disease under study. However reliance on literature-based evidence alone can run the risk of over-interpreting tenuous links between genes. This could be a particular problem in the case of poorly understood diseases, where unknown pathways would largely fail to register as a form of rationale. This issue is an argument to support a truly investigative approach to candidate gene identification. The candidate should be in the right place at the right time; beyond this further assumptions may be misleading. Data presented by tools such as Ensembl and the UCSC HGB can provide solid evidence which can help to identify genes which are at least expressed in the tissues affected by the disease. Obviously in the case of bipolar disorder, we are most interested in genes which show evidence of expression in the brain. This will inevitably include a large number of genes. In an analysis of the expression profiles of >33,000 genes, Su *et al.* (2002), found that on average any individual tissue expresses approximately 30–40% of known genes. For candidate gene studies, this implies that 30–40% of genes are likely to be candidates on the basis of expression in the disease tissue (assuming the disease affects only one tissue).

9.6.1 Analysis of Gene Expression

Four tracks in the UCSC HGB provide information about the tissue expression profiles of genes (Figure 9.6). The simplest level of information is provided by ESTs, each is implicitly a measure of gene expression, as each is derived from a specified tissue source. The UniGene track clusters all ESTs which map to a gene. Viewing ESTs from a UniGene record will confirm the expression of a gene in a particular tissue. The number of ESTs represented in each tissue will also give a *very* rough idea of the expression levels of the

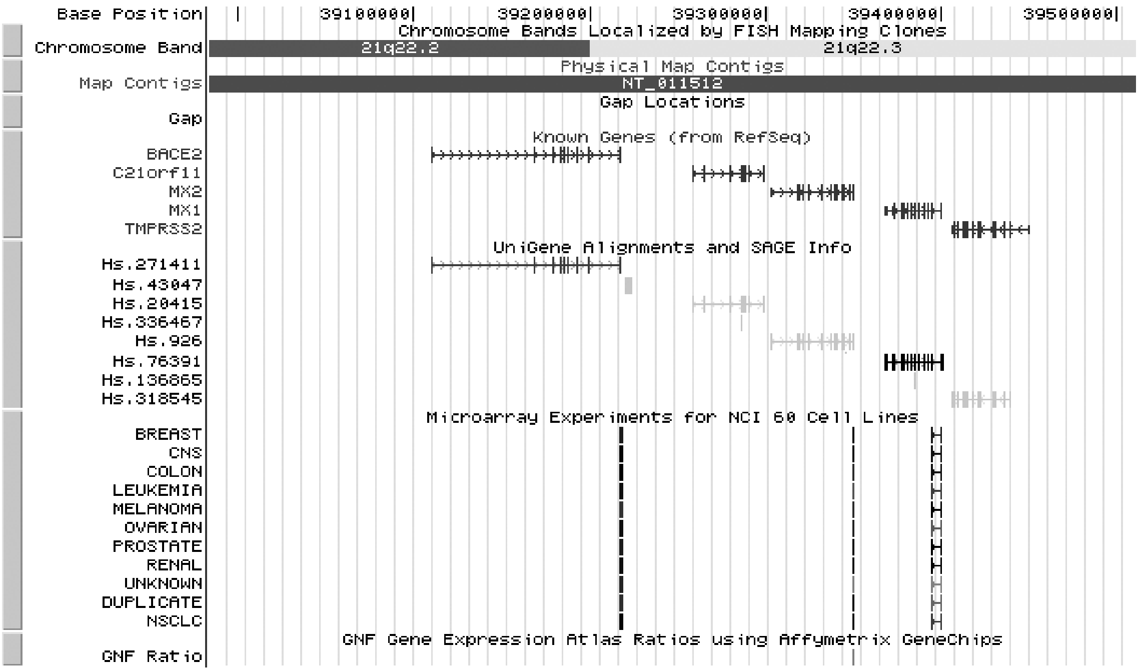


Figure 9.6 Using the UCSC human genome browser to evaluate gene expression across a locus. Four tracks provide information on gene expression. ESTs implicitly measure gene expression as each is derived from a specified tissue source. UniGene clusters link to SAGE expression profiles drawn from the SageMap project. Finally data from two genome-wide microarray projects are presented, the NCI60 cell line project and GNF gene expression atlas ratios (See Colour Plates).

gene, but it will not confirm the absence of a gene in a tissue. The most comprehensive measure of gene expression is also linked from the UniGene track in the HGB interface. Clicking on a UniGene cluster links to SAGE expression profiles drawn from the SageMap project (Lash *et al.*, 2000). Finally the HGB interface also provides two specialist tracks containing data from two genome-wide microarray projects, the NCI60 Cell Line Project (Ross *et al.*, 2000) and GNF Gene Expression Atlas Ratios (Su *et al.*, 2002; Table 9.3).

9.6.2 Serial Analysis of Gene Expression (SAGE)

The HGB UniGene track links to SAGE data for eight of the 10 genes in the D21S1245–D21S1852 interval. SAGE is a quantitative measure of gene expression based on tags in the $3'$ UTR of genes (see Chapter 16 for a detailed explanation of this technique). Clicking on a UniGene cluster returns a table of SAGE data for every Unigene cluster contained in the browser window. Selecting the Unigene cluster name will display the SageMap page for the cluster, leading to a so-called ‘Electronic Northern’, which might more appropriately be called an ‘electronic dot-blot’, as there is no element of transcript sizing. The summary data from the SageMap of the genes in the region is presented in a tabular form, so if you would rather see a graph across the region, it is quite easy to export this data to a spreadsheet to plot the expression profiles of each gene. SAGE expression data is available in a selected range of brain and neuronal libraries for eight of the genes across the locus. This data identifies expression of BACE2, PAPPA, Novel3, C21orf11, MX2 and MX1 in some of the normal brain tissues in the SAGE libraries. BACE2, Novel3 and MX1 show high expression in a wide range of normal brain tissues, which suggests that these genes may warrant priority as candidates. The TMPRSS1 and Novel2 genes show no evidence of brain expression and so it may be reasonable to reduce the priority of these genes.

The integration of SAGE data across the genome within the HGB interface, makes SAGE data one of the most comprehensive and convenient measures of gene expression across a genetic locus, allowing the user to quickly identify most genes across the locus which are expressed in a range of common tissues. The only real limit to this method is in the number and type of tissues, although these are extensive and growing in numbers (see http://www.ncbi.nlm.nih.gov/SAGE for a list of available tissues).

TABLE 9.3 Comparison of Public Domain Genome-Wide Expression Datasets

Dataset	SAGE	NCI60 Cell Line Project	GNF Gene Expression Atlas
Technology	Serial analysis of gene expression	cDNA microarray (Incyte)	Affymetrix U95a GeneChip microarray
No. of genes	>100 K tags[*]	∼8000	∼33 K
Tissues	37	—	33
Cell lines	4	—	—
Induced cell lines	3	—	2
Tumour material (incl. cell lines)	52	60	13
Total tissues	96	60	48
Reference	Lash *et al.* (2000)	Ross *et al.* (2000)	Su *et al.* (2002)

[*] SAGE tags are redundant across genes.

9.6.3 Microarray Data Tracks

9.6.3.1 The NCI60 Cell Line and GNF Gene Expression Atlas Ratios

The UCSC Human Genome Browser hosts two public domain microarray data tracks. The NCI60 cell line track presents data from a cDNA microarray experiment to assay the expression of more than 8000 genes among 60 tumour-derived cell lines used in the National Cancer Institute's (NCI) anti-cancer drug screens (Ross *et al.*, 2000; http://genome-www.stanford.edu/nci60/). The GNF track shows expression data generated from 46 human tissues and cell lines by the Genomics Institute of the Novartis Research Foundation (GNF) using a u95A Affymetrix GeneChip (Su *et al.*, 2002; http://expression.gnf.org/).

Both the NCI60 and GNF tracks are presented in a similar format, although the experimental details differ. Clicking on a transcript in either track, will bring up a tabular view of all genes in the current browser view, in which each column of coloured boxes represents the variation in transcript levels for a given cDNA across all of the array experiments and each row represents the measured transcript levels for all genes in a sample (Figure 9.7A). The variation in transcript levels for each gene is represented by a colour scale, in which red indicates an increase in transcript levels, and green indicates a decrease in transcript levels. These relative transcript levels are measured in a slightly different way between NCI60 and GNF tracks. In the NCI60 track expression levels are relative to a reference sample of 12 pooled tumour cell lines. In the GNF track the expression levels are relative to the signal of the probe in the particular tissue compared to the median signal of all experiments for the same probe. The saturation of the colour corresponds to the magnitude of transcript variation. A black colour indicates an undetectable change in expression and a grey box indicates missing data (see Su *et al.* (2002) and Ross *et al.* (2000) for a more detailed explanation of this method).

As the NCI60 data focuses on tumour-derived cell lines, it is not well suited for the determination of expression in normal tissues, although obviously this data would be very valuable for studies of cancer genetics. However, the GNF data track presents some very valuable information for complex disease genetics, including a breakdown of gene expression across different regions of the brain. This data is very valuable for candidate prioritization, as certain regions of the brain may have a more significant role in bipolar depression than others. For example, functional neuroimaging studies of bipolar patients have identified the thalamus as a key component of the main neuroanatomic circuitries which are altered in psychiatric illnesses, such as bipolar disorder (Soares and Mann, 1997). This information indicates that expression in the thalamus could help to prioritize candidate genes for analysis.

Only one gene from our core region, MX2, is represented in the GNF dataset, this shows low level expression throughout the different brain regions, with strongest expression in whole blood (data not shown). However if we expand the D21S1245–D21S1852 interval by 1 Mb on either side to include other flanking genes, much more data becomes available. Figure 9.7A shows a view of a selection of the available tissues, including all neuronal tissues for 15 genes across the wider locus. One gene, Purkinje cell protein 4 (PCP4–no. 37576), is immediately apparent with a high level of expression in a wide range of neuronal tissues, including the thalamus. By clicking on the PCP4 gene number in the expression view in HGB, a detailed expression profile of the gene in all available tissues is launched into a new window (Figure 9.7B). This shows that the expression of this gene is primarily limited to the brain, thyroid and prostate glands, with highest levels of expression in the caudate nucleus and thalamus. Obviously this makes PCP4 an interesting candidate gene.

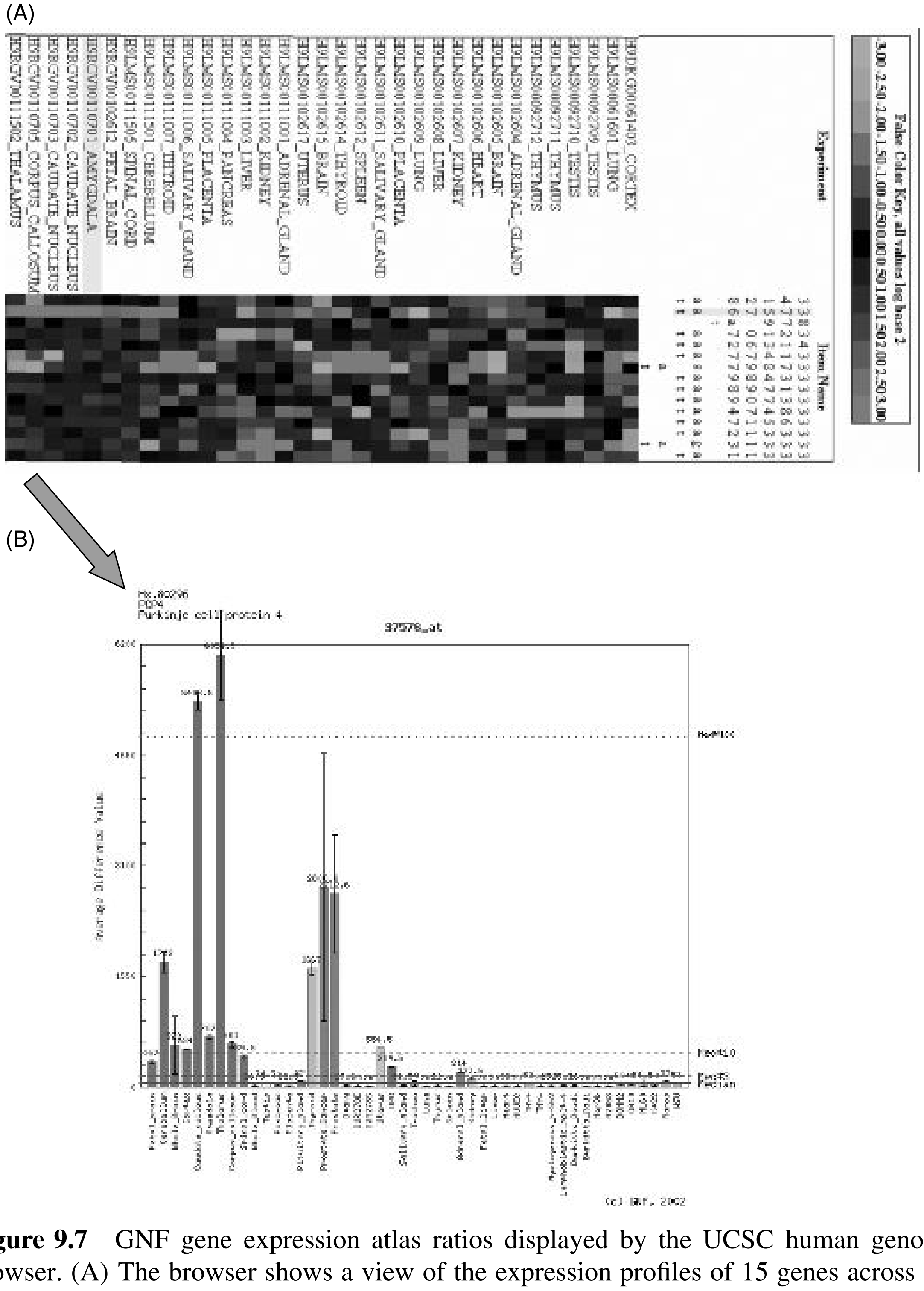


Figure 9.7 GNF gene expression atlas ratios displayed by the UCSC human genome browser. (A) The browser shows a view of the expression profiles of 15 genes across the wider locus. One gene, Purkinje cell protein 4 (PCP4, no. 37576), shows high expression in a wide range of neuronal tissues, including the thalamus. (B) Detailed gene expression profile for the PCP4 gene (See Colour Plates).

The relative cost of microarray technology was a cause of major concern for the academic research community, prompting fears that microarray data would be the preserve of cash-rich industry and biotech. However public domain projects like the NCI60 and GNF gene expression atlas should in part allay these fears. Although these microarray projects currently provide a somewhat limited coverage of human genes, they are complemented by other technologies such as SAGE and both types of data are constantly expanding in the public domain.

9.7 CASE STUDY 5: KNOWN AND NOVEL MARKER IDENTIFICATION (RECOMMENDED TOOLS: ENSEMBL, HGB, MAP VIEW, SNPPER)

Now that we have identified the genes in our locus and established some biological rationale to prioritize them for study, we need to ascertain which markers are available to complete this study. The human genome is a very convenient framework for organization of polymorphism data and so genome viewers are probably the best tools for identifying these polymorphisms.

All the genome viewers maintain SNP annotation across the human genome. Exact numbers of SNPs reported may differ between tools, for example, comparison of the D21S1245–D21S1852 locus between Ensembl, UCSC HGB, Map View and SNPper (see below), identifies 901, 903, 903 and 876 SNPs respectively. These minor discrepancies are a likely result of the different SNP mapping and repeat masking parameters between the tools. Missing a SNP or two across a locus may not be a problem for a large-scale analysis but if candidate gene analysis is the objective, it may be important to identify all variation to enable accurate construction of haplotypes or identification of potentially functional variants.

9.7.1 Identification of Potentially Functional Polymorphisms

Aside from the ordered convenience that genome browsers bring to SNP data, they also place a SNP into a full and diverse genomic context, giving information on nearby genes, transcripts and promoters. Both Ensembl and HGB show genome conservation between human and mouse, while HGB also includes tetradon and fugu (fish) genome conservation. Genome conservation between vertebrates is generally restricted to genes (including undetected genes) and regulatory regions (Aparicio *et al.*, 1995). Hence this is a simple but powerful method for identifying SNPs in regions which are potentially functionally conserved. Figures 9.8A shows a detailed HGB view of the BACE2 gene. After close viewing of the locus it is possible to assess the functional context and genomic conservation of the region surrounding each SNP. Where an overlap is unclear it is possible to zoom in to a resolution of just a few hundred base pairs to determine exact locations and overlaps between SNPs and gene features (Figure 9.8B). At the simplest level, identification of potentially functional SNPs is a matter of identifying SNPs which overlap highly conserved regions or putative gene or regulatory features. The UCSC browser presents some detailed information on putative promoter regions, including golden triangle and transfac analyses (see Chapter 12). Once identified, the impact of different alleles can be evaluated by running the alleles through the tool originally used to predict the sequence feature. These could include tools for promoter prediction, splice site prediction or gene prediction. In Chapters 12–14 we describe these analysis approaches in detail. One final point on the functional characterization of SNPs is that we currently know very little about the functional regions of the genome; but we do know that our tools are very limited, and so it is almost impossible to conclude that a SNP is *not* functional. All that we can do is make our best guess from the available evidence.

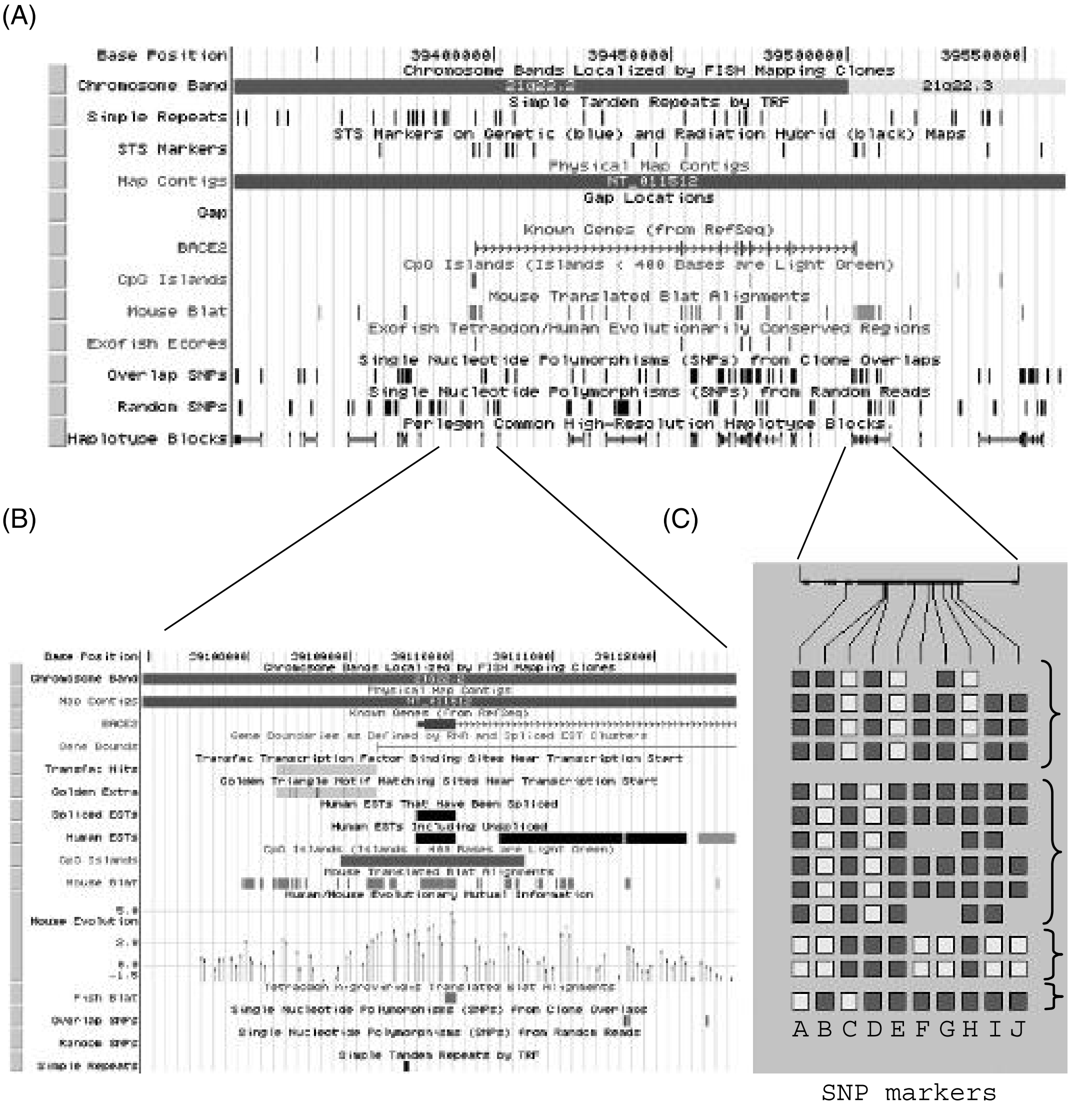

(A)
Base Position
39400000
39450000
39500000
39550000
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
21q22.2
21q22.3
Simple Tandem Repeats by TRF
Simple Repeats
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Physical Map Contigs
Map Contigs
NT_011512
Gap Locations
Gap
Known Genes (from RefSeq)
BACE2
CpG Islands (Islands < 400 Bases are Light Green)
CpG Islands
Mouse Translated Blat Alignments
Mouse Blat
Exofish Tetraodon/Human Evolutionarily Conserved Regions
Exofish Ecores
Single Nucleotide Polymorphisms (SNPs) from Clone Overlaps
Overlap SNPs
Single Nucleotide Polymorphisms (SNPs) from Random Reads
Random SNPs
Perlegen Common High-Resolution Haplotype Blocks
Haplotype Blocks
(B)
Base Position
Chromosome Band
Map Contigs
BACE2
Gene Bounds
Transfac Hits
Golden Extra
Spliced ESTs
Human ESTs
CpG Islands
Mouse Blat
Mouse Evolution
Fish Blat
Overlap SNPs
Random SNPs
Simple Repeats
Known Genes (from RefSeq)
Transfac Transcription Factor Binding Sites Near Transcription Start
Golden Triangle Motif Matching Sites Near Transcription Start
Human ESTs That Have Been Spliced
Human ESTs Including Unspliced
Mouse Translated Blat Alignments
Human/Mouse Evolutionary Mutual Information
Single Nucleotide Polymorphisms (SNPs) from Clone Overlaps
Single Nucleotide Polymorphisms (SNPs) from Random Reads
Simple Tandem Repeats by TRF
(C)
W
X
Y
Z
HAPLOTYPES
A B C D E F G H I J
SNP markers

Figure 9.8 SNP visualization in the UCSC human genome browser. (A) A detailed view of the BACE2 gene allows the user to view a range of information, including SNP haplotype data. (B) Close viewing of the BACE2 locus allows the user to assess the functional context and genomic conservation of the region surrounding each SNP. (C) A detailed view of a Perlegen haplotype (See Colour Plates).

9.7.2 Identifying Novel Microsatellites in Sequence Data

The polymorphism data we have identified so far, may be sufficient for a SNP scan across our locus, however in Chapter 8 we reviewed some data to suggest that LD may be detected over greater distances with STRs (Koch *et al.*, 2000). Inclusion of STRs spaced at intervals of 50–100 kb in this locus may assist in narrowing a critical interval to a distance of a few hundred kb in a case–control association study. Known genetic marker maps identify three polymorphic STRs across the 0.84-Mb, D21S1245–D21S1852 interval. To generate a sufficiently dense map, we need 10 to 15 extra markers. Potentially polymorphic STRs can easily be detected across a given region by using Tandem Repeat Finder (Benson, 1999; http://c3.biomath.mssm.edu/trf.html). The UCSC HGB interface already presents output from the Tandem Repeat Finder in the ‘simple repeats’ track. Figure 9.8B shows an example of this output — a simple repeat in the promoter region of BACE2 is identified by this tool. Use of the Tandem Repeat Finder interface identifies tandem repeats in a submitted sequence over a user-defined repeat unit size range (from 1 to 500 bp). Perfect or close to perfect tandem repeats of greater than 12 repeat units tend to be polymorphic (see Fondon *et al.* (1998) for a review of these methods).

9.7.3 Exporting SNP and Microsatellite Data

If comprehensive SNP coverage is a priority then results from both Ensembl and HGB can be compared and collated. The easiest way to do this is to compare both exported SNP sets in a Microsoft Excel spreadsheet. Both Ensembl and HGB have facilities to export SNP and microsatellite information across a defined locus. In Ensembl the user needs to select the 'export' menu above the detailed analysis window and select 'SNP list'. This allows the user to retrieve information about SNP accessions, golden path location and gene region directly into Microsoft Excel. The UCSC HGB is slightly more complicated, the user needs to select 'TABLES' and then separately select 'Random SNPs' and then repeat with 'Overlap SNPs' from the menu. This produces two similar tab-delimited files. Both files can be loaded into Excel and sorted by golden path location to obtain a non-redundant SNP map across the region.

9.7.4 Construction of Marker Maps

To complete an LD-based association scan across this region, we need to define a sufficiently dense set of markers to detect LD across the region. In the absence of knowledge of the haplotypic diversity of the interval in question, accurate selection of an optimal marker set is not possible. However a framework of markers spaced at 10–30-kb intervals might be appropriate; this is accepted as a reasonable assumption of LD in northern European populations (Ardlie *et al.*, 2002). Informative SNP markers for association studies need to be carefully selected, ideally with an allele frequency of around 25%, and generally no lower than 5%. Lower frequencies would require very large sample sizes to reach a sufficient power to detect association (Johnson *et al.*, 2001; see Chapter 8). Unfortunately this creates a technical problem, most SNPs from dbSNP, are 'candidate SNPs' with no available frequency information (see Chapter 3 for a discussion of this issue). Marth *et al.* (2001) determined the frequency of a large number of candidate SNPs from dbSNP and found that on average, 50% of SNPs assayed showed a frequency of >10%. Considering this success rate it may be necessary to identify SNP markers at 5–15-kb intervals across our region for frequency determination (20 chromosomes should be sufficient to identify the majority of SNPs with a minor allele frequency >10%), before defining a final marker map.

Use of an evenly spaced marker map is a pragmatic approach which assumes evenly spaced LD across the study region. Inevitably this does not always occur; instead LD extends over variable distances. The optimal, but time-consuming approach to map construction across a region is to first determine common haplotypes and use this data to define a minimal set of SNPs ('haplotype tags') that define these common haplotypes. In this study we have one major advantage, the UCSC Human Genome Browser presents information on common SNP haplotypes across the whole of chromosome 21. This data is derived from a study by the Biotech company, Perlegen, Inc. In their study, Patil *et al.* (2001) identified 25,000 SNPs with a frequency >10%, by sequencing 20 haploid copies of chromosome 21 derived from a cell line. They used these SNPs to directly determine common haplotypes. This data is visible in the 'haplotype blocks' track in Figure 9.8A.

Each haplotype block is represented by a blue horizontal line with taller vertical blue bars at the first and last SNPs of each block. The shade of the blue indicates the minimum number of SNPs required to discriminate between haplotype patterns which account for at least 80% of genotyped chromosomes, darker colours indicate fewer SNPs are necessary. Individual SNPs are denoted by smaller black vertical bars. This information is also available at the Perlegen website (http://www.perlegen.com/haplotype). Several haplotype blocks extend over the BACE2 gene region. Clicking on the haplotype bar opens a window displaying the structure of the selected haplotype. In Figure 9.8C we show an example of one of the haplotypes across the BACE2 gene. In this case two SNPs from a total of 10 SNPs (A–J) define four haplotypes (W,X,Y and Z). In this relatively simple haplotype, it is fairly easy to identify the SNP pairs that would 'tag' or distinguish the four haplotypes. However, the HGVbase website also has a tool, 'Tag 'n Tell', to automatically identify haplotype tags (http://hgvbase.cgb.ki.se/). This is particularly useful for identifying tags in larger, complex haplotypes. In the case of the haplotype in Figure 9.8C, we can evaluate the four haplotypes defined by the 10 SNP markers. To format this haplotype for analysis by 'Tag 'n Tell', we need to list the marker IDs (A B C D E F G H I J) on the first line separated by spaces. Then, we introduce each haplotype, defined by a list of alleles (one for each marker) followed by the haplotype name preceded by ':'. All four haplotypes are input on separate lines. Thus the input to the tool is as follows:

INPUT HAPLOTYPES (from Figure 9.8C):

```
A B C D E F G H I J
1 1 2 1 2 1 1 2 1 1 : W
1 2 1 2 1 1 1 1 1 1 : X
2 2 1 1 1 2 2 1 2 2 : Y
2 1 2 1 1 1 1 1 1 1 : Z
```

The 'Tag 'n Tell' program identifies two alternative sets of two tags which will distinguish all four haplotypes. These are SNPs A and B or SNPs A and C:

```
A B              A C
1 1 : W          1 2 : W
1 2 : X    OR    1 1 : X
2 2 : Y          2 1 : Y
2 1 : Z          2 2 : Z
```

This convincingly demonstrates the power and potential economies afforded by use of haplotype data. Unfortunately similar data is not yet available for all chromosomes (LD maps have also been published for chromosomes 19 and 22; see Chapter 7). A publicly funded genome-wide haplotype determination project is in progress so this situation should change fairly quickly.

9.7.5 Identifying 'Candidate SNPs'

Just as the candidate gene approach can complement the locus analysis approach, a complementary approach to regular marker map construction is to screen candidate SNPs across the locus. A candidate SNP is any SNP with a potential for functional effect, this could include non-synonymous SNPs, SNPs in regulatory regions or other functional regions. In fact uncertainty over functional prediction could stretch the definition of a candidate SNP to any SNP within 10 kb of gene. There are many ways to select such SNPs (see Chapters 12 and 13 for details). At the most detailed level, it is possible to identify all available SNPs in genes and putative regulatory regions by eye using human genome browsers to identify overlap between SNPs and features such as promoter regions or comparative genome conservation. This can be a useful and manageable approach for small loci, however, this is not always a practical approach for larger loci. Larger analyses can be facilitated by using the feature 'export facilities' in the human genome browsers. The coordinates of most features can be exported as tab-delimited data, these can be compared in a spreadsheet.

9.7.6 Moving from SNPs to Assays

By using both Ensembl and the UCSC HGB we have been able to identify and export a comprehensive list of SNPs across our locus in a convenient tab-delimited form. But one hurdle stands between our data and a large-scale SNP genotyping experiment. The problem is that the human genome browsers give us a list of SNP accession numbers, but not a primer design-ready list of SNP sequences. Until very recently the only web based approach to obtaining these sequences was to access each SNP individually! Fortunately both dbSNP and SNPper, now incorporate features which allow the user to export a list of SNPs with flanking sequence. SNPper (Riva and Kohane, 2001) maps RefSNPs and genes to the golden path in a very similar way to the genome browsers (but without a graphical interface), allowing SNP searching by gene or SNP name or by golden path position. This makes SNPper completely compatible with the coordinates generated by Ensembl and HGB (and so this tool can also be used at an earlier stage for SNP data mining). SNPper also produces a very effective gene report. The 'Annotated' link in the gene report displays a very informative SNP report which positions SNPs in the context of introns, exons and other gene features, including a mark-up of non-synonymous SNPs across a gene.

The great strength of SNPper lies in its data export and manipulation features. At the SNP report level, SNPs can be sent directly to automatic primer design through Primer3 (which allows the entry of multiple SNP sequences). At a whole gene level or even at a locus level, SNP sets can be defined and refined and e-mailed to the user in an Excel spreadsheet with SNP names in the first column and flanking sequences in the second, ready for primer design. This is a very useful function which is not currently offered by any other tool.

9.8 CASE STUDY 6: GENETIC/PHYSICAL LOCUS CHARACTERIZATION AND MARKER PANEL DESIGN (RECOMMENDED TOOLS: ENSEMBL, HGB AND MAP VIEW)

The wealth of data presented so far in this chapter has enabled us to define our locus to a level of detail that would allow us to complete a quite effective genetic study. However, before finalizing the requirements for this study it may be useful to spend some time characterizing the actual genetic and physical characteristics of the locus.

Ideally we would like to establish a detailed LD map across the D21S1245–D21S1852 interval. The Perlegen haplotype data presented in the HGB interface goes a long way towards this objective. However the haplotype coverage across this region is not particularly complete. Extended haplotypes only cover 40–50% of the interval, leaving large gaps across the region (Figure 9.9). These gaps may simply reflect insufficient coverage

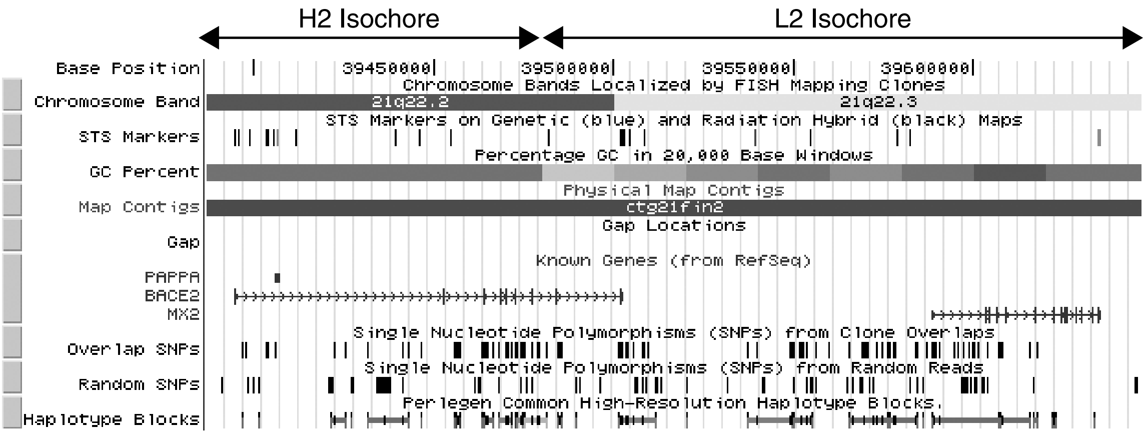

H2 Isochore
L2 Isochore
Base Position
39450000
39500000
39550000
39600000
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
21q22.2
21q22.3
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Percentage GC in 20,000 Base Windows
GC Percent
Physical Map Contigs
Map Contigs
ctg21fin2
Gap Locations
Gap
Known Genes (from RefSeq)
PAPPA
BACE2
MX2
Single Nucleotide Polymorphisms (SNPs) from Clone Overlaps
Overlap SNPs
Single Nucleotide Polymorphisms (SNPs) from Random Reads
Random SNPs
Perlegen Common High-Resolution Haplotype Blocks.
Haplotype Blocks

Figure 9.9 A putative isochore boundary in the BACE2 gene.

of the region by Perlegen SNPs. Alternatively LD across the gaps may actually be limited. A number of factors are known to influence LD (see Ardlie *et al.* (2002) for an excellent review). One of the most influential is the recombination frequency across the region. As we showed in Chapter 8, comparison of the physical and genetic distances between markers can give a direct measure of recombination frequency. In a comparison of Marshfield maps with the golden path, Yu *et al.* (2001) found that the genome-wide genetic/physical distance ratio ranged between 0 to 9 cM per Mb. They used this ratio to infer recombination rates and identified several chromosomal regions up to 6 Mb in length with very low or high recombination rates. They termed these recombination ‘deserts’ and ‘jungles’, respectively. LD was much more extended in recombination ‘deserts’ than in ‘jungles’ as higher rates of recombination reduced the extent of LD. This is an interesting approach, although its major drawback is the low resolution of genetic maps, this makes it very difficult to draw accurate conclusions about recombination over ranges of less than 1 Mb. There are only two genetic markers in the D21S1245–D21S1852 interval, so it is not possible to draw conclusions on recombination rate across this locus, analysis would need to encompass a much wider region.

9.8.1 Analysis of GC Ratio and Identification of Isochore Boundaries

Beyond the analysis of genetic and physical ratios, even simpler measures can give clues to the nature of recombination in a locus. GC content across a locus also has a weak influence on recombination rates. Lower GC ratios generally correspond to lower recombination rates (Yu *et al.*, 2001). There is cytogenetic evidence for this phenomenon, analyses have shown that meiotic crossovers are seen more frequently in GC-rich R and T bands than in GC-poor G bands (Holmquist, 1992). This observation directly relates recombination frequency with the gross Giemsa banding of chromosomes. These bands were believed to be made up of tracts of DNA with homogeneous GC content, known as isochores. Isochores are divided into two classes the GC-rich H2 and H3 isochores and the GC-poor L1, L2 and H1 isochores (Bernardi, 2000). Interestingly our region between D21S1245–D21S1852 spans the 21q22.2–q22.3 cytoband. The region also shows a clear shift from high GC (average 50%) to low GC (average 40%) in the region of the cytoband boundary (high GC is indicated by dark grey in Figure 9.9, lower GC is indicated by lighter shades of grey). This region is a putative isochore boundary, between an H2 and L1 isochore. It is difficult to determine if there is a significantly different extent of

LD between the putative L1 and H2 isochore regions. The markers within the first eight exons of the BACE2 gene do not generally show LD over a greater distance than 5 kb, while markers in the L1 region near the MX2 gene show LD over longer distances up to 10–15 kb (Figure 9.9). In a study of the NF1 gene, Eisenbarth *et al.* (2000), found a marked reduction in LD, which coincided with an L1 to H2 isochore boundary in the NF1 gene.

This analysis of the D21S1245–D21S1852 interval, may seem somewhat esoteric. We are just starting to understand how the physical properties of chromosomes affect their genetic properties. Undoubtedly, our understanding of these issues is still limited, but a pragmatic approach to marker map design across this region might be to establish a baseline marker density across the entire region (say at 1SNP/10 kb), once this has been reached within the budget of the project, then supplemental markers could be placed in regions with a higher predicted recombination rate.

9.9 CONCLUSIONS

In this chapter we have reviewed the key steps in the design and construction of genetic studies using genomic sequence as a template. When sequencing of the human genome is finally complete and as studies of the genome become more and more precise, much of genetics as we know it today may become an increasingly *in silico* process. Ten years ago who might have believed that the details of the genetic study process would have changed so dramatically (although the principles remain the same). As the genome wave continues to roll towards us, we may be looking at much more intelligently designed genetic studies, with maps which account for local recombination, LD and a detailed knowledge of genes and regulatory regions. And 10 years further on, perhaps we will look back again and marvel again at how much things have changed?

REFERENCES

Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, *et al.* (1995). Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc Natl Acad Sci USA* **92**: 1684–1688.

Ardlie KG, Kruglyak L, Seielstad M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Rev Genet* **3**: 299–309.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11**: 1005–1017.

Benson G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

Bernardi G. (2000). Isochores and the evolutionary genomics of vertebrates. *Gene* **241**: 3–17.

Eisenbarth I, Vogel G, Krone W, Vogel W, Assum G. (2000). An isochore transition in the NF1 gene region coincides with a switch in the extent of linkage disequilibrium. *Am J Hum Genet* **67**: 873–880.

Fondon JW III, Mele GM, Brezinschek RI, Cummings D, Pande A, Wren J, *et al.* (1998). Computerized polymorphic marker identification: experimental validation and a predicted human polymorphism catalog. *Proc Natl Acad Sci USA* **95**: 7514–7519.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, *et al.* (2001). Haplotype tagging for the identification of common disease genes. *Nature Genet* **29**: 233–237.

Holmquist GP. (1992). Chromosome bands, their chromatin flavors, and their functional features. *Am J Hum Genet* **51**: 17–37.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, *et al.* (2002). The Ensembl genome database project. *Nucleic Acids Res* **30**: 38–41.

Koch HG, McClay J, Loh EW, Higuchi S, Zhao JH, Sham P, *et al.* (2000). Allele association studies with SSR and SNP markers at known physical distances within a 1-Mb region embracing the ALDH2 locus in the Japanese, demonstrates linkage disequilibrium extending up to 400 kb. *Hum Mol Genet* **9**: 2993–2999.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, *et al.* (2000). SAGEmap: a public gene expression resource. *Genome Res* **10**: 1051–1060.

Marth G, Yeh R, Minton M, Donaldson R, Li Q, Duan S, *et al.* (2001). Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genet* **27**: 371–372.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, *et al.* (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.

Riva AA, Kohane IS. (2001). A web-based tool to retrieve human genome polymorphisms from public databases. *Proc AMIA Symp* 558–562.

Rogic S, Mackworth AK, Ouellette FBF. (2001). Evaluation of gene-finding programs on mammalian sequences. *Genome Res* **11**: 817–832.

Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, *et al.* (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet* **24**: 227–235.

Schuler GD. (1998). Electronic PCR: bridging the gap between genome mapping and genome sequencing. *Trends Biotechnol* **16**: 456–459.

Soares JC, Mann JJ. (1997). The anatomy of mood disorders — review of structural neuroimaging studies. *Biol Psychiatry* **41**: 86–106.

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, *et al.* (2002). Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci USA* **99**: 4465–4470.

Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, *et al.* (2001). Comparison of human genetic and sequence-based physical maps. *Nature* **409**: 951–953.

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray
Copyright © 2003 John Wiley & Sons, Ltd.
ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 10

SNP Discovery and PCR-based Assay Design: From *In Silico* Data to the Laboratory Experiment

ELLEN VIEUX[1], GABOR MARTH[2] AND PUI KWOK[3]

[1] *Washington University School of Medicine, St. Louis, MO, USA*

[2] *National Center for Biotechnology Information, Bethesda, MD, USA*

[3] *Cardiovascular Research Institute, University of California, San Francisco, USA*

10.1 INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most abundant form of DNA sequence variation in the human genome. It is widely believed that a significant fraction of SNPs contribute to our susceptibility to various diseases. In order to identify the SNPs associated with diseases, however, many groups are pursuing a case–control mapping strategy that requires a large number of SNP markers distributed throughout the human genome. Once a set of genes is implicated in a disease (either by genetic mapping or by obtaining biological evidence), the candidate genes are scanned for sequence variations that are likely to alter the genes' function. Therefore, identifying single base-pair changes, in a global or targeted fashion, is extremely important in genome research.

The central public polymorphism database, dbSNP (Sherry *et al.*, 2002), serves as an archival repository of nucleotide sequence variations. An important subset of these data, nearly 100,000 SNPs in transcribed regions, were found by analysing clusters of expressed sequence tags (ESTs) (Buetow *et al.*, 1999, 2001; Irizarry *et al.*, 2000) or by aligning ESTs to the human reference sequence (Marth *et al.*, 1999). The vast majority of genomic SNPs (single base pair variations found by analysing genomic sequence clones without regard to whether they represent exonic DNA) were discovered in sequences from restricted genome representation libraries (Altshuler *et al.*, 2000), random shotgun reads aligned to genome sequence (Sachidanandam *et al.*, 2001), and in the overlapping sections of the large-insert clones (mainly bacterial artificial chromosome, or BAC) that make up the public human reference genome (Tallion-Miller *et al.*, 1998). Because most sequences of these comparisons involved a small number of chromosomes (typically two), this collection of SNPs is enriched for common variants. Experimental characterization of these polymorphisms demonstrates that many of them occur at a high frequency in independently chosen samples, and often segregate in all or most human populations (Marth *et al.*, 2001). By the same argument, many rare polymorphisms, including those that cause noticeable but rare phenotypic effects, are likely to be absent from this set. The identification of rare phenotypic mutations will require significantly higher sample sizes and may only be possible by the cross-comparison between large samples of affected patients and those of controls (see Halushka *et al.* (1999) for an example of such a study).

Because the numbers are extremely large and the need for identifying SNPs in a timely fashion is great, computer tools are indispensable in the SNP discovery process. Fundamentally, one identifies a SNP by comparing two or more sequences from the same region on the chromosome. This can be done quite easily if the DNA sequence quality is high and the sequence data are derived from cloned DNA because each clone comes from a single copy of one of the two chromosomes in the diploid human cell. There are no unambiguous bases in regions where the data quality is high. In the case of identifying SNPs in targeted regions in the genome, one amplifies genomic DNA by PCR and sequences the PCR products derived from different individuals. In this situation, SNP discovery is complicated by the fact that the same regions on both chromosomes in the diploid cell are amplified by PCR and some bases will be heterozygous in one or more individuals. A good computer tool will be able to identify a SNP even when only heterozygotes and homozygotes of just one of the two alleles are present in the samples sequenced. This is not a trivial problem because the commonly used dye terminator-based DNA sequencing methods yield peaks of uneven heights at the polymorphic sites and the base-calling algorithm will frequently miscall the base at these sites in the sequences of heterozygous individuals.

In this chapter, we will survey the computer tools used in global and targeted SNP discovery and PCR-based assay design. Instead of describing the mechanics of how to use these bioinformatic tools, we refer the reader to the primary literature and the excellent documentations of these tools and concentrate on explaining the approaches these tools take and the limitations (if any) they may have.

10.2 SNP IDENTIFICATION

Computational discovery of polymorphisms in sequence data usually follows a four-step procedure. First, sequences of high similarity in multiple individuals are identified, usually with a BLAST (Altschul *et al.*, 1990) similarity search. To avoid spurious similarity due to known human repeats, sequences are masked for high copy number repetitive elements with REPEATMASKER (Arian Smit, unpublished data). Still, the possibility exists that the sequences originate from regions of as yet uncharacterized chromosomal duplications (Lander *et al.*, 2001). Inclusion of a second, paralogue-filtering step into the procedure can reduce false positive SNP predictions arising from comparing paralogous sequence copies. Following this step, false predictions due to paralogy were as low as 0.2% of the data collected through pooled SNP characterization in the Kwok laboratory (unpublished data).

The third step is the construction of a base-wise multiple alignment of the sequences. In the general case, this is a computationally expensive task. Aligning expressed sequences is even more complicated because of exon–intron punctuation and possible alternative splice variants. In the case of human data one can organize fragmentary sequences on top of the nearly complete reference sequence (Lander *et al.*, 2001). This approach was shown to work well for discovering SNPs in clusters of cDNA sequences (Marth *et al.*, 1999).

In the fourth (and final) step, sequences in the precise, base-to-base multiple alignment are scanned for nucleotide differences. Because of the possibility of sequencing errors, not every mismatch is a polymorphic site. Discrimination between true polymorphism and sequencing error uses statistical tools based on measures of sequence accuracy, or base quality values (Ewing and Green, 1998; Ewing *et al.*, 1998). Each SNP prediction is accompanied by a measure of confidence. Accurate confidence values permit one to use the highest number of candidates with an acceptable false positive rate.

Both commercial and academically developed programs are available for use in SNP detection. Some methods use sequence quality data to eliminate false positives due to poor sequencing quality. Others incorporate expected mutation rates to distinguish true SNPs. The most prominent methods of detecting SNPs are PolyBayes, PolyPhred, and Sequencher. Other methods incorporate neighbourhood quality standard (NQS) generated by Phred (Ewing *et al.*, 1998) to determine the quality of the data surrounding the SNP (Altshuler *et al.*, 2000; Mullikin *et al.*, 2000). PolyPhred and PolyBayes are freely available to academic groups, while Sequencher is produced by Gene Codes Corporation (URL: www.genecodes.com). Other companies have developed software based on the same principles. Typically, these products either offer a built-in graphical interface, or use an external, licensable interface program (such as CONSED). They can be used for both comparison of short known regions, or long shotgun regions, and are extremely useful when searching known regions of interest for novel SNPs.

10.2.1 PolyBayes

The POLYBAYES program was developed for *de novo* SNP discovery in non-ambiguous (clonal) sequence data (Marth *et al.*, 1999). The SNP detection algorithm employs a Bayesian approach to combine prior knowledge (such as average polymorphism rate or expected transition to transversion ratio) with the base calls and base quality values of the sequences in the multiple sequence alignment. Each SNP prediction comes with a

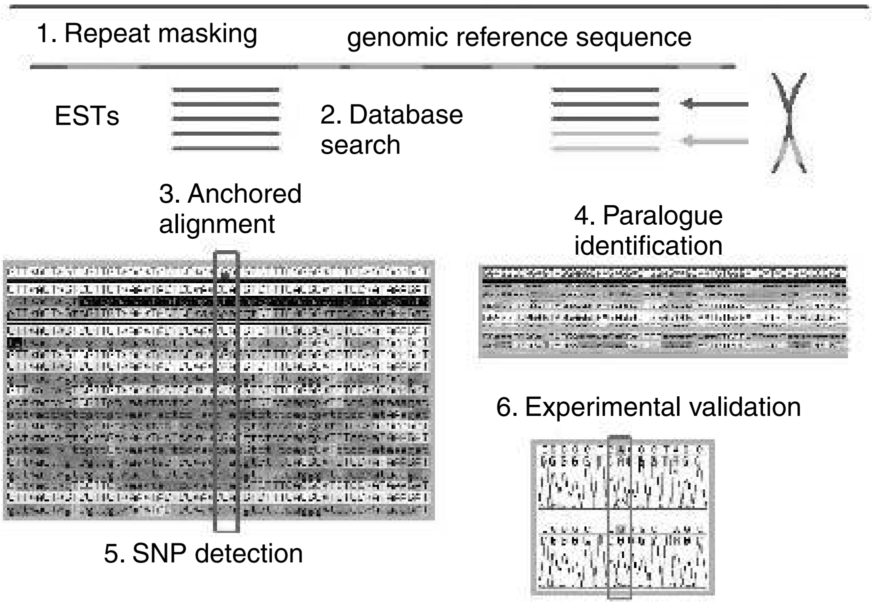

1. Repeat masking
genomic reference sequence
ESTs
2. Database search
3. Anchored alignment
4. Paralogue identification
6. Experimental validation
5. SNP detection

Figure 10.1 The POLYBAYES SNP discovery tool applied to EST-Mining. (1) The genome sequence (BAC clone or assembled sequence contig) is masked for known, large copy number human repeats. (2) Using the BLAST similarity search tool, expressed sequences in the public database (dbEST) that match the genomic sequence are identified. (3) Matching ESTs are aligned to the genomic sequence using an anchored alignment approach. (4) Possible paralogous ESTs are identified and discarded. (5) The multiple alignment is scanned for polymorphic sites. (6) Candidates are validated by sequencing in independent, population-specific DNA pools.

predicted true positive rate, or 'SNP score', which have been shown to be accurate (Marth *et al*., 2001). Figure 10.1 illustrates an example of using POLYBAYES for SNP discovery in ESTs aligned to genome sequence. POLYBAYES has been used to discover SNPs in overlapping regions of human BACs (Marth *et al*., 1999), in *C. elegans* (Wicks *et al*., 2001) and Drosophila (Berger *et al*., 2001).

10.2.2 PolyPhred

PolyPhred was developed to be used with Phred, Phrap, and CONSED to identify candidate SNPs in sequence trace data (Ewing and Green, 1998; Gordon *et al*., 1998; Table 10.1). In Consed, coloured marks in the sequence alignment are used to indicate candidate SNPs as well as confidence in the variations base call. The accuracy of the calls by PolyPhred has been tested using previously screened mitochondrial DNA. The results show that this software exhibits over 95% accuracy depending on the quality of the sequence traces (Nickerson *et al*., 1998). The primary use of PolyPhred is to identify SNPs in PCR-amplified data as it can detect heterozygous sequence peaks, and is thus widely employed in sequence-based genotyping applications.

10.2.3 Sequencher

Sequencher is a tool developed by GeneCodes for sequence alignment, annotation, editing and mutation identification. Although it is a commercial product, a free demo version is available (www.genecodes.com/features/html). Sequencher can be used with automated sequencers such as ABI, Pharmacia/ALF, LI-COR and VISTRA. GeneCodes continues to

TABLE 10.1 Tools and Related Resources for Primer Design

SNP detection tools	
Sequencher	http://www.genecodes.com/features.htm
PolyPhred	http://droog.mbt.washington.edu/PolyPhred.html
POLYBAYES	http://www.genome.wustl.edu/gsc/polybayes/
Repeat masking tools	
RepeatMasker	www.genome.washington.edu/uwgc/analysistools/repeatmask.html
MaskerAid	http://sapiens.wustl.edu/maskeraid/
Primer design tools	
Primer3	http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
TSC primer db	ftp://snp.cshl.org/pub/SNP/.
Primer design tips	http://www.alkami.com/primers/refdsgn.htm.
Tools for sequence extraction and manipulation	
SNPper	http://bio.chip.org:8080/bio/.
UCSC HGB	http://genome.ucsc.edu/index.html

add new tools and functionality to the software. The most appealing part of the software is the graphical interface, which is intuitive and easy to use on multiple platforms.

10.2.4 Non-sequencing Methods

Several groups have explored non-sequencing methods for SNP discovery. Among the most promising of these techniques is the use of high density DNA chips (Dong *et al.*, 2001). Variations of this method have been used to scan for genome wide SNPs (Wang *et al.*, 1998), in mitochondrial DNA (Chee *et al.*, 1996), and to scan all of chromosome 21 (Patil *et al.*, 2001). Methods for SNP scanning using DNA chips vary considerably in design (for a review see Draghici *et al.*, 2001).

10.3 PCR PRIMER DESIGN

A large number of candidate SNPs exists in public databases. Key to taking advantage of this resource is the ability to design PCR assays to amplify these loci uniquely and SNP genotyping assays for genetic studies under standardized conditions. Genetic researchers wanting to validate and assay SNPs are faced with the need for high throughput primer design. Manual picking of primers is time consuming, and some automated tools only allow for submission of one sequence at a time. There are many tools available over the web as well as software. In addition, some commercial companies offer genotyping assay design by order for their customers and a few assays are available through public databases.

10.3.1 Tools

Currently there is no standard method for calculating the annealing temperature (TM) of primers. Although many tools have been developed to determine the annealing temperature, their results vary. Furthermore, many of these programs use different entropy and enthalpy tables in their TM calculations, leading to further discrepancies (Owczarzy *et al.*, 1997). Despite these variances most of these tools will work and one program that has become a standard is Primer3 (http://www-genome.wi.mit.edu/genome_software/other/primer3.html). A comprehensive review of primer picking and TM predicting tools can be found at http://www.alkami.com/primers/refdsgn.htm.

Primer3 is a standard because it is freely available and easy to use. It is particularly useful for high throughput design because it can determine primers for multiple sequences at once. Some of its particular strengths are its many useful and well-documented options, its easily parsed output and its simple command line interface. Primer3 can be used for the design of both PCR primers and internal sequencing primers. Although Primer3 allows for individual SNP position targets and target lengths to be set for each sequence, if the data is highly varied in position and length it is possible to avoid setting parameters for each SNP by pre-formatting the data. The SNP sequence retrieval option on SNPper (see below) is a tool that can provide this uniformly formatted flanking sequence.

10.3.2 Custom Primer Design Services

Although primer design may be carried out in-house, many companies as well as public databases are offering high throughput design as part of their product support. Sequenom is one such example. Sequenom is in the process of making primers available through a site called RealSNP (www.RealSNP.com). Applied Biosystems is another company providing primer design through their 'Assay by Design^{TM}' Genomic Assay Service. The researcher provides the sequence, while Applied Biosystems designs and test all assays. These designs are optimized for Taqman^{TM} assays (http://www.appliedbiosystems.com/). There is no charge if an assay cannot be designed. Perkin-Elmer will also be providing SNP-specific assays through their website.

10.3.3 Public Databases

Primers generated by The SNP Consortium (TSC) Allele Frequency Project are available via ftp (Table 10.1). These primers have been released, by some of the groups, to the public with the assistance of TSC. It should be noted that these primers have been generated by separate groups via different methods and for specific experimental conditions. The NCBI's dbSNP database also contains primer designs for some of the SNP entries, but these have not been specifically designed for SNP validation (Sherry *et al.*, 2002). In addition to these public databases the Kwok laboratory has over 980,000 assays designed for sequencing of PCR products, for the specific purpose of pooled allele frequency determination. The Kwok laboratory also maintains over 1,400,000 SNP genotyping assays designed for single base extension using fluorescence polarization detection (available at http://snp.wustl.edu).

10.4 BROADER PCR ASSAY DESIGN ISSUES

SNP assay methods have three major components: (1) allelic discrimination methods, (2) reaction formats and (3) detection methods. Each area presents different challenges during SNP assay design. The most important consideration for assay design is the method of allelic discrimination. These methods vary greatly. For example four main methods of allelic discrimination are allele-specific hybridization, primer extension (includes single base extension), ligation and invasive cleavage. The reaction formats are either homogeneous reactions or solid phase reactions and the detection methods currently use product light emissions, product–mass measurements and electrical property changes in the product (Kwok, 2001).

In some cases the critical parameters that apply to one technique will not apply to others. However, almost all SNP genotyping assay techniques use PCR to amplify DNA. In order to design the correct primers one must first determine the method of assay. However, there are some basic guidelines used when designing primers for genomic sequence. All designs require obtaining sequence, repeat masking, setting experimental and design parameters, picking primers and formatting the information.

10.4.1 Obtaining Sequence

The flanking sequences for each SNP can be obtained from a variety of sources. For known SNPs two public databases, dbSNP and SNPper (Riva and Kohane, 2001; Table 10.1) provide a method for obtaining sequence. SNPper is run by Harvard's Children's Hospital Informatics Program (CHIP). Both dbSNP and SNPper offer batch query modes and return sequence in FASTA format. In the case of single SNP analysis, SNPper provides a link to the Primer3 website which will import the retrieved sequence into Primer3 and analyse it using default values. For SNPs that can be uniquely mapped SNPper can provide up to 1000 bases on either side of the SNP. It should be noted that at this time SNPper does not contain the most recent uniquely mapped SNPs in dbSNP.

10.4.2 Repeat Masking

A large amount of the genome consists of repeated regions or low complexity DNA. It is important to avoid selecting primers from these regions in order to avoid amplification of multiple products. Masking the repeats or making repeated sequence unavailable to the automated primer-picking programs prevents most unwanted amplification. A commonly used program for masking is RepeatMasker (see Table 10.1). A new resource that improves upon RepeatMasker is MaskerAid (Table 10.1), which increases the speed of masking more than 30-fold (Bedell *et al*., 2000). Default parameters in RepeatMasker will mask known repeat regions with Ns. RepeatMasker accepts FASTA files, and returns the sequence in the same format. Ready masked sequence can also be obtained from some of the public databases. dbSNP provides sequence in FASTA format with low-complexity sequence in lower case, while the University of California Santa Cruz Human Genome Browser (UCSC HGB) has options to save repeats as either Ns or lower case. However, this format is problematic when trying to represent the start and stop exons and introns on UCSC HGB, because lower case can also be used to represent introns. In some cases two files may be required to represent the masked and unmasked forms of sequence.

Masking repeats can only be accomplished in known repeat regions with current resources. However, there remain repeat regions of the genome that have not yet been identified. By using pooled sequencing, it is possible to identify regions that have duplicated and subsequently diverged. These can be identified by the presence of a large number of apparent SNPs that are all 50% in frequency, as shown in Figure 10.2. When designing SNP specific primers within PCR products, for example for a single base extension assay, the RepeatMasking stage is not necessary.

10.4.3 Setting Experimental and Design Parameters

If a large number of SNP candidates are to be assayed it is more efficient to eliminate the experiments that are less likely to be successful *in vitro* during the *in silico* design stage. Stringent design parameters allow for a first level of screening when designing primers.

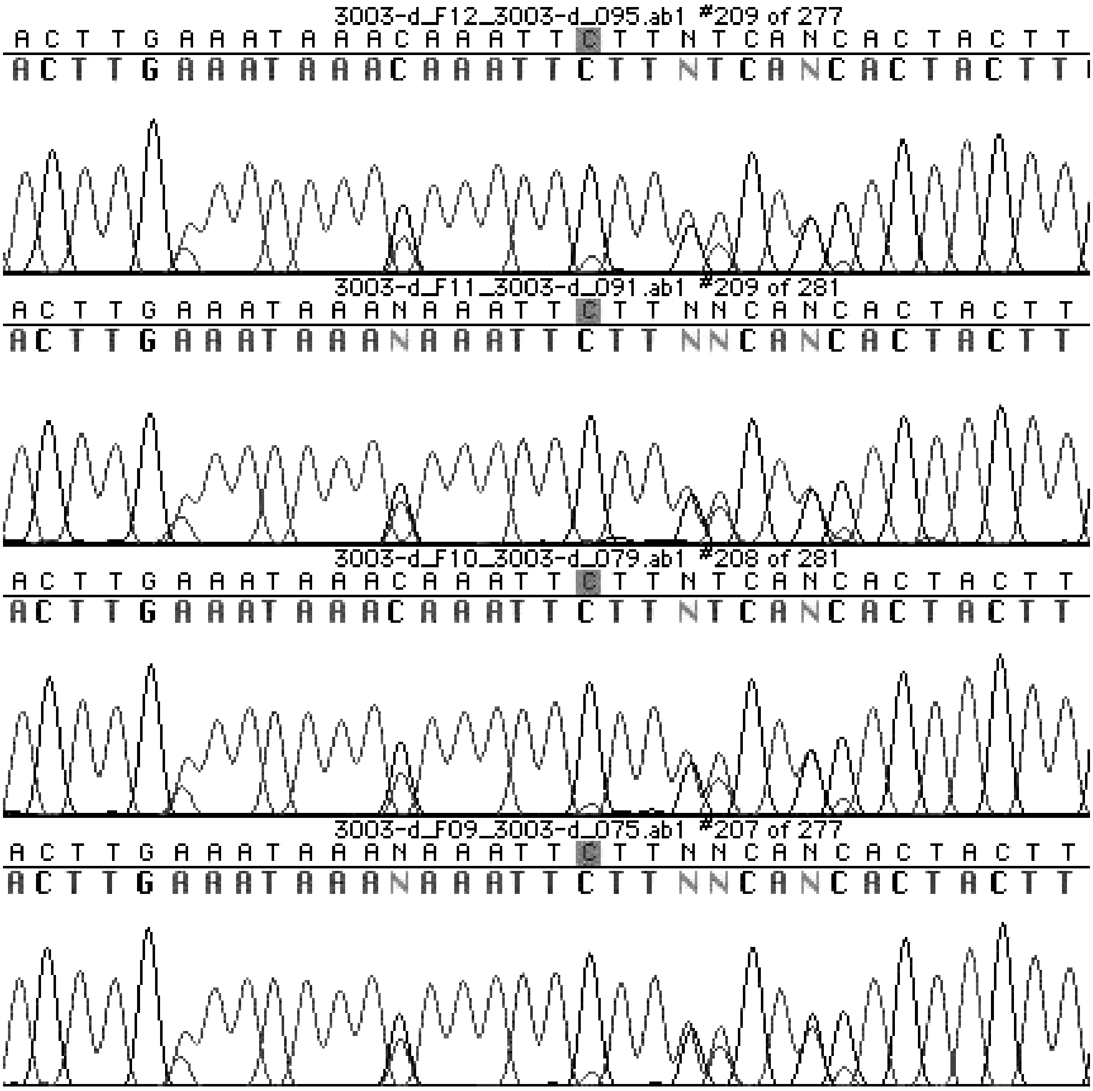


Figure 10.2 Duplicated and diverged regions in genome characterized by multiple 'SNPs' at 50% frequency in each pool.

Primer design programs such as Primer3 allow for input of both experimental parameters and primer structure parameters. The second level of screening can be done after candidate primers are chosen by primer selection programs to determine if the primers are likely to work.

Some suggestions for optimizing design parameters for the best experimental results can be found in *PCR Applications* (Beasley *et al.*, 1999). In general most primer design methods work (see http://gsu.med.ohio-state.edu/primer_design/sld001.htm for a detailed guide). However, under stringent experimental conditions optimized design parameters can decrease the level of experimental failures.

10.5 PRIMER SELECTION

In most design programs primer picking is preformed by the software. The primers are picked according to the specified parameters. If more than one primer set is returned some post processing will be required to select the most appropriate pair. Post processing can also be necessary for techniques such as pooled sequencing, where selection of a sequencing primer from the PCR primers is required.

10.5.1 Design Specific to Pooled Sequencing

Pooled sequencing uses sequencing to observe the frequency of a SNP in a group of individuals in one reaction. The candidate SNP and its flanking sequence is amplified from pools of DNA each containing individuals and a single reference individual. After sequencing of the PCR products is preformed using fluorescent dye-terminators, the sequence traces are aligned, allowing the allele frequencies to be estimated (Kwok *et al.*, 1994). Pooling DNA in this way prior to PCR amplification and estimating allele frequencies by subsequent quantitation of trace peak heights yields considerable time and cost savings.

There are several steps to designing pooled sequencing reactions. This method of design is carried out on a UNIX-based system, using RepeatMasker and Primer3. Repeats are masked before choosing PCR primers. Sequence that is not masked is retained for post processing. The input for Primer3 is set according to the optimized parameters (Beasley *et al.*, 1999) with a few optimizations. The optimizations are most important in the placement of the primers relative to the SNP. The primers are not allowed closer than 25 bases to the SNP, but are close enough to use one of the PCR primers for sequencing. After running Primer3 the results are processed to select for the best sequencing primer based on criteria to optimize experimental performance. These criteria are (1) the sequencing primer should be 100 bases from the target and (2) there should be no poly As or Ts greater than eight bases and no poly GTs or CAs greater than 10 pairs between the primer and the SNP. This design has been shown to work with less than 3% experimental failure and allows for the primers to be far enough from the SNP that the sequence is of high quality around the target as shown in Figure 10.3. During the design process as many as half of the SNPs fail to meet the design criteria, but this failure is at far lower cost than laboratory-based trial and error (Vieux *et al.*, 2002).

10.5.2 Design Specific to Single Base Extension (SBE) Reactions

SBE requires a primer that abuts the SNP under test. The primer is then extended by a single base, usually a labelled ddNTP (Hsu *et al.*, 2001). By using two different labels for the ddNTPs representing the two possible alleles, the allelic state of the SNP can be determined. SNP-specific SBE primer design can be undertaken using many of the same tools as pooled sequencing primer design. Both require repeats to be masked before designing PCR primers. The SNP-specific primers are chosen using non-masked sequence. The PCR product sizes can be smaller than for sequencing for all of the single base extension reactions. The SBE primer should not hang over the end of the PCR product and the PCR primers should not overlap with the SBE primer. In Primer3 the primers can overlap the target so it is important to give a SNP a large enough target area to prevent the overlap of primers. When choosing parameters and methods for SBE primers it is important to remember that different methods can have different primer requirements.

We have found that picking the shortest primer from 16–40 bases which has a TM between 60–65 degrees works well. In order to calculate TM for a small number of SBE primers it is possible to use free tools on the web. For high throughput design the best option is to solve TM equations after determining which set of entropy and enthalpy tables work best for the relevant method (Owczarzy *et al.*, 1997), and picking the shortest primer in the defined range. Further optimization can be achieved by picking the SBE primer with the least amount of secondary structure, and fewest runs of poly As and Ts.

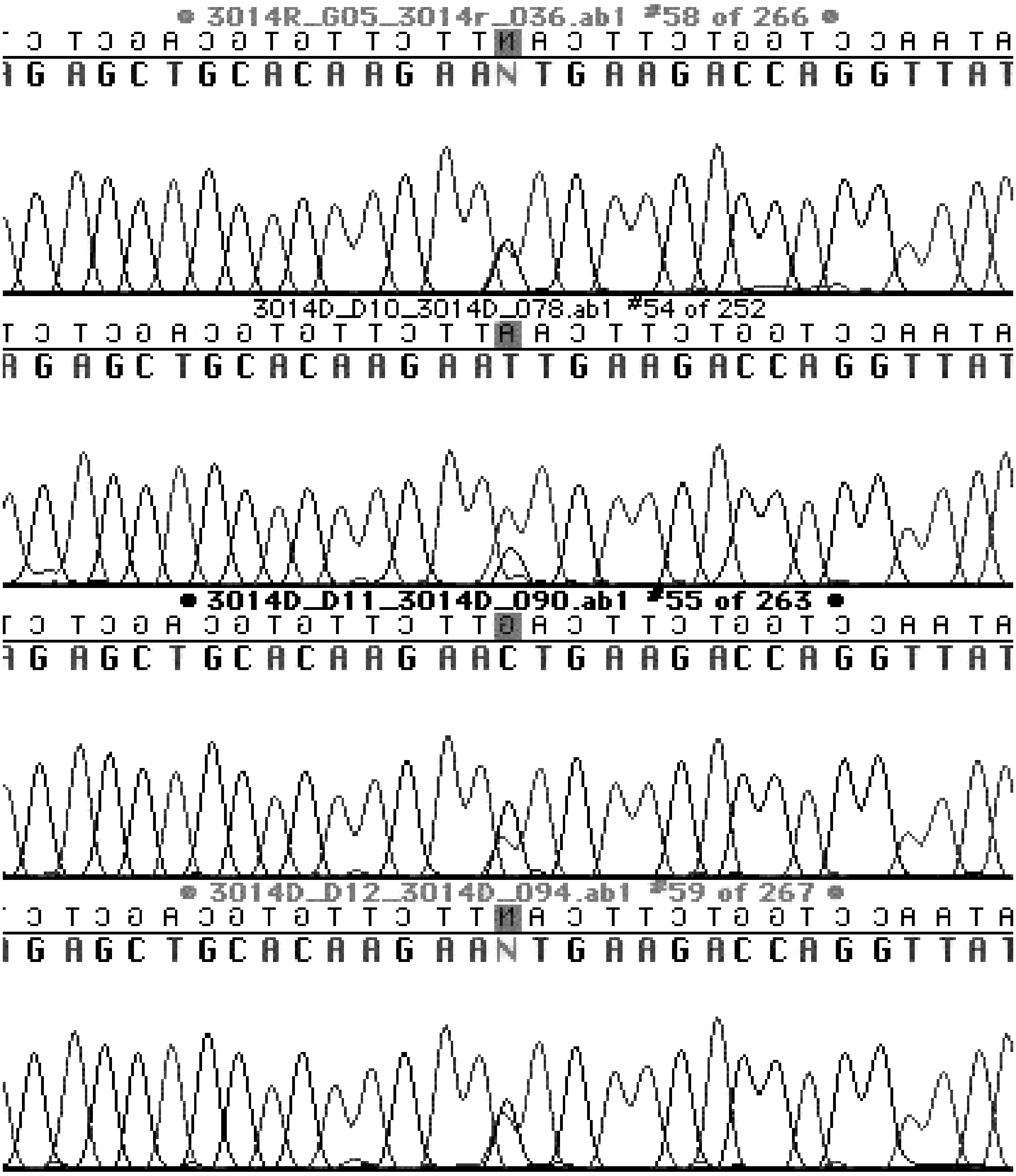


Figure 10.3 A clean pooled sequencing assay result shows a clear polymorphic site at the position of the candidate SNP.

10.6 PROBLEMS RELATED TO SNP ASSAY VALIDATION

As with any experimental design, assays for validation of candidate SNPs require attention to detail. Problems arise that are not always obvious or clearly stated in the documentation associated with the tools being used. Some problems are easy to overcome, while others cannot yet be solved.

With the completion of a final version of the human genome assembly a number of problems will be resolved, while inherent challenges will remain. There are many errors due to incorrect physical map order, gaps in physical map data and incorrect assembly (DeWan *et al*., 2002). These errors lead to SNPs mapping to multiple locations, incorrect haplotypes and difficulty in identifying paralogues. However, SNP locations are continually amended as assemblies are progressively corrected. Map locations will continue to change until the Human Genome Project is complete. This can cause difficulties in the analysis of data and obtaining guide sequence. Another difficulty with the unfinished map is unidentified paralogues. A SNP can appear to map to a unique position, when it is actually an artefact generated from an unknown paralogue of the original reference sequence. An example of such an artefact generated by pooled sequencing data is shown in Figure 10.2.

Guide sequence is provided for known SNPs through dbSNP, TSC and SNPper. The first two sites only provide a small amount of flanking sequence in their database for any given SNP. This can lead to failure in the design of PCR primers due to limited sequence information. SNPper provides far more flanking sequence by mapping the SNP location and retrieving guide sequence from the human genome assembly at the UCSC (Table 10.1).

Other problems are inherent when working with DNA and the current technologies. Long runs of a single nucleotide can cause sequencing reactions to fail, while insertion/deletion events can cause problems with sequencing and with SNP allelic discrimination methods such as allele-specific hybridization, primer extension (including SBE), ligation and invasive cleavage. These problems may only be solved with new technologies for SNP characterization.

10.7 CONCLUSION

Given the large number of SNPs in the human genome and the potential for large-scale experimentation, bioinformatics tools are essential for SNP discovery and genotyping assay development. The tools for comparing cloned (and hence homozygous) sequences are well developed and have proven useful. However, tools for comparing genomic sequences amplified by PCR, which are often heterozygous, still have room for computational and technical improvements. Following SNP discovery, there are many assay methods for genotyping, but none can satisfy all requirements. The basic methods for assay design are well defined, but specific optimizations are different for each method. With technology improvements, some of the current problems in SNP assay design will be solved, resulting in a reduction in the number of SNPs that are refractory to successful assay design. But for now design optimization using currently available tools and careful interpretation of subsequent results will provide assays and allele frequencies for a large portion of the SNPs currently available.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, *et al.* (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**: 513–516.

Beasley EM, Myers RM, Cox DR, Lazzeroni LC. (1999). Statistical refinement of primer design parameters. In Michael DHG, Innis A, Sninsky JJ. (Eds), *PCR Applications*. Academic Press: New York, pp. 55–71.

Bedell J, Korff I, Gish W. (2000). MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **16**: 1040–1042.

Berger J, Suzuki T, Senti K, Stubbs J, Schaffner G, Cickson BJ. (2001). Genetic mapping with SNP markers in Drosophila. *Nature Genet* **29**: 475–481.

Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, *et al.* (2001). High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc Natl Acad Sci USA* **98**: 581–584.

Buetow KH, Edmonson MN, Cassidy AB. (1999). Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genet* **21**: 323–325.

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, *et al.* (1996). Accessing genetic information with high density DNA arrays. *Science* **274**: 610–626.

DeWan AT, Parrado AR, Matise TC, Leal SM. (2002). The map problem: a comparison of genetic and sequence-based physical maps. *Am J Hum Genet* **70**: 101–107.

Dong S, Wang E, Hsie L, Cao Y, Chen X, Gingeras TR. (2001). Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation. *Genome Res* **11**: 1418–1424.

Draghici S, Kulin A, Hoff B, Shams S. (2001). Experimental design, analysis of variance and slide quality assessment in gene expression arrays. *Curr Opin Drug Discov Devel* **4**: 332–337.

Ewing B, Hillier L, Wendl MC, Green P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**: 175–185.

Ewing B, Green P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**: 186–194.

Gordon D, Abajian C, Green P. (1998). Consed: a graphical tool for sequence finishing. *Genome Res* **8**: 195–202.

Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, *et al.* (1999). Patterns of screening of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet* **22**: 239–247.

Hsu TM, Chen X, Duan S, Miller RD, Kwok PY. (2001). Universal SNP genotyping assay with fluorescence polarization detection. *Biotechniques* **31**: 560, 562, 564–568, *passim.*

Irizarry K, Kustanovich V, Li C, Brown N, Nelson S, Wong W, *et al.* (2000). Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nature Genet* **26**: 233–236.

Kwok P-Y. (2001). Methods for genotyping single nucleotide polymorphisms. *Annu Rev Genomics Hum Genet* **2**: 235–258.

Kwok P-Y, Carlson C, Yager TD, Ankener W, Nickerson DA. (1994). Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics* **23**: 138–144.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, *et al.* (1999). A general approach to single-nucleotide polymorphism discovery. *Nature Genet* **23**, 453–456.

Marth G, Yeh R, Minton M, Donaldson R, Li Q, Duan S, *et al.* (2001). Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genet* **27**, 371–372.

Mullikin JC, Hunt SE, Cole CG, Mortimore BJ, Rice CM, Burton J, *et al.* (2000). An SNP map of human chromosome 22. *Nature* **407**: 516–520.

Nickerson DA, Rieder MJ, Taylor SL, Tobe VO. (1998). Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Res* **26**: 967–973.

Owczarzy R, Vallone PM, Gallo FJ, Paner TM, Lane MJ, Benight AS. (1997). Predicting sequence-dependent melting stability of short duplex DNA oligomeres. *Biopolymers* **44**: 217–239.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, *et al.* (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.

Riva AA, Kohane IS. (2001). A web-based tool to retrieve human genome polymorphisms from public databases. *Proc AMIA Symp* 558–562.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, *et al.* (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.

Sherry ST, Ward MH, Kholdov M, Baker J, Phan L, Smigielski EM, *et al.* (2002). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308–311.

Tallion-Miller P, Gu Z, Li Q, Hiller L, Kwok PY. (1998). Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res* **8**: 748–754.

Vieux EF, Kwok P-Y, Miller RD. (2002). Primer design for PCR and sequencing in high-throughput analysis of SNPs. *Biotechniques* (in press).

Wang DG, Fan J-B, Siao C-J, Berno A, Young P, Sapolsky R, *et al.* (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077–1081.

Wicks SR, Yeh RT, Gish WR, Waterston RH, Plasterk RHA. (2001). Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. *Nature Genet* **28**: 160–164.

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray
Copyright © 2003 John Wiley & Sons, Ltd.
ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 11

Tools for Statistical Analysis of Genetic Data

ARUNA BANSAL[1], PETER R. BOYD[2] and RALPH MCGINNIS[1]

[1]*GlaxoSmithKline, Population Genetics*
New Frontiers Science Park (North)
Third Avenue, Harlow, Essex CM19 5AW, UK

[2]*GlaxoSmithKline, Population Genetics*
Medicines Research Centre, Gunnels Wood Road
Stevenage, Herts SG1 2NY, UK

11.1 INTRODUCTION

The focus of this chapter is on methods that aid in the identification of genetic variants that influence a trait of interest. The trait may be a biological measurement, possibly indicating risk of disease or it may be the response to an environmental stimulus such as a drug. Techniques such as linkage analysis and association analysis are central to the process. These methods are described and corresponding software is reviewed, with worked examples to show how they can be applied. The majority of tools covered may be downloaded, together with full documentation, by following links at http://linkage.rockefeller.edu. Web addresses for the few exceptions are provided in the text. Almost all are available free of charge.

11.2 LINKAGE ANALYSIS

Linkage analysis is applied in the early stages of gene localization and is one means by which an initial, often broad, chromosomal interval of interest is defined. It is a process of tracking the inheritance pattern of genetic markers with the inheritance pattern of a disease or trait. Disease linkage manifests as a marker allele being inherited in diseased individuals more often than would be expected under independent assortment.

Linkage analysis may be parametric to test whether the inheritance pattern of the trait fits a specific model of inheritance or it may be non-parametric (model-free). The former is more powerful under a correctly specified model and is most informative for large, multiply affected pedigrees. The latter is more powerful when the mode of inheritance is unknown, as in complex trait analysis for which small pedigrees are often ascertained.

11.2.1 Parametric Linkage Analysis

By the parametric approach (and in certain non-parametric cases), evidence of linkage is measured by the LOD score (Morton, 1955). It proceeds by an assessment of the recombination fraction, often denoted by theta (θ). Theta is the probability of a recombination event between the two loci of interest and as such it is a function of distance. Two unlinked loci are given by $\theta = 0.5$ and the closer a pair of loci, the lower their recombination fraction. The LOD may be expressed as follows, using L to denote likelihood.

$$LOD = \log_{10} \frac{L(\theta = \hat{\theta})}{L(\theta = 0.5)}$$

The likelihood in the numerator is based upon the maximum likelihood estimate of the recombination fraction, derived from the data. It is compared to that calculated under the null hypothesis of no linkage ($\theta = 0.5$). A high LOD score is thus consistent with the presence of linkage. Due to the computational complexity of the likelihood calculation, software for exact parametric linkage analysis is constrained either by pedigree size or by the number of markers included in the calculation.

The software VITESSE (O'Connell and Weeks, 1995) allows rapid, exact parametric linkage analysis of very extended pedigrees. At the expense of some speed, an alternative, FASTLINK (Cottingham *et al.*, 1993), allows the analysis of large pedigrees that also contain loops (marriages between related individuals). Both VITESSE and FASTLINK are based on an earlier program, LINKAGE (Lathrop *et al.*, 1984) and are available for UNIX, VMS and PC(DOS) systems. Using these pieces of software, analysis is typically conducted by means of a sliding window of one, two or four markers along the chromosome, although larger windows are also possible.

Parametric linkage analysis in more moderately-sized pedigrees is commonly carried out using the software GENEHUNTER (Kruglyak *et al.*, 1996). It is written in C, to be run on UNIX and uses a command-line interface. A major feature of this program is that it allows the rapid, simultaneous analysis of dozens of markers (often an entire chromosome) in a multipoint fashion, thereby providing increased power over single-marker analyses when map positions are known (Fulker and Cardon, 1994; Holmans and Clayton, 1995; Olson, 1995). In order to accommodate uncertainty in marker ordering, an option to perform single marker tests is also available. On most platforms, pedigrees up to size $2n - f = 16$ may be analysed by GENEHUNTER, where n is the number of non-founders (those with parents included in the pedigree), and f is the number of founders. This limit is important to consider, because larger pedigrees are automatically trimmed until they fall within it, leading to possible information loss. Results are stored graphically in postscript files for easy interpretation and presentation.

11.2.2 Non-parametric (Model-free) Linkage Analysis

Non-parametric linkage (NPL) analysis does not allow direct estimation of the recombination fraction, but one source of multiple testing — that derived from examining multiple models — is removed. The general principle is that relatives who share similar trait values will exhibit increased sharing of alleles at markers that are linked to a trait locus (see Holmans (2001) for a review of the method).

Allele sharing may be defined as identical by state (IBS) or identical by descent (IBD). Two alleles are IBS if they have the same DNA sequence. They are IBD if, in addition to being IBS, they are descended from (and are copies of) the same ancestral allele (Sham, 1998). A statistical test is performed to compare the observed degree of sharing to that expected under the assumption that the marker and the trait are not linked. While the test statistic may take the form of a chi-squared, normal or F statistic, often it is transformed to allow it to be expressed in LOD units.

NPL analysis often examines IBD or IBS allele sharing in sets of affected sib-pairs (ASPs), in which both siblings exhibit the trait of interest. In the absence of linkage, ASPs are expected to share zero, one or two alleles IBD, with probabilities 0.25, 0.5 and 0.25 respectively. The presence of linkage to a tested marker leads to a departure from these proportions which may be detected by means of a χ^2 test (Cudworth and Woodrow, 1975). Another model-free test, the mean test, tests the null hypothesis that the proportion of IBD allele-sharing equals 0.5. The latter is implemented in the programs SAGE (1999) and SIBPAIR (Terwilliger, 1996), allowing for larger sibships and cases where IBD status cannot be determined unequivocally.

MAPMAKER/SIBS (Kruglyak and Lander, 1995) is a piece of software widely used to test for linkage in sibling data. It was originally written as a stand-alone program, but its functionality and commands have now also been fully incorporated into GENEHUNTER (Kruglyak *et al.*, 1996) whose algorithms are similar. It accommodates both qualitative and quantitative data for either autosomal or sex-linked chromosomes and again, it allows large numbers of markers to be examined jointly.

For dichotomous trait data, a likelihood ratio (LR) test, analogous to the LOD score above is constructed in MAPMAKER/SIBS. The LR is a test for comparing two models in which the parameters of one model (the reduced model), form a subset of the parameters of the other (the full model). It has many genetic applications and may be expressed as follows, where L denotes likelihood.

$$LR = 2\log_e \frac{L_{full}}{L_{reduced}}$$

It is asymptotically distributed as a χ^2, with degrees of freedom equal to the difference in the number of parameters between the two models. In the current context, the numerator is calculated under maximum likelihood estimates of allele sharing proportions and the denominator is calculated assuming random segregation (Risch, 1990a, b). This LR test is also implemented in other software including SPLINK (Holmans and Clayton, 1995), and ASPEX (Hinds and Risch, 1996).

In the case of quantitative trait (QT) data, a test based on the Wilcoxon rank-sum test is available in MAPMAKER/SIBS. It is broadly applicable, as it makes no assumptions concerning the distribution of phenotypic effects. Alternatively, if the sib-pair QT differences are normally distributed, then the original Haseman–Elston method (1972), also implemented, may be applied with greater power. In this test, the squared QT differences between pairs of siblings are regressed on the proportion of alleles that each pair is estimated to share IBD. It is also implemented in SIBPAL2, part of SAGE (1999).

For pedigrees larger than sibships, there is an ‘NPL’ option in GENEHUNTER, but it was shown to be conservative (Kong and Cox, 1997). Alternatives include the modified version, GENEHUNTER-PLUS (Kong and Cox, 1997) and MERLIN (Abecasis *et al.*, 2002), which also incorporates this modification. The latter is a C++ program for UNIX, again with a command-line interface. It offers further improvements in computational speed and reduction in memory constraints, making it more suited to very dense genetic maps. It has the attractive properties of incorporating error detection routines to improve power, and simulation routines to estimate p-values. Graphical output is not however, currently provided.

For normally distributed quantitative traits (or those capable of being transformed to normality), variance component analysis represents a powerful approach to the study of pedigrees of any size (Amos, 1994; Blangero and Almasy, 1996; Goldgar, 1990). The variance component approach to linkage analysis assumes that the joint distribution of the data for a family depends only on means, variances and covariances. The variance of the phenotype is decomposed into (a) components due to linkage to individual marker locations and (b) residual polygenic and environmental components. Familial covariances are modelled in terms of a maximum of two parameters: an additive genetic-variance component and a dominant genetic-variance component, each estimated from the data. The method is implemented in SOLAR (Blangero and Almasy, 1996), in which the size of each effect may be estimated and tested by an LR test. This is a powerful approach and a major advantage is its scope for incorporating into models the effects of covariates, epistasis and gene–environment interaction. For highly complex problems, Markov Chain Monte Carlo Methods are also available, as implemented for example in LOKI (Heath, 1997) and BLOCK (Jensen *et al.*, 1995). When the parameter set is large however, the computational burden of these methods can be prohibitive.

11.2.3 Example: MAPMAKER/SIBS (Kruglyak and Lander, 1995)

11.2.3.1 Data Import

The current example follows a format originally designed for MAPMAKER/SIBS, but now also accommodated by GENEHUNTER. The input files match rather closely what has become known as ‘LINKAGE format’ due to the software in which it was first introduced (Lathrop *et al.*, 1984; Terwilliger and Ott, 1994). Two files are required, namely a pedigree file and a map file. In the current example, a genetic trait has been simulated for 200 sibships, and the files have been named *regionA.ped* and *regionA.loc* respectively. For the analysis of a quantitative trait, a third file is also required, called for our purposes, *test.pheno*.

The file *regionA.ped* takes the following form where, for simplicity, only a single marker, genotyped in two families has been presented.

```
70   8699      0      0   2   0   0   0
70   8698      0      0   1   0   0   0
70   2230   8698   8699   2   2   1   2
70   2231   8698   8699   2   2   2   2
75   8787      0      0   2   0   0   0
75   8786      0      0   1   0   0   0
75   2238   8786   8787   2   2   2   2
75   2239   8786   8787   2   2   2   2
```

The columns are as follows: kindred ID, individual ID, father’s ID, mother’s ID, sex (1 = male, 2 = female), affection status (1 = unaffected, 2 = affected), genotype. In practice, multiple (paired) columns of genotypes would be included, in map order, for each individual. Missing values are denoted by a zero.

This file therefore provides pedigree structure information, genotypes and, in the case of dichotomous traits, phenotype. For liability class data, an additional liability class column may be included after the affection status column and this is described in more detail in the manual.

For the current example, quantitative trait data is loaded separately using *test.pheno* (not shown). This file contains, on the first line, a count of the number of traits in the file. All subsequent lines take the space-separated form: kindred ID, individual ID and phenotype(s). Only sibling phenotypes should be included.

Lastly, the file *regionA.loc* lists the marker details in map order. Here, the ‘internal’ format is described, but LINKAGE format is also supported. The first line provides a count of the number of markers in the file, and is followed by a blank line. Subsequent lines are in six line blocks as follows. The first line has the marker name and number of alleles; the second has the allele labels; the third has the allele frequencies for each label; the fourth is blank; the fifth is the distance to the next marker; the sixth is blank. If the distances are all below 0.5, they are assumed to be recombination fractions, otherwise they are assumed to be distances in cM. The following is an example of a map file, say *regionA.loc* with just the first two markers shown for the sake of brevity.

```
29

MARKER1 6
1 2 3 4 5 6
.01 .95 .01 .01 .01 .01

5.1

MARKER2 10
1 2 3 4 5 6 7 8 9 10
.114988 .110626 .070579 .218874 .250991 .141158 .028549
.062649 .000793 .000793
```

Note that the program tends to crash if inter-marker distances less than 0.1 cM are provided. This should therefore be used as the lower bound even in the case of apparently recombinationally inseparable markers.

11.2.3.2 NPL Analysis for a Quantitative Trait

The following sequence of UNIX commands may be used.

```
load markers regionA.loc
prepare pedigrees regionA.ped
y
test.pheno
increment step 10
scan
p
nonparametric
1
np.out
np.ps
q
```

The process is as follows. The first step is to import the marker and pedigree data that are stored, respectively in *regionA.loc* and *regionA.ped*. You are then asked whether you wish to import additional phenotypic data. Upon typing *y* (yes), you are prompted for a filename, in this case, *test.pheno*. Increment step 10 specifies that linkage is to be assessed at 10 equally spaced points in each marker interval.

The *scan* command computes the full multi-point probability that two sibs share zero, one or two alleles identical by descent (IBD) with the given map and allele frequencies. You are asked whether to include affected (*a*) or phenotyped pairs (*p*). The latter (*p*) allows NPL analysis to follow. Non-parametric linkage analysis is to be applied to trait 1, with numerical output to be piped to *np.out* and graphical output to be stored in *np.ps* (Figure 11.1, below). Note that if only one trait exists in the phenotype file then the 1 above is not required.

As shown in Figure 11.1, a Z-score provides the measure of linkage and in this case evidence peaks close to marker 22. Localization cannot however be assumed to be precise and separation of at least 10 cM may be seen between studies (Hauser and Boehnke, 1997). It is therefore usual to construct a support interval around a strong linkage signal (Conneally *et al*., 1985). For example, having converted to LOD units, a 1-unit support interval is the interval that includes all (possibly disjoint) map positions with LOD score less than 1 LOD unit below the peak score. A conservative approach is to adopt a 1.5 to 2 LOD support interval. All points within the support interval are considered to be of interest.

A determination of information content in MAPMAKER/SIBS allows a representation of the amount of IBD information extracted by the genotype data, as plotted along the chromosome. Dips in the graph allow regions to be highlighted in which the typing of additional markers could be beneficial. The following commands are applied.

```
scan
a
infomap
info.out
info.ps
```

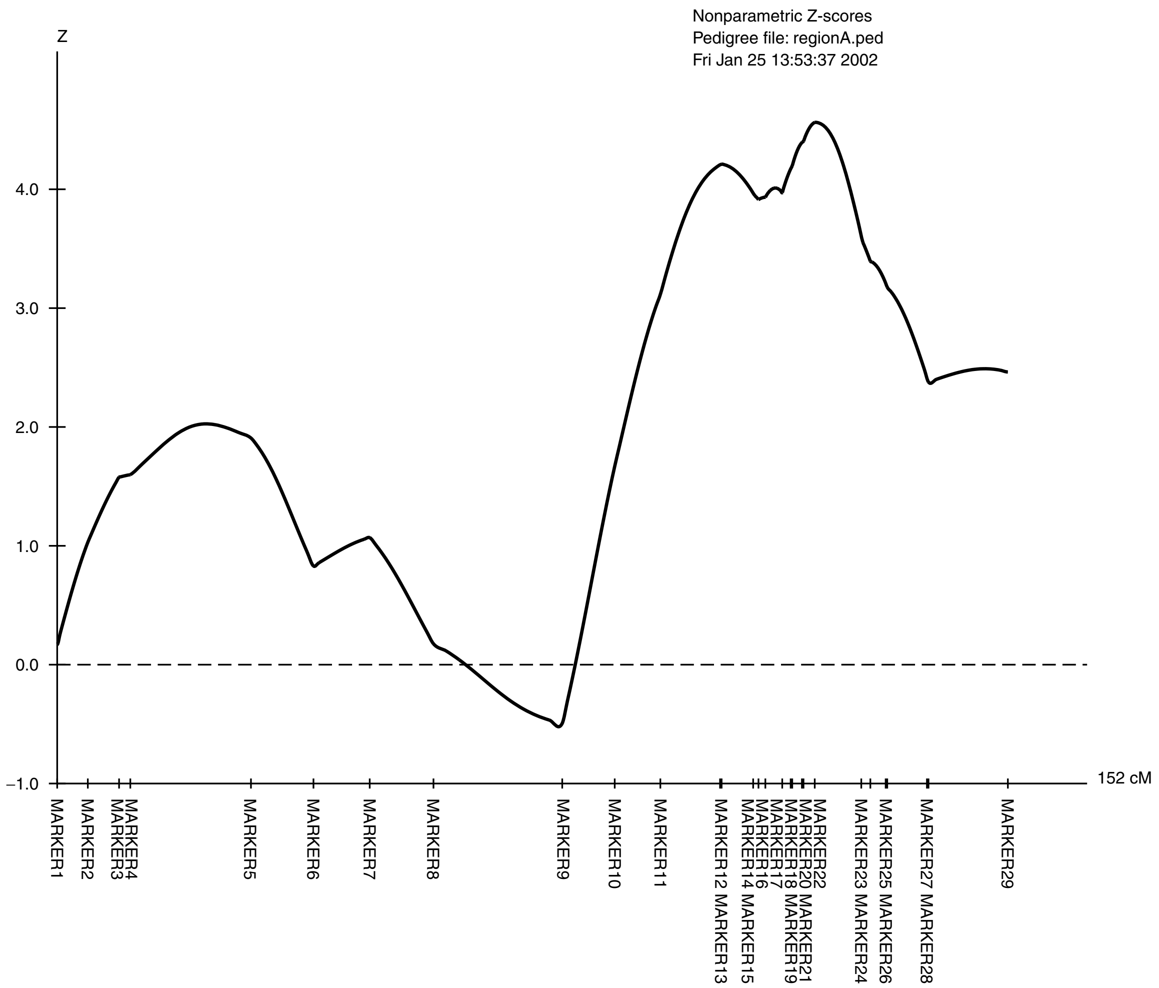


Figure 11.1 Postscript output from MAPMAKER/SIBS. This is *np.ps* from the example run.

A *scan* of affected pairs (*a*) is conducted and *infomap* is requested. The filenames ensure that numerical output is stored in *info.out* and a graphical representation is saved as *info.ps* (Figure 11.2). In this example the large gaps between markers 4 and 5 and between markers 8 and 9 manifest as troughs in the Information Content graph.

Another useful option (not shown) is that the IBD distribution can be output as a text file using the command *dump ibd.* This is a very rapid means of generating IBD probabilities for sibships and, after re-formatting, the output may be used to generate input files for other software such as QTDT (Abecasis *et al.*, 2000), to be discussed later. Another piece of software, SimWalk2 (Sobel and Lange, 1996) will generate IBD probabilities for a wider range of family structures, but in the case of sibships it is slower than MAPMAKER/SIBS.

11.3 ASSOCIATION ANALYSIS

Association analysis may be regarded as a test for the presence of a difference in allele frequency between cases and controls. A difference does not necessarily imply causality in disease, as many factors, including population history and ethnic make-up may yield this effect. In a well-designed study, however, evidence of association provides a flag

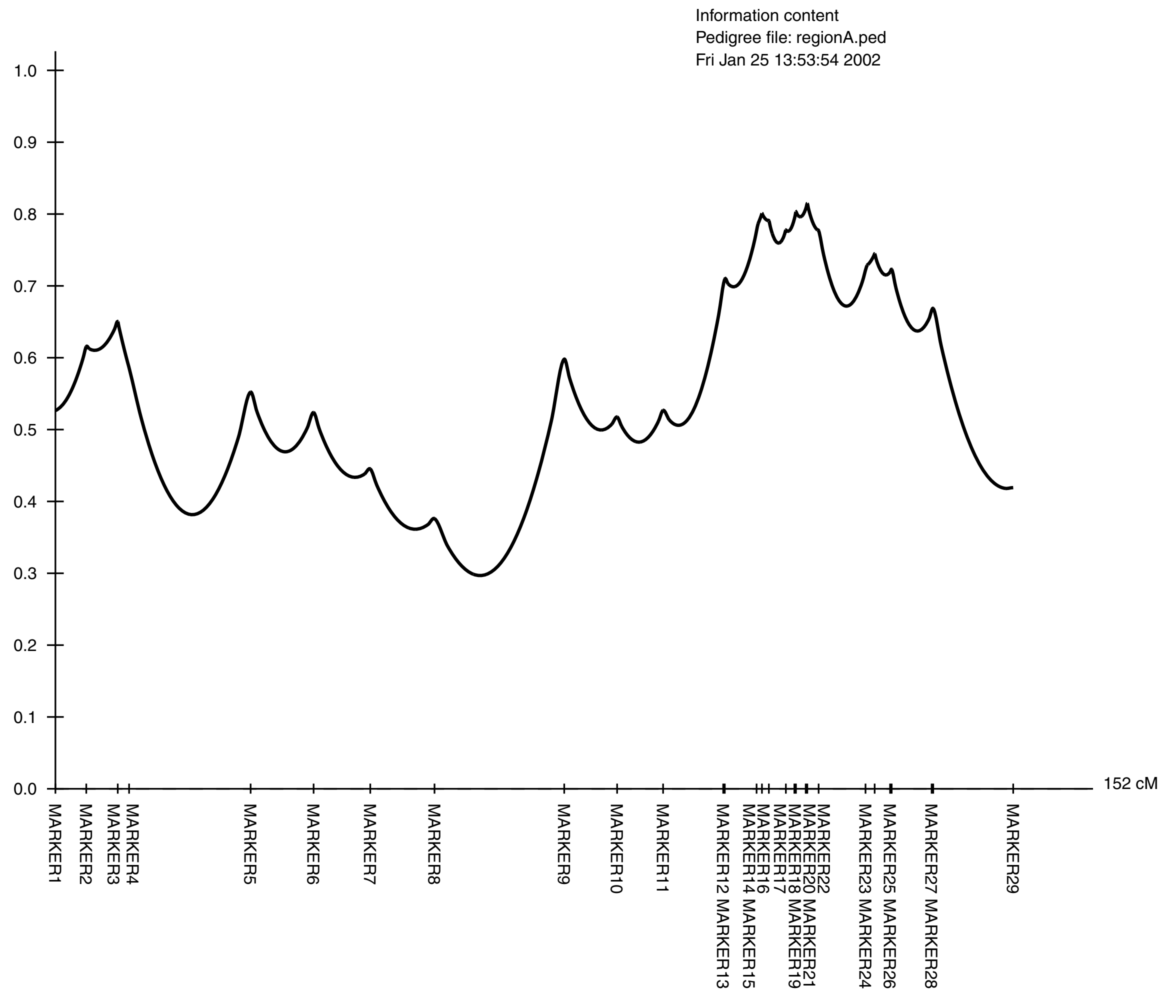


Figure 11.2 Postscript output from MAPMAKER/SIBS. This is *info.ps* from the example run.

for further study. In some instances it is due to the marker being physically close to the causal variant.

Association testing for case–control or population data is often carried out using general (non-genetic) statistical software packages, such as SAS or S-PLUS. A χ^2 test is applied to a contingency table, in which case/control status is tabulated by frequencies of either genotypes or alleles. The test takes the usual form,

$$\chi = \square \frac{(Obs - Exp)^2}{Exp}$$

where *Obs* and *Exp* are the observed and expected frequencies respectively, and the sum is taken over all cells in the table. The number of degrees of freedom is $(r - 1)$ $(c - 1)$, where r is the number of rows, and c is the number of columns in the table. Equivalently, logistic regression can be applied, using disease status as the dependent variable and alleles or genotypes as the independent variables (see Clayton (2001) for a detailed review of the method). The remaining sections of this chapter all involve applications and extensions of the traditional association test.

11.3.1 Transmission Disequilibrium Tests

In recent years, there has been an upsurge in interest in family-based testing owing to the concern that ethnic mismatching of non-family cases and controls (population stratification) can sometimes yield false positive evidence of association. In particular, the transmission/disequilibrium test or TDT (Spielman *et al.*, 1993) has gained prominence as a test of linkage in the presence of association that does not give false evidence of linkage due to population stratification. The TDT is applied by counting alleles transmitted from heterozygous parents to one or more affected children in nuclear families. The alleles *not* transmitted to affected children may be regarded as control alleles, perfectly ethnically matched to the 'case' alleles seen in the affected children. The test takes the form of a McNemar's test, which, under the null hypothesis of no linkage, follows a χ^2 distribution with one degree of freedom. The TDT is also a valid test for association, but only when applied to alleles transmitted from heterozygous parents to just one affected child per family.

Assuming a diallelic locus, let b denote the counts of heterozygous parent-to-offspring transmissions in which allele 1 goes to an affected child, while allele 2 is not transmitted. Let c denote the counts of transmissions the other way around, in which allele 2 is inherited in an affected child, while allele 1 is not transmitted. The test takes the following form:

$$\chi_1^2 = \frac{(b - c)^2}{(b + c)}$$

A number of groups have focused on generalizing the TDT to quantitative traits or to designs in which parental genotypes are not available. The sib-TDT or S-TDT (Spielman and Ewens, 1998) does not use parental genotypes and, like the original TDT, it is not prone to false positives due to population stratification. For association testing, the S-TDT requires that the data in each family consist of at least one affected and one unaffected sibling, each with different marker genotypes. This test and the original TDT are widely implemented, for example in the Java-based program TDT/S-TDT (Spielman and Ewens, 1996, 1998).

Multi-allelic markers may be tested using ANALYZE (Terwilliger, 1995). This has the advantage of taking LINKAGE format files as input and so provides a natural follow-up to a genome scan. It does however require that LINKAGE (Lathrop *et al.*, 1984) be installed on your system. Other software able to handle multi-allelic markers includes ETDT (Sham and Curtis, 1995) and GASSOC (Schaid, 1996).

For quantitative traits, a major development was the release of QTDT (Abacasis *et al.*, 2000), software which allows TDT testing under a variance components framework. It is applicable to sibships with or without parental genotypes and incorporates a broad range of quantitative trait tests — those proposed by Rabinowitz (1997), Allison (1997), Monks *et al.* (1998), Fulker *et al.* (1999) and Abecasis *et al.* (2000). It is written in C++, to be run on UNIX and has a command-line interface. Its input files are based on LINKAGE format, but in addition, one input file of IBD probabilities must be prepared in advance. QTDT assumes the IBD format generated by the programs SimWalk2 (Sobel and Lange, 1996) and MERLIN (Abecasis *et al.*, 2002). Covariates may also be modelled, but should be kept to a minimum in order to maintain performance.

11.4 HAPLOTYPE RECONSTRUCTION

A haplotype is a string of consecutive alleles lying on the same chromosome. Each individual therefore has a pair of haplotypes for any chromosomal interval — one inherited from the paternal side and one inherited maternally. In statistical genetics, their importance lies in the fact that tests of association may be applied to haplotypes instead of single loci. This may yield increased power if the variant of interest is not being tested directly or if adjacent loci are contributing to a single effect (see Clark *et al.*, 1998; Nickerson *et al.*, 1998). Haplotypes can be inferred from the genotypes of parents or other family members (Weeks *et al.*, 1995) or by laboratory methods (Clark 1990; Nickerson *et al.*, 1998). Often, however, they are estimated by means of the Expectation–Maximization (EM) algorithm (Dempster *et al.*, 1977; Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Little and Rubin, 1987; Long *et al.*, 1995).

The EM algorithm is a method that aims to provide maximum likelihood parameter estimates in the presence of incomplete data. In the case of haplotype frequency estimation, it proceeds as follows (Schneider *et al.*, 2000).

1. An initial set of plausible haplotype frequencies is assigned — for example the product of the relevant allele frequencies may be used.
2. The E-step: assuming Hardy–Weinberg equilibrium, the haplotype frequencies are used to estimate the expected frequencies of ordered genotypes.
3. The M-step: the expected genotype frequencies are used as weights to produce improved estimates of haplotype frequencies.
4. Steps 2 and 3 are repeated until the haplotype frequencies reach equilibrium.

Note that, as with other iterative techniques, it is wise to compare the results of multiple starting points as the EM algorithm may converge to a local, rather than global, optimum. It is not always reasonable to assume that the maximum likelihood haplotype configuration has been reached.

Software written specifically for haplotype analysis includes EHPLUS (Zhao *et al.*, 2000), a reworked and extended version of the earlier program EH (Xie and Ott, 1993). It is written in C and is available in both UNIX and PC versions. EHPLUS can be applied to either case–control data or data assumed to come from a random-mating population. It accommodates large numbers of haplotypes and incorporates a companion program, PMPLUS, which will reformat genotype data ready for use. Estimated haplotypes and their frequencies are output and may be subjected to association tests. Permutation features allow the calculation of empirical p-values for these.

Further software for sophisticated haplotype analysis is available from ftp://ftp-gene.cimr.cam.ac.uk/software/clayton/. Resources include SNPHAP, a program that uses the EM algorithm to estimate haplotype frequencies for large numbers of diallelic markers using genotype data. Another program, TDTHAP (Clayton and Jones, 1999) allows the TDT to be applied to extended haplotypes. STATA routines to aid SNP selection by haplotype tagging (Johnson *et al.*, 2001) are available in ftp://ftp-gene.cimr.cam.ac.uk/software/clayton/stata/htSNP/.

Haplotype reconstruction from family data can be achieved by using SimWalk2 (Sobel and Lange, 1996). The derived haplotypes may then be imported to a pedigree-drawing package such as Cyrillic (Chapman, 1990) for viewing recombinants in positional cloning for example. MERLIN (Abecasis *et al.*, 2002) and GENEHUNTER (Kruglyak *et al.*, 1996) also output haplotypes estimated from family data. Another piece of software, TRANSMIT (Clayton, 1999) allows association testing of family-based haplotypes. All of these programs allow for missing parental genotypes.

11.4.1 Example: EHPLUS and PMPLUS (Zhao *et al.*, 2000)

11.4.1.1 Data Import

PMPLUS requires two input files, namely a parameter file and a data file. The data file contains for each individual, subject ID, subject status (0 = control, 1 = case) and genotypes listed as either pairs of numbered alleles or as numerical genotype codes. A data file with three markers takes the following form:

```
[Subject ID] [Status] [1a] [1b]   [2a] [2b]   [3a] [3b]
   or
[Subject ID] [Status] [1] [2] [3]
```

where *[1a]* and *[1b]* are the alleles of the first genotype or, alternatively, *[1]* alone represents the first genotype. Currently, the compiled limits are 15 alleles, 30 markers and 800 subjects. Note also that whereas subject IDs with a decimal point (e.g. ‘20.1’) work well, more complex IDs containing several dashes and decimal points may lead to erroneous output.

The parameter file consists of five lines of space-delimited integer values, and it defines the tests to be carried out. The following parameter file *(hapfrest.par)* may be used to estimate haplotype frequencies:

```
3 0 0 0
2 2 2
0 0
1 1 1
1 1 1
```

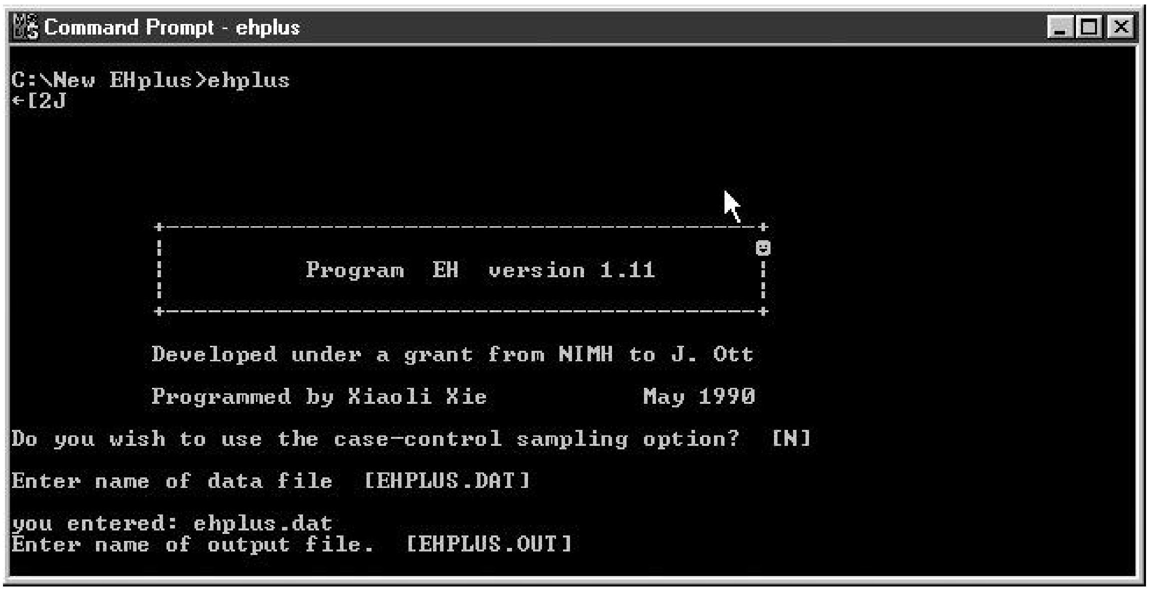
The four values on line 1 specify the number of markers in the data file, whether to perform a marker–marker or case–control analysis (0 or 1, respectively), whether case–control status is to be permuted (0 = no, 1 = yes) and the number of permutations to perform. Line 2 gives the number of alleles for each marker in the data file. The first value on line 3 specifies whether genotypes in the data file are pairs of alleles or numbered genotypes (0 or 1, respectively), while the second value specifies whether screen output is suppressed or shown (0 or 1). Line 4 has a 1 for each marker to be included in the analysis; zero otherwise. Line 5 assigns each marker to one of two blocks (0 or 1), if required in a marker–marker analysis.

11.4.2 Estimating Haplotype Frequencies

Firstly, PMPLUS is run by typing the following:

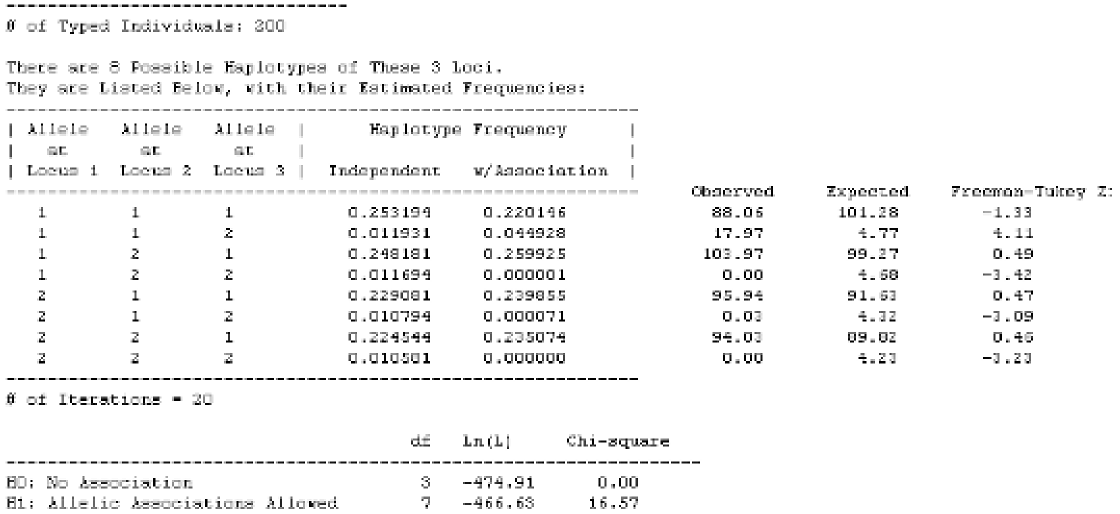
```
>pmplus hapfrest.par hapfrest.dat hapfrest.out
```

Here *hapfrest.out* is an output file named by the user and created by PMPLUS to record chi-squared statistics and associated p-values for the specified analysis. A second output file named *ehplus.dat* is also generated, in which the contents of *hapfrest.dat* have been converted into EHPLUS format ready for estimation of haplotypes.



```
C:\New EHplus>ehplus
←[2J


          +--------------------------------------------+
          |                                            ☻
          |          Program   EH   version 1.11       |
          |                                            |
          +--------------------------------------------+

          Developed under a grant from NIMH to J. Ott

          Programmed by Xiaoli Xie           May 1990

Do you wish to use the case-control sampling option?  [N]

Enter name of data file  [EHPLUS.DAT]

you entered: ehplus.dat
Enter name of output file.  [EHPLUS.OUT]
```

Figure 11.3 EHPLUS interface, as used for estimating haplotype frequencies.



```
-----------------------------------
# of Typed Individuals: 200

There are 8 Possible Haplotypes of These 3 Loci.
They are Listed Below, with their Estimated Frequencies:
-----------------------------------------------------------
| Allele   Allele   Allele  |      Haplotype Frequency      |
|   at       at       at    |                               |
| Locus 1  Locus 2  Locus 3 |  Independent   w/Association |
-----------------------------------------------------------      Observed     Expected    Freeman-Tukey Z:
    1        1        1        0.253199      0.220146              88.06       101.28          -1.33
    1        1        2        0.011931      0.044928              17.97         4.77           4.11
    1        2        1        0.248181      0.259925             103.97        99.27           0.49
    1        2        2        0.011694      0.000001               0.00         4.68          -3.42
    2        1        1        0.229081      0.239855              95.94        91.63           0.47
    2        1        2        0.010794      0.000071               0.03         4.32          -3.09
    2        2        1        0.224544      0.235074              94.03        89.82           0.46
    2        2        2        0.010501      0.000000               0.00         4.23          -3.23
-----------------------------------------------------------
# of Iterations = 20

                                           df    Ln(L)      Chi-square
-------------------------------------------------------------------
H0: No Association                          3    -474.91       0.00
H1: Allelic Associations Allowed            7    -466.63      16.57
```

Figure 11.4 Haplotype frequency output from EHPLUS — named *ehplus.out*.

Haplotype estimation is carried out by typing >*ehplus* to invoke the program and then pressing the <*CarriageReturn*> three times to accept the default options provided. The process, as seen using the PC(DOS) version, is shown in Figure 11.3. The output file, *ehplus.out*, shown in Figure 11.4, contains the estimated haplotype frequencies (see column labelled *w/Association*) as well as log likelihoods for the null and alternative hypotheses of *No Association* between the markers and *Allelic Associations Allowed* between the markers. Such inter-marker association is termed linkage disequilibrium and shall be the topic of the next section of this chapter.

11.4.3 Haplotype-based Association Testing

In order to test 2-point haplotypes for association to a disease, the parameter file, *hapfrest.par* was modified to produce the following parameter file (*cscntcom.par*):

```
3 1 1 100
2 2 2
0 0
1 1 0
0 0 0
.01 0 1 1
```

The second, third and fourth entries on line 1 of this parameter file now specify that a case–control analysis is to be performed and significance determined by permuting case–control status 100 times. Furthermore, other entries specify that only the first two markers are to be included (line 4), and that genotypes are *not* to be permuted (line 5). This time a sixth line is included, applicable only to case–control analyses. This line contains four possibly non-integer values that define a model of the mode of inheritance of the trait. The first value specifies the assumed disease allele frequency; the following three are penetrant estimates, resulting from zero, one or two copies of the disease allele respectively. In the current example, the model assigns a 0.01 allele frequency and a fully penetrant, pure dominant mode of inheritance. The new data file (*cscntcom.dat*) contains the simulated genotypes of both cases and controls, formatted as described previously. PMPLUS is executed as follows:

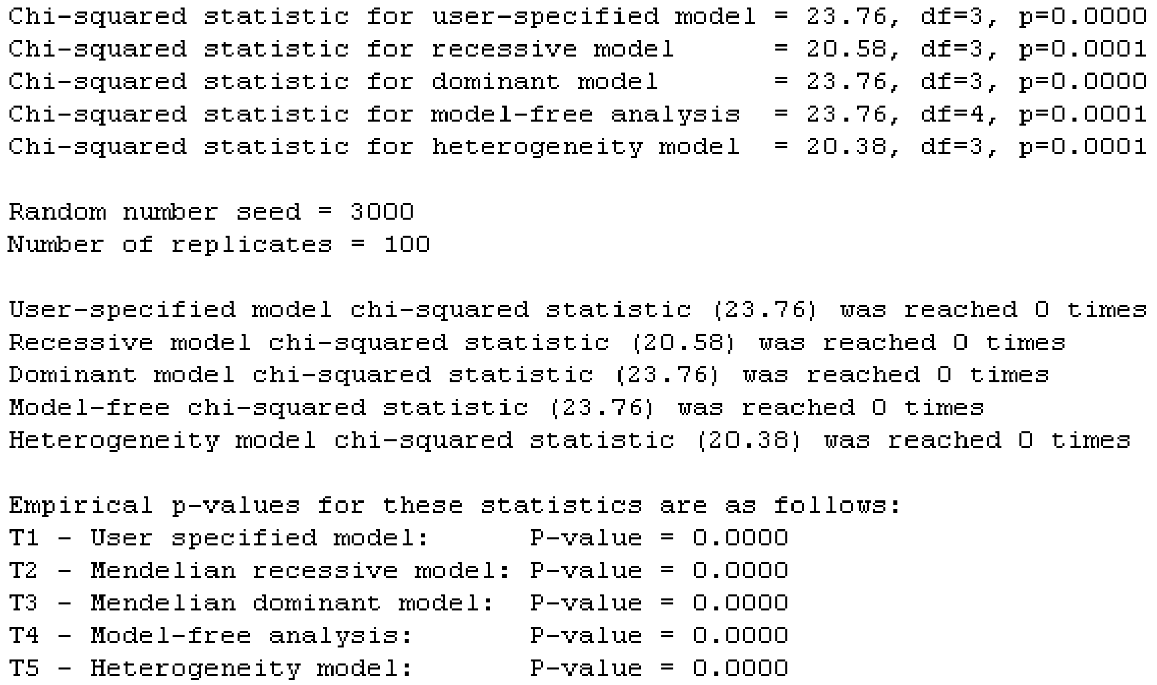
```
>pmplus cscntcom.par cscntcom.dat cscntcom.out.
```

When PMPLUS is instructed to perform permutations, EHPLUS is automatically invoked at the end of each PMPLUS run (i.e. following each permutation of the dataset) and thus program control passes back and forth between the two programs until the permutations are complete. Since PMPLUS permutes the data via the *ehplus.dat* input file, the final EHPLUS output file (*ehplus.out*) does not have meaningful haplotype frequency estimates. These are based on permuted, rather than real data.

The output produced by PMPLUS (in this case *cscntcom.out*) contains the key analysis results. These are the χ^2 values and permutation-derived p-values obtained under five sets of assumptions as follows: (1) under the user-specified disease model, (2) under a Mendelian recessive model, (3) under a Mendelian dominant model, (4) by maximizing the log likelihood ratio over multiple disease models and (5) by a non-parametric 'homogeneity' test, to compare log likelihoods calculated from pooling cases and controls and considering them separately. The fifth test is completely non-parametric, while the others are constrained by the population prevalence of disease implied by the user-specified disease model. Figure 11.5 shows the results of evaluating the 2-point haplotype for association with the simulated disease. Note that p-values below 0.0001 are rounded down to zero.

11.5 LINKAGE DISEQUILIBRIUM

Linkage disequilibrium (LD) is a lack of independence, in the statistical sense, between the alleles at two loci. LD exists between two linked loci when particular alleles at these loci occur on the same haplotype more often than would be expected by chance alone. This phenomenon can provide valuable information in locating disease variants from marker data, as a marker in LD with the causal variant provides a flag for its location. LD information also provides a means by which the efficiency of high-density marker maps can be increased. If markers are in strong LD with each other, there is an argument for genotyping only a subset of them.



```
Chi-squared statistic for user-specified model = 23.76, df=3, p=0.0000
Chi-squared statistic for recessive model      = 20.58, df=3, p=0.0001
Chi-squared statistic for dominant model       = 23.76, df=3, p=0.0000
Chi-squared statistic for model-free analysis  = 23.76, df=4, p=0.0001
Chi-squared statistic for heterogeneity model  = 20.38, df=3, p=0.0001

Random number seed = 3000
Number of replicates = 100

User-specified model chi-squared statistic (23.76) was reached 0 times
Recessive model chi-squared statistic (20.58) was reached 0 times
Dominant model chi-squared statistic (23.76) was reached 0 times
Model-free chi-squared statistic (23.76) was reached 0 times
Heterogeneity model chi-squared statistic (20.38) was reached 0 times

Empirical p-values for these statistics are as follows:
T1 - User specified model:       P-value = 0.0000
T2 - Mendelian recessive model: P-value = 0.0000
T3 - Mendelian dominant model:  P-value = 0.0000
T4 - Model-free analysis:        P-value = 0.0000
T5 - Heterogeneity model:        P-value = 0.0000
```

Figure 11.5 Output of haplotype-based association testing in EHPLUS.

The extent of pair-wise LD may be measured by the value D, as follows (Lewontin, 1964). Assume two diallelic loci are linked and let p_{ij} be the proportion of chromosomes that have allele i at the first locus and allele j at the second locus. For example, p_{12} is the frequency of the haplotype with allele 1 at the first locus and allele 2 at the second locus. The disequilibrium coefficient D is the difference between the observed haplotype frequency p_{12} and the haplotype frequency expected under linkage equilibrium, the latter being the product of the two allele frequencies, say p_{1+} and p_{+2}. It may be written as follows:

$$D = p_{12} - p_{1+}p_{+2}$$

Another commonly quoted measure of LD is D' (Lewontin, 1964). This is a normalized form, with numerator equal to D and denominator equal to the absolute maximum D that could be achieved given the allele frequencies at the two loci. Many other valid measures of pair-wise LD exist and have been reviewed elsewhere (Devlin and Risch, 1995; Hedrick, 1987).

As noted above, EHPLUS can perform tests of LD among a group of markers. The complete set of pair-wise tests for the group, together with D and D' values, can be achieved in a single step using software such as Arlequin (Schneider *et al.*, 2000). This is a C++ program available for PC(Win), Linux and MacOS systems. The statistical significance of observed LD is estimated for phase known (haplotype) data by means of a Fisher's Exact Test. For phase unknown data, a likelihood ratio test is applied. An alternative tool is GDA (Lewis and Zaykin, 2001), the PC(Win) companion program to the book, *Genetic Data Analysis II* (Weir, 1996). Both are well documented and perform a broad range of population genetic tests.

The software, GOLD (Abecasis and Cookson, 2000), available for PC(Win), is another program that will calculate D and D', and it is noteworthy in that it can output them in graphical form. For each marker pair, the pair-wise disequilibrium statistics are colour coded (bright red to dark blue) and plotted. The output is valuable for presentation purposes and provides a useful summary of the properties of dense maps. The software takes haplotype estimates as input and, in the case of family data, these must be reconstructed using software such as SimWalk2 (Sobel and Lange, 1996) prior to use. Case–control data is not well supported by GOLD, which relies for this purpose upon a limited interface to the software, EH (Xie and Ott, 1993).

Other methods of estimating LD include the Moment Method, applicable to newly-formed populations under certain assumptions concerning the evolutionary process (Hastabacka *et al.*, 1992; Kaplan *et al.*, 1995; Lehesjoki *et al.*, 1993). Maximum likelihood methods have also been explored (Hill and Weir, 1994; Kaplan *et al.*, 1995). Composite likelihood methods were proposed to evaluate the information from multiple pairs of loci simultaneously. Examples of software for the composite likelihood approach include DMAP (Devlin *et al.*, 1996) and ALLASS (Collins and Morton, 1998). The latter uses the Malecot isolation by distance equation and has the advantage of accommodating multiple founder mutations. Each method however relies upon population assumptions and may suffer reduced power when these are not met.

11.5.1 Example: Arlequin (Schneider *et al.*, 2000)

11.5.1.1 Data Import

Arlequin categorizes data into five groups, namely DNA sequences, RFLP data, microsatellite data, allele frequency data and standard data. The latter assumes that different alleles are mutationally equidistant from each other, as is the case with SNP data. Data can be loaded in two ways, by importing a project file, or by using the Project Wizard, to guide you through the creation of a project. Figure 11.6 shows the Arlequin interface in Windows NT, having selected the import screen. As shown, a number of data formats may be read in, and converted by selecting Arlequin as the Target format. LINKAGE format is not however, supported.

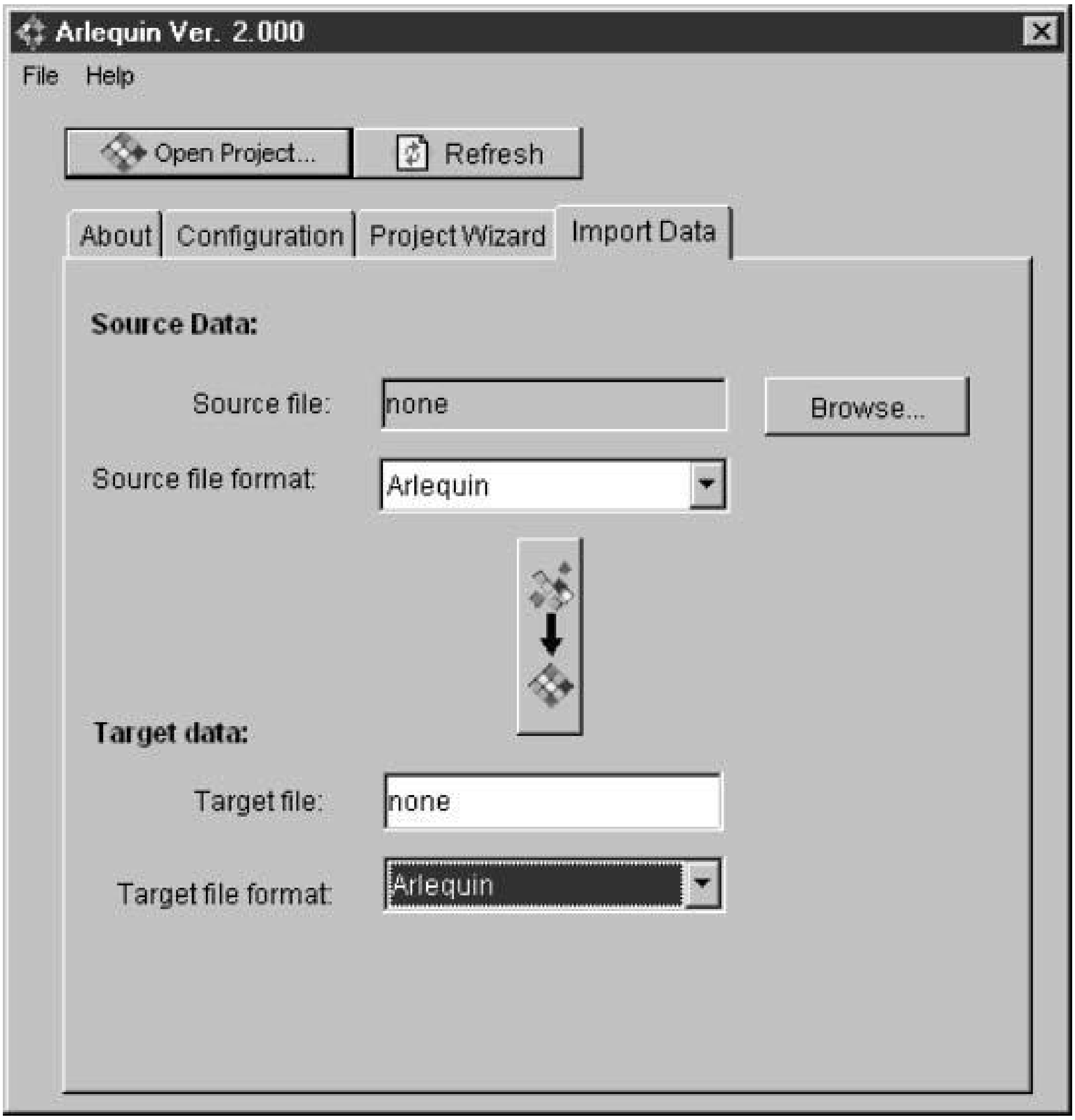
With the objective of testing for LD between five markers, the current example may be regarded as a Standard data project. The data and the parameters of the project are shown below in an Arlequin format, for which the filename extension *.arp* is required. The first *[Profile]* section describes the data before it is listed in the second, *[Data]* section. Comments are included, preceded by '#' and these are ignored by the program.

```
[Profile]            # first describe the data for this project

Title = 'Simulated data for five genetic markers'
      NbSamples = 1    # Number of study populations in the project.
  DataType = STANDARD
      GenotypicData = 1   # 1= yes; 0 = no (i.e. haplotypic)
  LocusSeparator = WHITESPACE
      GameticPhase = 0    # 1 = yes; 0 = no (i.e. phase unknown)
      RecessiveData = 0   # 1 = yes; 0 = no (i.e. codominant alleles)
      RecessiveAllele = null # because RecessiveData = 0
      MissingData  = '.'  # the missing data code
[Data]                    # next list the data points

[[Samples]]

         SampleName = 'Simulation 1'
         SampleSize = 200 #200 individuals are in the study set
         SampleData = {
```

```
CONFIG1 34 1 1 1 1 2 # The first genotype combination is labelled CONFIG1
           2 1 2 1 2 # 34 individuals have this set of five genotypes
CONFIG2 14 2 1 1 1 2
           2 1 1 1 2
CONFIG3 9  1 1 1 1 2 # 9 individuals have this set of five genotypes
           1 2 2 1 2
```

Subsequent lines of data follow the same paired format and the final line consists of a '}' symbol. This project file is specific to the problem in hand, namely phase-unknown genotype data. Variations exist for other data types and are described in detail in the user manual. It can be seen that genotypes are written with one allele directly below the other allele. This allows a mechanism for inputting phase-known data, for which each line represents a haplotype. In our case, the phase is unknown, so the relative orderings of the alleles are ignored.

Upon successful import, a 'Project' is created by Arlequin. It is remembered by the system and can be recalled at a later date. Its details can be viewed by selecting the menu



Source Data:

Source file: none

Source file format: Arlequin

Target data:

Target file: none

Target file format: Arlequin

Figure 11.6 Arlequin Screen. Initiating an analysis run.

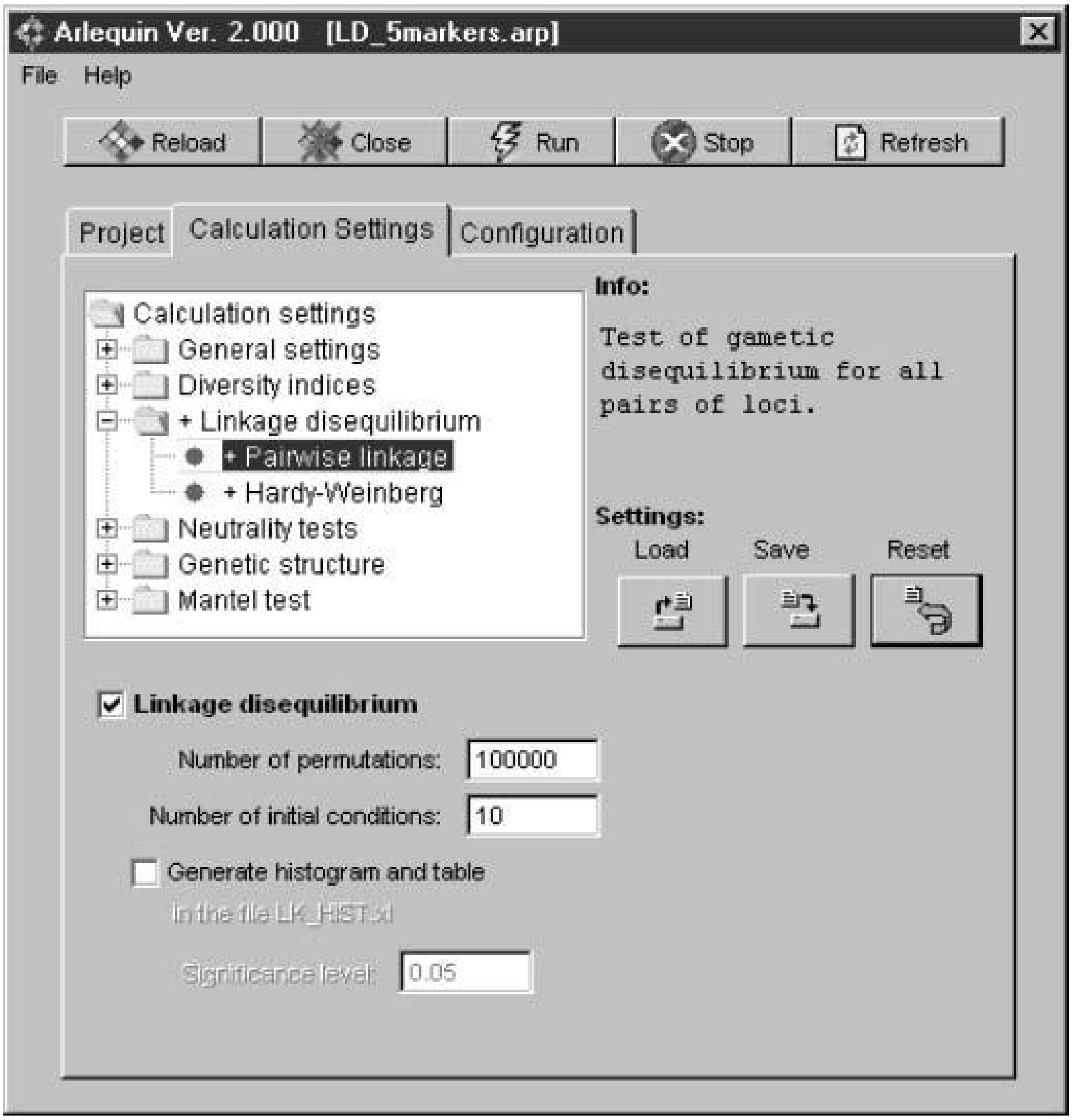
items Project>View Project Info. The following analyses are then performed by making selections in the launch pad dialogue box.

11.5.2 Linkage Disequilibrium Analysis of Genotypes with Unknown Phase

An LR test statistic, denoted by S, is used to test for LD between a pair of loci when phase is unknown (Slatkin and Excoffier, 1996). It compares the likelihood of a model assuming linkage equilibrium to that of a model allowing linkage disequilibrium. Asymptotically, this statistic follows a χ^2 distribution, but to allow for small sample size or the study of markers with large numbers of alleles, Arlequin also uses a permutation procedure to test for significance.

The analysis screen is given in Figure 11.7. The procedure is as follows:

1. Click on the *Calculation Settings* tab
2. Click to the left of the folder *Linkage disequilibrium*



Info:

Test of gametic disequilibrium for all pairs of loci.

☑ **Linkage disequilibrium**

Number of permutations: 100000

Number of initial conditions: 10

□ Generate histogram and table

In the file LK_HIST.txt

Significance level: 0.05

Figure 11.7 Arlequin Screen. Setting up LD analysis of phase-unknown genotype data.

3. Select *Pairwise linkage*
4. Select the *Linkage Disequilibrium* box below the settings window. A plus sign will appear
5. Input parameter values. Default values are given in Figure 11.7. However in the manual, it is recommended that 16,000 permutations be conducted to establish significance and the EM be applied to at least 100 initial conditions
6. Click on the *Run* button

Detailed output is written to an HTML result file, in a sub-directory of that containing the input. First, parameter settings are stated, then for each locus pair, there is a listing of the log-likelihoods under the null and alternative hypotheses, a p-value determined by permutation, the χ^2 test statistic and its corresponding (asymptotic) p-value. Lastly, a table is provided, in which a '+' sign denotes nominal evidence of a departure from linkage equilibrium. This allows the results to be scanned rapidly by eye. Samples of the output are shown below.

```
Pair(0, 1)
     LnLHood LD: -302.71677      LnLHood LE: -319.36838
     Exact P = 0.00000 +- 0.00000 (16002 permutations done)
Chi-square test value = 33.30322 (P  = 0.00000, 1 d.f.)
Pair(0, 2)
     LnLHood LD: -420.27411      LnLHood LE: -420.70319
     Exact P = 0.36164 +- 0.00381 (16002 permutations done)
Chi-square test value  = 0.85815 (P  = 0.35426, 1 d.f.)
```

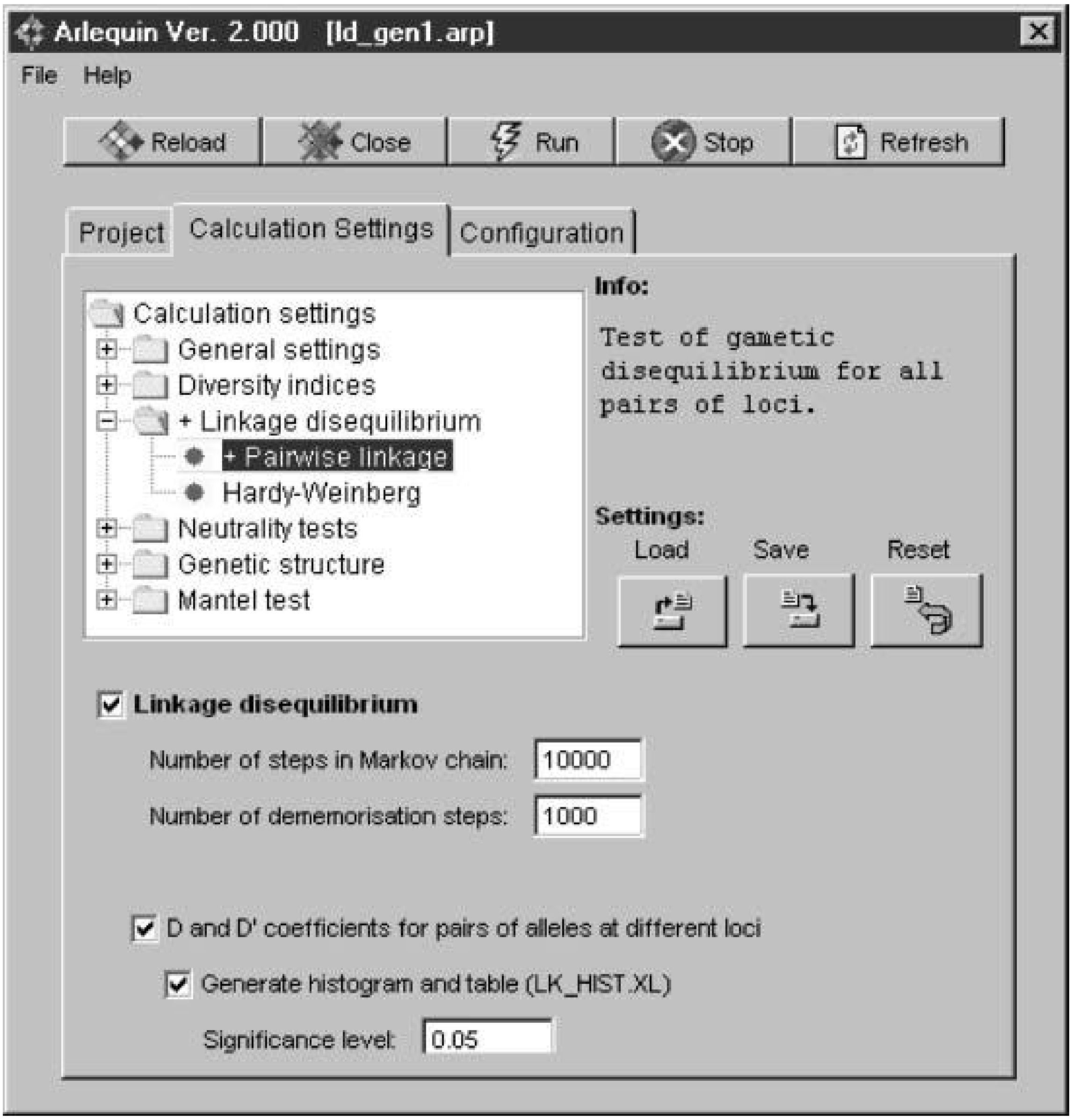
(and so on)

```
Table of significant linkage disequilibrium (significance
level  = 0.0500):

Locus # |   0|  1|  2|  3|  4|
----------------------------------------------
     0|    *    +    -    -    -
     1|    +    *    +    -    -
     2|    -    +    *    -    -
     3|    -    -    -    *    -
     4|    -    -    -    -    *
```

11.5.3 Linkage Disequilibrium Analysis of Haplotypes

Arlequin uses a modified Fisher's Exact test, as opposed to the LR test, to examine LD in haplotype data. Such data is given by $GameticPhase = 1$. The program employs Markov Chain Monte Carlo sampling to explore the space of different possible contingency tables rather than enumerating all the possible contingency tables. In this case, the LD measures, D and D' may also be generated. The analysis screen reflects these additional options as shown in Figure 11.8.

The process of initiating the analysis is very similar to that described above. This time, the number of steps in the Markov chain must be specified, together with the number of de-memorization steps. Again, the default values are lower than those suggested in the manual, which mentions values of 100,000 and 'a few thousand' respectively. If the D and D' boxes are selected, all pair-wise values are tabulated and output in HTML format as well as in a file called *LD_DIS.XL*, ready for inputting to MS Excel.



Info:

Test of gametic disequilibrium for all pairs of loci.

☑ Linkage disequilibrium

Number of steps in Markov chain: 10000

Number of dememorisation steps: 1000

☑ D and D' coefficients for pairs of alleles at different loci

☑ Generate histogram and table (LK_HIST.XL)

Significance level: 0.05

Figure 11.8 Arlequin Screen. Setting up LD analysis of haplotype data.

11.6 QUANTITATIVE TRAIT LOCUS (QTL) MAPPING IN EXPERIMENTAL CROSSES

In contrast to human studies, in which variances of phenotypic differences are used to establish the presence of linkage, QTL mapping in experimental crosses involves comparing means of progeny inheriting specific parental alleles. This is simpler and more powerful (Kruglyak and Lander, 1995). It can be achieved by any of a number of standard statistical methods, such as *t*-tests, analysis of variance (ANOVA), Wilcoxon rank-sum and regression techniques. Again, missing data can be accommodated by an application of the EM algorithm.

Of the very broad array of possible diploid crosses, the following are particularly common. They are derived from a pair of divergent inbred lines in which the genotypes at the majority of loci are homozygous and distinct, say *aa* and *bb* for a particular locus in the two lines respectively. The filial F_1 generation results from crossing these two lines to produce individuals with heterozygous genotype *ab*. In the backcross (BC) design, F_1 is crossed with one of the parent strains. For example, in the case of a cross with the *aa* parent, half the offspring produced are *ab* and half are *aa*. In the filial F_2 design, the F_1 is selfed, or two F_1 individuals are crossed so that offspring are *aa, ab,* and *bb* in the ratio 1:2:1. Lastly, in the recombinant inbred line (RIL), each F_2 enters individually a single seed descent-inbreeding programme so that all progeny are homozygous for the chosen allele.

The original statistical framework for QTL mapping in experimental crosses was based upon a marker-by-marker analysis. Of particular relevance to sparse maps however, simple interval mapping (IM or SIM) allows the evaluation of any position within a marker interval. The maximum likelihood approach to IM proceeds by calculation of a LOD score (Lander and Botstein, 1989). Similarly, and with lower computational burden, least squares regression achieves the same goal (Haley and Knott, 1992; Martinez and Curnow, 1992). IM may be carried out using a range of software, including MAPMAKER/QTL (Lander *et al*., 1987). This may appeal to regular users of MAPMAKER/SIBS or GENEHUNTER, as the syntax is similar. It relies upon data pre-processing in MAPMAKER/EXP (Lander *et al*., 1987) and allows simple graphical output.

Two newer and related methods are Composite Interval Mapping (CIM) and Multiple QTL Mapping (MQM). Both involve performing a genome scan by moving stepwise along the chromosome and testing for the presence of the QTL using a pre-defined set of markers as co-factors (Jansen 1992, 1993; Jansen and Stam, 1994; Kao *et al*., 1999; Zeng, 1993, 1994; Zeng *et al*., 1999). In other words, in the sparse map case, interval mapping is combined with multiple regression on markers. This approach allows you to control, to some extent, for effects of other QTLs. Software such as QTL Cartographer (Basten *et al*., 1994, 1997) and PLABQTL (Utz and Melchinger, 1996; http://probe.nalusda.gove:8000/otherdocs/jqtl/) allow the selection of such co-factors by stepwise regression. These programs offer options that will automatically include or exclude background markers according to user-defined criteria.

Lastly, Bayesian methods allow the consideration of multiple QTLs, QTL positions and QTL strengths (Jansen, 1996; Satagopan *et al*., 1996; Sillanpaa and Arjas, 1998; Uimari *et al*., 1996). The software Multimapper (Sillanpaa, 1998), for example, allows the automatic building of models of multiple QTLs within the same linkage group. It is designed to work as a companion program to QTL Cartographer (Basten *et al*., 1994, 1997) and allows a more detailed follow-up of regions of interest. As with other Markov Chain Monte Carlo methods, however, this approach is computer intensive and may suffer from problems of convergence to a local, rather than global, optimum or of lack of convergence if run for a short time.

Ten of the most prominent pieces of software for QTL mapping are reviewed in greater detail by Manley and Olson (1999). The majority will perform IM and CIM for backcross, filial F_2 and recombinant inbred lines. Cordell (2002) provides worked examples of the usage of three of them, MAPMAKER/QTL, QTL Cartographer and another piece of software, MapQTL (van Ooijen and Maliepaard, 1996a, b).

A major limitation of QTL mapping using inbred lines is the broad, ill-defined nature of the resulting linkage peaks, which typically span tens of centiMorgans even if large numbers of progeny are analysed (for example see Farmer *et al*., 2001). This is a consequence of the multifactorial nature of quantitative traits, which results in an inability to identify unequivocal recombinants that precisely delineate a critical genetic interval, in contrast with monogenic phenotypes. Subsequent attempts to narrow a locus by, for example, successive rounds of backcrossing are often frustrated by the dilution or loss of unlinked genetic co-factors that are required for trait manifestation. In the future, QTL mapping using genetically heterogeneous stocks may gain in prominence (Mott *et al*., 2000). Talbot *et al*. (1999) were able to achieve a mapping resolution of less than 1 cM by the study of heterogeneous stocks from eight known inbred mouse progenitor strains that had been intercrossed over 30–60 generations. The group has released software called HAPPY (Mott *et al*., 2000) which requires knowledge of the ancestral alleles in the inbred founders, together with the genotypes and phenotypes in the final generation. It will then apply variance component methods to test for linkage to the QTL.

11.6.1 Example: Map Manager QTX (Manley *et al*., 2001)

Map Manager QTX is available for both MacOS and PC(Win). It has no licence fee and was selected here due to the usefulness of its graphic user interface. It has both IM and CIM capability and can reformat data for use in other important software such as QTL Cartographer. Interval mapping is based on the Haley and Knott (1992) procedure, and CIM is achieved by adding background loci. Significance can be assessed by permutation (Churchill and Doerge, 1994).

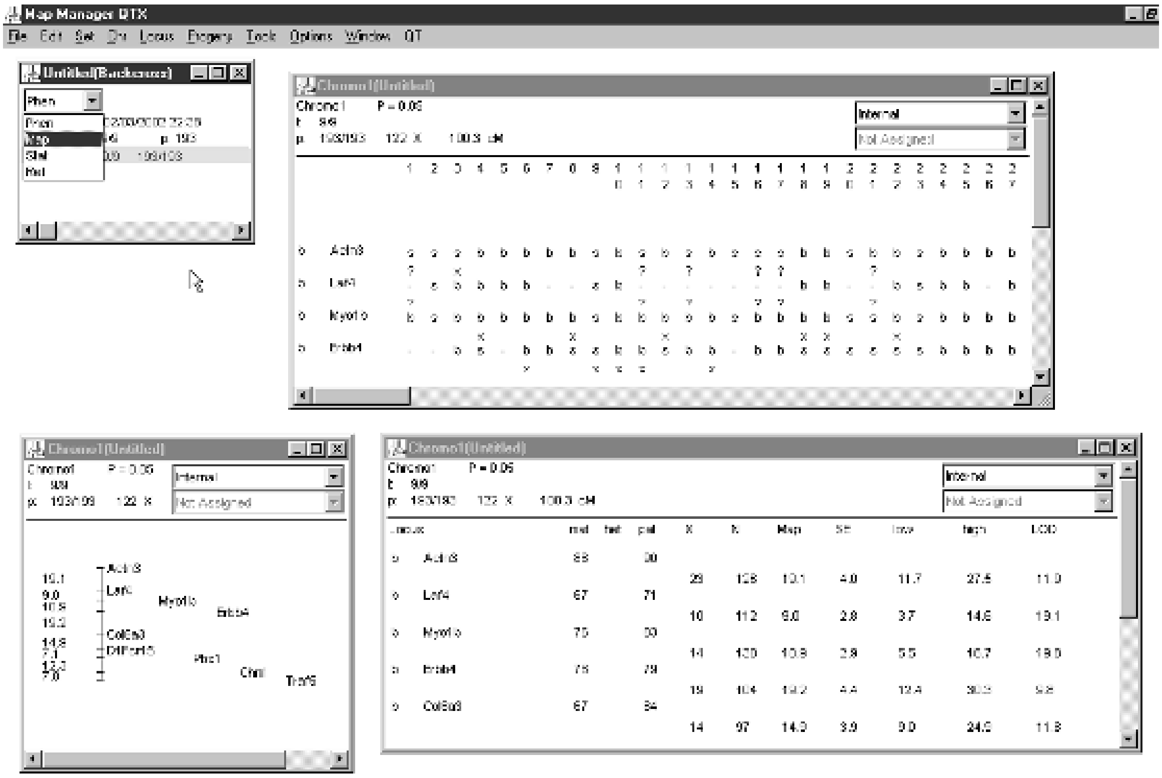
The genotype data may derive from inbred or non-inbred stock and options are provided for a variety of experimental designs. Extensive documentation can be downloaded in either pdf or Hypertext formats. The *Tutorial* is especially helpful; but readers should be aware that its files are somewhat inconspicuously tucked in with *Sample Data* files, rather than being included in the Map Manager QTX Manual.

For the current example, genotype data was downloaded from the Mouse Genome Database (2001). Specifically, it consists of mouse chromosome 1 genotypes from the Copeland–Jenkins backcross, and a selected subset of 10 markers spanning the entire □ 100-cM length of the chromosome. Marker *En1* is located near the middle of the chromosome, between markers *Col6a3* and *D1Fcr15*, and it was used to simulate the quantitative trait (QT) for the 193 backcross mice. Homozygotes (denoted as *b*) at *En1* received a QT value of 50 ± 20 (mean ± SD) while heterozygotes (*s*) at *En1* received a QT value of 100 ± 20. *En1* was then removed from the dataset and Map Manager QTX was used to analyse QT association with the remaining nine markers as shown below.

11.6.1.1 Data Import

Map Manager QTX is launched by a mouse click on the Map Manager icon (*QTXb13.exe*), thus opening the main menu. The genotype data (alternatively termed 'Phenotype data' by Map Manager QTX) is imported by selecting *File>Import>Text*. The name of each marker and the genotypes (phenotypes) of the cross progeny are imported as a single line of text. The marker name is separated from the genotypes by a tab character but the genotypes, each represented as above by a single letter, can be given as either an unbroken string of characters or space-separated. In our case, the first two lines of input therefore took the following form (with missing genotypes given by a hyphen):

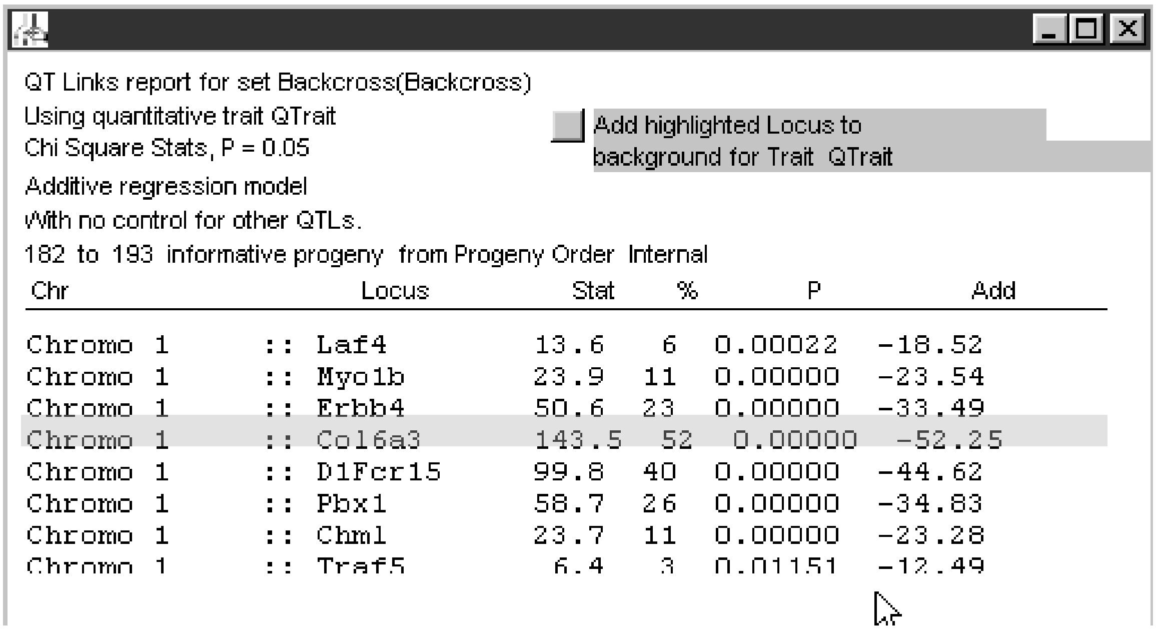
```
Actn3<tab>sssbbbbsbsbsbsssbbsbbsbbbbssbsbbsbsb-bbb-ssss
  <CarriageReturn>
Laf4<tab>-sbbbb--sb-------bb--bsbb-bbsb--s-bbbbbsbssb-
  bs<CarriageReturn>
```



Chromo1 P = 0.05
t 9/9
p 193/193 122 X 100.3 cM

Locus	mat	het	pat	X	N	Map	SE	low	high	LOD
o Actn3	88		90							
				23	128	10.1	4.0	11.7	27.5	11.0
o Laf4	67		71							
				10	112	9.0	2.8	3.7	14.6	19.1
o Myo1b	75		90							
				14	120	10.9	2.8	5.5	16.7	18.0
o Erbb4	76		79							
				19	104	14.2	4.4	12.4	30.3	9.8
o Col6a3	67		84							
				14	97	14.0	3.9	9.0	24.0	11.8

Figure 11.9 Screens in Map Manager QTX. The dataset window (upper left), the Phenotype window (upper right), the Map window (lower left) and the Statistics window (lower right). Genotypes with permission from Mouse Genome Database (2001).

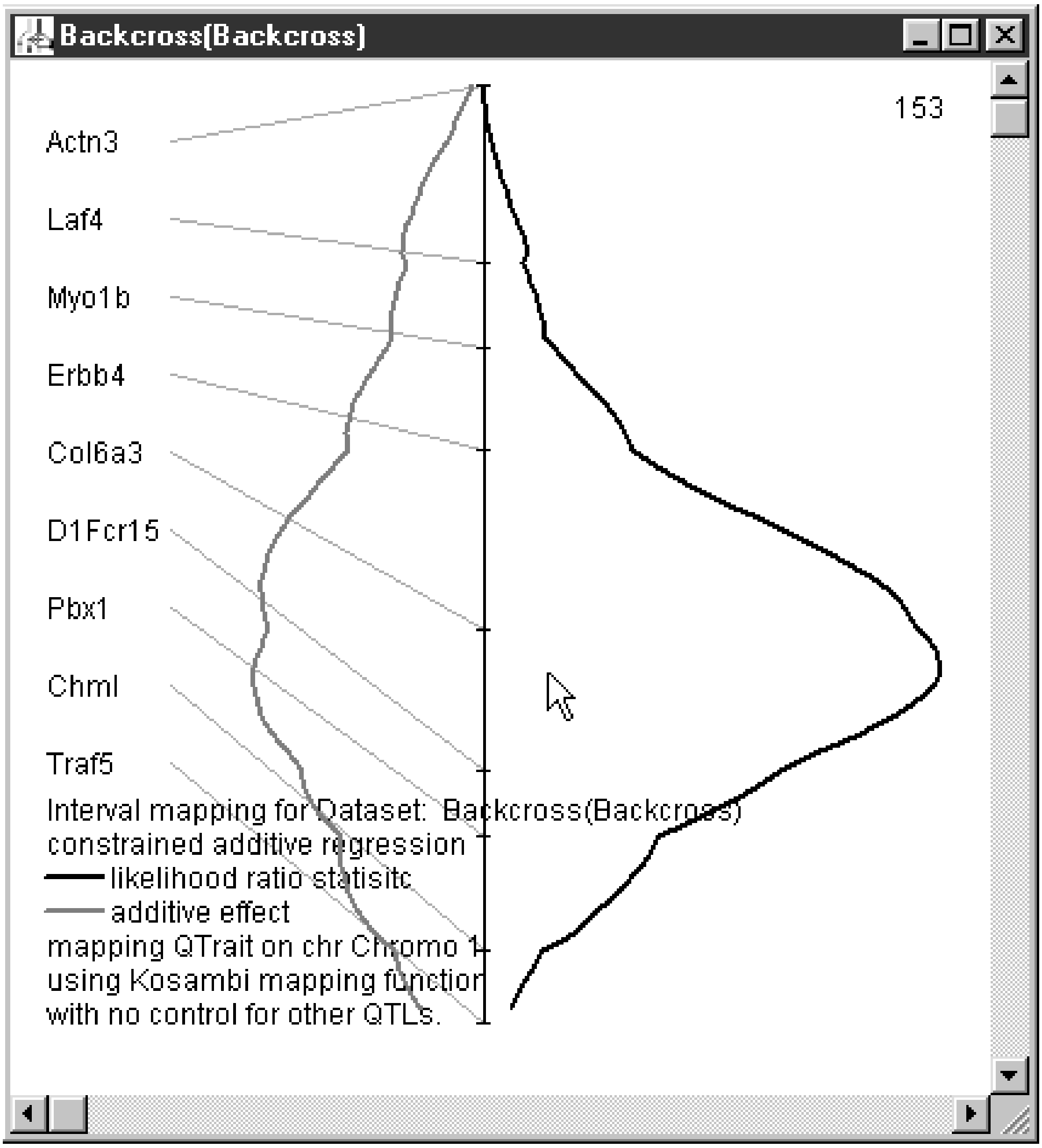


```
QT Links report for set Backcross(Backcross)
Using quantitative trait QTrait
Chi Square Stats, P = 0.05
Additive regression model
With no control for other QTLs.
182 to 193 informative progeny from Progeny Order Internal
Chr                      Locus          Stat     %        P          Add
Chromo 1      :: Laf4         13.6    6   0.00022   -18.52
Chromo 1      :: Myo1b        23.9   11   0.00000   -23.54
Chromo 1      :: Erbb4        50.6   23   0.00000   -33.49
Chromo 1      :: Col6a3       143.5   52   0.00000   -52.25
Chromo 1      :: D1Fcr15      99.8   40   0.00000   -44.62
Chromo 1      :: Pbx1         58.7   26   0.00000   -34.83
Chromo 1      :: Chml         23.7   11   0.00000   -23.28
Chromo 1      :: Traf5         6.4    3   0.01151   -12.49
```

Figure 11.10 Output from single marker association testing in Map Manager QTX: The ‘Links Report’. ‘Add’ denotes the additive regression coefficient for the association. Genotypes with permission from Mouse Genome Database (2001).

Quantitative trait data are then read in from a second text file via *File>Import>Trait Text*. The format is almost identical, except that the name of the trait replaces marker name and the trait value for each mouse must be separated from adjacent values by at least one space. Again, the name of the quantitative trait and all of the values for cross progeny must be in a single line of text.

Successful import of a text genotype file produces a small pop-up window (the *dataset* window), as shown in Figure 11.9, top left. Within it is a menu allowing selection of *Phen, Map, Stat* or *Ref.* Selecting one of these options and double-clicking on a chromosome name in the *dataset* window, produces the chosen window as shown in Figure 11.9. The *Phenotype* window (top right) displays the marker names on the left side of the window, with one column for each member of the progeny. The body of the *Phenotype* window shows the genotype at each locus and also indicates locations of recombination



Interval mapping for Dataset: Backcross(Backcross)
constrained additive regression
—— likelihood ratio statisitc
—— additive effect
mapping QTrait on chr Chromo 1
using Kosambi mapping function
with no control for other QTLs.

Figure 11.11 Output from Map Manager QTX. Results of interval mapping across nine markers. Genotypes with permission from Mouse Genome Database (2001).

events with an X. Pairs of question marks denote the possible locations of crossovers whose more precise location cannot be specified due to missing genotype data. The *Map* window (bottom left), shows a genetic map with estimated cM distance between markers, and the *Statistics* window (bottom right) summarizes useful numerical information, such as the number of recombination events between adjacent markers and LOD evidence for linkage.

11.6.1.2 Single marker association

Testing for association between an individual marker and a quantitative trait is accomplished by first selecting a p-value cut-off in the *Main* menu under *Options>Search&Linkage criteria*, and then choosing *QT>Links Report* in the *Main* menu. This produces a window allowing the user to select both the name of the quantitative trait to test and the background QTLs to be included in the analysis.

Figure 11.10 shows the table or *Links Report* that was produced by testing each of the nine markers in our panel for association with the simulated trait. Note that only eight markers appear in the table, as one marker did not meet the $p < 0.05$ criterion. Note also that marker *Col6a3* is highlighted as giving the strongest association and therefore as being the best marker to include as a background QTL in analyses of other chromosomal loci.

11.6.1.3 Simple Interval Mapping

Simple interval mapping of a QT across a series of markers is accomplished by choosing *QT>Interval Mapping* from the *Main* menu. This produces a window which again allows the user to specify the trait to be analysed and whether any background QTLs are to be included in the analysis. Once options in this window are specified, Map Manager QTX produces a table and a figure displaying the Interval Mapping results. Figure 11.11 shows the result of interval mapping our simulated trait across the nine markers on mouse chromosome 1. As indicated by the position of the cursor, the peak of the likelihood ratio statistic falls very close to the true location of the simulated QT locus, between markers *Col6a3* and *D1Fcr15*.

ACKNOWLEDGEMENTS

The authors wish to thank Heather Cordell, Dmitri Zaykin, Clive Bowman, Meg Ehm and Leonid Kruglyak for helpful discussions, advice and support during the writing of this chapter.

REFERENCES

Abecasis GR, Cardon LR, Cookson WO. (2000). A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* **66**: 279–292.

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.* **30**: 97–101.

Abecasis GR, Cookson WOC. (2000). GOLD—Graphical Overview of Linkage Disequilibrium. *Bioinformatics* **16**: 182–183.

Allison DB. (1997). Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet* **60**: 676–690.

Amos CI. (1994). Robust variance components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* **54**: 535–543.

Basten C, Weir BS, Zeng Z-B. (1994). Zmap — a QTL cartographer. In Smith C, Gavora JS, Benkel B, Chesnais J, Fairfull W, Gibson JP, Kennedy BW, Burnside EB. *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production: Computing Strategies and Software* vol. 22. pp. 65–66. (On-line publication)

Basten C, Weir BS, Zeng Z-B. (1997). *QTL Cartographer: A Reference Manual and Tutorial for QTL Mapping*. Department of Statistics, North Carolina State University: Raleight, NC. (http://statgen.ncsu.edu/qtlcart/).

Blangero J, Almasy L. (1996). *SOLAR: Sequential Oligogenic Linkage Analysis Routines*. Technical notes no. 6, Population Genetics Laboratory, Southwest Foundation for Biomedical Research: San Antonio, TX.

Chapman CJ. (1990). A visual interface to computer programs for linkage analysis. *Am J Med Genet* **36**: 155–160.

Churchill GA, Doerge RW. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–971.

Clark AG. (1990). Inference of haplotypes from PCR amplified samples of diploid populations. *Mol Biol Evol* **7**: 111–122.

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, *et al.* (1998). Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* **63**: 595–612.

Clayton D. (1999). A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* **65**: 1170–1177.

Clayton D. (2001). Population association. In Balding DJ, Bishop M, Cannings C. (Eds), *Handbook of Statistical Genetics*. John Wiley: Chichester, pp. 519–540.

Clayton D, Jones HB. (1999). Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet* **65**: 1161–1169.

Collins A, Morton NE. (1998). Mapping a disease locus by allelic association. *Proc Natl Acad Sci USA* **95**: 1741–1745.

Conneally PM, Edwards JH, Kidd KK, Lalouel J-M, Morton NE, Ott J, *et al.* (1985). Report of the committee on methods of linkage analysis and reporting. *Cytogenet Cell Genet* **40**: 356–359.

Cordell HJ. (2002). Diabetes in the NOD mouse. In Camp N, Cox A. (Eds), *Quantitative Trait Loci: Methods and Protocols*. Humana Press: pp. 165–198.

Cottingham RW Jr, Idury RM, Schaffer AA. (1993). Fast sequential genetic linkage computation. *Am J Hum Genet* **53**: 252–263.

Cudworth AG, Woodrow JC. (1975). Evidence for HLA-linked genes in 'juvenile' diabetes mellitus. *Br Med J* **3**: 133–135.

Dempster AP, Laird NM, Rubin DB. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc* **B39**: 1–38.

Devlin B, Risch N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**: 311–322.

Devlin B, Risch N, Roeder K. (1996). Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics* **36**: 1–16.

Excoffier L, Slatkin M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12**: 921–927.

Farmer MA, Sundberg JP, Bristol IJ, Churchill GA, Li R, Elson CO, *et al.* (2001). A major quantitative trait locus on chromosome 3 controls colitis severity in IL-10-deficient mice. *Proc Natl Acad Sci USA* **98**: 13820–13825.

Fulker DW, Cardon LR. (1994). A sib-pair approach to interval mapping of quantitative trait loci. *Am J Hum Genet* **54**: 1092–1103.

Fulker DW, Cherny SS, Sham PC, Hewitt JK. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet* **64**: 259–267.

Goldgar DE. (1990). Multipoint analysis of human quantitative genetic variation. *Am J Hum Genet* **47**: 957–967.

Haley CS, Knott SA. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *J Hered* **69**: 315–324.

Haseman JK, Elston RC. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* **2**: 3–19.

Hastabacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E. (1992). Linkage disequilibrium mapping in isolated founder populations: Diastrophic dysplasia in Finland. *Nature Genet* **2**: 204–211.

Hauser ER, Boehnke M. (1997). Confirmation of linkage results in affected-sib-pair linkage analysis for complex genetic traits. *Am J Hum Genet* **61**: A278.

Hawley ME, Kidd KK. (1995). HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered* **86**: 409–411.

Heath S. (1997). Markov chain segregation and linkage analysis for oligogenic models. *Am J Hum Genet* **61**: 748–760.

Hedrick PW. (1987). Gametic disequilibrium measures: Proceed with caution. *Genetics* **117**: 331–341.

Hill WG, Weir BS. (1994). Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am J Hum Genet* **54**: 705–714.

Hinds D, Risch N. (1996). The ASPEX package: affected sib-pair mapping ftp://lahmed.stanford.edu/pub/aspex.

Holmans P. (2001). Nonparametric linkage. In Balding DJ, Bishop M, Cannings C. (Eds), *Handbook of Statistical Genetics*. John Wiley: Chichester, pp. 487–505.

Holmans P, Clayton D. (1995). Efficiency of typing unaffected relatives in an affected sib-pair linkage study with single locus and multiple tightly-linked markers. *Am J Hum Genet* **57**: 1221–1232.

Jansen RC. (1992). A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics* **85**: 252–260.

Jansen RC. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**: 205–211.

Jansen RC. (1996). A general Monte Carlo method for mapping multiple quantitative trait loci. *Genetics* **142**: 305–311.

Jansen RC, Stam P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**: 1447–1455.

Jensen CS, Kong A, Kjaerulff KM. (1995). Blocking–Gibbs sampling in very large probabilistic expert systems. *Int J Hum Computer Studies* 647–666.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, *et al.* (2001). Haplotype tagging for the identification of common disease genes. *Nature Genet* **29**: 233–237.

Kao CH, Zeng Z-B, Teasdale RD. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**: 1203–1216.

Kaplan N, Hill WG, Weir BS. (1995). Likelihood methods for locating disease genes in nonequilibrium populations. *Am J Hum Genet* **56**: 18–32.

Kong A, Cox NJ. (1997). Allele sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet* **61**: 1179–1188.

Kruglyak L, Lander ES. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* **57**: 439–454.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* **58**: 1347–1363.

Lander ES, Botstein D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185–199.

Lander ES, Green P, Abrahamson J, Barlow A, Daly M, Lincoln SE, *et al.* (1987). MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**: 174–181.

Lathrop GM, Lalouel JM, Julier C, Ott J. (1984). Strategies for multilocus linkage analysis in humans. *Proc Natl Acad Sci USA* **81**: 3443–3446.

Lehesjoki A-E, Koskiniemi M, Norio R, Tirrito S, Sistonen P, Lander E, *et al.* (1993). Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: Linkage disequilibrium allows high resolution mapping. *Hum Mol Genet* **2**: 1229–1234.

Lewis PO, Zaykin D. (2001). Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c). Free program distributed by the authors over the internet from http://lewis.eeb.uconn.edu/lewishome/software.html.

Lewontin RC. (1964). The interaction of selection and linkage I. General considerations; heterotic models. *Genetics* **49**: 49–67.

Little RJA, Rubin DB. (1987). *Statistical Analysis with Missing Data*. Wiley: New York.

Long JC, Williams RC, Urbanek M. (1995). An E-M algorithm and testing strategy for multiple locus haplotypes. *Am J Hum Genet* **56**: 799–810.

Manly KF, Olson JM. (1999). Overview of QTL mapping software and introduction to map manager QT. *Mamm Genome* **10**: 327–334.

Manly KF, Cudmore RH, Meer JM. (2001). Map Manager QTX, cross-platform software for genetic mapping. *Mamm Genome* **12**: 930–932.

Martinez O, Curnow RN. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor Appl Genet* **85**: 480–488.

Monks SA, Kaplan NL, Weir BS. (1998). A comparative study of sibship tests of linkage and/or association. *Am J Hum Genet* **63**: 1507–1516.

Morton NE. (1955). Sequential tests for the detection of linkage. *Am J Hum Genet* **7**: 277–318.

Mott R, Talbot C, Turri M, Collins AC, Flint J. (2000). A new method for fine mapping quantitative trait loci in outbred animal stocks. *Proc Natl Acad Sci USA* **97**: 12649–12654.

Mouse Genome Database (MGD). (2001). Mouse Genome Informatics Web Site, The Jackson Laboratory, Bar Harbor, Maine. World Wide Web (URL: http://www.informatics.jax.org/).

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, *et al.* (1998). DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet* **19**: 233–240.

O'Connell JR, Weeks DE. (1995). The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genet* **11**: 402–408.

Olson JM. (1995). Multipoint linkage analysis using sib pairs: an interval mapping approach for dichotomous outcomes. *Am J Hum Genet* **56**: 788–798.

Rabinowitz D. (1997). A transmission disequilibrium test for quantitative trait loci. *Hum Hered* **47**: 342–350.

Risch N. (1990a). Linkage strategies for genetically complex traits II. The power of affected relative pairs. *Am J Hum Genet* **46**: 229–241.

Risch N. (1990b). Linkage strategies for genetically complex traits: III. The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet* **46**: 242–253.

SAGE (1999). *Statistical Analysis for Genetic Epidemiology*, Release 4.0. Department of Epidemiology and Biostatistics, Rammelkamp Center for Education and Research, MetroHealth campus, Case Western Reserve University: Cleveland, OH.

Satagopan JM, Yandell BS, Newton MA, Osborn TC. (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144**: 805–816.

Schaid DJ. (1996). General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* **13**: 423–449.

Schneider S, Roessli D, Excoffier L. (2000). *Arlequin ver. 2.000: A software for population genetics data analysis*. Genetics and Biometry Laboratory, University of Geneva: Geneva, Switzerland.

Sham PC. (1998). *Statistics in Human Genetics*. Arnold Publishers: London; John Wiley and Sons Inc.: New York.

Sham PC, Curtis D. (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet* **59**: 323–336.

Sillanpaa MJ (1998). Multimapper Reference Manual. http://www.RNL.Helsinki.F1/□mjs/.

Sillanpaa MJ, Arjas E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**: 1373–1388.

Slatkin M, Excoffier L. (1996). Testing for linkage disequilibrium in genotypic data using the EM algorithm. *Heredity* **76**: 377–383.

Sobel E, Lange K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *Am J Hum Genet* **58**: 1323–1337.

Spielman RS, Ewens WJ. (1996). The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* **59**: 983–989.

Spielman RS, Ewens WJ. (1998). A sibship test for linkage in the presence of association: the sib-transmission/disequilibrium test. *Am J Hum Genet* **62**: 450–458.

Spielman RS, McGinnis RE, Ewens WJ. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. *Am J Hum Genet* **52**: 506–516.

Talbot CJ, Nicod A, Cherny SS, Fulker DW, Collins AC, Flint J. (1999). High-resolution mapping of quantitative trait loci in outbred mice. *Nature Genet* **21**: 305–308.

Terwilliger JD. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am J Hum Genet* **56**: 777–787.

Terwilliger JD. (1996). Program SIBPAIR — sib pair analysis on nuclear families. ftp://linkage.cpmc.columbia.edu.

Terwilliger JD, Ott J. (1994). *Handbook of Human Genetic Linkage*. Johns Hopkins: Baltimore.

Uimari P, Thaller G, Hoeschele I. (1996). The use of multiple markers in a Bayesian method for mapping quantitative trait loci. *Genetics* **143**: 1831–1842.

Utz HF, Melchinger AE. (1996). PLABQTL: a program for composite interval mapping of QTL. *J Quant Trait Loci* **2**, http://probe.nalusda.gove:8000/otherdocs/jqtl/.

van Ooijen JW, Maliepaard C. (1996a). MapQTL version 3.0: software for the calculation of QTL positions on genetic maps. Plant Genome IV abstracts. http://probe.nalusda.gov:3000/otherdocs/pg/pg4/abstracts/p316.html.

van Ooijen JW, Maliepaard C. (1996b). MapQTL version 3.0: software for the calculation of QTL positions on genetic maps. CPRO-DLLO: Wageningen, ISBN90-73771-23-4.

Weeks DE, Sobel E, O'Connell JR, Lange K. (1995). Computer programs for multilocus haplotyping of general pedigrees. *Am J Hum Genet* **56**: 1506–1507.

Weir BS. (1996). *Genetic Data Analysis II*. Sinauer Associates Inc Publishers: Sunderland, MA, USA.

Xie X, Ott J. (1993). Testing linkage disequilibrium between a disease gene and marker loci. *Am J Hum Genet* **53**: 1107.

Zeng Z-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc Natl Acad Sci USA* **90**: 10972–10976.

Zeng Z-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**: 1457–1468.

Zeng Z-B, Kao CH, Basten CJ. (1999). Estimating the genetic architecture of quantitative traits. *Genet Res* **74**: 279–289.

Zhao JH, Curtis D, Sham PC. (2000). Model-free analysis and permutation tests for allelic associations. *Hum Hered* **50**: 133–139.