# 1

# Linkage Disequilibrium and Association Mapping

*An Introduction*

**Andrew R. Collins**

**Summary**

The basis for recent developments on the characterization of the linkage-disequilibrium structure of the genome and the application of association mapping to genes for common human diseases is described. Patterns of linkage disequilibrium are now understood, for a number of human populations, in unprecedented detail. This information not only provides a vital resource for the design and execution of powerful association-mapping studies, but opens new avenues of research into the genetic history of human populations and the effects of natural selection, mutation, and recombination on the genomic landscape.

**Key Words:** Recombination; linkage mapping; linkage disequilibrium; haplotype blocks; selection; case–control study; candidate regions; genome-wide association.

## 1. Introduction

### *1.1. Linkage and Linkage Disequilibrium*

#### *1.1.1. Recombination and Linkage*

The human genome comprises 23 chromosomes of which 22 are autosomes and 1 is a sex chromosome (either X or Y). Sperm or egg cells contain one copy of the genome (haploid) and the fertilized egg contains two copies (diploid) with a copy derived from each parent. Thus, a diploid cell contains 22 homologous pairs of autosomal chromosomes and a pair of sex chromosomes (XY in males and XX in females). The production of haploid sperm or egg cells (gametes) requires the process of meiosis, which takes place in the testes or ovaries. This reduces the genome from diploid to haploid through two cycles of cell division (meiosis I and II). The divisions are preceded by chromosome duplication after which each homolog comprises two sister chromatids that are

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

joined at the chromosome centromere. The maternally and paternally derived homologs, each with two chromatid strands, align to form a bivalent, which is a four strand structure. The bivalent later undergoes two rounds of meiotic division to yield four gametes. However, the most important process for our purposes takes place in the bivalent when pairs of homologous chromatids (derived from paternal and maternal chromosomes) connect to form chiasmata (the plural of chiasma). The process of meiotic recombination takes place in these locations and exchanges genetic material between the two chromosomes. The arms of each chromosome usually have at least one chiasma and certain chromosome regions typically show fewer (centromeric regions) and more (subtelomeric regions) chiasmata, respectively. There is also a sex difference so that, when meiosis is taking place in females, there are more chiasmata overall and, therefore, a greater amount of genetic recombination than in male meiosis. There is also a somewhat different distribution of chiasmata between the sexes. The consequence of this process is that gametes receive a recombined chromosome from both maternally and paternally derived homologs. An understanding of the distribution and frequency of chiasmata, as reflected in a genetic-linkage map, is critical for genetic mapping. The genome is peppered with polymorphisms, including short DNA sequences and single base variants, and different individuals may have different forms (alleles) at these locations. A tiny proportion of these polymorphisms are the disease genes that are being so actively searched for. Where they are informative, polymorphisms are used as markers to determine the locations of meiotic chromosome breaks (**Fig. 1**). The key to exploiting the information provided by marker polymorphisms along a chromosome is provided by linkage: markers are linked if they do not recombine freely. Recombination can take place between any two markers along the chromosome but those that are far apart will have a higher frequency of recombination than those that are close together. Recombination frequencies can be converted into a genetic distance (expressed in morgans or centimorgans). One centimorgan corresponds to 1% recombination. The computation of map distances enables the construction of whole-chromosome and genome-wide genetic-linkage maps. To convert recombination fractions into genetic distances a mapping function is required. The simplest assumes that, for three ordered polymorphisms A,B,C, recombination fractions are additive: A–C= A–B + B–C. However, this supposes there is only one crossover between adjacent loci, which is unlikely to be true for markers separated by a large distance on the chromosome. A further complication is chiasma interference, which reduces the probability of a second chiasma forming in proximity to one that is already in place. The Kosambi mapping function is most frequently used in genetic-mapping programs but probably underestimates the strength of chiasma interference generally ***(1)***. Linkage maps comprise ordered markers and the genetic (centimorgan)

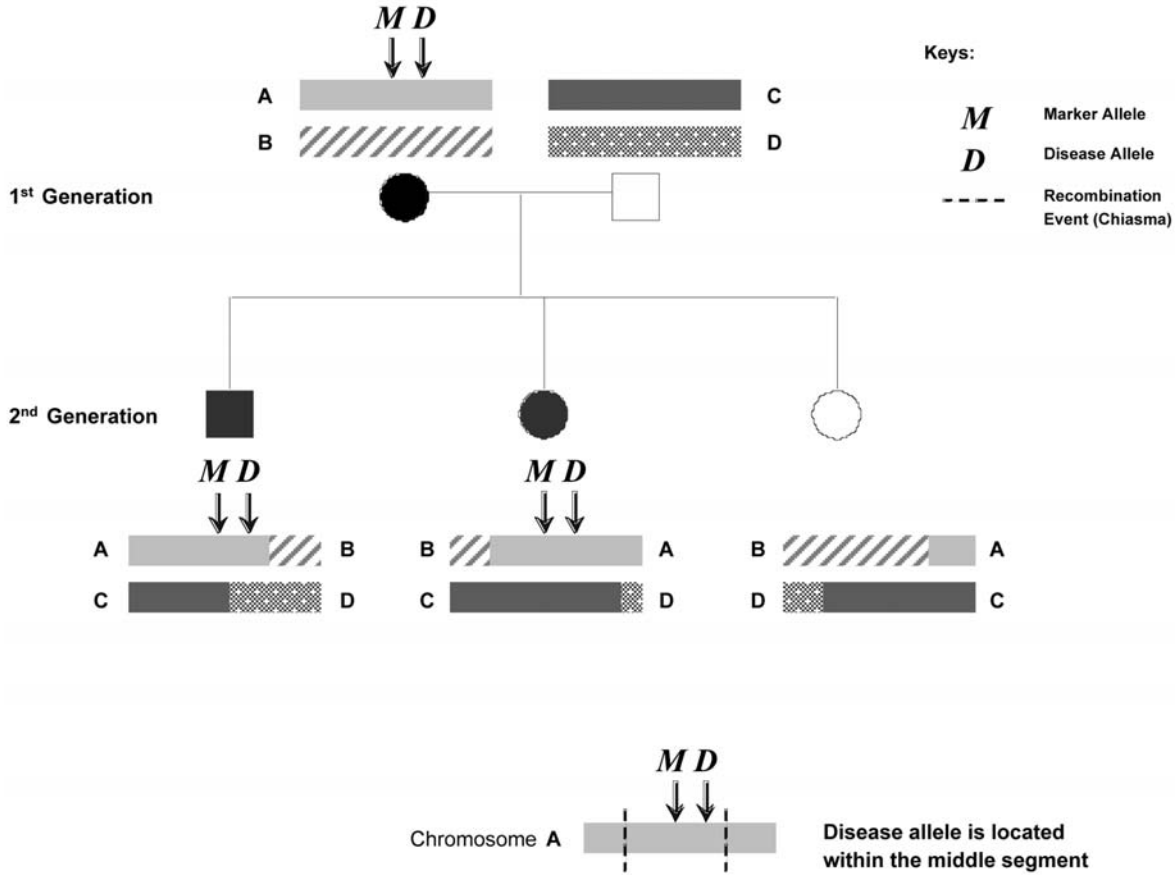



Fig. 1. Linkage mapping: evidence for the location of a disease-causing allele (D) can be obtained by coinheritance with a marker allele (M) in a pedigree. Transmitted recombined haplotypes help to reduce the target region to search for the disease-causing gene.

distances between them. Because of the nonuniformity of recombination breaks, the linkage map shows "hot" and "cold" regions where the recombination rate is locally high and low, respectively.

The identification of many disease genes has been achieved through linkage mapping. Linkage mapping relies on the coinheritance of marker polymorphisms and disease phenotype in pedigrees. **Figure 1** shows how recombination reduces the size of a target chromosome segment containing a putative disease allele (D). Exploiting the process of meiotic breakage and consequent reduction of the size of DNA segment to be searched for disease causing variant(s) is the basis of both linkage and also linkage disequilibrium (LD) mapping. Linkage mapping requires family material so the coinheritance of markers and disease can be tracked. However, relatively few meiotic breaks occur over the small number of generations available in most pedigrees. This limits the extent to which a disease can be fine mapped. At best linkage mapping can be expected to map a gene to a region no smaller than about one megabase. However, this has been more than adequate to enable the positional cloning of a great many "major" disease genes that are relatively rare but have a large phenotypic effect.

### *1.2. Linkage Disequilibrium*

LD is the association between alleles at two linked loci that reflects, in part, their proximity and the correspondingly low probability of recombination breaking the haplotype on which they are found. The strength of allelic association depends, for a population, on the number of founding individuals (and therefore the number of founding haplotypes), the time since founding (and therefore the number of generations over which recombination has driven the decay of LD), along with a number of other factors such as mutation, drift, and selection. Some of these issues are discussed in detail in Chapters 2, 3, and 5. Association between marker polymorphisms and disease phenotype can be exploited as a powerful tool for disease mapping and is the focus of enormous effort worldwide to map the genes involved in common human diseases, such as asthma, heart disease, and diabetes. **Figure 2** classifies disease-influencing genes in terms of their phenotypic effect. Many major genes, which have low population frequencies but severe phenotypic consequences, for example genes for Huntington's chorea ***(2)*** and cystic fibrosis ***(3)***, have been mapped in families segregating the disease by linkage. Rare mutations in single genes are involved in this class of disease-causing variants and their mode of inheritance is typically Mendelian with high penetrance, facilitating mapping. Oligogenes contribute to the majority of common human disease and are possibly relatively frequent, but phenotypic consequences depend on a complex array of genetic and environmental influences. This class of genes is the focus of efforts to develop and

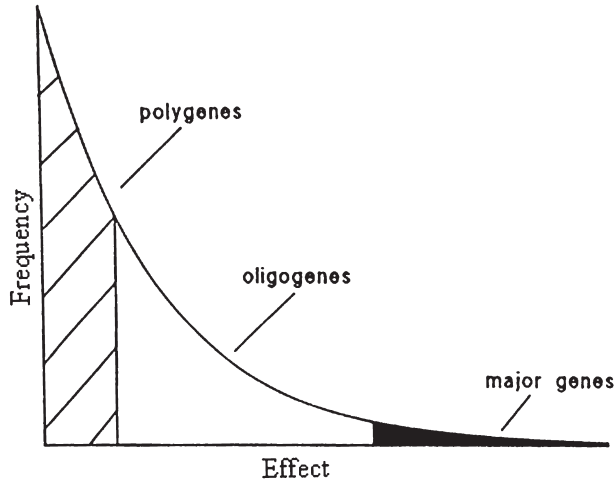



Fig. 2. Disease gene classes by phenotypic effect and frequency. Major genes are rare but have a severe phenotypic effect, oligogenes that contribute to common human disease may be more frequent but have variable phenotypic effects influenced by complex genetic and environmental factors.

apply association mapping but the complex pattern of inheritance presents many challenges. However, the exact nature of the genetic basis of complex traits is a topic of much discussion (*see* Chapter 6). Polygenes are most numerous and have the smallest individual phenotypic effects. These are essentially undetectable by current methods but also of less immediate interest because polygenic forms of disease will be far less tractable to therapeutic intervention.

**Figure 3** illustrates the rationale behind the LD-mapping approach. A population originating from a relatively small number of founding individuals, and their corresponding set of haplotypes, is exposed to recombination over many (*n*) generations to create the larger set of recombined haplotypes in a present-day population. An allele (M) at a particular marker polymorphism shows a statistical association with a disease-influencing allele (D) because of the proximity of marker and disease loci and the relatively infrequent recombination between them. Because the location of the disease allele is, in reality, not known its location can be inferred by measuring the association between disease phenotype and marker allele. Although both linkage mapping and LD (association) mapping exploit the breaks in haplotypes created by recombination to determine the genomic location for a disease gene, they differ in that association mapping exploits the recombination over a great many generations and does not require the collection of pedigrees. Association mapping, therefore, offers substantial advantages including much finer scale mapping (down to tens of kilobases or less) by

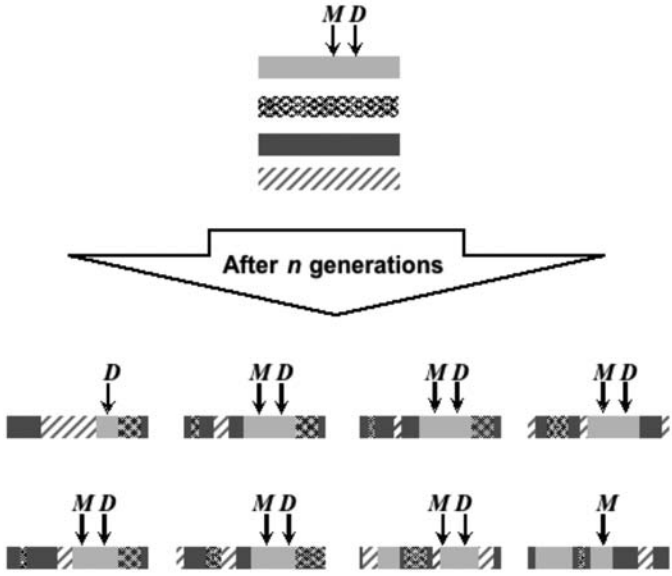



Fig. 3. Association mapping: a mutation disease-causing allele (D) in a founding haplotype is coinherited with a marker allele (M) over many generations. D can be mapped by exploiting the residual association between M and the disease phenotype because of the linkage disequilibrium between the disease and marker alleles.

exploiting the large number of historical meiotic breaks and the use of more easily collected, genotyped, and analyzed samples through, for example, case–control studies. Furthermore, unlike single-gene disorders, the genes involved in common diseases do not show simple Mendelian patterns of inheritance in pedigrees but reflect the combined effects of multiple genes and incomplete penetrance, which further limits the utility of pedigree-based samples in this context.

Most of the current focus of research involves the use of panels of single-nucleotide polymorphisms (SNPs) as markers for association mapping. These are the most common type of polymorphism in the genome and are typically biallelic with the two alleles most commonly A or G (substitution of purines) or T or C (substitution of pyrimidines). Other types of polymorphism, including microsatellites (tracts of short repeated DNA sequences) have been invaluable for linkage mapping but are less useful for association mapping because they are far less abundant, more difficult to analyze because of multiple alleles, and are less amenable to automated "array" genotyping approaches (*see* Chapter 7).

### *1.3. Pairwise LD Metrics*

The computation of the degree of association between pairs of SNP markers from observed or estimated haplotype frequencies underpins the construction

**Table 1**
**Haplotype Frequencies in Single Nucleotide Polymorphisms**

| | | $SNP_2$ alleles | | |
|---|---|---|---|---|
| | | **$B_{(1)}$** | **$b_{(2)}$** | Total |
| **$SNP_1$ alleles** | **A** | **AB** | **Ab** | **Q** = $(n_{11} + n_{12}) / N$ |
| | (1) | $n_{11}$<br>$\pi_{11} = QR + D$ | $n_{12}$<br>$\pi_{12} = Q(1-R) - D$ | |
| | **a** | **aB** | **ab** | **1 - Q** = $(n_{21} + n_{22}) / N$ |
| | (2) | $n_{21}$<br>$\pi_{21} = R(1-Q) - D$ | $n_{22}$<br>$\pi_{22} = (1-R)(1-Q) + D$ | |
| **Total** | | **R**<br>= $(n_{11} + n_{21}) / N$ | **1 - R**<br>= $(n_{12} + n_{22}) / N$ | **N** = $n_{11} + n_{21} + n_{12} + n_{22}$ |

of LD maps using the LDMAP program (*see* Chapters 3 and 4). **Table 1** gives haplotype frequencies for a pair of SNP markers. The SNPs have alleles A/a and B/b, respectively, from which the four possible haplotypes are designated AB, Ab, aB, and ab have population counts $n_{11}$, $n_{12}$, $n_{21}$, and $n_{22}$, respectively. Alleles A, a, B, and b have, respectively, allele frequencies Q, 1–Q, R, and 1–R. At equilibrium (whereby there is no allelic association and therefore disequilibrium D = 0) the haplotype frequencies ($\pi_{11}$, $\pi_{12}$, $\pi_{21}$, $\pi_{22}$) are simply products of the appropriate allele frequencies. However, in the presence of LD, the distortion in the haplotype-frequency distribution is reflected in the covariance D. The various standard metrics for pairwise association ***(4)*** are given in **Table 2**. For pairs of SNPs (but not for quantifying the association with disease), the ρ metric (association) equates to the absolute value of the D' metric ***(5)***. This has been shown ***(6)*** to have the greatest efficiency for modeling the exponential decline of association with distance and is used in LDMAP.

The degree of association between marker and disease (phenotype) is quantified in a somewhat different way. For major genes it is possible to determine the "case" haplotypes that carry a disease-causing variant and "control" haplotypes that do not. Collins and Morton ***(7)*** describe the computation of the pairwise association between markers and disease in these circumstances while accounting for the enrichment of the case sample. For complex traits, however, disease haplotypes cannot be assigned and the disease-allele frequency cannot be computed. Maniatis et al. ***(8,9)*** and Chapter 8, have developed an alternative metric (Z) in these circumstances.

**Table 2**
**Alternative Measures of Pairwise Association**

| Definition | Symbol | Estimate $\hat{\psi} = D/C$ |
|---|---|---|
| Covariance | D | $D=\left\|\pi_{11}\pi_{22}-\pi_{12}\pi_{21}\right\|$ |
| Association | ρ | $D/Q(1–R)$ |
| Correlation | r | $D/\sqrt{Q(1-Q)R(1-R)}$ |
| Regression | b | $D/R(1–R)$ |
| Frequency difference | f | $D/Q(1–Q)$ |
| Delta | δ | $D/Q(1–R–Q+RQ+D)$ |
| Yule | y | $D/[2Q(1–Q)R(1–R)+D(1–2Q)(1–2R)+2D^2]$ |

## 2. The LD Structure of the Human Genome and the Relationship to Recombination Patterns

Toward the end of the 1990s it was recognized that the LD structure of the human genome had to be characterized before association-mapping studies could be efficiently devised. The understanding of the LD structure has increased remarkably since the somewhat alarming findings of Kruglyak ***(10)*** that implied, through coalescent simulation neglecting population bottlenecks, that LD was unlikely to be extensive enough to exploit association in a cost-effective way to predict the location of disease-influencing variants. Subsequently many papers have shown that in reality, even in outbred heterogeneous populations, LD is rather extensive (on average ~50 kb). In principal it is only necessary to genotype a proportion of the SNPs across a region or chromosome to recover the majority of the information about association as there are many "redundant" SNPs in complete or very strong LD with each other. The discovery that a substantial proportion of the genome comprises "blocks" of low haplotype diversity spawned the International HapMap Project (http://www.hapmap.org/) ***(11)*** that has determined a reduced set of SNPs that "tag" haplotypes and are intended to recover the majority of the information encoded in the 15 million or so SNPs in the genome ***(12)***. The effectiveness of tagging has been questioned however (*see* for example **refs.** ***13*** and ***14***). The block regions are, in effect, recombinationally suppressed segments punctuated by recombination hot spots, as demonstrated by Jeffreys et al. ***(15)***. The ability to construct maps that are analogous to the linkage map from LD data has shown the overwhelmingly dominant role of recombination in determining LD structure. **Figure 4** illustrates the variations in the extent of LD across chromosome 22. Regions with extensive LD (the flatter regions of the LD map curve) coincide strongly with recombination-suppressed regions and, conversely, regions with low

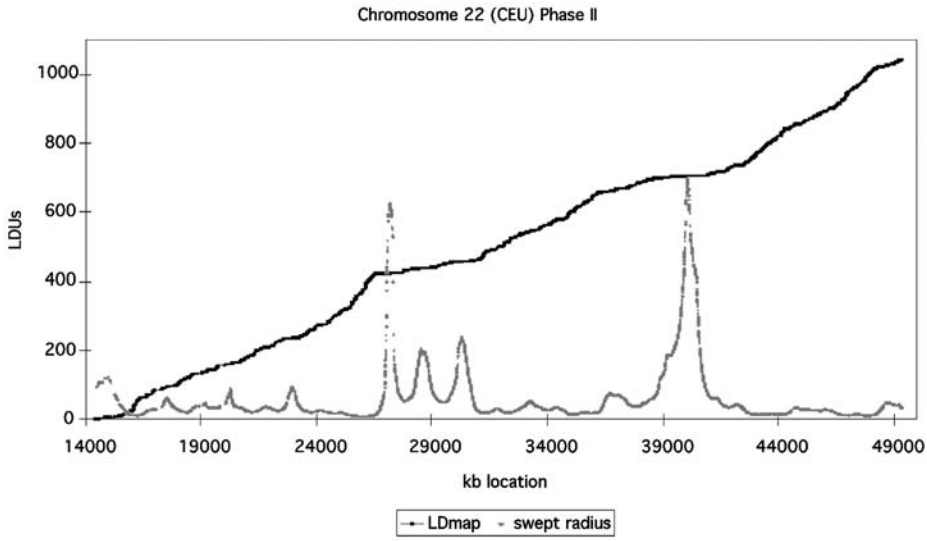



Fig. 4. A linkage-disequilibrium (LD) map of chromosome 22 from the CEPH (CEU) HapMap population. The LD map (upper curve) shows broad, more vertical regions that reflect high-recombination intensity and flatter "block"-like regions of strong LD. The "swept radius" shows variation in the average extent of LD and regions with high swept radii will require a lower marker density for effective screening.

LD (the steeper part of the LD map curve) coincide with recombination-intense regions. Because recombination hot spots are colocalized in all human populations and, seemingly, hot-spot locations are conserved over the time-scales represented by human populations, the concept of a "cosmopolitan" LD map has been advanced ***(16)***, which has a number of advantages including applicability to all human populations after appropriate linear scaling.

Isolated populations have more extensive LD because of smaller founding population sizes and recent founding (and therefore reduced time for recombination to break up haplotypes). Service et al. ***(17)*** examined 11 isolated populations and found that, as well as requiring approx 30% fewer markers for adequate coverage, these populations have far fewer "holes": very narrow regions that appear to coincide with particularly intense recombination hot spots within and around which there is very little LD. Any disease-influencing variants in the vicinity of holes will be extremely difficult to map using LD. The much higher numbers of holes in outbred populations reflects the large number of generations over which recombination events have accumulated in these recombinogenic regions. The reduced duration for many isolates means the intensity of the break up of LD is reduced and the fewer holes presents further justification for gene-mapping studies in isolated populations.

### 2.1. Selection

LD maps *(1)* are useful for the identification of genomic regions that have undergone differential selection (between populations) and related phenomena. Genomic regions where an inversion has attained a high frequency will show a distorted pattern of recombination, which can be mistaken for selection if the DNA sequence is not determined. New mutations that are beneficial increase in frequency in a population through natural selection. This reduces the haplotype diversity in the region around the beneficial allele, thereby increasing the amount of LD *(18)*. Local reduction in variability is termed a selective sweep *(19)*. There has been much theoretical and some empirical work in this area in recent years. The high-density genotyping in the data generated by the HapMap Project enables screening of the whole genome. The identification of selected regions is of interest to both evolutionary studies and for identification of genes involved in disease. The utility of many approaches for detecting selective sweeps is limited by simplifying demographic assumptions about the population (population at equilibrium and of constant size, no population subdivision or gene flow), problems of ascertainment bias in the sample *(20)*, and also neglect of the underlying recombination-dominated LD structure. However, recent genome-wide, high-density SNP genotyping permits further progress and a review of the fascinating topic is presented in Chapter 5.

## 3. Association-Mapping Strategies

### 3.1. Family-Based or Case–Control

Successful characterization of the underlying LD structure underpins efforts to apply association mapping of disease genes. However, the design of a particular study is also critical and a fundamental aspect is the nature of the DNA resources to be sampled. Associations with a quantitative trait might be successfully examined in random samples of individuals. However, associations with disease are more usually studied in sample-based case–control studies (Chapter 8) or using family-based material (Chapter 10). In case–control studies, sampling can be effectively targeted to maximize power. Risch and Zhang *(21)* described the use of extreme discordant sib pairs for quantitative trait (linkage) mapping. This study examined the increased power achieved by contrasting, for a particular trait, the most severely affected individuals with "hyper-normal" individuals. Such a strategy is easier for association mapping where unrelated individuals are being analyzed in case–control sampling. Other examples of selective case–control samples highlighted by Morton and Collins *(22)* include contrasting early onset cases with normal, elderly controls, and affected individuals with normal individuals who have had intense environmental exposure to factors known to be associated with the trait. The power benefits of using

familial cases and a set of unrelated controls are described in Chapter 11 as a strategy to identify low-penetrance disease alleles. The sampling designs that can be devised to maximize power using case–control samples should be contrasted with the analysis of families that cannot achieve equivalent samples except through selection from a much larger sample. Family-based procedures for association mapping include the transmission-disequilibrium test ***(23)***. The test was developed in recognition of the concern that association between a disease and a marker can arise as an artifact of population structure. The transmission-disequilibrium test formally tests for linkage between a disease and marker locus showing population association and is not affected by biases from population stratification. It considers parents who are heterozygous for an allele associated with disease and determines the frequency with which that allele is transmitted to an affected offspring. Hence, the test is typically conducted in trios, although there are many variants and extensions. Morton and Collins ***(22)*** showed that the test is only one-sixth as efficient as case-hypernormal control pairs and argued that only in exceptional cases does avoidance of stratification problems justify the family-based design for association mapping. However, developments such as pseudomarker analysis and the family-based association test (Chapter 10), have extended these earlier methods to handle general pedigrees and are particularly useful for the analysis of family material (much of which is derived from earlier linkage studies) and for samples known to have potential population stratification issues.

### *3.2. Genotyping Strategies*

For a genome-wide association study the genotyping of, potentially, hundreds of thousands of SNPs poses obvious difficulties. The first choice is the selection of SNPs with the aim of achieving "complete" coverage of the genome with budgetary considerations in mind. Uniform spacing of SNPs on the linkage-disequilibrium unit scale is optimal (*see* Chapters 3, 8, and 13) and it has been suggested that SNPs spanning a range of allele frequencies within each LD unit will maximize coverage ***(24)***. However, on current panels, which includes gene chips (arrays of up to 500,000 SNPs for genome scanning), SNPs are not optimally spaced but their locations are determined by the location of small restriction fragments. This issue and alternative panels are described in detail in Chapter 7. DNA-pooling experiments have well-developed theory (Chapter 12) and strategies that combine pooling, and chips hold promise to reduce genotyping costs.

### *3.3. Analysis in Candidate Regions*

Candidate regions are typically chromosome segments of at most a few megabases identified either from some evidence of relationship to disease (such

as from a linkage study) or based on prior knowledge of gene content (a region around a gene that has some evidence of association with disease). Alternative approaches to analysis are numerous, ranging from the assessment of association at many single SNPs, through model fitting to determine evidence for a specific location in the region to coalescent methods that model shared ancestry of all of the chromosomes in a sample (Chapter 9). Testing the association with many single SNPs across the region is clearly a poor approach given issues of multiple testing and consequent loss of power (Chapter 14). Furthermore, because only a small proportion of the SNPs in the candidate region are likely to have been sampled it is unlikely that the SNP with the lowest *p*-value is actually a/the causal variant. This is well illustrated in the case of the CYP2D6 region *(9)* where two SNPs with the highest association flank the causal site. Successful approaches, therefore, always model the pattern of association with disease across the region. **Figure 5** illustrates the substantial gain in power and precision of localization in the CYP2D6 region when association is modeled using an LD-unit map to represent the underlying LD structure, rather than a sequence-based (kilobase) map.

### 3.4. Toward Genome-Wide Association

The ability to screen the whole genome for variants that influence the common human diseases has at last arrived. Genotyping strategies and SNP-screening panels are improving rapidly so that the whole genome can be effectively screened and all of the important disease loci identified. Many large-scale studies are already yielding exciting and important results. The analysis of genome-wide data poses particular challenges because of the statistical issues involved in making huge numbers of tests that results in the identification of many false-positives (type-I errors). Methods to control for type-I errors are described in detail in Chapter 14. The problem was highlighted by Risch and Merikangas *(25)* who assumed a panel of 1 million tested SNPs to obtain high power for a relatively low risk ratio. If SNPs are tested individually, a nominal significance level $p = 5 \times 10^{-8}$ would be required to achieve an acceptably low genome-wide type-I error. The sample sizes needed to achieve such low *p*-values are clearly not practical for most studies. But this is the worst-case scenario because there is little utility in conducting single-SNP tests and greater power can be achieved by modeling association within a region using fewer degrees of freedom, assuming the genome is divided into a set of regions each containing a number of typed SNPs. However, the definition of regions is arbitrary and the estimation of a small number of parameters to represent association within the region requires simplifying assumptions (such as only one causal "site"). Further complications arise from the distribution of the test statistic that is strongly influenced by the nonindependence, because of LD, between SNP markers. This makes the generation

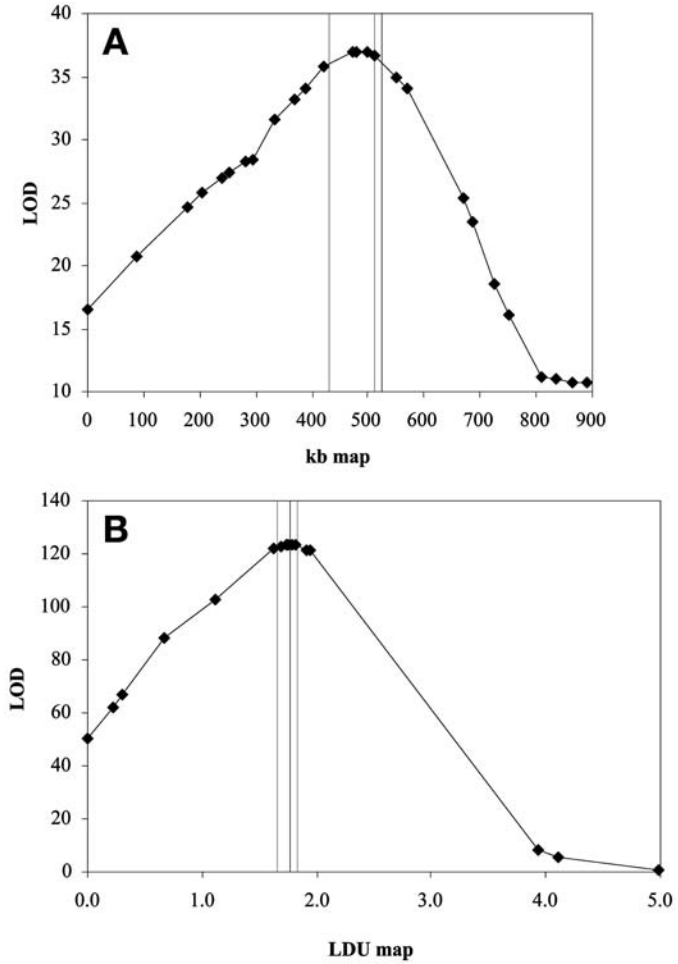



Fig. 5. Logarithm of odds (LOD) curves for association mapping of the poor drug-metabolizing phenotype in the CYP2D6 locus (*9*). The known location of the causal polymorphism is at 525.3 kb (shown as a vertical line in both plots). Modeling the pattern of association on the kilobase map **(A)** places the location (inferred by linkage disequilibrium from other markers in the region) at >50 kb from the correct location and is outside of the support interval (gray vertical lines). Mapping on the linkage-disequilibrium unit scale localizes the causal site to within a few kilobases and with much greater power, as reflected in the LOD.

of many permutation samples (where replicate tests are undertaken using a randomized "shuffled" phenotype) important. Such an approach is computationally very intensive but may enable appropriate transformations that make the use of techniques such as the false-discovery rate (*26*) invaluable to establish genome-wide significance.

However the analytical problems of dealing with genome-wide data are addressed follow-up studies from independent samples are essential to confirm any apparent association. Sequencing and functional tests are typically undertaken given strong evidence. Success in mapping genes involved in many common diseases is already very evident including, for example, asthma (*see* Chapter 15) and type II diabetes (*see* Chapter 16). Given currently available powerful combinations of new genotyping technologies, advances in knowledge of LD structure, novel statistical approaches, and parallel computing platforms there is little doubt that the important disease-influencing polymorphisms will all be determined within a relatively short space of time and the long process of developing novel therapeutic agents can begin in earnest.

## References


1 Tapper, W., Collins, A., Gibson, J., Maniatis, N., Ennis, S., and Morton, N. E. (2005) A map of the human genome in linkage disequilibrium units. *Proc. Natl. Acad. Sci. USA* **102,** 11,835–11,900.

2 Gusella, J. F., Wexler, N. S., Conneally, P. M., et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306,** 234–238.

3 Kerem, B., Rommens, J. M., Buchanan, J. A., et al. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science* **245,** 1073–1080.

4 Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P. Y., and Collins, A. (2001) The optimal measure of allelic association. *Proc. Natl. Acad. Sci. USA* **98,** 5217–5221.

5 Lewontin, R. (1988) On measures of gametic disequilibrium. *Genetics* **120,** 849–852.

6 Collins, A., Lonjou, C., and Morton, N. E. (1999) Genetic epidemiology of single nucleotide polymorphisms. *Proc. Natl. Acad. Sci. USA* **96,** 15,173–15,177.

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

8 Maniatis, N., Collins, A., Gibson, J., Zhang, W., Tapper, W., and Morton, N. E. (2004) Positional cloning by linkage disequilibrium. *Am. J. Hum. Genet.* **74,** 846–855.

9 Maniatis, N., Elahi, E., Gibson, J., et al. (2005) The optimal measure of linkage disequilibrium minimizes error in positional cloning of affection status. *Hum. Mol. Genet.* **14,** 145–153.

10 Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22,** 139–144.

11 International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* **437,** 1299–1320.

12 Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33,** 228–237.

13 Zhang, W.**,** Collins, A., and Morton, N. E**.** (2004). Does haplotype diversity predict power for association mapping of disease susceptibility? *Hum. Gen.* **115,** 157–164.

14 Terwilliger, J. D. and Hiekkalinna, T. (2006) An utter refutation of the 'Fundamental Theorem of the HapMap'. *Eur. J. Hum. Gen.*, in press.
15 Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29,** 217–222.
16 Gibson, J., Tapper, W., Zhang, W., Morton, N., and Collins, A. (2005) Cosmopolitan linkage disequilibrium maps. *Hum. Gen.* **2,** 20–27.
17 Service, S., Deyoung, J., Karayiorgou, M., et al. (2006). Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat. Genet.* **38,** 556–560.
18 Barton, N. H. (1998) The effect of hitch-hiking on neutral genealogies. *Genet. Res.* **72,** 123–133.
19 Sabeti, P. C., Reich, D. E., Higgins, J. M., et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837.
20 Nielson, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005) Genomic scans for selective sweeps using SNP data. *Genome Res.* **15,** 1566–1675.
21 Risch, N. and Zhang, H. (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268,** 1584–1589.
22 Morton, N. E. and Collins, A. (1998) Tests and estimates of allelic association in complex inheritance. *Proc. Natl. Acad. Sci. USA* **95,** 11,389–11,393.
23 Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993) The transmission test for linkage disequilibrium: the insulin gene and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52,** 506–516.
24 Collins, A., Lau, W., and De La Vega, F. M. (2004) Mapping genes for common diseases: the case for genetic (LD) maps. *Hum. Heredity* **58,** 2–9.
25 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517.
26 Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Nat. Acad. Sci.* **100,** 9440–9445.

# 2

# A History of Association Mapping

**Newton E. Morton**

**Summary**

The current exciting developments in association mapping are founded on theory, which has been developed since the beginning of the last century. I hereby review these developments in their historical context.

**Key Words:** Neutral theory; linkage analysis; Malecot model; composite likelihood; genome scan.

## 1. Introduction

Any attempt to write the history of a science encounters a discovery so momentous that preceding events are only a prologue, whereas subsequent events seem the inevitable consequence of a single act of creation. Mendel provided this watershed for genetics. The Human Genome Project was the genesis of association mapping as we know it, with convenient centenary punctuation.

## 2. The 20th Century

A theory for association between two diallelic loci was first given by Robbins *(1)*, assuming constant allele frequencies subject to no pressure except recombination. Under this model, association decreases steadily from an assumed bottleneck to an asymptote at zero. This created a dilemma as the few known polymorphisms were studied: how could they remain in apparently stable equilibrium? Ford *(2)* provided one explanation: "genetic polymorphism is the occurrence together in the same locality of two or more discontinuous forms of a species in such proportions that the rarest of them cannot be maintained by recurrent mutation." His hypothesis encountered two difficulties: (1) there is no frequency that cannot in principle be maintained by recurrent mutation *(3)*; (2) as advances in biochemical genetics were made, the number of

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

known polymorphisms increased hyperexponentially and exceeded credible limits of selection. Neutral theory became increasingly attractive as it was shown to be consistent with genetic diversity and rates of evolution *(4)*. Allowance for effective population size $N_e$ and linear pressure because of mutation and long-range migration, $v$, ultimately provided an evolutionary theory for association to increase or decrease, depending on selection or changes in $N_e$ and $v$ *(5)*.

Freed from the incubus of balanced polymorphism, association of closely linked genes gradually became an object of research in its own right. The G6PD locus is closely linked to the OPN1MW locus for deutan color blindness on the X chromosome. An excess of coupling in Sardinia *(6)* led to the speculation of gene interaction, but this was contradicted by repulsion excess in Iraqi and Kurdish Jews *(7)*. Finally, random association in African Americans indicated that these apparently discordant results originated from a small number of founders who introduced into local populations the G6PD allele that is protective against malaria and by chance had or transmitted an excess or deficiency of alleles for color blindness *(8)*. Lewontin *(9)* introduced four D' measures of allelic association in random pairs of diallelic markers, the largest of which is identical with the association probability ρ that has an evolutionary theory, has been generalized to nonrandom samples of cases and controls, and is used today for association mapping *(10)*.

A method to estimate haplotype frequencies for pairs of diallelic markers and diplotypes was introduced by Bennett *(11)* and Hill *(12)*. Hill and Robertson *(13)* developed a theory for linkage disequilibrium in a finite population. They used the squared correlation $r^2$ that has no probabilistic interpretation in the interval between 0 and 1, and they assigned a value of zero when drift led to monomorphism for either diallelic marker. Therefore in time $r^2$ declined to zero under their convention. Sved *(14)* used theory of Malecot *(15)* to develop a probability model for identity by descent of a marker conditional on identity by descent for another marker. This had no direct application to linkage disequilibrium or association mapping, but was in the path to the Malecot model for those phenomena.

During this period there was an explosive development of linkage analysis that succeeded in identifying many major genes with effects on disease susceptibility or drug response, but genes with small effects were more difficult; initial claims often were not confirmed, or at best established a broad confidence interval in which the causal locus might later be identified. It became obvious that markers within or near a disease locus are associated to an extent that tends to decrease with physical distance. This led Chakravarti *(16)* to infer "hot spots" of recombination from associations in the HBB locus, thereby anticipating later evidence for interspersed blocks and steps in allelic association *(17)*. Kerem et al. *(18)* used allelic association to map and then characterize the locus for cystic

fibrosis (CFTR). These data were used by Terwilliger ***(19)*** to begin the development of composite likelihood for association mapping. Devlin and Risch ***(20)*** introduced the δ metric for case–control samples, now superseded by the general and more powerful model that allows ρ to vary with ascertainment ***(5)***. Finally Risch and Merikangas ***(21)*** proposed association mapping based on diallelic single-nucleotide polymorphisms (SNPs), each tested separately. An alternative is composite likelihood that gives $\Lambda = -2\text{lnlk} = \Sigma_i K_i (\hat{\psi}_i - \psi_i)^2$, where $\psi_i$ is a predicted measure of association between disease-related marker *i* and a measure of disease such as affection status or a quantitative trait, the circumflex signifies an estimate from the data allowing for the population frequency of affection, and $K_{\psi i}$ is the information about $\psi_i$. This is the basis for both methods of association mapping in common use, the Malecot model ***(10)***, and coalescent theory ***(22)***, although the latter is not explicit about the relation between linkage and association.

## 3. The 21st Century

These and other harbingers of association mapping came to fruition with the Human Genome Project that provided for the first time a physical map at nucleotide resolution ***(23)***. The stage was set for modern association mapping, but gaps in the genome sequence continued to be filled in and incorrect sequences to be corrected. A practical limit was set by the fact that the public genome is only one among many billions that might have been sequenced, and it does not represent substitutions, deletions, insertions, or rearrangements from this arbitrary standard that other approaches now address. The Human Genome Project was the first great achievement of the 21st century, but it was only the last step in preparation for the genome revolution. Like the busy week of Genesis, it was not the beginning of the end, but the end of the beginning.

In response to these developments, an International HapMap Project was initiated in 2001 with emphasise on haplotypes that was muted in the progress report 4 yr later ***(24)*** when the number of typed SNPs far exceeded the original target. Consonant with the Human Genome Project that produced a single representative map, application of their product was left to the decision of each user, conforming to the general rule that inventors and exploiters of a novelty are rarely identical. The first step was to establish that a measure of association ρ for haplotypes and γρ for diplotypes has greater relative efficiency than other metrics ***(5)***. The next step was to define a linkage disequilibrium unit (LDU) as $\Sigma_i \varepsilon_i d_i = 1$, where $d_i$ is the physical distance in the $i^{th}$ interval and $\varepsilon_i$ is a Malecot parameter, and to show that a map in LDU generally has substantially greater power for association mapping than a kilobase map ***(25,26)***. The final step is to extend association mapping to what are loosely called genome scans, although single and multiple chromosomes pose the same problems. This process involves

**Table 1**
**Mapping of Disease Susceptibility**

| Stage | Linkage utility | SNP density | Rare SNPs | Functional tests |
|---|---|---|---|---|
| 1. Genome scan | + | + | – | – |
| 2. Candidate region | ± | ++ | – | – |
| 3. Candidate locus | – | +++ | + | + |
| 4. Causal SNP | – | all SNPs | ++ | ++ |

**Table 2**
**Choices for a Genome Scan**

1. More SNPs (e.g., 500,000) vs fewer tagged SNPs.
2. The Malecot model vs alternatives.
3. LDU maps vs linkage maps for association mapping.
4. Single markers vs composite likelihood.
5. Nominal significance vs real significance.

several steps in addition to the initial scan (**Table 1**), which itself requires several decisions (**Table 2**). One protocol divides the scan into nonoverlapping regions (perhaps at least 10 LDU in length, with at least 30 markers and without breaking an LD block). Composite likelihood is more powerful than single SNPs and requires a smaller correction for multiple tests. Any correction must be based on phenotype shuffling under the null hypothesis $H_0$ of no causal SNP in the region and must satisfy the $H_0$ moments of the test metric (usually $\chi^2_1$) and the uniform distribution of real significance levels expected for the false-discovery rate. It may take several years to settle these problems, but one observation cheers us on: association mapping in 3 yr has reached higher resolution than linkage mapping in 90 yr.

## References


1 Robbins, R. B. (1918) Some applications of mathematics to breeding problems. III. *Genetics* **3,** 375–389.

2 Ford, E. B. (1940) Polymorphism and taxonomy, in *The New Systematics* (Huxley, J., ed.), Clarendon Press, Oxford, UK, pp. 493-513.

3 Wright, S. (1931) Evolution in mendelian populations. *Genetics* **16,** 97–159.

4 Kimura, M. and Ohta, T. (1973) Mutation and evolution at the molecular level. *Genetics Supplement* **73,** 19–135.

5 Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P. Y., and Collins, A. (2001) The optimal measure of allelic association. *Proc. Natl. Acad. Sci. USA* **98,** 5217–5221.

6 Sinascalco, M., Bernini, L., Latte, B., and Motulsky, A. G. (1961) Favism and thalassemia in Sardinia and their relationship to Malaria. *Nature* **190,** 1179–1180.

7 Adam, A. (1961) Linkage between G-6-P-D deficiency and colour-blindness. Proc. 2nd Inter. Cong. Hum. Genet. Rome, Italy, pp. 565–567, Excerpta Med. E53.
8 Porter, I. H., Schulze, J., and McKusick, V. A. (1962) Genetical linkage between the loci for glucose-6-phosphate dehydrogenase deficiency and colour-blindness in American Negroes. *Ann. Hum. Genet.* **26,** 107–122.
9 Lewontin, R. (1988) On measures of gametic disequilibrium. *Genetics* **120,** 849–852.
10 Collins, A. and Morton, N. E. (1998) Mapping a disease locus by allelic association. *Proc. Natl. Acad. Sci. USA* **95,** 1741–1745.
11 Bennett, J. H. (1965) Estimation of the frequencies of linked gene pairs in random mating populations. *Amer. J. Hum. Genet.* **17,** 51–53.
12 Hill, W. G. (1974) Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33,** 229–239.
13 Hill, W. G. and Robertson, A. (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38,** 226–231.
14 Sved, J. A. (1970) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Pop. Biol.* **2,** 125–141.
15 Malecot, G. (1948) *Les mathématiques de l'hérédité*. Mason et Cie, Paris, France.
16 Chakravarti, A., Buetow, K. H., Antonarakis, S. E., Waber, P. G., Boehm, C. D., and Kazazian, H. H. (1984) Nonuniform recombination within the human beta-globin gene cluster. *Am. J. Hum. Genet.* **36,** 1239–1259.
17 Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.* **29,** 229–232.
18 Kerem, B., Rommens, J. M., Buchanan, J. A., et al. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science* **245,** 1073–1080.
19 Terwilliger, J. D. (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more marker loci. *Am. J. Hum. Genet.* **56,** 777–787.
20 Devlin, B. and Reisch, N. (20) A comparison of linkage disequilibrium measures for fine scale mapping. *Genomics* **29,** 311–322.
21 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517.
22 McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004) The fine scale structure of recombination rate variation in the human genome. *Science* **304,** 581–584.
23 International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945.
24 International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* **437,** 1299–1320.
25 Maniatis, N., Collins, A., Xu, C. F., et al. (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl. Acad. Sci. USA* **99,** 2228–2233.
26 Zhang, W., Collins, A., Maniatis, N., Tapper, W. J., and Morton, N. E. (2002) Properties of linkage disequilibrium (LD) maps. *Proc. Natl. Acad. Sci. USA* **99,** 17,004–17,007.

# 3

# Linkage Disequilibrium Maps and Location Databases

**William Tapper**


## Summary

Effective application of association mapping for complex traits requires characterization of linkage disequilibrium (LD) patterns that reflect the dominant process of recombination and its duration in addition to the more subtle influences of mutation, selection, and genetic drift. Maps expressed in linkage disequilibrium units (LDUs) reflect the influences of these factors with the use of a modified version of Malecot's isolation-by-distance model. As a result, LDU maps are analogous to linkage maps in so far as their provision of an additive metric that is related to recombination and facilitates association-mapping studies. However, unlike linkage maps, LDUs also reflect the partly cumulative effects of multiple historical bottlenecks that account for substantial variations in LD patterns between populations. This chapter provides an overview of the data requirements and methodology used to construct LDU maps, their applications outside association mapping, and their integration into location databases.


**Key Words:** Linkage disequilibrium unit; Malecot model; recombination; hot spot; SNP selection.

## 1. Introduction

There is considerable interest in describing patterns of linkage disequilibrium (LD) in the human genome in order to aid association mapping, extend the resolution of the linkage map, identify recombination hot spots, compare populations, infer their paleodemography, and detect selective sweeps and other events of evolutionary interest. Many alternative metrics have been used to measure LD between single-nucleotide polymorphisms (SNPs), for example covariance (D), association (ρ), correlation (r), regression (b), frequency difference (f), Yule metric (y), and population-attributable risk (δ). In 2001, Morton et al. ***(1)***, demonstrated that association (ρ) was the most efficient metric for modeling the relationship between association and distance in a large sample of haplotypes

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

compared with six alternatives (D, r, b, f, δ, y). At the same time, Collins et al. *(2)*, showed that LD hot and cold spots could be delineated by estimating epsilon (ε) in the Malecot equation from pairwise measures of association. In this manner, areas of low and high LD are defined by large and small values of ε that coincide with recombination hot and cold spots and form the basis of an LD map. The Malecot equation is given by $\rho = (1-L)Me^{-\Sigma \varepsilon_i d_i} + L$ where ρ is the probability of association, L is the residual association at large distance, M describes the amount of association at zero distance and is a measure of phylogeny where a value of 1 is consistent with monophyletic origin and are less than 1 otherwise, and $\varepsilon_i$ gives the exponential decline of association with physical distance $d_i$ in kilobases between the $_i$th pair of SNPs. Using the Malecot model, Maniatis et al. *(3)* constructed the first LD map with additive distances expressed in linkage disequilibrium units (LDUs). The LDU distance between the $_i$th pair of neighboring SNPs is given by $\varepsilon_i d_i$ where the product εd is equivalent to θt, where θ is a small frequency of recombination, and t is the effective number of generations over which recombination has accumulated after one or more population bottlenecks *(4)*. Because εd is not biased in favor of the linkage map and is more accurately known than θt, it is a more useful metric for LD. Although εd is primarily a function of recombination and time, it is also influenced by mutation, selection, and other evolutionary forces. One LDU corresponds to one swept radius, defined as the average extent of useful LD (the distance in kilobases at which disequilibrium has declined to $e^{-1}$ ~0.37 of its starting value), and so even spacing of SNPs on the LDU scale is optimal.

Comparison with sperm-typing data from six hot spots within a 216-kb region of 6p21.3 *(5)* confirmed the relationship between ε and recombination by demonstrating remarkable alignment between LDU steps and sites of meiotic recombination (**Fig. 1**; *[6]*). At the same time, plateaus in the LDU map were found to correspond with eleven blocks of low haplotype diversity *(6)* within a 617-kb region of 5q31 *(7)*. Because recombination has a dominant role in shaping patterns of LD, the LDU scale is analogous to the centimorgan (cM) scale of linkage maps. Unlike linkage maps, however, LDU maps are also influenced by gene conversion, genetic drift, selection, mutation, and the partly cumulative effects of multiple historical bottlenecks that inflate LD because of founder effects. As a result, although the contours of population-specific LDU maps are highly concordant, the magnitude of LDU steps differ according to population differences in duration. These observations led Lonjou et al. *(8)* and Gibson et al. *(9)* to develop a single "cosmopolitan" map which, when appropriately scaled, recovered 91–95% of the information from LDU maps constructed from population-specific samples. Their results suggest that recombination hot spots are colocalized in all populations, which accounts for the success of cosmopolitan LDU maps. These findings are perhaps contradicted by evidence

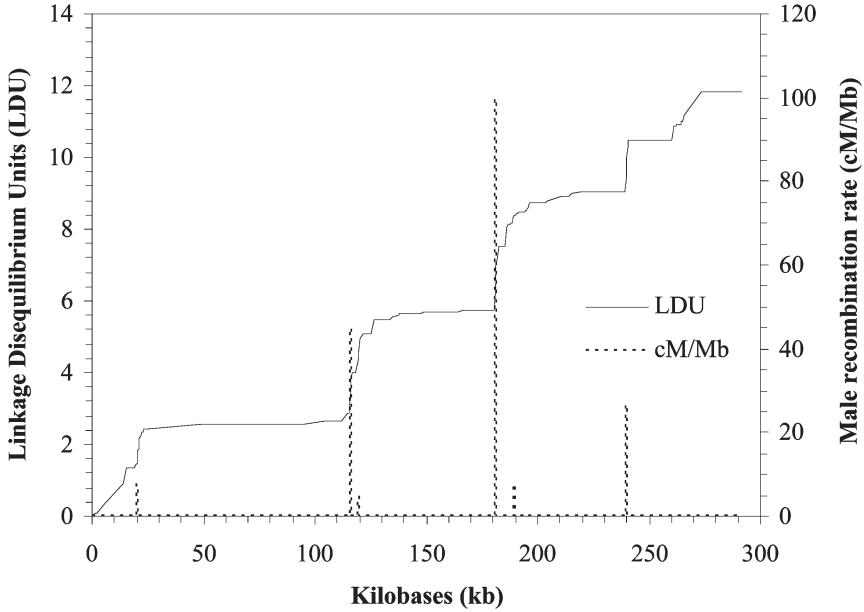



Fig. 1. Linkage disequilibrium unit steps and plateaus correspond with hot spots and haplotype blocks.

that suggests that hot spots may be prone to extinction by meiotic drive *(10)*, whereby alleles in heterozygotes that are active in recombination are lost by gene conversion, causing excess transmission of recombination suppressing alleles. However, at low resolution, the linkage and LDU maps are essentially identical *(11)* despite their representation of current and historical recombination patterns that might be expected to differ given this process of meiotic drive. The remaining differences between population-specific LDU maps may reflect regions of the genome that have been influenced by other processes such as mutation, selection, or drift that vary between populations. As a result, the identification and analysis of these regions is one aspect of currently active research into the application of LDU maps.

LDU maps require accurate physical (sequence) maps and high-density SNP genotype data to give correct distances of $d_i$ and allow reliable estimation of $\varepsilon_i$. Completion of the Human Genome Project *(12)* and genotypic data for a whole chromosome *(13)* led to the first whole-chromosome LDU map *(14)* and an updated location database *(15)*. The location database integrates cytogenetic, physical, linkage, and LD maps in order to provide truly integrated maps that give locations for all loci on all scales in tabular formats. This is more practical than several alternatives that simply connect alternative scales by a handful of shared markers and have graphical formats that do not facilitate the use of the

information in different analyses. As the information on one map increases in volume or reliability, the integration process may be repeated to add value to other maps until they also are revised from data unique to that map. Location databases that integrate cytogenetic, physical, linkage, and LD maps are extremely useful tools for gene mapping and for investigating the biological relationships between sequence and patterns of recombination and LD. Individually, the different maps have different properties, resolution, and applications. Although cytogenetic locations are coarse, this information is useful for regional assignment of chromosomal rearrangements. Linkage maps have proven invaluable for low-resolution mapping of genes predisposing to particular diseases with notable early successes such as localization of the Huntington's gene *(16)*. The resolution and accuracy of linkage maps have much improved in recent years *(17–19)* and their application to disease mapping continues *(20,21)*. Integration of these maps is essential as, for example, cytogenetic analyses may orientate linkage studies and association mapping is often directed by low-resolution linkage analyses, which identify large candidate regions (1–10 Mb) within which fine mapping is required to determine a causal locus. Once candidate regions have been identified, physical and genetic maps are required to determine regional gene content and gene function so that candidate genes can be prioritized for further investigation. Completion of phase I of the international HapMap *(22)* Project has enabled the creation of LDU maps for each chromosome *(12)*. Here, we review the construction and properties of LDU maps including use of the LDMAP program and its data requirements. We also describe the inclusion and preliminary use of such maps in a linkage disequilibrium database (LDDB) that is publicly available (http://cedar.genetics.soton.ac.uk/public_html) for association mapping, population comparisons, SNP selection, and identification of LD features and recombination hot spots.

## 2. Methods

### 2.1. Data Requirements for LDU Map Construction

The construction of LDU maps requires genotypic SNP data from unrelated individuals in the form of disomic genotypes (diplotypes) or haplotypes. Although both types of data can be used to construct highly concordant LDU maps, previous studies have shown that "real" haplotypes, as opposed to inferred, are on average 50% more informative than diplotypes *(4,23)*. However, this gain must be balanced against the extra cost and error involved in determining haplotypes from either somatic cell hybrids or inferring them from family material.

Prior to LDU map construction, the genotypic data are screened to remove SNPs with extreme deviations ($\chi^2_1$>10) from the Hardy–Weinberg test *(24)* and SNPs with minor allele frequencies less than 5%. Previous studies have shown that maps built from subsets of rare markers have shorter LDU lengths *(25)*.

Given the strong dependence of LD measures on allele frequencies and thus mutation age, this is expected because LD is higher between recent (rare) markers compared with older (more common) markers. This implies that the length of LDU maps constructed from studies with nonrandomly ascertained SNPs, such as the HapMap Project that focuses on common markers, will be somewhat different in corresponding maps derived from samples with different allele-frequency distributions.

LDU maps are largely insensitive to marker density as their profiles determined at various SNP densities from 1/2–1/23 kb are highly concordant *(12,14,25)*. In contrast, methods to detect blocks of low haplotype diversity, such as the Gabriel approach *(26)* the D' threshold technique *(27)*, and the four-gamete test *(28)*, are sensitive to marker density and there is little correspondence between blocks identified by different methods *(25)*. Specific markers often deviate markedly from pairwise LD trends, presumably because of factors other than recurrent recombination such as the age of mutations, genetic drift, mutation rate differences, and gene conversion. LDU maps are more robust to this sort of deviation, as the model fitting and simultaneous use of multiple intervals (*see* **Subheading 2.2.**) provides a degree of smoothing that seems to remove much of the marker density effects that plagues other fine-scale descriptions of LD. However, LDU maps may contain a small proportion of intervals with indeterminate values of $\varepsilon_i$, also known as holes, which are assigned maximum values of 3 LDU *(14)*. This value was established by using maximum likelihood to evaluate the fit at different limits but has been reduced to 2.5 LDU when using high-density data *(12)*. Using a low-density LDU map of chromosome 22 (1 SNP per 23 kb), Tapper et al. *(14)* showed that holes are correlated with regions of elevated recombination ($r = 0.11$, $p = 0.0003$). Further studies with high-density (1 SNP per 4 kb) genome-wide LDU maps *(12)* constructed from the phase-I HapMap data *(22)* identified a significant relationship between the density of holes per Morgan and markers per Morgan (**Fig. 2**) that suggested that the number of holes will decline as the HapMap Project progresses. However, the factors that determine holes are not only complex and dominated by recombination, but also include SNP distribution, kilobase width of holes, the criteria to declare a hole, and errors in estimation of $\varepsilon_i$. Therefore, the number of holes in future databases cannot be estimated reliably but is likely to decrease to a nonzero limit as LD maps evolve. It is attractive to assume that reduced numbers of holes in response to increased marker density also account for the small reductions in LDU length and suggest that the length of holes is generally exaggerated. However, comparing low- and high-density LDU maps of a 10-Mb region of 20q12-13.2 showed that low-density holes are resolved in an unpredictable fashion that may increase or decrease their LDU length when markers are added. Identification of holes is extremely useful to define regions

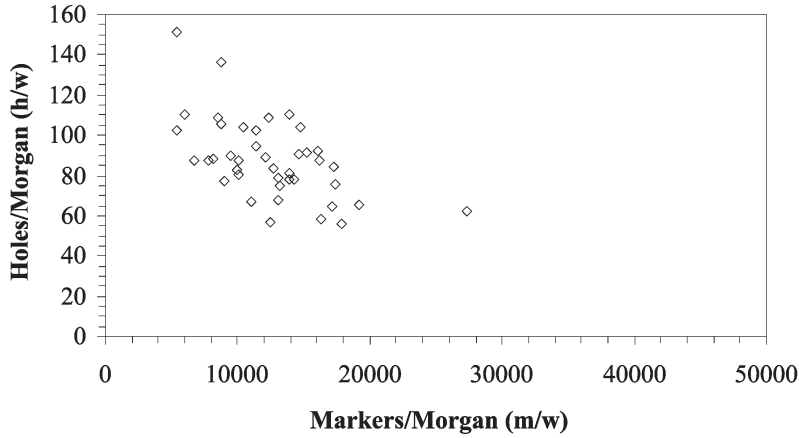



Fig. 2. The declining density of holes with marker density among chromosome arms.

that require increased marker density to refine map length, LD structure, and perform efficient positional cloning. As a result, although SNP densities as low as 1/20 kb are adequate for map construction, densities closer to phase I and phase II of the HapMap Project (1/5 kb and 1/1 kb) yield more precise maps. Consequently, using appropriately scaled cosmopolitan maps derived from high-density HapMap data *(9,10)* may be more reliable than low-density LDU maps constructed from study-specific data alone.

Although the contours of population-specific LDU maps are highly concordant, the magnitude of their LDU steps varies according to population differences in duration as a result of the partly cumulative effect of multiple bottlenecks such as migration, wars, famine, and disease. Stochastic pressures such as genetic drift, mutation, and selection also cause differences between populations but these are likely to be more subtle. The population(s) from which data originates should, therefore, be carefully defined in order to accurately interpret patterns of LD. LDU maps have demonstrated robustness toward sample size with few differences observed between maps constructed from 45 to 200 individuals *(29)*. However, as the number of individuals falls below a threshold the resulting map will have fewer steps and shorter total LDU length as the diversity of genotypic data is reduced because rare recombinant haplotypes may not be observed *(30,31)*. The number of genotyped individuals required for LD map construction should, therefore, be maximized and maps constructed from less than 20 individuals be treated with caution.

### 2.2. Using LDMAP to Construct LDU Maps

Having obtained a quality controled dataset with sufficient numbers of individuals and markers, LDU maps can be constructed by the LDMAP program *(4)*

also available from http://cedar.genetics.soton.ac.uk/public_html/. The first step is to create an intermediate file, from either haplotype or diplotype data, containing pairwise association probabilities ρ ***(2)***, information $K_\rho$, and the kilobase size of intervals $d_i$. The association probability (ρ) is given by $\rho = D/Q(1 - R)$, where D is the absolute value of the difference between a haplotype frequency and its equilibrium value as the product of allele frequencies ***(32,33)***. For marker-by-marker association in unrelated individuals, ρ equates to the absolute value of D' ***(34)***. Under the null hypothesis that D = 0, the information ($K_\rho$) is given by $N\,Q(1 - R)/R(1 - Q)$ for N haplotypes or diplotypes. Under the alternative hypothesis the information from haplotypes is a closed form in D ***(2)***, but the information from diplotypes requires inversion of the $3 \times 3$ information matrix for Q, R, and D ***(35)***.

Having produced an intermediate file from either haplotype or diplotype data, the Malecot model, given by $\rho = (1 - L)Me^{-\Sigma \varepsilon_i d_i} + L$, is used to estimate values of epsilon ($\varepsilon_i$) for each marker interval that are then used to construct LDU maps. Large maps with high density, such as those constructed for whole chromosomes using HapMap data, can be constructed in segments of 500–1000 markers with little loss of information, whereas smaller maps can be made in one piece. Values of ε are estimated by a multiple pairwise algorithm; for example, Collins et al. ***(36)*** consider a map with five SNPs and four intervals (**Fig. 3**), a value of ε for the first interval is calculated using all of the pairwise measures of association that include that interval (i.e., the pairwise association between 1–2, 1–3, 1–4, and 1–5). These measures are combined by weighting them according to their information ($K_\rho$) so that interval 1–2 is given the biggest weight and 1–5 the smallest weight. The Malecot model is fitted to this data to determine values of ε by (composite) maximum likelihood. In this example, the first interval measures 0.5 LDU that is given by $\varepsilon_i d_i$ where $\varepsilon_i$ is equal to 0.05 and $d_i$ is equal to 10 kb. This process is repeated for each interval to produce an LDU map.

Unlike ε, single values of the M and L Malecot parameters are estimated for the entire sample. Because L is the residual association at a large distance, erroneous estimates may occur for small regions and/or high-density data because of the predominance of block structures with high association. It is, therefore, advisable to use predicted L ($L_p$), given by the weighted mean deviation for a normal distribution ***(7)*** when dealing with these types of data. Convergence of parameter estimates is achieved by maximizing the composite likelihood given by $\exp\left[-\sum K_\rho(\hat{\rho}_i - \rho_i)^2 / 2\right]$ where $\hat{\rho}_i$ is an estimate with prediction $\rho_i$, and the summation is over n pairs of SNPs used for LD analysis within a given window containing r SNPs. To compare the fit of the pairwise data to the kb and LDU maps we calculate the error variance ***(10,32)*** as $V_{kb} = -2\ln lk/(N - 2)$, where ε and M are estimated $V_{LDU} = -2\ln lk/(N - m - 1)$, where N is the number of pairs, and m is the number of intervals in which ε has been estimated. The

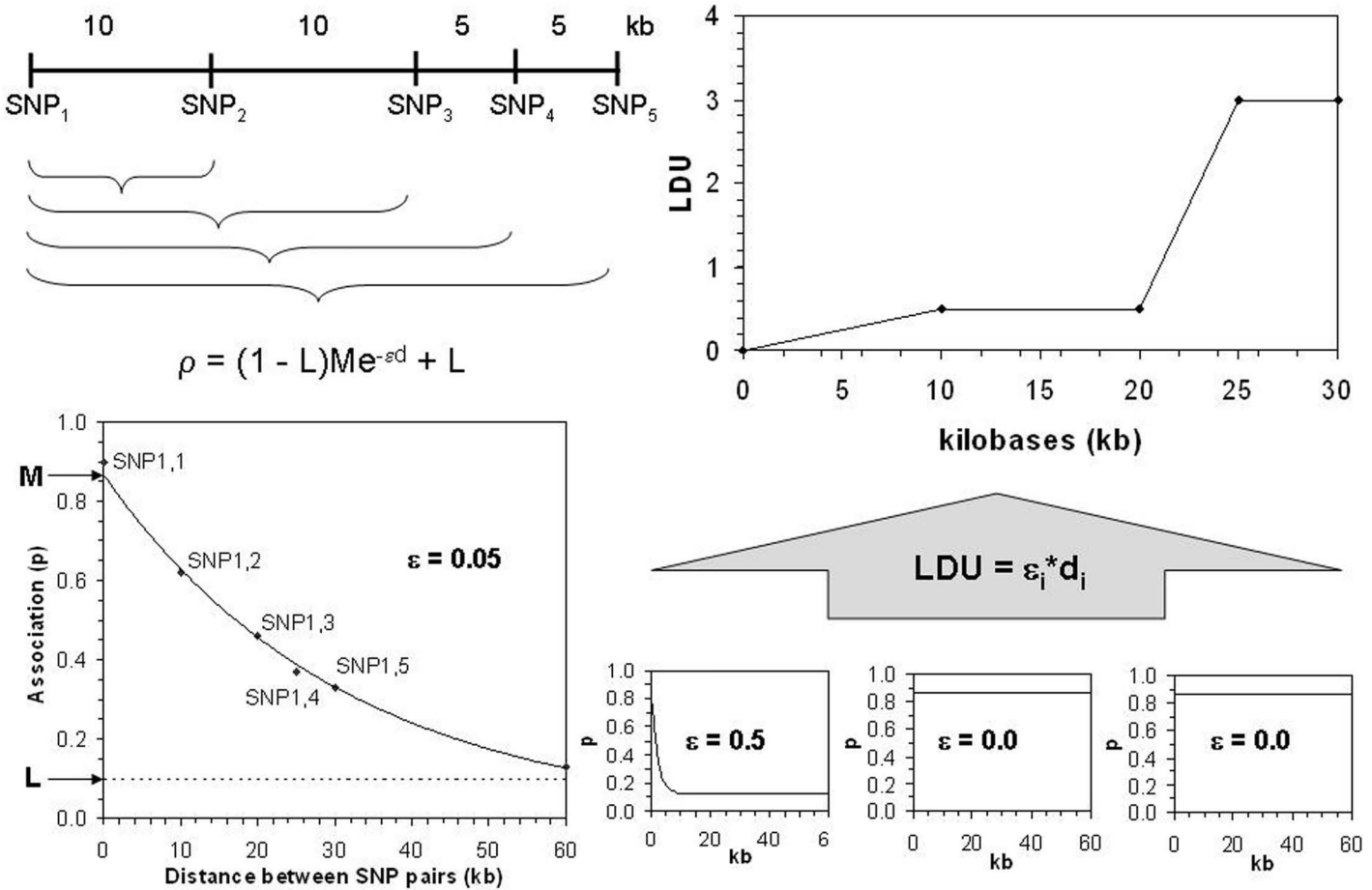



Fig. 3. Construction of linkage disequilibrium unit maps.

default parameters of the LDMAP program (maximum distance between any SNP pair = 500 kb, maximum number of intervals between SNP pairs = 100, maximum number of iterations = 20,000, stop iterating in an interval when $\varepsilon_i = 0$, markers per segment = 1000, and markers overlapping adjacent segments = 25) are designed to give efficient construction of high-quality LDU maps but can be adjusted according to time and computing constraints.

Cosmopolitan maps, which combine data from several populations to form a single map applicable to all populations, may be produced by pooling haplotype counts *(8,9)* following the algorithm described by Hill *(35)* or calculating the mean LDU location between populations. Although the pooled haplotype method provides good maps, it is computationally intensive and slow for large samples. In comparison, mean LDU locations can be quickly determined and perform equally well. Irrespective of these methods, the allele coding must be consistent between populations in order to construct meaningful cosmopolitan LDU maps.

### 2.3. Construction of a LDDB

The LDDB (http://cedar.genetics.soton.ac.uk/public_html/LDB2000/release.html) is constructed by a bioinformatic approach consisting of automated data retrieval and organization so that large amounts of data from internet sources can be integrated and efficiently revised when sequence, linkage, or LD data is updated. Database construction begins with a comprehensive cross-referenced list of genes and polymorphic markers obtained from the Human Genome Browser (http://genome.ucsc.edu/), Ensembl (http://www.ensembl.org/), UniGene (http://www.ncbi.nlm.nih.gov/), linkage data *(19)*, and HapMap (http://www.hapmap.org/). This list is submitted to Entrez Gene (http://www.ncbi.nih.gov/) and GDB (http://gdbwww.gdb.org/) via their CGI programs to extract aliases and primer sequences associated with each locus using a java-based webpage reader. Sequence locations for known and predicted genes are obtained from the Human Genome Browser and Ensembl, whereas SNP locations are taken from HapMap and microsatellites are localized by electronic PCR (ePCR). The ePCR searches are made by a stand-alone version of BLAST *(37)* against finished whole-chromosome sequence assemblies (NCBI build 34, UCSC hg 16 July 2003) that are split and indexed into 1-Mb segments. Sequence locations presented in LDDB are geometric points representing the midpoint of genes and markers using the p telomere as the origin.

The multiple names used for specific genomic segments pose some difficulties for map integration. The nomenclature committee supported by HUGO *(38)* is concerned with standardizing names of expressed sequences for which something is known about function. However, for the vast majority of map objects (ESTs, STSs, SNPs, and microsatellites) there are no nomenclature standards

and the result has been a proliferation of symbols and alternative names. Many symbols are, therefore, difficult to trace and there is no single resource that maintains the complete set of synonyms. This problem is tackled by organizing all symbols retrieved from the java-based web searches and sequence locations into an "alias-map." Clusters are formed by comparing individual entries in the alias map representing a single locus from the starting list with all other entries by cross-referencing loci; matching entries are then combined to form a cluster. An alias-cluster map is produced by repeating this procedure for all entries. The clustering process is checked by calculating the variance in physical locations in each cluster, clusters with large variances are examined to determine the source of errors (typically these are nomenclature problems), and incorrect clusters, which do not represent a single-expressed sequence or set of synonyms, are prevented from forming. The clustering process is designed to avoid redundant entries that describe the same locus. A class field is used in the summary maps that indicates the type of locus (Gene G or amplifiable marker A), as this is not immediately obvious from the name and allows quick and easy organization and extraction from the database.

Map integration is defined as the process whereby locations on different scales (LD, linkage, physical, and cytogenetic band), derived from a number of sources, are represented in a summary map. Missing sex-specific centimorgan (mcM, fcM), LDU and cytogenetic band locations are inferred by linear interpolation from sequence locations for all loci in the alias-cluster map, resulting in a fully integrated map within which all scales agree with sequence order. If a, c represent flanking loci with sequence locations Sa, Sc, and genetic locations (cM or LDU) Ga, Gc, and Sb is the sequence location for a locus without a genetic location, and $Sa < Sb < Sc$, then the interpolated genetic location (Gb) is $Gb = (Gc - Ga)/(Sc - Sa)(Sb - Sa) + Ga$. The integration process is completed by assigning each locus to a cytogenetic band on the basis of its physical location and the physical location of cytogenetic bands ***(39)***. The precision of the physical locations of band borders is limited by the resolution of fluorescence *in situ* hybridization to between 0 and 5 Mb (1–1.5 bands).

A summary map is built by taking a single "primary" name and its location from each cluster in the alias-cluster map. Links are maintained between primary names and their associated aliases from the alias map, so that when locus searching is performed hits are generated using both aliases and primary names. Primary names for genes are chosen using the following hierarchy: HUGO nomenclature committee > Entrez Gene > Human Genome Browser. D-numbers and dbSNP names are prioritized as primary names for STRs and SNPs although in their absence other symbols are used. High-resolution sex-specific linkage maps and population-specific LD maps can be accessed through hypertext

links and a search interface. Exonic structure and other features are not represented as these require a high level of sequence annotation and can be obtained from the Human Genome browser and Ensembl.

## 3. Methods

### 3.1. Genome-Wide LDU Maps

Following completion of the phase-1 HapMap Project *(22)*, the first genome-wide LDU maps *(11)* were updated to increase their density and incorporate three populations that were previously excluded. The genotypic data for these maps was downloaded from the 16a release of the phase-1 HapMap data (http://www.hapmap.org/) and consisted of approx 1 million SNPs for each of 60 CEPH (CEU), 60 Yoruban (YRI), 45 Chinese (CHB), and 44 Japanese (JPT)-unrelated individuals. Between 24 and 36% of these genotypes, accounting for over 250,000 SNPs from each population, were removed when screening for rare SNPs and SNPs with significant deviations from the Hardy–Weinberg test *(24)*. This procedure reduced the density of available SNPs from approx 1 SNP per 3 kb to approx 1 SNP per 4 kb. All of these SNPs have nucleotide positions relative to the July 2003 freeze of Golden Path database (http://genome.ucsc.edu/). Genome-wide LDU maps were constructed following the described methodology (*see* **Subheading 2.2.**). The resulting LDU maps contain between 3000 and 5000 holes that account for 1–2% of the physical sequence and between 11 and 23% of total LD lengths (**Table 1**). The majority of these holes are small, with average sizes from 8 to 10 kb, but some reach up to 500 kb in length. Identification of holes is extremely useful to define regions that require increased marker density to refine map length, block structure, and perform efficient positional cloning. The number of holes will decline as the number of SNPs genotyped by the HapMap Project increases and would decline more rapidly if holes were specifically targeted. However, a proportion of these holes are expected to persist because of the presence of intense hot spots causing extremely low LD in which it will be very difficult to detect association. As a result of the holes in the LDU maps, the map length estimates presented here, although encouragingly similar to other whole-chromosome LD maps *(11,14,40)*, must be regarded as approximate until density is increased. The LDU maps were analyzed to determine the influence of recombination on patterns of LD; whether LDU maps can identify recombination hot spots and enhance the resolution of linkage maps; determine the similarity between population-specific LDU; and calculate effective bottleneck times (t). The statistical analysis divided the LDU maps into chromosome arms and deciles of these arms between the first and last physical location shared by the linkage *(19)* and LDU map, omitting heterochromatic, centromeric, and pseudoautosomal regions.

**Table 1**
**Linkage Disequilibrium Unit Length and Hole Characteristics**

| Sample | LDU map length | No. holes | % of sequence length | Mean hole size (bp) | Range (bp) | % of LDU map length |
|---|---|---|---|---|---|---|
| CEU | 56247 | 2911 | 0.8 | 8,176 | 50–447,953 | 15.3 |
| JPT | 56656 | 3731 | 1.3 | 10,417 | 40–498,834 | 19.5 |
| CHB | 62686 | 4879 | 1.6 | 9,842 | 43–498,834 | 23.1 |
| YRI | 79499 | 2978 | 0.9 | 8,599 | 44–407,285 | 11.1 |

### *3.2. Recombination and LDU Maps*

Because LDUs are the product of recombination and distance ($\varepsilon_i d_i$), they are negatively correlated with LD and positively correlated to recombination. Therefore, regions with high LD have few LDUs, appearing as plateaus on the LDU map with low recombination, whereas areas of low LD have many LDUs that form discrete steps and have high levels of recombination. The extent to which LDU maps are influenced by recombination was examined in the 41 chromosome arms and 410 deciles by regressing LDUs on Morgans through the origin and weighting by Morgans. Despite the low resolution of the linkage map and stochastic variation and selection acting solely on the LDU map, recombination accounts for 99% of the LDU variance in chromosome arms and 98% in their deciles (**Fig. 4**). Furthermore, the close correspondence between several sites of meiotic recombination determined by sperm typing and specific LDU steps (**Fig. 1**) suggests that LDU maps can be used to identify novel hot spots and enhance the resolution of linkage maps.

Inference of the location and perhaps intensity of novel hot spots by genome-wide LDU maps is demonstrated by plotting the LDU rate (LDU/Mb) for CEU and YRI populations within 5-kb windows for a 1-Mb region of chromosome 1 (**Fig. 5**) that incorporates a 206-kb region with sperm-typing data ***(41)***. Despite the low resolution of these LDU maps (1/4 kb) compared with sperm typing (1/1 kb), they clearly identify the eight sperm typed hot spots as narrow peaks with high LDU rates that are shared between the CEU and YRI populations. Additionally, this search intimates several novel hot spots including, notably, three proximal to the sperm-typed region that appear to be more active. Two sperm-typing studies that cover a combined region of approx 0.5 Mb ***(6,41)*** suggest that recombination hot spots are only 1- to 2-kb wide and occur roughly every 30 kb. If representative of the genome, this implies a total in the region of 100 thousand hot spots that corresponds well with the number of steps in the LDU maps (**Table 2**). However, because LDU maps reflect patterns of LD that are influenced by factors such as genetic drift, selection, and mutation in addition to

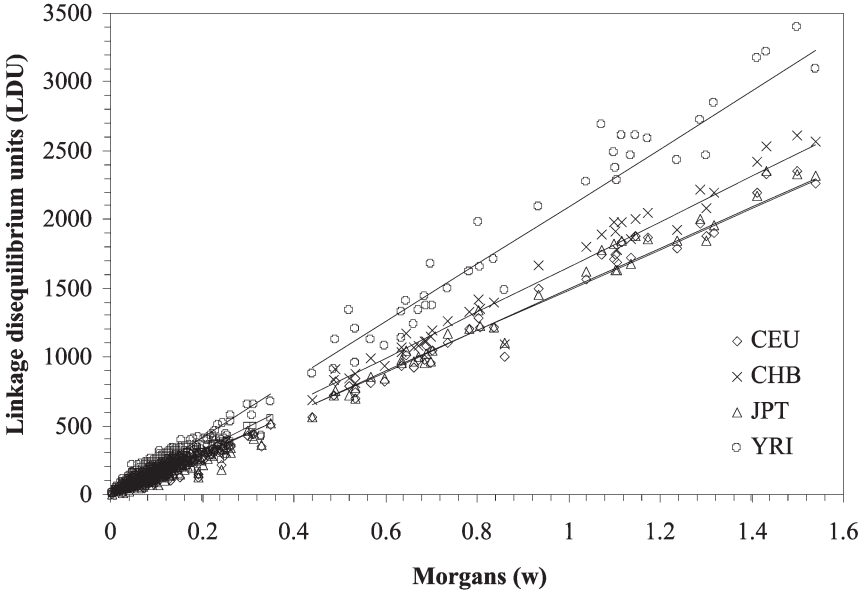



Fig. 4. The relationship in the human genome between linkage disequilibrium unit and linkage in chromosome arms and deciles.

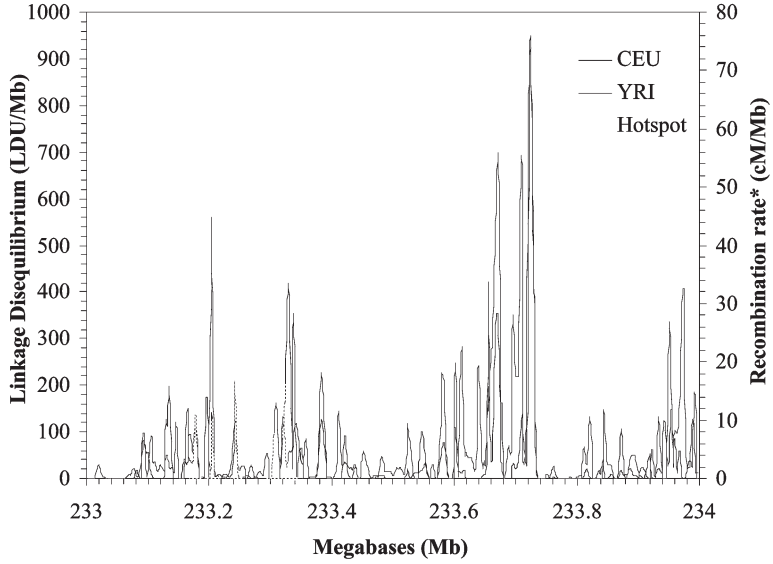



Fig. 5. Identification of recombination hot spots.

**Table 2**
**Genomic Characteristics of Linkage Disequilibrium Blocks and Steps Defined From the Linkage Disequilibrium Unit Maps**

| Feature | Population | No. | Mean size (bp) | Size range (bp) | % of map length |
|---|---|---|---|---|---|
| Steps | CEU | 119,131 | 8,577 | 1–3,239,101 | 36.0 |
| | JPT | 103,991 | 10,553 | 1–3,239,101 | 38.7 |
| | CHB | 107,014 | 10,501 | 1–3,247,974 | 39.6 |
| | YRI | 141,671 | 8,274 | 1–3,241,216 | 41.3 |
| | All* | 46,185 | 18,975 | 31–3,241,216 | 30.9 |
| Blocks | CEU | 119,147 | 15,246 | 1–843,179 | 64.0 |
| | JPT | 103,997 | 16,729 | 1–1,022,120 | 61.3 |
| | CHB | 107,021 | 16,012 | 1–1,512,117 | 60.4 |
| | YRI | 141,681 | 11,757 | 1–698,823 | 58.7 |

All represents steps that are in common with the four populations and therefore have stronger support for containing hot spots.

the prevailing role of recombination, some of these steps may not be related to recombination. Steps with high LDU rates that are shared between all populations are more likely to reflect hot spots, whereas small steps, not present in all populations, may be unrelated to recombination or refer to rare and/or ancestral recombination events. We have, therefore, identified 46,185 LDU steps that are strong candidates for harboring novel hot spots by virtue of their presence in all populations (**Table 2**). The discovery of hot spots by LD may be hampered because of sperm-typing evidence that suggests that some hot spots may have evolved recently whereas others may be prone to extinction by meiotic drive ***(10,41)***. A proportion of hot spots identified by LDU maps could, therefore, reflect extinct hot spots and contain no active hot spots, whereas relatively young hot spots may be hidden in regions of high LD and missed by LDU analysis. Hot spot discovery by LDU maps should prove successful however as the strong concordance between linkage and LDU maps ***(11)*** and the similarity between population-specific maps ***(8,9)*** suggests that the evolution of hot spots is constrained by an undefined factor(s) such as access to the DNA sequence or the sequence itself that restricts hot spots to narrow regions. Clearly further studies are required to elucidate the characteristics of hot spots and LD analyses should contribute toward this aim.

Linkage maps are unique in their ability to identify sex-specific rates of recombination. However, linkage maps suffer from low resolution because the number of meioses that can be determined from multisib families is limited by cost. Although sperm typing offers the highest resolution recombination data currently available, the technique is only feasible for very small regions of male chromosomes and does not reflect historical events. The LDU map reflects the combined effects of recombination events in both males and females over many

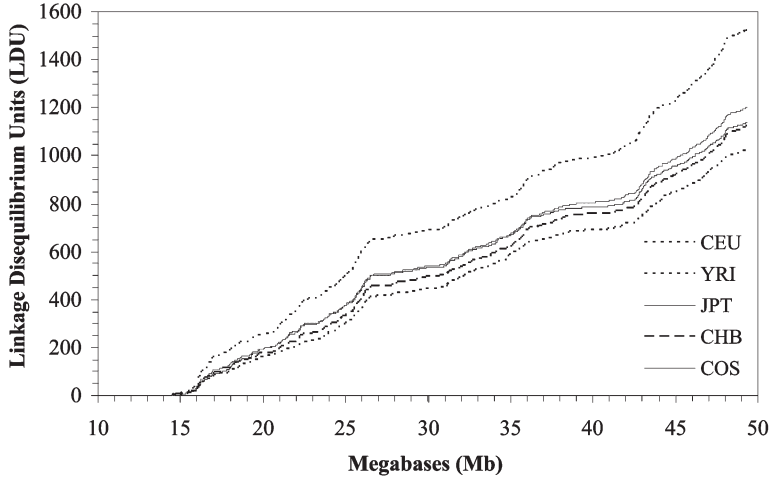



Fig. 6. Comparison of linkage disequilibrium unit maps of chromosome 22 from four populations.

generations. Although this pattern is distorted to a degree by evolutionary and stochastic pressures, the LDU map can be used to locally increase the resolution of sex-specific linkage maps by linear interpolation. In this process, we conserve the locations of framework loci on the linkage map and compute linkage locations for markers in the LD map that lie between adjacent framework loci. A framework locus is declared if the distance to the preceding framework locus in the linkage map is non-zero. In this manner the sex-specific linkage framework is maintained in the high-resolution linkage maps.

### *3.3. LDU Maps in Human Populations*

Plotting the population-specific LDU maps clearly shows that fine-scale patterns of LD, consisting of distinct plateaus and steps, are shared between populations (**Fig. 6**). As sample sizes are large, the cosmopolitan LDU maps were constructed by calculating the mean LDU location between populations. The similarity between population-specific and cosmopolitan LDU maps was assessed by comparing their error variances (**Subheading 2.2.**) as previously described ***(9)***. When appropriately scaled, the cosmopolitan maps recovered between 83 and 92% of the information from LDU maps constructed from population-specific samples that demonstrates their strong concordance as previously shown ***(8,9,40)***. Cosmopolitan maps constructed from pooled haplotype counts and mean LDU locations were found to perform equally well when comparing these maps for chromosome 22. Typically, cosmopolitan maps have fewer holes and higher resolution than population-specific maps as they are derived from larger data samples with more individuals and markers.

Although patterns of LD are conserved between populations, variations in their demographic histories generate considerable variability in the extent of LD. The predominant source of variation between population-specific LDU maps is, therefore, confined to the magnitude of LDU steps that increase in height to give the following ranking: CEU<JPT<HCB<YRI. As a result, the mean swept radius, the extent of useful LD, is far greater in Caucasians (114 kb) compared wtih Africans (74 kb). This is consistent with previous studies *(8,9)* and results from population differences in duration since the last major population bottleneck. Estimates of the M parameter, which reflect association at the bottleneck, support these observation since they are lower in Africans (0.74) compared with Caucasians (0.88). The effects of duration are most apparent in isolated populations that have been recently founded, such as the Finnish subisolate of Kuusamo that have a swept radius nearly twice as large as Caucasians (203 vs 114 kb; *[29]*). Ignoring stochastic variation and selection in the LDU map and errors in estimating the linkage map, the Malecot model predicts that the ratio of corresponding distances in the LDU map and linkage map, in Morgans, estimates the effective bottleneck time t in generations that is constant between chromosomes. The genome-wide LDU maps estimate t as 1472, 1483, 1648, and 2073 generations for CEU, JPT, HCB, and YRI populations, respectively, implying bottleneck times of between approx 36,800 and 51,825 yr, assuming a maximum of 25 yr per generation. Presumably, this low estimate reflects the partly cumulative effect of bottlenecks since the out-of-Africa migration (~100,000 yr). Although recombination accounts for 99% of the LDU variance, estimates of t (LDU per morgan) vary with respect to chromosome arm length (Mb per Morgan) so that smaller autosomes have significantly lower estimates of t (**Fig. 7**). This phenomenon is attributed to the linkage map *(19)* that uses the Kosambi function to account for chiasma interference despite studies showing that interference exists at levels greater than this function provides for and appears to vary between chromosomes *(42–44)*. The linkage map therefore exaggerates Morgan lengths for all chromosomes especially shorter ones that support evidence that interference intensifies with decreasing chromosome size as in the mouse *(45)*. The X chromosome appears as an outlier with exceptionally high LD (LDU/w), despite correcting for the absence of recombination in males by multiplying the female linkage by two-thirds. This finding is consistent with more rapid selection against deleterious mutations when the X chromosome is monosomic in males, and to a lesser extent, under random inactivation (lyonization) in females *(45,46)*.

### *3.4. SNP Selection for Association Mapping*

LD maps are constructed in such a way that one LDU corresponds to the average extent of useful LD or "swept radius," the distance in kilobases at

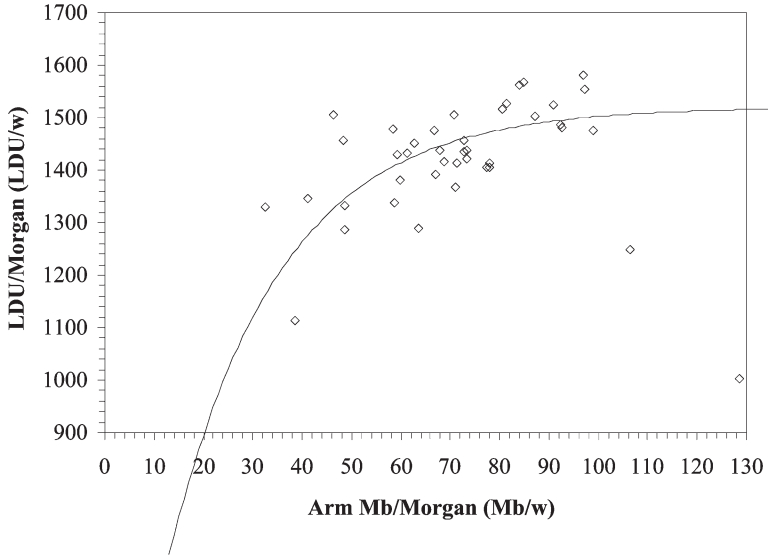



Fig. 7. A graph showing the effects of interference on estimates of the effective bottleneck time t (LDU/Morgan). Nonlinear regression was used to describe the relationship between estimates of t and chromosome arm size (Mb/Morgan) for the CEU population as shown previously *(11)*.

which disequilibrium has declined to $e^{-1}$ ~0.37 of its starting value. Even spacing of markers on the LDU scale should, therefore, be optimal. A minimum density of 1 SNP per LDU is required and several markers per LDU, spanning a range of allele frequencies, are likely to be required to maximize power for association mapping. Increases in power will be correlated with increasing density per LDU *(5)* until a threshold is reached over which further increases in density will have little influence on power. Given the genome LDU lengths (**Table 1**), a genome screen would require at least, and probably some multiple of, 55,500 SNPs for Caucasian populations and 79,500 SNPs for an African population. Low-density screens might usefully identify a number of candidate regions *(47,48)* but further typing at higher density, followed by functional analyses, will be required to identify and verify causal mutations.

### 3.5. Location Database

The LDDB website presents integrated summary maps, population-specific LD maps, and high-resolution sex-specific linkage maps. The text-based (rather than graphical) representation of the integrated maps is designed to deliver the maximum information content in the form of maps for multilocus disease gene mapping. Selecting any locus in the map, or searching for a locus using its primary

name or alias, launches a CGI program that reports aliases, splice variants, and detailed location information associated with that locus. In this manner, the left and right physical, LDU, and cytogenetic locations for each locus in the summary map can be obtained. The aliases in the additional information screen serve as links to other websites including Entrez Gene, the Human Genome Browser, Ensembl, GenBank, the Genome Database, and OMIM. This information can be used to determine the splice variants, disease association, tissue expression, and genomic background of each locus in the summary map.

The integrated maps can also be sampled through a variety of search facilities offering various regional views of the maps, such as a region between two sequence locations or two loci and a physical distance around a locus. Other search options allow new physical locations to be integrated into the summary or LD map by interpolation. The search interface can also be used to determine new physical locations, by ePCR, of primers that employs BLAST against whole-chromosome sequences. Once physical locations have been determined by ePCR they can be interpolated into the integrated summary map to give new linkage, LDU, and cytogenetic locations. The sequence associated with an ePCR localization or between two physical locations can also be extracted from the whole-chromosome assemblies via the search facility. The LDU maps can be graphically represented via the graph function that uses ChartDirector software (http://www.advsofteng.com) to plot LDUs against kilobase locations between specified kilobase locations. The graph function also presents LDU maps as rates, plotting LDU/Mb against kilobase locations, which can be used to infer the location and intensity of novel recombination hot spots. Population-specific LDU maps can also be compared by the graph function that plots the ratio between LDU rates (LDU/Mb) from two populations. In this manner, regions with similar LDU rates are identified by small ratios, whereas discrepant regions have large ratios that may be indicative of positive selection.

The following algorithm has been implemented in LDDB to allow uniform selection of SNPs on the LDU scale from any of the four populations subject to a minimum minor allele frequency threshold. The user specifies a region of interest, density of SNPs/LDU, population, and threshold minor allele frequency. The desired density of SNPs defines a mean LDU width that is used to choose SNPs. The SNP nearest the p telomere is taken as the "starting SNP," whereas the SNP closest to the mean LDU width is chosen as the next SNP. The chosen SNP then becomes the new "starting SNP" and the process continues along the whole map. The LDU length of the region is then divided by the number of selected SNPs to calculate the new average density over the region. Depending on whether this density is greater or less than the desired density, the mean LDU length is altered accordingly and the process repeated until the desired mean density is created. Given a discrete candidate region, a greater multiple of SNPs/LDU should be

economically feasible but efforts to identify causal mutations may also be aided by incorporating SNPs of known functional consequence. Within genes these include nonsynonomous coding mutations, putative splice site variants, and SNPs found in conserved noncoding regions between genes that may be more likely to effect phenotype risk by moderating expression levels than other SNPs.

## 4. Discussion

Effective association studies require LD maps that describe the combined affects of recombination, population bottlenecks, mutation, selection, and genetic drift. Fine-scale maps of recombination *(49)* may therefore be inefficient for association mapping as they do not reflect these demographic factors. In addition, coalescent models make unrealistic assumptions that populations are at equilibrium and that unique founder haplotypes exist that have never recombined with any of the other lineages of the coalescence. Furthermore, the coalescent scale requires arbitrary smoothing and scaling to conform with sex-averaged linkage maps that misrepresent interference and minimize the differences between population-specific patterns of LD *(11)*. In comparison, LDU maps describe LD patterns that reflect demographic factors, as well as recombination, making them ideal to detect functional variants underlying common disorders by means of LD to nearby SNPs. The significantly different estimates of effective bottleneck time *(4)* between African (YRI) and European (CEU) or Asian (JPT, CHB) populations point toward their different demographic histories and demonstrates the importance of duration on patterns of LD. Although abstract in nature because of the partly cumulative effect of multiple bottlenecks and other factors such as the size and expansion rate of founding populations, estimates of effective bottleneck time are relative and therefore permit useful comparisons between populations. Furthermore, these differences have important implications for the SNP density and resolution of association studies. For example, young population isolates with extensive LD, such as Finland, may be ideal for whole-genome association studies as they are likely to require fewer markers than outbred populations but may identify larger candidate regions.

Because LDUs define the extent of useful LD across the genome, evenly spacing SNPs on the LDU scale should be optimal for association studies and ensure coverage of a region. The number of SNPs required for whole-genome association studies is therefore anticipated as some multiple of the number of LDUs in the genome that incorporates a range of allele frequencies. The various techniques of haplotype annotation fail to characterize interblock regions that may account for up to approx 30% of the genome and do not provide estimates of the number of SNPs required for whole-genome association.

Once a genetic determinant of disease susceptibility is localized to a specific region of the genome by linkage or LD, the identity and function of the gene(s)

responsible may be established. This process requires accurate maps of the human genome that convey physical and genetic distances, gene content and function, and location of polymorphic markers. The integration of LDU maps with genetic, physical, and cytogenetic maps, therefore, makes the LDDB an invaluable source for locating disease genes by linkage and association. The LDDB provides truly integrated maps that present locations for all loci on all scales that agree in physical order. The organization of the database into flat files is designed to deliver the maximum information content for disease gene mapping. The bioinformatic procedure of automated data retrieval from the internet and its organization into alias-cluster maps has increased the information content and ease of updating the database when sequence, LD, or linkage data improves. Functional data for each locus in the database can be obtained via links to various databases that detail expression, genomic background, and disease association through the list of aliases. The new and improved search tools have increased the functionality of the database. Notable additions include applications that allow for SNP selection, population comparison, and graphical representation of LD patterns. The ability to identify recombination hot spots and regions of atypical LD, such as extreme steps, blocks, or discrepancies between populations makes the LDDB an ideal starting point for studying the general characteristics of these regions and identifying any sequence motifs associated with them. Other search functions enable new physical locations to be pasted as a list, uploaded from a file, or determined by ePCR from primers to be interpolated into the integrated maps. A separate function has also been added to allow sequence extraction.

Early completion of the phase-I HapMap Project (~1 million SNPs genotyped, February 2005) has lead to the production of a new phase II release that aims to genotype approx 5 million SNPs. This will increase the density of SNPs across the genome from the current average of 1 every 3000 bases to about 1 every 600 bases and motivate renewal of the LD maps held in the LDDB along with probable updates to the physical map. This resolution should be sufficient to eliminate many of the holes present in the phase-I LD maps. However, if regions requiring more SNPs are identified by the kilobase scale rather than an LD-based unit, the phase-II data will resolve fewer holes including those (44–88) that occur in intervals less than 600 bp long. Updates to linkage maps are also likely especially in light of evidence that suggests that current maps underestimate interference especially for small chromosomes *(**12**)*.

## References

1 Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P. Y., and Collins, A. (2001) The optimal measure of allelic association. *Proc. Natl. Acad. Sci. USA* **98,** 5217–5221.

2 Collins, A., Ennis, S., Taillon-Miller, P., Kwok, P. Y., and Morton, N. E. (2001) Allelic association with SNPs: metrics, populations, and the linkage disequilibrium map. *Hum. Mutat.* **17,** 255–262.
3 Maniatis, N., Collins, A., Xu, C. F., et al. (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl. Acad. Sci. USA* **99,** 2228–2233.
4 Zhang, W., Collins, A., Gibson, J., et al. (2004) Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc. Natl. Acad. Sci. USA* **101,** 18,075–18,080.
5 Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29,** 217–222.
6 Zhang, W., Collins, A., Maniatis, N., Tapper, W., and Morton, N. E. (2002) Properties of linkage disequilibrium (LD) maps. *Proc. Natl. Acad. Sci. USA* **99,** 17,004–17,007.
7 Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.* **29,** 229–232.
8 Lonjou, C., Zhang, W., Collins, A., et al. (2003) Linkage disequilibrium in human populations. *Proc. Natl. Acad. Sci. USA* **100,** 6069–6074.
9 Gibson, J., Tapper, W., Zhang, W., Morton, N., and Collins, A. (2005) Cosmopolitan linkage disequilibrium maps. *Hum. Genomics* **2,** 20–27.
10 Jeffreys, A. J. and Neumann, R. (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat. Genet.* **31,** 267–271.
11 Tapper, W., Collins, A., Gibson, J., Maniatis, N., Ennis, S., and Morton, N. E. (2005) A map of the human genome in linkage disequilibrium units. *Proc. Natl. Acad. Sci. USA* **102,** 11,835–11,839.
12 Collins, F. S., Green, E. D., Guttmacher, A. E., and Guyer, M. S. (2003) A vision for the future of genomics research. *Nature* **422,** 835–847.
13 Dawson, E., Abecasis, G. R., Bumpstead, S., et al. (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418,** 544–548.
14 Tapper, W. J., Maniatis, N., Morton, N. E., and Collins, A. (2003) A metric linkage disequilibrium map of a human chromosome. *Ann. Hum. Genet.* **67,** 487–494.
15 Ke, X., Tapper, W., and Collins, A. (2001) LDB2000: sequence-based integrated maps of the human genome. *Bioinformatics* **17,** 581–586.
16 Gusella, J. F., Wexler, N. S., Conneally, P. M., et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306,** 234–238.
17 Matise, T. C., Sachidanandam, R., Clark, A. G., et al. (2003) A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *Am. J. Hum. Genet.* **73,** 271–284.
18 Kong, A., Gudbjartsson, D. F., Sainz, J., et al. (2002) A high-resolution recombination map of the human genome. *Nat. Genet.* **31,** 241–247.
19 Kong, X., Murphy, K., Raj, T., He, C., White, P. S., and Matise, T. C. (2004) A combined linkage-physical map of the human genome. *Am. J. Hum. Genet.* **75,** 1143–1148.

20 Collins, F. S. (1992) Positional cloning: let's not call it reverse anymore. *Nat. Genet.* **1,** 3–6.
21 Collins, F. S. (1995) Positional cloning moves from perditional to traditional. *Nat. Genet.* **9,** 347–350.
22 Altshuler, D., Brooks, L. D., Chakravarti, A., Collins, F. S., Daly, M. J., and Donnelly, P. (2005) A haplotype map of the human genome. *Nature* **437,** 1299–1320.
23 Thompson, E. A., Deeb, S., Walker, D., and Motulsky, A. G. (1988) The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes. *Am. J. Hum. Genet.* **42,** 113–124.
24 Gomes, I., Collins, A., Lonjou, C., et al. (1999) Hardy-Weinberg quality control. *Ann. Hum. Genet.* **63,** 535–538.
25 Ke, X., Hunt, S., Tapper, W., et al. (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet.* **13,** 577–588.
26 Gabriel, S. B., Schaffner, S. F., Nguyen, H., et al. (2002) The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229.
27 Phillips, M. S., Lawrence, R., Sachidanandam, R., et al. (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat. Genet.* **33,** 382–387.
28 Wang, N., Akey, J. M., Zhang, K., Chakraborty, R., and Jin, L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* **71,** 1227–1234.
29 Service, S., DeYoung, J., Karayiorgou, M., et al. (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association. *Nat. Genet.* **38,** 556–560.
30 Teare, M. D., Dunning, A. M., Durocher, F., Rennart, G., and Easton, D. F. (2002) Sampling distribution of summary linkage disequilibrium measures. *Ann. Hum. Genet.* **66,** 223–233.
31 Tenesa, A., Wright, A. F., Knott, S. A., et al. (2004) Extent of linkage disequilibrium in a Sardinian sub-isolate: sampling and methodological considerations. *Hum. Mol. Genet.* **13,** 25–33.
32 Collins, A. and Morton, N. E. (1998) Mapping a disease locus by allelic association. *Proc. Natl. Acad. Sci. USA* **95,** 1741–1745.
33 Collins, A., Lonjou, C., and Morton, N. E. (1999) Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl. Acad. Sci. USA* **96,** 15,173–15,177.
34 Lewontin, R. C. (1964) The interaction of selection and linkage. I. General considerations. *Genetics* **49,** 49–67.
35 Hill, W. G. (1974) Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33,** 229–239.
36 Collins, A., Lau, W., and De La Vega, F. M. (2004) Mapping genes for common diseases: the case for genetic (LD) maps. *Hum. Hered.* **58,** 2–9.
37 Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.
38 White, J. A., McAlpine, P. J., Antonarakis, S., et al. (1997) Guidelines for human gene nomenclature (1997). HUGO Nomenclature Committee. *Genomics* **45,** 468–471.

39 Cheung, V. G., Nowak, N., Jang, W., et al. (2001) Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409,** 953–958.
40 De La Vega, F. M., Isaac, H., Collins, A., et al. (2005) The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. *Genome Res*. **15,** 454–462.
41 Jeffreys, A. J., Neumann, R., Panayi, M., Myers, S., and Donnelly, P. (2005) Human recombination hot spots hidden in regions of strong marker association. *Nat. Genet*. **37,** 601–606.
42 Rao, D. C., Morton, N. E., Lindsten, J., Hulten, M., and Yee, S. (1977) A mapping function for man. *Hum. Hered*. **27,** 99–104.
43 Collins, A., Teague, J., Keats, B. J., and Morton, N. E. (1996) Linkage map integration. *Genomics* **36,** 157–162.
44 Broman, K. W., Rowe, L. B., Churchill, G. A., and Paigen, K. (2002) Crossover interference in the mouse. *Genetics* **160,** 1123–1131.
45 Charlesworth, B., Borthwick, H., Bartolome, C., and Pignatelli, P. (2004) Estimates of the genomic mutation rate for detrimental alleles in Drosophila melanogaster. *Genetics* **167,** 815–826.
46 Giannelli, F. and Green, P. M. (2000) The X chromosome and the rate of deleterious mutations in humans. *Am. J. Hum. Genet*. **67,** 515–517.
47 Ozaki, K., Ohnishi, Y., Iida, A., et al. (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet*. **32,** 650–654.
48 Klein, R. J., Zeiss, C., Chew, E. Y., et al. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* **308,** 385–389.
49 McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* **304,** 581–584.

# 4

# LDMAP

## *The Construction of High-Resolution Linkage Disequilibrium Maps of the Human Genome*

**Tai-Yue Kuo, Winston Lau, and Andrew R. Collins**

**Summary**

The precise characterization of the linkage disequilibrium (LD) landscape from high-density single-nucleotide polymorphism (SNP) data underpins the association mapping of diseases and other studies. We describe the algorithm and implementation of a powerful approach for constructing LD genetic maps with meaningful map distances. The computational problems posed by the enormous number of SNPs typed in the HapMap data are addressed by developing segmental map construction with the potential for parallelization, which we are developing. There is remarkably little loss of information (1–2%) through this approach, but the computation times are dramatically reduced (more than fourfold for sequential map assembly). These developments enable the construction of very high-density genome-wide LD maps using data from more than 3 million SNPs in HapMap. We anticipate that a whole-genome LD map will be useful for disease gene mapping, genomic research, and population genetics.

**Key Words:** Linkage disequilibrium maps; human genome; computational load; relative efficiency; segmental assembly.

## 1. Introduction

### *1.1. Linkage Disequilibrium Maps*

Linkage disequilibrium (LD, or allelic association), describes the statistical association between polymorphisms, such as single-nucleotide polymorphisms (SNPs), and between markers and genes contributing to disease. The existence of LD reflects transmission over many generations of short segments of ancestral haplotypes comprising closely linked markers. Allelic association is evident because haplotype frequencies are not simply the products of the appropriate

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

allele frequencies, hence "disequilibrium." LD is present because recombination, which destroys LD, is infrequent over small distances, whereas other processes, such as genetic drift and population bottlenecks, act to create LD over a number of generations. A thorough understanding of the extent and structure of LD is essential for association mapping of the polymorphisms that contribute to human diseases. Given the availability of substantial bodies of high-resolution SNP data (for example, from the International HapMap Project, http://www.hapmap.org/; ***[1]***), it is now possible to characterize LD patterns genome wide. Once the structure is characterized there are likely to be substantial payoffs from increased resolution and power for localization of disease genes ***(2)***, and for identifying genomic regions subject to selection ***(3)***.

It is known that LD extends for tens of kilobases, on average, in the human genome. This is true even for large heterogeneous human populations and not just isolates ***(4)***, suggesting that the genome might be screened with reduced numbers of SNPs because close association implies some redundancy. This is the main motivation behind the HapMap Project, which aims to identify "tag" SNPs to represent a particular haplotype with little loss of power, a strategy relying on recognition that some parts of the genome contain regions (blocks) of low haplotype diversity ***(5)***. However, much of the genome is more complex, reflecting the combined effects of intense recombination hot spots, more randomly distributed recombination events, and other phenomena. Furthermore, the definition of block boundaries and the instability of blocks defined with different marker densities poses difficulties ***(6,7)***. It is also evident that a "haplotype map" ***(8)***, although providing annotation, is not a genetic map with meaningful distances that describe LD structure.

A successful alternative strategy is to represent LD patterns in the form of a metric map with additive "linkage disequilibrium unit" (LDU) distances ***(9)***. The low-resolution features of LD maps resemble the linkage map in pattern but there are important differences, which reflect population history. A whole-chromosome LD map of chromosome 22 ***(6)*** shows a close correspondence between areas of extensive LD with low recombination and areas of low LD with intense recombination. LD maps have already been used for multilocus disease gene mapping using locations on the LDU scale as the association mapping analog of the linkage map for localizing major genes ***(9,10)***. LD units are analogous to centimorgans (cM) in that locations increase monotonically with physical distance but, although linkage map length is related to recombination in one generation, the LDU map length reflects accumulated recombination over many generations. The ratio of the LDU map length to the linkage map in Morgans estimates the effective number of generations over which recombination has occurred (the "effective bottleneck time;" ***[11]***), with some distortion in the LD map because of selection and systematic errors in estimating interference in the linkage map.

Algorithms to construct LD maps have been developed and evaluated by Maniatis et al. *(9)* and Lonjou et al. *(4)*. The LDMAP program (http://cedar.genetics.soton.ac.uk/public_html/), described here, implements and extends these algorithms. We describe an approach for the construction of a genome-wide LD map at very high density by addressing the particular computational difficulties posed by the analysis of huge numbers of markers.

### 1.2. Overview of the Basic Algorithm

The population genetics theory behind LD map construction is described by Morton et al. *(12)*. The decline of LD, modeled as association ρ as a function of distance d, in kilobases, is $\rho = (1–L)Me^{-\varepsilon d} + L$, in which the L parameter reflects residual association at large distance not from linkage, M is the intercept, the association at zero distance, and ε is the exponential decline of LD as the product of recombination θ and number of generations t. The model has the same form as that developed by Malecot *(13)* to describe genetic isolation by distance but has different parameters.

LD map construction estimates ε in each map interval between adjacent SNPs. For any pair of SNPs the association probability ρ and the information $K_\rho$ form the data for LD map construction. Pairs that span a given interval contain information about association in that interval, but pairs at large distances are uninformative. The estimation of the ε vector requires the iterative substitution of distance d in the Malecot equation with distances in LDUs. These are defined, for the $i^{th}$ interval between adjacent SNPs, as $\varepsilon_i d_i$ with locations by summation over preceding intervals *(9)*. The LDU locations, when plotted against kb, typically show a pattern of steps where LD is breaking down and plateaus or blocks of high LD.

## 2. Methods

### 2.1. Model Implementation

The raw data comprise SNP genotypes (diplotypes) from unrelated individuals with alleles coded 0 (missing), 1, and 2. Alternatively, where known with a high degree of reliability, SNP haplotypes are used. The physical location, in kilobases from an origin closest to the p telomere for each SNP is obtained from the latest human genome sequence release.

The genotypic data are reduced to pairwise association and the corresponding information *(14,15)*. Informative SNP pairs are selected subject to two constraints, of which the minimal set is used in the analysis. The first is the maximum distance in kilobases between any pair of SNPs, defaulted to 500 kb. This eliminates pairs separated by a distance that greatly exceeds the range of LD in most human populations, although for isolated populations, certain genomic regions and for building LDU maps of other organisms this constraint

may not be appropriate. For sub-Saharan African populations, and genomic regions with a high recombination rate, the 500-kb distance is excessive but inclusion of these pairs only impacts on computation time. However, at the SNP densities available in the HapMap data this constraint is much less important than the second constraint, which restricts the number of map intervals between any pair of SNPs. To compute ε for a given interval between adjacent SNPs, a pair that spans that interval is potentially informative but the information approaches zero if the number of intervals between the pair is large. To reduce the computational load the default maximum number of intervals, between a pair of SNPs informative for a given intervals is 100. Therefore, for the computation of ε, there is a sliding window that encompasses all the informative pairs that span the interval. When the maximum number of intervals constraint is operating (and no pairs are eliminated by the maximum distance constraint) the total number of pairs used (N) in a map of n SNPs is:

$$N = \frac{n(n-1)}{2} - \frac{(n-s-1)(n-s-2)}{2}$$

To compute ρ for SNP pairs from diplotype data we apply the expectation maximization (E.M.) algorithm of Hill ***(16)***, which iteratively reduces a $3 \times 3$ table of genotypic counts to four haplotype frequencies. These are converted to counts and a file that specifies the SNP pair, and the sequence locations in kilobases, together with the four counts, is produced. Because no rearrangement of the $2 \times 2$ table has taken place at this point the four counts correspond to the 11, 12, 21, and 22 haplotypes from the marker pair. This file can be concatenated with corresponding files from other populations and counts summed for shared marker pairs, assuming alleles are labeled consistently. The summed counts have been used to compute ρ for construction of "cosmopolitan" maps ***(4,17)***.

Rare SNPs with minor allele frequencies less than 0.05 are eliminated, as are any that show strong deviation from Hardy–Weinberg equilibrium ***(18)***. The association probability ρ is obtained by rearranging the $2 \times 2$ table (**Table 1**) to ensure that Q is the minimal allele frequency (Q < R, 1–R, and 1–Q) and that products of haplotype frequencies give ad > bc. Conforming to this rearrangement requires the relabeling of SNPs ($SNP_1$ becoming $SNP_2$ and vice versa) and/or relabeling of the SNP alleles. To achieve Q < R, markers are interchanged by switching b and c, which has the effect of exchanging Q with R and 1–Q with 1–R; for Q < 1–R markers are interchanged by switching a and d, which has the effect of exchanging Q with 1–R and 1–Q with R; for Q<1–Q alleles are interchanged (a with c and b with d) which switches Q with 1–Q. Finally, to conform to ad > bc, alleles are interchanged, a with b and c with d, which switches R with 1–R. Columns are also interchanged in the special case that disequilibrium D is zero, where b > a. The "intermediate" file used by the

**Table 1**
**Haplotype Frequencies (a, b, c, and d) for a Pair of SNPs**

| | | $SNP_2$ alleles | | |
|---|---|---|---|---|
| | | 1 | 2 | |
| | 1 | a | b | Q |
| $SNP_1$ alleles | | | | |
| | 2 | c | d | 1–Q |
| | | R | 1–R | |

The table is ordered such that: Q, 1–Q are allele frequencies at $SNP_1$, where Q < (1–Q, R, 1–R) and R, 1–R are allele frequencies at $SNP_2$ and ad > bc. D = ad–bc; $\rho = D/Q(1–R)$; $K_\rho = mQ(1–R)/R(1–Q)$, where m is the sample size for the pair of SNPs; $\chi^2 = \rho^2 K_\rho$.

program specifies the SNP pair, sequence locations (Kb), ρ, $K_\rho$, $\chi^2$, sample size m, Q, R, D, and the pair selection criteria (maximum number of intervals, maximum window size in kilobases).

### 2.2. Fitting Data to the Kilobase Map

From the intermediate file the fit of the pairwise data to the kilobase map under the Malecot model is established. Pairwise data enter composite log likelihood as:

$lnlk = -\Sigma K_\rho(\hat{\rho} - \rho)^2/2$, where the summation is over informative pairs (i = 1, N), ρ is the observed association between the $i^{th}$ pair (**Table 1**), and $\hat{\rho}$ are the fitted values. Function minimization is achieved using the variable metric method implemented in the subroutine *dfpmin* (*see* Chapter 10; ***[19]***). Parameter estimation for ε, L, and M is controlled through a script, which allows testing of hypotheses such as deviations from L = 0 or M = 1. In general, two models (A and B) can be compared as $\chi^2_n = (-2lnlk_A - -2lnlk_B)/V_B$, where model B has one or more additional parameters estimated than the simpler model A. $V_B$ is the error variance of model B defined as $V_B = -2lnlk_B/(N-g)$, where N is the number of pairs and g is the number of parameters estimated.

Morton et al. ***(12)*** defined a predicted value for the L parameter (Lp), which is equal to the $K_\rho$-weighted mean of $\sqrt{2/\pi K_\rho}$ where $K_\rho$, the information about ρ per marker pair, is proportional to sample size. Lp depends only on the mean value of ρ for markers at large distances such that the expected value of disequilibrium D is zero.

### 2.3. Construction of a LD Map

The Malecot parameters from the kilobase map provide starting values for construction of the LD map. The iterative process implemented in LDMAP estimates ε, for intervals between adjacent SNPs, following the "interval" method described by Maniatis et al. ***(9)***. Briefly, let $S_{hk} = \Sigma \varepsilon_i d_i$ where i is an interval between adjacent SNPs and summation is over all intervals contained between

SNPs h and k and $\rho_{hk}=(1–L)Me^{-S_{hk}}+L$ using trial values for M, L, and $\varepsilon_i$ as described previously. The estimate of $\varepsilon_i$, at iteration t, is given by:

$$\varepsilon_i^{(t)}= \varepsilon_i^{(t-1)}+(U_i/K_i)^{(t-1)}, \text{ where } U_i = \sum\left(\frac{\partial \ln lk}{\partial \rho_{hk}}\right)\left(\frac{\partial \rho_{hk}}{\partial \varepsilon_i}\right) \text{ and } K_i = \sum K_{\rho hk}\left(\frac{\partial \rho_{hk}}{\partial \varepsilon_i}\right)^2.$$

At convergence each revised estimate $\varepsilon_i$ contributes toward a "global" iteration, which is a complete update of the ε vector and the computation of the global composite likelihood, which is maximized iteratively. The M parameter is assumed constant for all intervals and is updated periodically at global iterations 25, 50, 100, 200, 400, 800, 1600, and so on. At these points the composite log likelihood for the LD map (–2lnlk) is obtained. This updating procedure accelerates convergence. Experience with LD map construction has shown that the estimated L exceeds Lp in small samples. This might be attributed to the local effect of block structure, which can distort L ***(4)***, therefore Lp is used throughout in LDMAP.

Typically there are small numbers of intervals where $\varepsilon_i d_i$ exceeds 3, termed "holes" ***(6)***. In high-density maps holes are associated with a locally high recombination rate, and segments requiring local increases in marker density can thus be identified ***(17)***. When an estimate $\varepsilon_i$ bounds at zero (consistent with "complete" LD), that estimate is fixed at zero and no further iteration takes place, with a consequent reduction in computation time. We have found that removing these intervals from further iteration has very little effect on the final map, suggesting that most estimates remain at the zero limit once reached. The same applies to holes, and these intervals are also dropped from further iteration. However, the constraints are not applied until a "burn-in" period corresponding to 50 global iterations has taken place. Convergence is declared when a difference in global composite likelihood between two consecutive iterations is less than 0.01.

### *2.4. Toward a Genome-Wide LD Map*

In a map of n loci there are n–1 estimates ε, achieved through maximizing the composite likelihood for which the computation time may be substantial. The computation time depends on a number of factors, but particularly the number of pairs used in map construction. Exclusion of pairs that contain no significant information about a given interval is one approach to reducing computation time. However in maps with many tens of thousands of loci the exclusion of these pairs is inadequate for constructing maps within an acceptable time frame. We have examined a number of alternatives to reduce computation time, including the construction of maps at adaptively increasing densities. However, this was found to offer only modest speed enhancements. The assembly of maps in overlapping sections, with distances averaged in the overlap region, is

**Table 2**
**Maps of Chromosome 22 Constructed Using Different Numbers of Segments**

| Number of segments | Loci per segment | Map length LDU | Error variance | Computation time, minutes |
|---|---|---|---|---|
| 1 | 13,959 | 1017 | 0.841 | 2471 |
| 2 | 6980 | 1017 | 0.842 | 2348 |
| 6 | 2327 | 1022 | 0.842 | 859 |
| 14 | 997 | 1024 | 0.843 | 760 |
| 20 | 698 | 1037 | 0.847 | 733 |
| 40 | 349 | 1040 | 0.853 | 509 |
| 60 | 233 | 1055 | 0.851 | 409 |
| 100 | 140 | 1054 | 0.870 | 408 |
| 200 | 70 | 1089 | 0.886 | 177 |

much more promising and we here examine the impact of this approach on the quality of the map. For this evaluation the September 2004 release of the HapMap data for chromosome 22 was used. The Centre d'Etudes du Polymorphisme Humain (CEPH) sample comprises 9658 loci from 60 unrelated individuals of western European ancestry. We constructed nine LD maps of chromosome 22 for a range of numbers of segments between 1 (the complete map constructed in one piece) and 200 pieces (**Table 2**). We computed the error variance for each map by testing the fit of a "standard" set of pairwise data for each chromosome (with default settings of 500 kb as the maximum window size and 100 as the number of intervals). This enabled direct comparison of the relative efficiency of alternative numbers of map segments and the relationship to computing time. For each map the error variance, V, was computed and the efficiency of each map was computed relative to the map constructed in one piece. We also looked at the relative computation time and relative resulting map lengths in the same way.

## 3. Results

The LDU map length (**Table 2**) is rather stable showing a maximum increase in length of approx 7% over the range of tests (maps constructed in 1–200 segments). The error variances are similarly stable varying over the range 0.841 to 0.886. The relative efficiency (**Fig. 1**) is therefore high across the range with a maximum loss of information of only approx 5% when the map is constructed in 200 segments and much reduced losses for maps built in fewer segments (for example, the loss of information is less than 1% for maps built from segments of 698 loci). **Figure 2** plots the contour of the LDU maps constructed as one piece in contrast to the LDU map constructed from 200

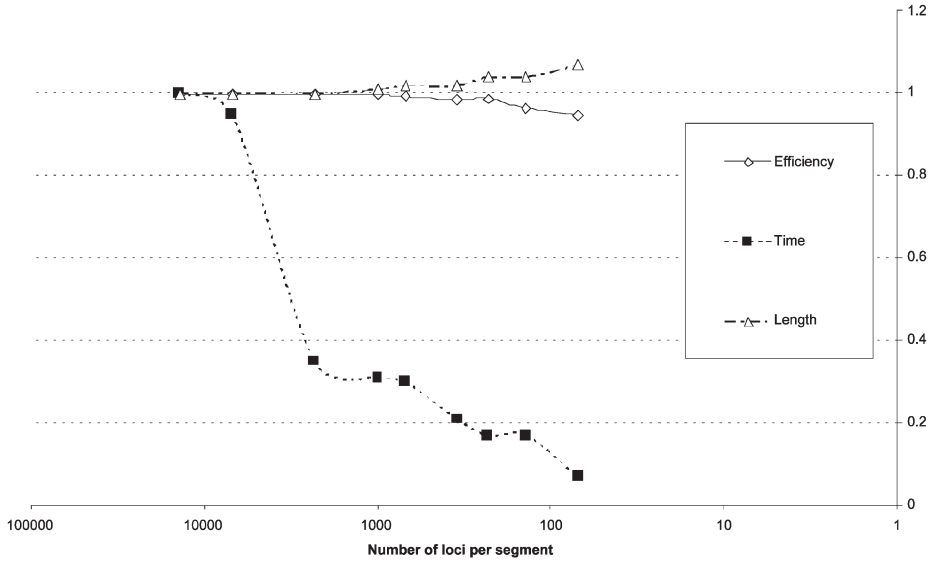



Fig. 1. For maps constructed in segments there is a modest loss of information but dramatic savings on computing time, suggesting that segments of 100–1000 loci might be optimal.

segments. Both contours show details of the LD structure of chromosome 22 including two large regions (plateaus) of at least 5 Mb at around 30 and 40 Mb along the chromosome where there is extensive LD. There are also three to four regions of rather intense recombination, the most striking of which is around 26 Mb along the chromosome. The contour for the two maps is strikingly similar and the small difference in map length appears to be spread over the whole length of the map. This suggests that the segment approach slightly exaggerates map length because of the loss of information at the ends of segments where there are no flanking SNPs. It is important to balance computational feasibility and number of segments for map construction. Although the LDU maps are stable and the loss of information is minimal, the computation times of a SUN V440 server vary enormously over the range. The computing time for the map when constructed in one piece is 14-fold greater than for the map constructed from 200 segments. It is evident that a good compromise between optimal computational times and minimizing information loss in the map is achieved in the 100 to 1000 loci per segment range. The use of approx 500 loci per segment seems justified for map construction generally and the construction of maps of the largest chromosomes becomes feasible, even at the higher SNP densities of the later HapMap releases.

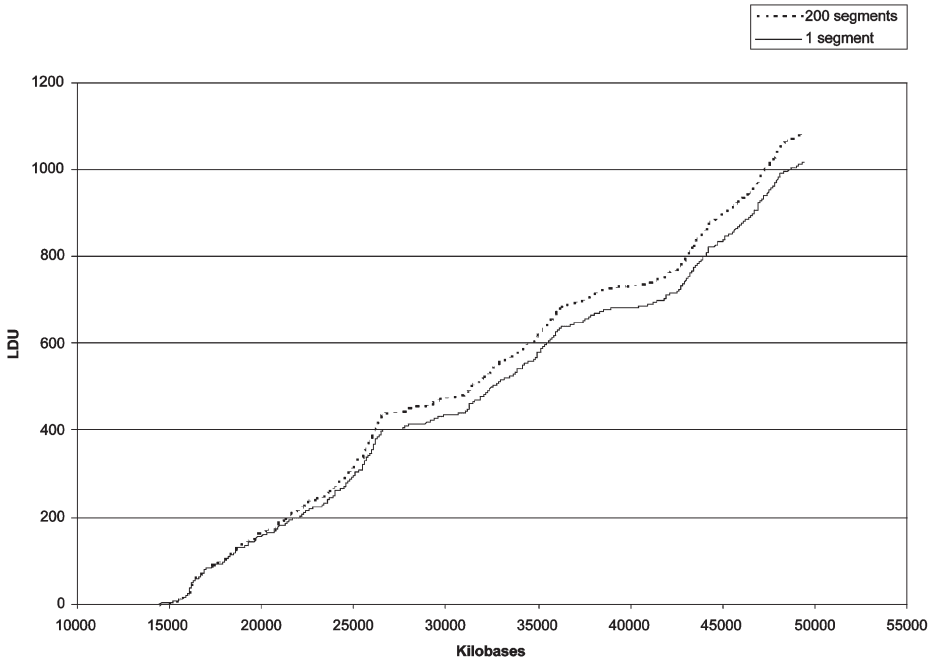



Fig. 2. The linkage disequilibrium maps of chromosome 22 are closely similar, whether constructed in one piece or in many segments. The latter shows modestly longer linkage disequilibrium unit maps.

## 4. Discussion

The results of the analysis show that LD maps are robust to segmental assembly of LD maps with very little loss of information even for the smallest segment sizes. This justifies the construction of a genome-wide LD map using selected pairs and through segmental map assembly. The results demonstrate that genome-wide LD maps are achievable even at the highest marker densities including the HapMap data with greater than 3 million SNP genotypes for a range of populations. The mean extent of LD in the CEPH sample is approx 50 kb, implying a mean spacing of approx 3 kb in a 1 million SNP map. A pair of SNPs separated by 100 intervals will span approx 300 kb and, therefore, imposing this limit will result in little or no loss of information at HapMap densities. However, when applied to a map of 50,000 SNPs, which will be exceeded for the larger chromosomes, there will still be more than 2.5 million pairs for analysis. Because of the computational load and the modest loss of information we have demonstrated, map assembly using overlapping segments is a practical approach.

We have considered the performance of sequential map assembly using segments. However, maps can be constructed in parallel with consequent dramatic

increases in the speed of assembly. Map assembly in overlapping segments is ideally suited to GRID computing (for example using the Condor program, http://www.cs.wisc.edu/condor/) and might be profitably achieved through a world wide web-based tool. A GRID-enabled version is currently being tested.

## References


1 International HapMap Consortium. (2003) The International HapMap Project. *Nature* **426,** 789–796.

2 Maniatis, N., Morton, N. E., Gibson, J., Xu, C. F., Hosking, L. K., and Collins, A. (2005) The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. *Hum. Mol. Genet.* **14,** 145–153.

3 Sabeti, P. C., Reich, D. E., Higgins, J. M., et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837.

4 Lonjou, C., Zhang, W., Collins, A., et al. (2003) Linkage disequilibrium in human populations. *Proc. Natl. Acad. Sci. USA* **100,** 6069–6074.

5 Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.* **29,** 229–232.

6 Tapper, W. J., Maniatis, N., Morton, N. E., and Collins, A. (2003) A metric linkage disequilibrium map of a human chromosome. *Ann. Hum. Genet.* **67,** 487–494.

7 Ke, X., Hunt, S., Tapper, W., et al. (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet.* **13,** 577–588.

8 Dawson, E., Abecasis, G. R., Bumpstead, S., et al. (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418,** 544–548.

9 Maniatis, N., Collins, A., Xu, C. F., et al. (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl. Acad. Sci. USA* **99,** 2228–2233.

10 Maniatis, N., Collins, A., Gibson, J., Zhang, W., Tapper, W., and Morton, N. E. (2004) Positional cloning by linkage disequilibrium. *Am. J. Hum. Genet.* **74,** 846–855.

11 Zhang, W., Collins, A., Gibson, J., et al. (2004) Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc. Natl. Acad. Sci. USA* **101,** 18,075–18,080.

12 Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P. Y., and Collins, A. (2001) The optimal measure of allelic association. *Proc. Natl. Acad. Sci. USA* **98,** 5217–5221.

13 Malecot, G. (1948) *Les Mathematiques de l'Heredite*, Maison et Cie, Paris, France.

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

15 Collins, A., Lonjou, C., and Morton, N. E. (1999) Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl. Acad. Sci. USA* **96,** 15,173–15,177.

16 Hill, W. G. (1974) Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33,** 229–239.

17 Gibson, J., Tapper, W., Zhang, W., Morton, N., and Collins, A. (2005) Cosmopolitan linkage disequilibrium maps. *Hum. Genomics* **2,** 20–27.
18 Gomes, I., Collins, A., Lonjou, C., et al. (1999) Hardy-Weinberg quality control. *Ann. Hum. Genet.* **63,** 535–538.
19 Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1994) *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK.

# 5

# Linkage Disequilibrium as a Tool for Detecting Signatures of Natural Selection

**Sarah Ennis**

**Summary**

Natural selection has been theoretically and empirically proven to alter patterns of linkage disequilibrium (LD). Reciprocally, recombination, the driving force behind LD, modifies the signature of natural selection by reintroducing variation in a punctuate manner across the genome. To date, efforts to identify genes that have been subjected to historical selective pressure by examining polymorphic variation and allelic association have frequently fallen short of unambiguously distinguishing selection from other biological mechanisms. Contemporary genetic maps that describe LD in fine detail represent a much needed tool that can be exploited by researchers aiming to tease apart these opposing signals.

**Key Words:** Linkage disequilibrium; selection mapping; neutral theory; LD maps.

## 1. Introduction

The genetic code holds the historical record of the adaptive evolution of all living things. From a human perspective, deciphering the clues held by DNA both within and between humans and other species provides a wealth of information on our origins and the adaptive changes to our molecular make-up over the intervening time.

Identification of genomic regions having undergone selection, whether these represent genes or other fundamental untranslated code, can give particular insight into our adaptive divergence from our closest ancestor the chimpanzee, and help discern the natural mechanisms of disease resistance. Characterization of genes involved in adaptive changes to pathogens can provide information for: (1) accurate diagnoses/prognoses; (2) early (prenatal) detection of disease susceptibility genes; (3) more informed drug/vaccine design; and (4) customized pharmacogenetics.

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

As the predominant force governing the pattern of linkage disequilibrium (LD), recombination represents the singularly greatest biological mechanism that shapes the architecture of the genome. On the back of the human genome-sequencing project and massive biotechnological advances, researchers are uniquely poised to annotate the human genome with the specific patterns of LD on the kilobase scale. The potential to map more subtle disturbances caused by selection will be greatly improved when executed on a platform of accurate genetic (LD) maps.

## 2. Neutral Theory

> "In the survival of favored individuals and races, during the constantly recurring struggle for existence, we see a powerful and ever-acting form of selection."
>
> Charles Darwin

Darwin's theory of natural selection entered a new phase in population genetics subsequent to the identification of DNA and the advent of molecular genetics, which permitted the unambiguous connection between phenotype and genotype. Building on the seminal work of Haldane and Wright, Kimura formulated his random drift theory of molecular evolution–the neutral theory *(1)*. Kimura controversially asserted "the very high rate of nucleotide substitution which I have calculated can only be reconciled…..by assuming that most mutations produced by nucleotide replacement are almost neutral in natural selection." His theory, also known as the neutral mutation–random drift hypothesis, assumes: (1) the vast majority of new mutation is selectively neutral or nearly neutral; (2) the majority of these mutants, including those with a small selective advantage are quickly lost from the population and only a fractional remainder randomly drift to fixation; (3) the probability of eventual fixation of a neutral allele corresponds to 1/(2N) (where N represents the actual population size) and time taken for complete fixation is calculated as $4N_e$ generations (where $N_e$ represents the number of reproducing individuals per generation); and (4) if new alleles appear at a site at rate v, then the average length of time between consecutive substitutions in the population is 1/v generations *(2)* (**Fig. 1**).

Kimura further deduced that a reduction in observed amino acid substitution in the hemoglobin gene was because of the necessary elimination of deleterious alleles at the locus and argued that such selective constraint of functionally important molecules and not positive selection was the primary force governing natural selection. These observations laid the groundwork for subsequent detection of selection whereby selective neutrality observed at noncoding regions could be used as the null hypothesis against which tests for evidence of selection could be performed.

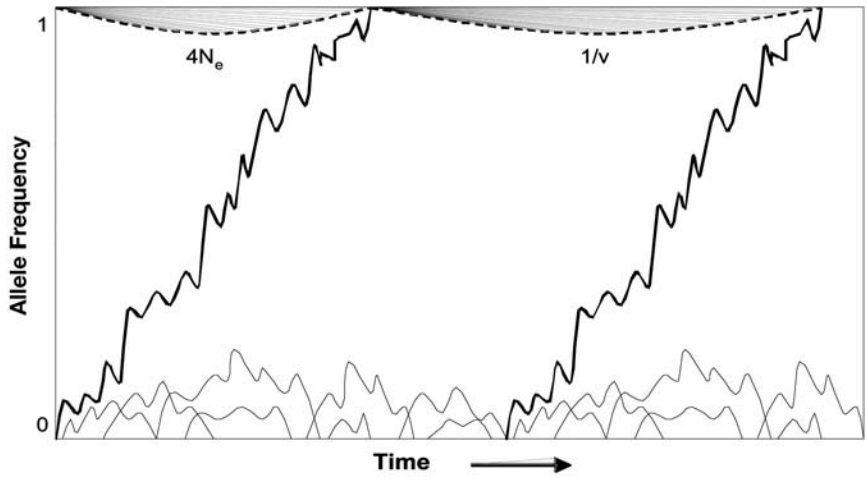



Fig. 1. Graphical representation depicting fluctuations in allele frequencies over time. The time taken for complete fixation is calculated as $4N_e$ generations (where $N_e$ represents the number of reproducing individuals per generation or *effective* population size). The average length of time between consecutive substitutions in the population is 1/v generations where v represents the mutation rate.

## 3. Forms of Selection

The direction of selection may be positive or negative. In order for a mutation to be driven out of the gene pool by negative selection, it must first represent a deleterious change at a functionally important site, most likely a coding region. However, even within the coding region of a gene, there are many forms of mutation without significant detrimental effects. Synonomous changes have no effect on which amino acid is finally translated into protein, and non-synonomous changes can also have little effect on fitness when they occur. Holmquist et al. ***(3)*** showed that for the region coding the helical loops of transfer RNA there is little selective constraint beyond that of maintenance of hydrogen bonding along the stem section of the final clover-leaf-shaped protein. However, despite the rarity of deleterious mutations compared with neutral mutations, there are still abundant opportunities for these variations to occur within functionally integral regions and these will be rapidly expunged from the gene pool. This was elegantly demonstrated when Kimura showed the evolutionary rate on the surface of the hemoglobin α protein was 10 times greater than that within the heme pocket ***(4)*** and argued that rather than this being because of differential mutation rates at the two sites, that the strong functional constraint at the heme pocket meant fewer mutations were likely to be neutral and randomly carried to fixation. Instead, those mutations that had occurred at

this critical site were more likely to compromise binding and so be eliminated at a rate proportional to their negative selective pressure. This represents a model where there is positive correlation between increasing sophistication and intricacy in protein function and the frequency of deleterious mutations. Elimination of neutral polymorphism as a result of negative (or purifying) selection against deleterious mutations at a linked site is referred to as background selection.

The very existence of a successfully reproducing organism indicates a necessarily high level of molecular adaptation to environment. As experimental evidence in support of neutral theory grew, it became increasingly apparent that positive adaptive selection made only a fractional contribution to evolutionary change. Even between distantly related species there is often negligible variation at functionally important sites indicating variants conferring a distinct advantage (and rising in frequency to fixation) are extremely rare. Haldane estimated that given one of these very rare advantageous mutants, the probability of it becoming fixed in a population was approximately twice its selection coefficient when the advantage was small (***5***), i.e., given a mutation conferring a 2% selective advantage, there is only a 4% chance it will become fixed in a population. As mutations with large effects tend to be deleterious, advantageous mutations are most likely to have small effects, meaning that of all the positive mutations that have occurred in our history, only a small proportion of them are now evident in our DNA. However, given a functionally less stringent genomic segment (such as a duplicated gene), the probability of a beneficial mutation occurring increases.

Kimura described polymorphisms as "a transient phase in molecular evolution" and for the most part this accurately reflects the path a new mutation will have to fixation (by random drift of neutral mutations or positive selective pressure), or extinction (the more common fate of neutral and deleterious alleles). However, it is possible to have both a positive and negative selective pressure acting on a mutation such that its frequency in the population is maintained between zero and one over many generations. This phenomenon is called balancing selection and infection by the malaria parasite is the key to one of the best studied examples in humans. A T>A polymorphism in the coding region of the hemoglobin β gene results in a glutamic acid-to-valine substitution at position 6 of the protein. Homozygotes for this HbS variant are susceptible to ischemia and infarction as their red blood cells polymerize abnormally and adopt a sickle shape rather than the normal discoid conformation upon deoxygenation. These sickled red blood cells obstruct capillaries, cause anemia, and reduce blood flow to vital organs. As life expectancy is often reduced with this disease, one might expect a strong negative selective coefficient to expunge the variant from the gene pool. Heterozygote carriers of the HbS allele have mild symptoms that

manifest under oxygen deprivation or severe dehydration. However, in regions where malarial infection by *Plasmodium falciparum* is endemic, carriers exhibit a heterozygote advantage in that they have increased resistance to infection by malaria as the mutant allele impairs parasite cell entry and growth. These opposing pressures mean the selective disadvantage apparent in homozygotes is countered by a survival advantage to carriers and the mutant allele is maintained at frequencies of up to 40% in sub-Saharan Africa, the Middle East, and parts of the Indian subcontinent ***(6)***.

## 4. Selective Pressure

Recent human history has been defined by rapid expansion since the spread of agriculture and animal breeding subsequent to the end of the last ice-age about 10,000 yr ago. This cultural transition from hunter-gatherer to farmer not only resulted in a change of diet and culture, but led to increasing population density and concomitant exposure to new pathogens. These combined alterations to our environment would have rendered our Palaeolithic genomic composition favorable to adaptive change.

A classic model of adaptive genetic change to environmental pressure is that of the lactase gene. The matching geographic distributions of lactase persistence and dairy farming has long been recognized ***(7)***. More recently, Bersaglieri et al. ***(8)*** identified a common haplotype (77% frequency in individuals with north-west European ancestry), which "extends largely undisrupted for >1 Mb." Population comparisons of the extent of LD across this long tract provided data to estimate positive selective pressure between 5000 and 10,000 yr ago, coincident with the integration of animal milk as a major protein and carbohydrate source in the human diet.

Conversely, the *L-gulonolactone oxidase* (*LGO*) gene emerged in early terrestrial vertebrates and functioned in the terminal pathway for synthesis of vitamin C—a vital antioxidant. However, this enzyme is no longer functional in humans. Jukes and King ***(9)*** proposed that the degeneration of this characteristic ability in humans, primates, and fruit bats resulted from a relaxation of the selective constraint when vitamin C became a common dietary component. An abundant environmental source of vitamin C would have rendered mutations in the enzyme selectively null and free to go to fixation. Interestingly, Nandi et al. ***(10)*** identified a comparative increase in superoxide dismutase (*SOD*) activity in mammals lacking the functional *LGO* gene. Whether this upregulation of *SOD* preceded *LGO* degeneration and contributed to the reduced selective constraint on *LGO* or was a consequence of *LGO* gene loss and increased pressure, *SOD* remains an unsolved chicken-and-egg conundrum.

Removal of selective pressure and degeneration of the genetic code is also evident at the human olfactory receptors. Analysis of redundancy at a number

of the bitter taste receptors (TAS2Rs) in humans estimates functional relaxation at these loci at 0.75 million years ago, which coincides with controlled use of fire by hominids. Significant detoxification of poisonous food by cooking and an increasingly carnivorous diet is likely to have reduced the necessity to detect at least some bitter-tasting toxins and render gene-degenerating mutations selectively null *(11)*. However, as the hominid diet changed over time, exposure to new toxins was also likely and evidence for positive selection at the *TAS2R16* gene, which facilitates taste of harmful cyanogenic glycocides, has recently been presented *(12)*. These fluctuating pressures at the human taste receptors provide intriguing examples of the dynamic nature of human adaptation.

## 5. Detecting Signatures of Selection

### *5.1. Common Techniques*

Over the past 50 yr many methods have been developed to detect selection from population genetic data. Under standard neutrality with a randomly mating population of constant size and without substructure, a summary of the observed allele frequencies of segregating mutations is known as the "frequency spectrum." Many tests have been developed to detect deviations from this distribution as expected under neutrality—the most popular of which is Tajima's D *(13)*. This test compares the difference between S, the number of segregating sites in a sample and $\pi$, the average pairwise difference in the number of nucleotides. An excess of low-frequency polymorphisms compared with equilibrium neutral expectations results in a negative D statistic, rejection of the null hypothesis, and is indicative of a recent selective sweep. Conversely, a skew in the frequency spectrum toward an excess of intermediate-frequency polymorphisms gives a positive D statistic and is indicative of balancing selection. However, inappropriate demographic assumptions may confuse the results as population bottlenecks/expansions and population substructure will give rise to positive and negative D statistics, respectively. Other popular tests of neutrality include $F_{st}$ *(14)* and its derivatives, Fu and Li's D, and F, *(15)* which apply coalescent theory and incorporate outgroups from related species but similarly to Tajima's D, their reliance on the frequency spectrum makes it difficult to unambiguously distinguish between selection and population demographics. Often applied for multiple loci, the HKA test *(16)* compares within-species polymorphism and between-species divergence and again assumes no recombination within loci. The McDonald–Kreitman *(17)* test is less sensitive to demographic changes and compares the ratio of nonsynonomous-to-synonomous polymorphisms within species to that between species. A comprehensive review of these methods is beyond the scope of this chapter but these are examined thoroughly elsewhere *(18,19)*.

Despite an arsenal of tests for selection, results are often contentious or inconsistent. Of particular note was the case for the *CCR5* gene, a deletion

allele that confers strong resistance to HIV infection. Evidence was presented from analyses of nonsynonomous mutation rates *(20)* and coalescent theory *(21)* supporting a rise in frequency of this Δ32 allele in Europeans approx 700 yr ago because of some strong selective agent such as Bubonic plague. Recent reevaluation of the data using more dense genetic maps has shown preliminary results may have been misleading because of inappropriate stratification of the data and that there is no detectable evidence of recent selection at the locus *(22)*. Other complexities in interpreting evidence for balancing selection have been highlighted by Kreitman et al. *(23)* where the misleading effects of single-nucleotide polymorphism (SNP)-ascertainment bias in coalescent models may have led to erroneous results.

### 5.2. LD and Hitchhiking

Maynard-Smith and Haigh *(24)* extended the population genetics theory on positively selected mutations by describing the "hitchhiking effect." Hitchhiking is the elimination of allelic variation at neutral sites linked to a selected site as the whole haplotype rises in frequency. When a new mutation occurs, it is in complete LD with all variants along the same chromosome. Over subsequent generations, as the advantageous mutation rises in frequency, there is limited opportunity for meiotic events to exchange genetic information. Only genomic variants sufficiently separated from the mutation by a recombination hot spot will escape elimination and maintain diversity. This reduction in diversity is further confounded when the time to fixation is sufficiently short that there is little opportunity for new variation in the form of new mutation to appear on the haplotype. For these reasons, positive selection and concurrent hitchhiking of linked sites has the effect of increasing LD while at the same time reducing local variability. The extent of this increased LD over the region will be a function of the local recombination rates.

The hitchhiking theory underpins the effects selection may have on LD; however, an often neglected issue in selection mapping is the reciprocal impact recombination, the principle force governing LD *(25)*, has on efforts to identify selected genes. Recombination promotes increased variation through recurrent exchanges between ancestrally diverse chromosomes. Methods to detect and understand selection that consider the form and frequency of polymorphic data and/or apply coalescent theory of genomic sequences must incorporate information on the underlying location and rates of recombination. Models that ignore this fundamental biological characteristic will be susceptible to ambiguous and/or misleading results.

### 5.3. LD-Based Methods for Detecting Selection

Linkage maps were groundbreaking tools of the 1980s and 1990s and indispensable in the quest to map major genes but their low resolution rendered them

less useful in providing the fine-scale information required to assess recombination within genes. Since the Human Genome Mapping Project, the potential to study selection as a function of LD had greatly increased. Accurate localization of abundant SNP markers combined with diminishing costs of technology to genotype SNPs in large data sets have made the necessary raw data more accessible. The subsequent availability of high-resolution LD data across the genome led to a number of new methods to detect signatures of selection.

In 2002, Sabeti et al. ***(26)*** introduced a method to detect recent positive selection in humans, which was dependent on the relationship between an allele's frequency and the extent of LD surrounding it. When a mutation first occurs it is in complete LD with linked markers but it is incredibly rare. Neutral mutations may rise in frequency to fixation, but this process is very slow and there is ample opportunity for both recombination to reduce the extent of LD and new variation to occur on the adjacent haplotype. However, under positive selection, an accelerated rise in frequency means there is little opportunity for either recombination or mutation to disrupt long-range LD on the haplotype bearing the beneficial mutation. The authors applied their model to *G6PD* and other genes implicated in malaria infection ***(27)***. The authors genotyped 11 SNP markers across the gene and identified 9 core haplotypes. Additional SNPs were added at increasing distances from the core haplotype and the extent of haplotype homozygosity (EHH) was assessed. EHH is defined as the probability that two randomly chosen chromosomes carrying a core haplotype of interest are identical by descent (as assayed by homozygosity at all SNPs ***[28]***) for the entire region from the core region to point x. Results showed that core haplotype 8 had a combination of high frequency and high EHH, as compared with other core haplotypes at the locus. This test controls for fluctuation in local recombination rates by using the various core haplotypes as internal controls but relies heavily on the accuracy of other programs for inferring haplotypes. Significance values were assessed using simulated data under various population demographic models. Earlier studies had confirmed that the variant unique to core haplotype 8 conferred a 50% reduction in risk of severe infection with malaria ***(29)***.

A method relying on similar principles to the EHH approach was developed by Wang et al. and applied to a whole-genome scan for positive selection ***(30)***. Designated the LD decay test, this method uses homozygote data only and calculates the fraction of inferred recombinant chromosomes adjacent to any given SNP. Differences in the fraction of inferred recombinant chromosomes between the major and minor alleles are compared with those averages across the genome and outliers beyond a 99.5% threshold taken to be indicative of selection. The method was applied to data for three ethnic groups and over 1.6% of the 1.6 million SNPs used in the analysis showed "highly unusual genetic architecture." As well as implicating many loci previously identified as candidates

for positive selection, this method also identified a large number of novel candidates that need confirmation by other methods. The authors suggest their method may lack discrimination between selection and other mechanisms because of "lack of acknowledgement of LD structure."

More recently, as genome-wide LD data become increasingly available, a new class of selection-mapping approaches that not only incorporate but are based upon LD structure is emerging. Serre et al. have recently compared genome-wide recombination rate estimates from three different human populations and suggested that although the ratio of rate estimates between any two populations are generally uniform across chromosomes, local deviations from the overall pattern can be used to detect polymorphic inversions or regions under local selection ***(31)***. Using a similar approach comparing population-specific LD maps of chromosome 17, the Genetic Epidemiology Group at the University of Southampton showed very high power to detect a common inversion under selection in Europeans ***(32)***. LD maps were created using HapMap data as described in Chapter 3. The resultant maps for the Yoruba in Ibadan (YRI) and for Utah residents with northern and western European ancestry (CEU) were aligned and divided into 100-kb nonoverlapping windows. For each window the number of LD units per megabase was calculated. A plot of the ratio of LD units per megabase in the YRI population over that in the CEU population reveals a sharp peak directly over the selected inversion (**Fig. 2**).

In order to extend such approaches to genome-wide analyses, further refinement is needed to assign appropriate significance levels to regions indicating more subtle effects of selection that do not correspond to large inversions. Comparison of large 100-kb windows are unlikely to be optimally informative for genes measuring just a few kilobases long and accurate estimates of significance will become increasingly necessary as the numbers of tests rise. Nevertheless, these efforts provide proof of principle that comparison of LD rates between populations can be used to identify regions having undergone recent selection.

Approaches using LD-rate comparisons have a number of attractive features. Although SNP ascertainment is not complete, HapMap has genotyped in excess of 3 million SNPs within each of four populations, these SNP sets may not be identical but ascertainment methodology is the same for each population. Beyond HapMap, technological advances in sequencing may make it possible to create LD maps from the set of all polymorphisms within the human genome for many more ethnic populations. Interethnic LD-rate comparisons do not rely on broad demographic assumptions and have the power to identify recent adaptations to the specific environmental pressures unique to each group.

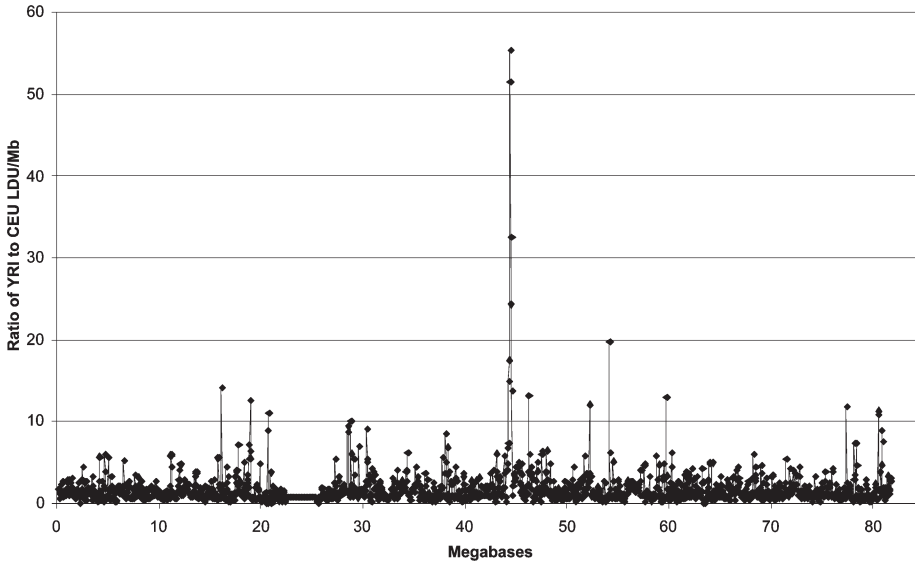



Fig. 2. Using population-specific linkage disequilibrium (LD) maps of chromosome 17, a comparison of LD rates (in 100-kb sliding windows) between Yoruban (YRI) and Caucasian (CEU) maps clearly depicts a rate difference coincident with a selected inversion *(32)* (result courtesy of Gibson, J. and Ennis, S., personal communication).

## 6. Conclusions

Understanding how, why, and upon what the forces of selection have been working over recent human history holds tantalizing promise for the fields of anthropology and medical genetics. Resounding steps have been made toward decoding the history book that is our genome but continued efforts combining the disciplines of molecular biology, mathematics, and genomic science are essential before we are in a position to decipher without ambiguity. Disentangling the effects of demographic expansions and bottlenecks from real selection; background selection from hitchhiking; incorporating recurrent selection and recurrent mutation; allowing for selection on standing genetic variation of appreciable frequency vs that on new rare variation *(33)*; understanding and correcting for the impact of polymorphism ascertainment bias; accounting for genetic drift and population substructure; integrating the highest order LD data; each of these aspects require further scientific endeavour before the full potential of understanding the dynamics of genomic adaptation is realized.

There is no doubt that the human genome continues to maintain reproductive fitness in response to both new and old environmental pressures. With an ever quickening pace of life, *some* of the effects of routine consumption of "fast-food and microwaved ready-meals" steeped in preservatives instead of sourcing

local fresh produce are evident in spiralling rates of childhood obesity and diabetes. In itself, western medicine has introduced a unique set of forces from widespread use of antibiotics and subsequent emergence of antibiotic-resistant "superbugs," common use of contraceptives, and pharmaceutical success in treating previously life-threatening diseases such as childhood cancers, chronic asthma, and neurological disorders. Perhaps of growing importance is our ability to detect relaxation of selective constraints as well as positive selective pressure. From the anthropologist interested in evolution to the pharmacologist designing effective clinical intervention therapies, a better understanding of natural response to environmental change will help us understand where our genome has been and even suggest where it is going.

## References


1 Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* **217,** 624–626.
2 Crow, J. F. and Kimura, M. (1970) *An Introduction to Population Genetics Theory*. Harper and Row, New York.
3 Holmquist, R., Jukes, T. H., and Pangburn, S. (1973) Evolution of transfer RNA. *J. Mol. Biol.* **78,** 91–116.
4 Kimura, M. and Ota, T. (1973) Mutation and evolution at the molecular level. *Genetics* **73,** 19–35.
5 Haldane, J. (1927) A mathematical theory of natural and artificial selection. Part V. Selection and mutation. *Proc. Camb. Phil. Soc.* **23,** 838–844.
6 Weatherall, D. J., Miller, L. H., Baruch, D. I., et al. (2002) Malaria and the red cell. *Hematology* (Am Soc Hematol Educ Program) **1,** 35–57.
7 Simoons, F. J. (1969) Primary adult lactose intolerance and the milking habit: a problem in biological and cultural interrelations. I. Review of the medical research. *Am. J. Dig. Dis.* **14,** 819–836.
8 Bersaglieri, T., Sabeti, P. C., Patterson, N., et al. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74,** 1111–1120.
9 Jukes, T. H. and King, J. L. (1975) Evolutionary loss of ascorbic acid synthesizing ability. *J. Hum. Evol.* **4,** 85–88.
10 Nandi, A., Mukhopadhyay, C. K., Ghosh, M. K., Chattopadhyay, D. J., and Chatterjee, I. B. (1997) Evolutionary significance of vitamin C biosynthesis in terrestrial vertebrates. *Free Radic. Biol. Med.* **22,** 1047–1054.
11 Wang, X., Thomas, S. D., and Zhang, J. (2004) Relaxation of selective constraint and loss of function in the evolution of human bitter taste receptor genes. *Hum. Mol. Genet.* **13,** 2671–2678.
12 Soranzo, N., Bufe, B., Sabeti, P. C., et al. (2005) Positive selection on a high-sensitivity allele of the human bitter-taste receptor TAS2R16. *Curr. Biol.* **15,** 1257–1265.
13 Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595.
14 Taylor, M. F., Shen, Y., and Kreitman, M. E. (1995) A population genetic test of selection at the molecular level. *Science* **270,** 1497–1499.

15 Fu , Y. X. and Li, W. H. (1993) Statistical tests of neutrality of mutations. *Genetics* **133,** 693–709.
16 Hudson, R. R., Kreitman, M., and Aguade, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116,** 153–159.
17 McDonald, J. H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351,** 652–654.
18 Kreitman, M. (2000) Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics Hum. Genet.* **1,** 539–659.
19 Nielsen, R. (2005) Molecular signatures of natural selection. *Annu. Rev. Genet.* **39,** 197–218.
20 Carrington, M., Kissner, T., Gerrard, B., Ivanov, S., O'Brien, S. J., and Dean, M. (1997) Novel alleles of the chemokine-receptor gene CCR5. *Am. J. Hum. Genet.* **61,** 1261–1267.
21 Stephens, J. C., Reich, D. E., Goldstein, D. B., et al. (1998) Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *Am. J. Hum. Genet.* **62,** 1507–1515.
22 Sabeti, P. C., Walsh, E., Schaffner, S. F., et al. (2005) The case for selection at CCR5-Delta32. *PLoS Biol.* **3,** e378.
23 Kreitman, M. and Di Rienzo, A. (2004) Balancing claims for balancing selection. *Trends Genet.* **20,** 300–304.
24 Smith, J. M. and Haigh, J. (1974) The hitch-hiking effect of a favourable gene. *Genet. Res.* **23,** 23–35.
25 Tapper, W., Collins, A., Gibson, J., Maniatis, N., Ennis, S., and Morton, N. E. (2005) A map of the human genome in linkage disequilibrium units. *Proc. Natl. Acad. Sci. USA* **102,** 11,835–11,839.
26 Sabeti, P. C., Reich, D. E., Higgins, J. M., et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837.
27 Ruwende, C. and Hill, A. (1998) Glucose-6-phosphate dehydrogenase deficiency and malaria. *J. Mol. Med.* **76,** 581–588.
28 Nei, M. (1987) Equation 8.4. *Molecular Evolutionary Genetics*. Columbia University Press, New York, NY.
29 Tishkoff, S. A., Varkonyi, R., Cahinhinan, N., et al. (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293,** 455–462.
30 Wang, E. T., Kodama, G., Baldi, P., and Moyzis, R. K. (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proc. Natl. Acad. Sci. USA* **103,** 135–140.
31 Serre, D., Nadon, R., and Hudson, T. J. (2005) Large-scale recombination rate patterns are conserved among human populations. *Genome Res.* **15,** 1547–1552.
32 Stefansson, H., Helgason, A., Thorleifsson, G., et al. (2005) A common inversion under selection in Europeans. *Nat. Genet.* **37,** 129–137.
33 Przeworski, M., Coop, G., and Wall, J. D. (2005) The signature of positive selection on standing genetic variation. *Evolution Int. J. Org. Evolution* **59,** 2312–2323.

# 6

# The Genetic Basis of Complex Traits

*Rare Variants or "Common Gene, Common Disease"?*

**Sudha K. Iyengar and Robert C. Elston**

## Summary

The goal of the Human Genome Project and the subsequent HapMap Project was to accelerate the pace at which genes for complex human traits were discovered. Elated by the early successes from cloning disease genes for monogenic disorders, the architects of the projects reasoned that complex human diseases were tractable to positional cloning methods. However, a schism emerged in the field, with hot debates regarding two competing hypotheses being publicly waged. These opposing hypotheses pertained to the anticipated allelic spectrum and frequency of disease variants associated with common, complex disease. The common disease, common variant hypothesis (CD/CV) stated that a few common allelic variants could account for the genetic variance in disease susceptibility, whereas the rare variant (CD/RV) hypothesis stated that DNA sequence variation at any gene causing disease could encompass a wide range of possibilities, with the most extreme being that each mutation is only found once in the population. The practical consequence of the debate can be broken into two parts. If the CD/CV hypothesis is true, then application of the positional cloning paradigm to map disease genes would be eminently more feasible, as a common allele would be easier to locate. Conversely, if rare variants cause common disease, then identifying these genetic susceptibility variants would be challenging. Whether a disease is caused by rare or common alleles will have an impact on clinical applications, such as designing prognostic assays, or planning therapeutic interventions; fewer susceptibility alleles will simplify assay design, and the associated reduction in costs would amortize if a universally applicable therapy can be deployed. A current review of the literature suggests that both these hypotheses are correct, depending on the gene and disease examined. Although the controversial debate is revived with the identification of each new disease gene, the time has come to integrate both hypotheses in a manner that best explains biological variation in natural populations. The allelic spectrum of variation in a particular gene may be better explained by one of the two hypotheses but, for a multifactorial trait, a composite encompassing all influential genes needs to be constructed.

**Key Words:** DNA sequence variation; allelic heterogeneity; phenotypic complexity; genetic architecture; attributable risk.

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

## 1. Introduction

Variation at the level of the DNA sequence is prevalent in natural human populations. Some variants are apparently neutral, whereas others influence specific phenotypes and may lead to disease. During the past 15 yr, more than 1800 disease loci have been identified that are believed to cause both rare and common human diseases *(1)*. In examining these loci at the sequence level, a broad spectrum of disease-causing variants have been identified, which vary in characteristics from gene to gene. For some disorders, the same or similar phenotype is attained through mutations that affect different parts of the same gene (allelic heterogeneity), with each mutation only constituting a small proportion of the allelic spectrum *(2)*. For other genes, variants recur at the same site or, as described later, may originate from a common ancestor, and these common variants comprise a greater proportion of the disease burden *(2)*. These observations from early studies regarding the frequency of the variants that account for disease were formulated into generic but competing hypotheses: the common disease common variant (CD/CV) and the multiple rare variant hypothesis (CD/RV).

Prototypic examples supporting both theories prevailed in the literature *(3–7)*, but empirical data especially supporting the CD/CV hypothesis was scarce. However, as the population aged and many complex diseases were observed to increase in prevalence, the CD/CV hypothesis garnered support among optimists as the panacea for genetic mapping woes *(8,9)*. The theoretical underpinnings supporting this opinion were explicated by Reich and Lander *(10)* in a position paper, and are described next in greater detail. However, the motivating factor for this stance was that the majority of the diseases under consideration, such as type 2 diabetes, asthma, neuropsychiatric diseases, and so on were common (>1% prevalence), and were refractory to previous methods of analysis because of their multifactorial basis. If the CD/CV hypothesis were true, then such diseases would become amenable to positional cloning approaches and through this mechanism to therapy. Put simplistically, if only a handful of genes of the approx 20,000–30,000 genes in the human genome caused a particular disease, and if the allelic spectrum of the disease was further limited to a few common variants, the prospect of identifying disease variants was achievable with a reasonable investment of resources. Although the justification was simple, it would not have been sufficient for the hypothesis to gain popularity had it not been bolstered by the convergence of several other transformations in the human genetic epidemiology scientific milieu. With the advent of the Human Genome Project *(11,12)*, the molecular technology was ramped up for larger studies, and model-free methods for linkage analysis of data gained acceptance *(13–15)*. On the heels of the completion of the Human Genome Project, and initiation of the International HapMap Project *(16–18)*, genotyping technology continued on an upswing, and there was a resurgence in

interest for association studies *(19)* fostered by development of new methods and technologies. All these elements together promoted the continued advancement of the CD/CV hypothesis, despite a public decry from some population geneticists *(20–22)* who were proponents of the alternative, the CD/RV theory.

### *1.1. Theoretical Considerations in Favor and Against the CD/CV Hypothesis*

The central premise of the CD/CV hypothesis is that variants that cause common diseases are reasonably frequent in the population, ranging from 1 to 10% *(8,9)*. Constraints on the model include lack of selection for or against these variants, and that the original mutation arose more than 100,000 yr ago. Evolutionary data demonstrating the swift proliferation of the human population from a small group of founders to the extant 6 billion plus was advocated as corroborating evidence in favor of the hypothesis. The proponents of the hypothesis contended that, in a small group of progenitors, the spectrum of mutant alleles at the same locus was likely to be narrow and, with the small group effect in play, a specific mutant could become quite common. Because of its higher frequency the "common" mutant would become preferentially propagated to the expanding population; newer, competing mutations that occurred would curtail the hegemony of the common allele very gradually during the expansion, with the expectation that the postexpansion period was not lengthy enough for these changes to overwhelm the common variant effect as yet *(23,24)*. The caveat remains that, although the data on human evolution and dispersion are growing, the data are by no means final. Two contending theories also cause a rift in this area of research (reviewed briefly in Doris *[25]*). Only one of these hypotheses, which are of a single origin for all modern humans, supports the CD/CV theory. The alternative theory of multiple events leading to the origin of modern humans, the so-called multiregional hypothesis, renders the CD/CV theory untenable.

The case against the CD/CV hypothesis comprised data from numerous sources, but was predominantly derived from observations in population genetics and late-onset Mendelian diseases (originally described in Terwilliger and Weiss *[20]*; reviewed in Wright and Hastie *[23]*). The first obstacle pertained to the role of environment in multifactorial diseases. The challengers of the CD/CV hypothesis asserted that environmental influences modulated complex disease prevalence far more than common variants. Correspondingly, rare rather than common variants were more prevalent in complex disease, and thus attributable risk, due to any particular variant was modest to insignificant. The second argument against the CD/CV hypothesis came from the population genetics paradigm that alleles not under selective pressure could easily become fixed in the population (as discussed previously, the common variants in the

founder populations are ostensibly free from selective pressure). In contrast, rare alleles that were restrained by selection would be maintained through a specific mechanism, and would account for a significant proportion of the composite genetic variance. The last reservation about the CD/CV hypothesis came from observations of Mendelian diseases with a late age at onset. Because late-onset diseases do not typically affect reproductive capability and hence genetic fitness, both rare and common variants are equally probable as the cause for these diseases. However, the vast majority of these diseases displayed a diverse repertoire of disease-associated allelic variants *(26)* that did not fit the projections of the classic CD/CV hypothesis.

### *1.2. Genetic Modifiers, Epistatic Interactions, and Phenotypic Buffers*

The CD/CV hypothesis suffered from another weakness. The role of genetic modifiers and phenotypic buffering was not explicitly modeled into its tenets. The assumptions underlying the model were similar to those for single gene disorders, where the expression of the trait was synonymous with the presence of the variant, with no role (or a very modest role) for cofactors modifying the one-to-one correspondence. Even monogenic disorders do not show complete one-to-one equivalence between the allelic state and the disorder (or phenotypic consequence). For example, in cystic fibrosis where the ΔF508 mutation at the *cystic fibrosis transmembrane conductance regulator* locus is extremely common and represents about 60–70% of mutations in Caucasian populations *(27,28)*, there is considerable variation in severity and onset of symptoms that cannot be attributed to the *cystic fibrosis transmembrane conductance regulator* ΔF508 variant alone *(29,30)*. A recent search led to the identification of variants in the *transforming growth factor* β *1* gene as modifying the phenotypic effects *(31)*. Thus, the phenotype of the ΔF508 variant, the prototypical example for the CD/CV hypothesis among single gene disorders, is buffered by variants in other genes and the environment.

That genes modify the effects of other genes is certainly not a new concept in genetics. The discovery of these phenomena dates as far back as the early 1900s. Both modifier effects and epistatic interactions are well discussed in a number of genetic textbooks, with examples from both the plant and the animal kingdom. Although the existence of gene-by-gene interactions is widely known, in an effort to comprehend the function of genes a reductionist paradigm of studying single genes in model organisms was adopted. These ideas stemmed from the study of single gene disorders, where knockouts or transgenics were anticipated to produce unambiguous, deterministic phenotypes. Many of these experiments gave unanticipated results *(32–35)* because of the phenotypic plasticity in the genomes of these animals, demonstrating that the genome was buffered against large phenotypic changes resulting from a single insult.

More importantly, phenotypic effects seen in one genetic background were mediated by moving the variant to another background, suggesting that gene-by-gene interactions were critical in mediating the final phenotype ***(34,36,37)***. In a review regarding buffering mechanisms in natural populations, Rutherford ***(38)*** has cited examples such as the *Drosophila* mutant *ey* that can be completely suppressed by variants in the background. The *ey* phenotype is caused by mutations in a key gene, *paired box 6* (*PAX6*), but the gene network supporting eye development can circumvent even highly penetrant mutations to produce a functional organ.

As in the examples given previously, if the genomic context of the network surrounding the common variant generates the eventual phenotype, the CD/CV hypothesis must be renovated and recast in the framework of systems biology (**Fig. 1**). This discipline of systematically building, assessing, and perturbing networks of genes is still immature, although some recent advances have been proposed ***(39)***. Thus, for a multifactorial trait (or disease) and the gene-discovery paradigm using association studies, these relationships may need to be modeled more explicitly to obtain incontrovertible results and consistency among studies.

### *1.3. When Can a Gene/Variant be Nominated as Fitting Either the CD/CV or the CD/RV Hypothesis?*

The field of gene discovery for common complex diseases has undergone many transformations in the past 10 yr. Initial studies were small in size and often administered from a single-investigator cottage-industry type of layout. Disappointed with the meager and inconsistent results, the field grappled with issues of sample size and effect size, and larger community-based projects mediated by teams of scientists in consortia were initiated. These issues remain at the forefront of the problems confronting genome-wide association studies ***(24)*** and are likely to plague the assignment of a variant into either the CD/CV or CD/RV paradigm.

Habitually, once a disease-associated gene is discovered, the debate about whether the mutant allele(s) meets the criteria for the CD/CV hypothesis resurfaces. The initial discovery may have led one to believe that the mutant allele is common and suitable for nomination to the CD/CV group. Frequently, replication studies seem unable to confirm these results. Either the mutant is not common in the replication studies, or the effects are dispersed across other mutations. The reason for the difference in the results is that the initial study was likely best suited for discovery of that variant. It was perhaps the top signal in a modest dataset. Exemplifying this type of predicament is the association of the *calpain 10* gene with type 2 diabetes ***(40)***. In this case, the original single-nucleotide polymorphisms (SNPs) associated with disease were discovered by sequencing 10 Mexican-American samples, after fine mapping of an interesting linkage signal ***(41)***. The majority of the replication studies for this

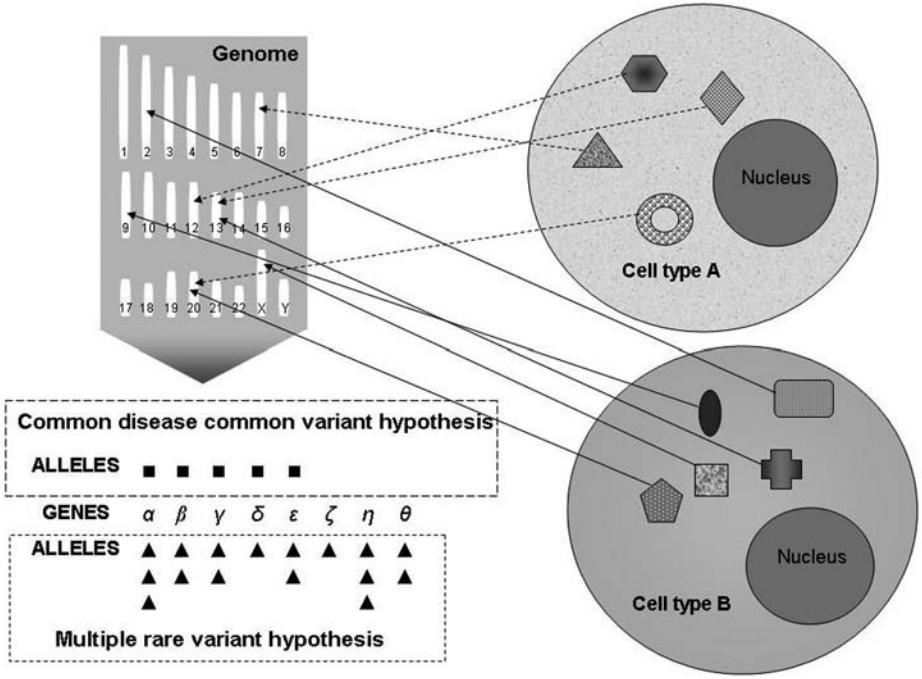



Fig. 1. This figure illustrates the complexities involved in genetic mapping. On the right, two cell types are presented that have a variety of molecules (not necessarily proteins) that are involved in the etiology of a disease. The location of these molecules is mapped to the genome on the left using arrows to point to the chromosomal location. On the bottom left are presented the contrasting theories regarding the frequency and effect of the allelic variants at each locus, with either a few common variants (black squares) or multiple rare variants (black triangles) predisposing to disease or the trait. Evidence from association experiments can be used to aggregate information about the individual components and potential interactions involved in the final pathway to disease pathogenesis. However, the results need to be substantiated with evidence from other types of experiments to gain a fuller perspective for the disease.

gene performed no sequencing for mutation discovery, but merely typed the SNPs that showed association in the original study as surrogates for the mutation(s) itself *(41–44)*. Therefore, if other mutations abounded, but the coverage of SNPs was not comprehensive enough to suffice as a proxy, these studies may have missed the association altogether. In a later study, the coverage of markers was shown to be adequate to represent the linkage disequilibrium in calpain 10, but the results remained equivocal in the replication datasets *(45,46)*.

In an effort to capture additional variation, larger scale sequencing efforts have been mounted for some projects *(47–50)*. Deeper resequencing will certainly lead

to discovery of less frequent variants that may provide additional information. However, the answer to how many samples should be sequenced once an association is uncovered is unclear. The rules for this stage of the process vary from study to study. Various sampling strategies have been employed to discover causative mutations, including sampling from tails of phenotypic distributions ***(51,52)*** to identification of linked families with informative members ***(41,44)***, to comprehensive sequencing of large numbers of samples and then referring back to the quantitative trait ***(47,50)***. For genome-wide association studies, because most study designs contrast cases with controls to obtain significant frequency differences between these groups, no guidelines for sequencing individuals to discover variants, similar to those in linkage studies, have been developed. One strategy for case–control studies would be to focus on sequencing haplotypes, either derived molecularly or predicted, that are associated with susceptibility or resistance to the trait of interest, if an obvious coding variant cannot be identified. This would be akin to the strategy applied by geneticists who study model organisms, where the variation is limited and is well characterized within specific strains. Comparison of the sequence of a "susceptible" vs "resistant" strain enables them to find key variants. The process in human samples may be more labor intensive and not as straightforward because some heterogeneity is anticipated and annotation of the variants is expected to be complex.

The coverage of the gene itself has also been variable, and is still dictated by rules formulated for Mendelian diseases. Most laboratories focus on exonic regions first with the hope of identifying a coding variant, followed by promoter or upstream regulatory regions. Focus on introns and extensive characterization of the regulatory domains through deep resequencing, as was done for the rearranged during transfection (RET) proto-oncogene and Hirschprung disease ***(53,54)***, comes only at later stages. The latter experiments were inspired by the observation that the established variants did not fully explain the linkage and association signals. In the interest of comprehensive testing, even replication studies should employ a sequencing strategy if the results do not fit the original conclusions. This area of association testing can benefit from further scientific investigation.

Others propose that meta-analyses be conducted to arrive at firm conclusions regarding a particular variant(s) and its association with disease ***(55–57)***. However, the data structure underlying the individual studies contained in the meta-analysis will influence the final inference. In the case of *calpain 10*, despite meta-analyses no firm conclusions can be drawn about it meeting criteria for CD/CV. Recent data affirm its role in the biology of type 2 diabetes ***(58–61)*** but the association evidence is tenuous, and hence more in line with the CD/RV hypothesis.

A recent victory for the CD/CV hypothesis has been the discovery of the much replicated association of the Y402H variant in complement factor H

(CFH) and age-related macular degeneration (AMD) ***(62–64)***. The gene resides under a highly replicated linkage peak ***(65)*** and was independently discovered by three groups. Since its discovery, the association has been confirmed in multiple samples ***(62–64,66–71)***, but there is some evidence that other mutations may also be involved ***(71)***. Similar to the case study for *calpain 10*, none of the replication studies have performed extensive sequence analysis, but the results with this variant have been remarkably consistent. Therefore, even if Y402H is not a causative variant in some samples, it appears to be a reasonably good proxy for causation in the majority of the analyses.

One more piece of data needs to be considered before a variant can be classified as common or rare. Its frequency in the general population needs to be estimated to enable computation of attributable risk. As described previously, based on the currently published reports the Y402H variant in CFH would meet criteria for membership in the CD/CV club. However, all the samples that have been genotyped are subject to ascertainment bias; its effect in the general population remains unknown. This last point is germane to the remainder of the genetic risk for AMD. Although the Y402H polymorphism is the most publicized AMD risk factor, over four other positional genes/loci have been proposed for AMD risk ***(67,72–75)***. Added together, the genetic risk at each of these loci from the current studies could easily exceed 100% of the total attributable fraction. Obviously, the ascertainment bias from these studies is inflating the risk estimates. Therefore, even a common variant needs to be placed in the context of other genes that cause the disease.

### *1.4. Multifactorial Disease: The Undiscovered Country*

The International HapMap Project has provided us with a glimpse of the extent of variation in the human genome across multiple ethnic groups ***(16–18)***. Between 9 and 10 million SNPs have already been discovered across these ethnic groups. Neither the linkage disequilibrium structure, nor the SNP content was identical across the four populations studied, although there were a number of similarities. The difficulties that trouble studies of common complex disease have been presented numerous times. These are small sample and effect sizes, allelic and locus heterogeneity, epistasis/hypostasis and modifiers, epigenetic effects, parent of origin effects, environmental correlates, and phenocopies. The idea that either genes or environments mediate disease (the nature vs nurture paradigm) has long been abandoned, and a composite view of complex disease has emerged. Data from numerous studies suggest that it may be necessary to similarly abandon the CD/CV vs the CD/RV schism, and develop composite theories set in a more appropriate modern framework. This would mean aggregating data across assorted ascertainment schemes and diverse platforms that cover genetics, structural and functional genomics, proteomics, and

population dynamics. Processing of these data will require a significant investment in bioinformatic and statistical tools. These investments should enable us to begin to examine larger questions. For example, if variants are modulated by networks, then the key question that remains is—is the variance in the genetic background among the populations sufficient to explain the differences in prevalence for complex diseases among and within these populations?

## 2. Conclusion

Following popular dogma acid, the CD/CV hypothesis has been unfairly invoked more often than is necessary when a positive association/linkage has been found. Whereas some genes can be categorized as following this mechanism, the alternative theory also has significant support. In the context of association mapping, mechanisms that disrupt the connection between genes of interest and the ultimate phenotype will lead to inconsistent results among studies. For a multifactorial trait, all viable scenarios may need to be considered simultaneously to comprehend the pathological basis for disease. Last, some genes that are often quoted as having common variants, such as apolipoprotein E, angiotensin-converting enzyme, and complement factor H, may have variants that affect multiple organ systems, i.e., have systemic effects. If prevention is the objective, then a measure such as attributable community risk as proposed by Wacholder ***(76)*** is perhaps more attractive than the popular population-attributable risk.

## Acknowledgments

This research was supported, in part, by US Public Health Service research grants EY015810 from the National Eye Institute; U01DK057292 from the National Institute of Diabetes and Digestive and Kidney Diseases Institute; GM28356, from the National Institute of General Medical Sciences Institute; and resource grant RR03655, from the National Center for Research Resources.

## References

1 McKusick–Nathans Institute for Genetic Medicine and Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information National Library of Medicine Bethesda MD. 2000 Online Mendelian Inheritance in Man.

2 Smith, D. J. and Lusis, A. J. (2002) The allelic structure of common disease. *Hum. Mol. Genet.* **11,** 2455–2461.

3 Corder, E. H., Saunders, A. M., Strittmatter, W. J., et al. (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261,** 921–923.

4 Roa, B. B., Boyd, A. A., Volcik, K., and Richards, C. S. (1996) Ashkenazi Jewish population frequencies for common mutations in BRCA1 and BRCA2. *Nat. Genet.* **14,** 185–187.

5 Dunning, A. M., Chiano, M., Smith, N. R., et al. (1997) Common BRCA1 variants and susceptibility to breast and ovarian cancer in the general population. *Hum. Mol. Genet.* **6,** 285–289.
6 Altshuler, D., Hirschhorn, J. N., Klannemark, M., et al. (2000) The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* **26,** 76–80.
7 Healey, C. S., Dunning, A. M., Teare, M. D., et al. (2000) A common variant in BRCA2 is associated with both breast cancer risk and prenatal viability. *Nat. Genet.* **26,** 362–364.
8 Lander, E. S. (1996) The new genomics: global views of biology. *Science* **274,** 536–539.
9 Collins, F. S., Brooks, L. D., and Chakravarti, A. (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8,** 1229–1231.
10 Reich, D. E. and Lander, E. S. (2001) On the allelic spectrum of human disease. *Trends Genet.* **17,** 502–510.
11 Lander, E. S., Linton, L. M., Birren, B., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.
12 Venter, J. C., Adams, M. D., Myers, E. W., et al. (2001) The sequence of the human genome. *Science* **291,** 1304–1351.
13 Haseman, J. K. and Elston, R. C. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2,** 3–19.
14 Schaid, D. J., Olson, J. M., Gauderman, W. J., and Elston, R. C. (2003) Regression models for linkage: issues of traits, covariates, heterogeneity, and interaction. *Hum. Hered.* **55,** 86–96.
15 Schaid, D. J., Elston, R. C., Tran, L., and Wilson, A. F. (2000) Model-free sib-pair linkage analysis: combining full-sib and half-sib pairs. *Genet. Epidemiol.* **19,** 30–51.
16 The International HapMap Project. (2003) *Nature* **426,** 789–796.
17 International HapMap Consortium (2004) Integrating ethics and science in the International HapMap Project. *Nat. Rev. Genet.* **5,** 467–475.
18 Thorisson, G. A., Smith, A. V., Krishnan, L., and Stein, L. D. (2005) The International HapMap Project Web site. *Genome Res.* **15,** 1592–1593.
19 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517.
20 Terwilliger, J. D. and Weiss, K. M. (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr. Opin. Biotechnol.* **9,** 578–594.
21 Terwilliger, J. D. and Hiekkalinna, T. (2006) An utter refutation of the 'Fundamental Theorem of the HapMap'. *Eur. J. Hum. Genet.* **14,** 426–437.
22 Weiss, K. M. and Terwilliger, J. D. (2000) How many diseases does it take to map a gene with SNPs? *Nat. Genet.* **26,** 151–157.
23 Wright, A. F. and Hastie, N. D. (2001). Complex genetic diseases: controversy over the Croesus code. *Genome Biol.* **2,** COMMENT2007.
24 Zondervan, K. T. and Cardon, L. R. (2004) The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* **5,** 89–100.

25 Doris, P. A. (2002) Hypertension genetics, single nucleotide polymorphisms, and the common disease: common variant hypothesis. *Hypertension* **39,** 323–331.
26 Pritchard, J. K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69,** 124–137.
27 Estivill, X., Bancells, C., and Ramos, C. (1997) Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. The Biomed CF Mutation Analysis Consortium. *Hum. Mutat.* **10,** 135–154.
28 Morral, N., Bertranpetit, J., Estivill, X., et al. (1994) The origin of the major cystic fibrosis mutation (delta F508) in European populations. *Nat. Genet.* **7,** 169–175.
29 Castaldo, G., Tomaiuolo, R., Vanacore, B., et al. (2006) Phenotypic discordance in three siblings affected by atypical cystic fibrosis with the F508del/D614G genotype. *J. Cyst. Fibros.* Epub ahead of print.
30 Corvol, H., Flamant, C., Vallet, C., Clement, A., and Brouard, J. (2006) Modifier genes and cystic fibrosis. *Arch. Pediatr.* **13,** 57–63.
31 Drumm, M. L., Konstan, M. W., Schluchter, M. D., et al. (2005) Genetic modifiers of lung disease in cystic fibrosis. *N. Engl. J. Med.* **353,** 1443–1453.
32 Bielsky, I. F., Hu, S. B., and Young, L. J. (2005) Sexual dimorphism in the vasopressin system: lack of an altered behavioral phenotype in female V1a receptor knockout mice. *Behav. Brain Res.* **164,** 132–136.
33 Bontekoe, C. J., McIlwain, K. L., Nieuwenhuizen, I. M., et al. (2002) Knockout mouse model for Fxr2: a model for mental retardation. *Hum. Mol. Genet.* **11,** 487–498.
34 Fukamauchi, F., Wang, Y. J., Mataga, N., and Kusakabe, M. (1997) Paradoxical behavioral response to apomorphine in tenascin-gene knockout mouse. *Eur. J. Pharmacol.* **338,** 7–10.
35 Shi, W., Wang, X., Shih, D. M., Laubach, V. E., Navab, M., and Lusis, A. J. (2002) Paradoxical reduction of fatty streak formation in mice lacking endothelial nitric oxide synthase. *Circulation* **105,** 2078–2082.
36 Johnson, K. R., Zheng, Q. Y., and Noben-Trauth, K. (2006). Strain background effects and genetic modifiers of hearing in mice. *Brain Res.* **1091,** 79–88.
37 Seidelmann, S. B., De, L. C., Leibel, R. L., Breslow, J. L., Tall, A. R., and Welch, C. L. (2005) Quantitative trait locus mapping of genetic modifiers of metabolic syndrome and atherosclerosis in low-density lipoprotein receptor-deficient mice: identification of a locus for metabolic syndrome and increased atherosclerosis on chromosome 4. *Arterioscler. Thromb. Vasc. Biol.* **25,** 204–210.
38 Rutherford, S. L. (2000) From genotype to phenotype: buffering mechanisms and the storage of genetic information. *Bioessays* **22,** 1095–1105.
39 Papin, J. A., Hunter, T., Palsson, B. O., and Subramaniam, S. (2005) Reconstruction of cellular signaling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.* **6,** 99–111.
40 Song, Y., Niu, T., Manson, J. E., Kwiatkowski, D. J., and Liu, S. (2004) Are variants in the CAPN10 gene related to risk of type 2 diabetes? A quantitative assessment of population and family-based association studies. *Am. J. Hum. Genet.* **74,** 208–222.

41 Horikawa, Y., Oda, N., Cox, N. J., et al. (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat. Genet.* **26,** 163–175.
42 Taillon-Miller, P., Saccone, S. F., Saccone, N. L., et al. (2004). Linkage disequilibrium maps constructed with common SNPs are useful for first-pass disease association screens. *Genomics* **84,** 899–912.
43 Clark, V. J., Cox, N. J., Hammond, M., Hanis, C. L., and Di, R. A. (2005) Haplotype structure and phylogenetic shadowing of a hypervariable region in the CAPN10 gene. *Hum. Genet.* **117,** 258–266.
44 Hayes, M. G., Del Bosque-Plata, L., Tsuchiya, T., Hanis, C. L., Bell, G. I., and Cox, N. J. (2005) Patterns of linkage disequilibrium in the type 2 diabetes gene calpain-10. *Diabetes* **54,** 3573–3576.
45 Cox, N. J., Hayes, M. G., Roe, C. A., Tsuchiya, T., and Bell, G. I. (2004) Linkage of calpain 10 to type 2 diabetes: the biological rationale. *Diabetes* **53,** S19–S25.
46 Weedon, M. N., Schwarz, P. E., Horikawa, Y., et al. (2003) Meta-analysis and a large association study confirm a role for calpain-10 variation in type 2 diabetes susceptibility. *Am. J. Hum. Genet.* **73,** 1208–1212.
47 Carlson, C. S., Aldred, S. F., Lee, P. K., et al. (2005) Polymorphisms within the C-reactive protein (CRP) promoter region are associated with plasma CRP levels. *Am. J. Hum. Genet.* **77,** 64–77.
48 Crawford, D. C., Akey, D. T., and Nickerson, D. A. (2005) The patterns of natural variation in human genes. *Annu. Rev. Genomics Hum. Genet.* **6,** 287–312.
49 Livingston, R. J., von Niederhausern N. A., Jegga, A. G., et al. (2004) Pattern of sequence variation across 213 environmental response genes. *Genome Res.* **14,** 1821–1831.
50 Crawford, D. C., Yi, Q., Smith, J. D., et al. (2006) Allelic spectrum of the natural variation in CRP. *Hum. Genet.* **119,** 496–504.
51 Cohen, J., Pertsemlidis, A., Kotowski, I. K., Graham, R., Garcia, C. K., and Hobbs, H. H. (2005) Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* **37,** 161–165.
52 Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R., and Hobbs, H. H. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305,** 869–872.
53 Burzynski, G. M., Nolte, I. M., Bronda, A., et al. (2005) Identifying candidate Hirschsprung disease-associated RET variants. *Am. J. Hum. Genet.* **76,** 850–858.
54 Grice, E. A., Rochelle, E. S., Green, E. D., Chakravarti, A., and McCallion, A. S. (2005) Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer. *Hum. Mol. Genet.* **14,** 3837–3845.
55 Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A., and Contopoulos-Ioannidis, D. G. (2001) Replication validity of genetic association studies. *Nat. Genet.* **29,** 306–309.
56 Salanti, G., Sanderson, S., and Higgins, J. P. (2005) Obstacles and opportunities in meta-analysis of genetic association studies. *Genet. Med.* **7,** 13–20.
57 Whitcomb, D. C., Aoun, E., Vodovotz, Y., Clermont, G., and Barmada, M. M. (2005) Evaluating disorders with a complex genetics basis: the future roles of meta-analysis and systems biology. *Dig. Dis. Sci.* **50,** 2195–2202.

58 Marshall, C., Hitman, G. A., Partridge, C. J., et al. (2005) Evidence that an isoform of calpain-10 is a regulator of exocytosis in pancreatic beta-cells. *Mol. Endocrinol.* **19,** 213–224.
59 Turner, M. D., Cassell, P. G., and Hitman, G. A. (2005) Calpain-10: from genome search to function. *Diabetes Metab Res. Rev.* **21,** 505–514.
60 Harris, F., Chatfield, L., Singh, J., and Phoenix, D. A. (2004) Role of calpains in diabetes mellitus: a mini review. *Mol. Cell Biochem.* **261,** 161–167.
61 Cox, N. J., Hayes, M. G., Roe, C. A., Tsuchiya, T., and Bell, G. I. (2004) Linkage of calpain 10 to type 2 diabetes: the biological rationale. *Diabetes* **53,** S19–S25.
62 Edwards, A. O., Ritter, R., 3rd, Abel, K. J., Manning, A., Panhuysen, C., and Farrar, L. A. (2005) Complement factor H polymorphism and age-related macular degeneration. *Science* **308,** 421–424.
63 Haines, J. L., Hauser, S. L., Schmidt, S., et al. (2005) Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308,** 419–421.
64 Klein, R. J., Zeiss, C., Chew, E. Y., et al. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* **308,** 385–389.
65 Fisher, S. A., Abecasis, G. R., Yashar, B. M., et al. (2005) Meta-analysis of genome scans of age-related macular degeneration. *Hum. Mol. Genet.* **14,** 2257–2264.
66 Okamoto, H., Umeda, S., Obazawa, M., et al. (2006) Complement factor H polymorphisms in Japanese population with age-related macular degeneration. *Mol. Vis.* **12,** 156–158.
67 Rivera, A., Fisher, S. A., Fritsche, L. G., et al. (2005) Hypothertical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. *Hum. Mol. Genet.* **14,** 3227–3236.
68 Souied, E. H., Leveziel, N., Richard, F., et al. (2005) Y402H complement factor H polymorphism associated with exudative age-related macular degeneration in the French population. *Mol. Vis.* **11,** 1135–1140.
69 Sepp, T., Khan, J. C., Thurlby, D. A., et al. (2006) Complement factor H variant Y402H is a major risk determinant for geographic atrophy and choroidal neovascularization in smokers and nonsmokers. *Invest. Ophthalmol. Vis. Sci.* **47,** 536–540.
70 Magnusson, K. P., Duan, S., Sigurdsson, H., et al. (2006) CFH Y302H confers similar risk of soft drusen and both forms of advanced AMD. *PLoS Med.* **1,** e5.
71 Hageman, G. S., Anderson, D. H., Johnson, L. V., et al. (2005) A common haplotype in the complement regulartory gene factor H (HF1/CFH) predisposes individuals to age-relted macular degeneration. *Proc. Natl. Acad. Sci. USA* **102,** 7227–7232.
72 Gold, B., Merriam, J. E., Zernant, J., et al. (2006) Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. *Nat. Genet.* **38,** 458–462.
73 Haines, J. L., Schnetz-Boutaud, N., Schmidt, S., et al. (2006) Functional candidate genes in age-related macular degeneration: significant association with VEGF, VLDLR, and LRP6. *Invest. Ophthalmol. Vis. Sci.* **47,** 329–335.

74 Jakobsdottir, J., Conley, Y. P., Weeks, D. E., Mah, T. S., Ferrel, R. E., and Gorin, M. B. (2005) Susceptibility genes for age-related maculopathy on chrsomosome 10q26. *Am. J. Hum. Genet.* **77,** 389–407.

75 Zareparsi, S., Buracynska, M., Branham, K. E., et al. (2005) Toll-like receptor 4 variant D299G asscoiated with susceptibility to age-related macular degeneration. *Hum. Mol. Genet.* **14,** 1449–1455.

76 Wacholder, S. (2005) The impact of a prevention effort on the community. *Epidemiology* **16,** 1–3.

# 7

# Linkage Disequilibrium Mapping for Complex Disease Genes

**Andrew DeWan, Robert J. Klein, and Josephine Hoh**

**Summary**

In this chapter we lay out some background information about gene mapping for human complex traits. The chapter covers issues such as study design, high-throughput genotyping technologies, statistical analysis, and others. Many of the materials are based on and related to our recent experiences in case–control-association studies.

**Key Words:** LD; SNPs; association; case–control; complex; design; high throughput; genome; power; admixture.

## 1. Introduction

Association mapping localizes genes affecting disease susceptibility by associating nearby genetic polymorphisms, such as single-nucleotide polymorphisms (SNPs). Association differs from linkage, although it is thought that an association study is a special form of a linkage study in which a population forms an extended family. Common diseases are presumably a consequence of some solitary and frequent DNA version(s), or a combination of variants in the population. These SNPs are likely to enter a family through multiple founders rather than passing down the lineage from a single ancestor. Thus, family-linkage studies, which trace one unique segregation path from an original founder to the current generation, may fail to detect links between the trait and the risk-conferring allele(s). In addition, family studies of common diseases may identify rare or even spurious associations that are not found confirmed when tested prospectively in patients outside the family group. These are the principal reasons that carefully matched randomly selected cases and controls offer the preferred design of studies aiming to disentangle the complexities of the genetics of common diseases.

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

It is easier to circumvent complex and hard-to-detect segregation patterns originating from multiple founders with the population-based association approaches, because genotype frequency differences between unrelated diseased and disease-free individuals can be compared directly, without accommodating the correlations among family members. Successful association studies share several key elements in their design; neglect of these requirements has impaired such studies in the past. As a result there have been so few successes and so many failures to replicate positive association findings that some experts have raised doubts about the advisability of the approach *(1–3)*. Common genetic diseases must have finite and frequent common variants as causes across populations.

To maximize the probability of success of the overall association approach, we should consider the following issues prior to embarking on the study:

1. Unambiguously characterize clinical features of the phenotype.
2. Carefully match cases and controls with respect to genetic, nongenetic, and environmental factors.
3. Avoid potential confounding effects from unmeasured (hidden) population substructure differences in cases and controls.
4. Include sufficient genetic markers and high-throughput genotyping capacity.

### 1.1. "Parsimonious" Study Design

We have applied the concepts in the "extreme discordant sib pair" design, a mapping strategy of high statistical power for quantitative trait loci proposed by Risch and Zhang *(4)*. **Table 1** excerpted from their paper shows the number of sib pairs required as the trait value is broken into 10 consecutive intervals, or deciles; the horizontal and vertical represent the trait deciles for two respective siblings; the upper and lower diagonal represent two levels ρ (ρ is higher in the lower diagonal) of other risk factors shared between the two siblings. Those that are extremely discordant pairs provide substantial power, reducing the sample size necessary over conventional designs of selecting pairs at random by as much as 50-fold.

One can then adapt this "extreme discordant sib pair" strategy to design powerful association studies in unrelated subjects. Ideally, we will select cases and controls based on a quantitative measure of phenotype for which all cases come from "decile 10" and controls from "decile 1" (i.e., one in the top 10% and the other in the bottom 10% of the distribution). In order to achieve meaningful ρ values, we "mimic" the sibling-shared environment in the case–control setting; one should collect the information of known risk factors and match them between cases and controls.

We applied this design principal to our age-related macular degeneration (AMD) study, choosing the drusen measure as the quantitative trait *(5)*.

**Table 1**
**Number of Sib Pairs Required; * > 999, ** > 9999, *** > 99999**

| decile | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | decile |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 478 | 693 | * | * | * | *** | * | * | 277 | 62 | 1 |
| | | * | * | * | * | *** | ** | * | 458 | 102 | 2 |
| 1 | 632 | | * | * | ** | *** | ** | * | 780 | 168 | 3 |
| 2 | * | * | | * | ** | *** | ** | * | * | 310 | 4 |
| 3 | * | * | * | | ** | *** | *** | ** | * | 747 | 5 |
| 4 | ** | * | * | * | | *** | *** | *** | *** | * | 6 |
| 5 | * | ** | * | * | * | | ** | ** | ** | ** | 7 |
| 6 | 878 | ** | ** | * | * | * | | * | * | * | 8 |
| 7 | 283 | * | * | *** | * | * | * | | 635 | 259 | 9 |
| 8 | 115 | 346 | * | * | *** | * | * | * | | 73 | 10 |
| 9 | 49 | 110 | 224 | 523 | * | ** | * | * | 539 | | |
| 10 | 19 | 31 | 48 | 77 | 136 | 302 | * | *** | 647 | 121 | |

Excerpted from **ref. *4***. (Copyright [1995] AAAS.)

Drusen, a deposition behind the retina and concentrated in and around the macula, are a hallmark feature of AMD. The extreme cases are defined as patients having >125 μm in diameter drusen and control drusen <63 μm in diameter; most of the controls had virtually no drusen at all. Moreover, we attained homogeneity in known lifestyle and environmental influencers of AMD by matching several intervening variables such as age, diet, and UV exposure in cases and controls.

### *1.2. Analysis Issues*

In the last 2 yr, large-scale sequencing and genotyping projects have been conducted in several places to uncover human genetic variation. It is estimated that there are 10–15 million SNPs with a minor allele frequency (MAF) more than 1% in the genome ***(6,7)***. Half of these SNPs have an MAF less than 10%, whereas the other half exhibits an MAF exceeding 10%.

We have benefited from both advanced genotyping technologies and increased genomic information in our whole-genome association study, which successfully discovered a major AMD susceptibility gene ***(5)***. But high-throughput genotyping technologies that generate huge amounts of high-dimensional data create a series of statistical challenges. What is the optimal strategy to correct for chance results based on testing more than 116,000 SNPs individually or in combination? Are statistically associated SNPs causal or surrogates in linkage

disequilibrium (LD) with the causal locus? How can haplotypes be consistently estimated in a manner that is "robust" to underlying model assumptions? Simple gene-by-gene analysis can be effective in finding a major susceptibility locus but may not suffice for discovering loci of lesser effects. Difficulties are also greatly compounded by the fact that individuals can be differently affected by the same genetic risk factors that depend on cofactors (epistasis; *see* **Subheading 4.5.**).

Proper analyses must be applied to minimize false-positive/negative errors *(8)*. Two excellent, comprehensive reviews about the statistical issues, treatments, and genomic information in association studies are given by Cordell and Clayton *(9)* and Palmer and Cardon *(10)*, respectively.

## 2. Human SNPs

The advent of high-throughput genotyping technologies has allowed for the production of millions of genotypes of single basepair mutations (SNPs) across the genome. In the population, these SNPs can have a wide range of MAF, ranging from 0.5 (both alleles equally frequent) to approx 0 (only one person in the population carries this mutation). Although there may be upwards of 7 million SNPs with a MAF greater than 5% *(7)*, it has been shown that the MAF values for numerous SNPs vary widely between ethnic groups *(11,12)*. In a survey of approx 1.6 million SNPs in Americans of European, African, and Asian ancestry 291,012 SNPs (these 18% are referred to as "private SNPs") were found to be segregating in only one of the three populations. Although most of these private SNPs have lower MAFs than other SNPs, 37% of these private SNPs have a MAF greater than 10% *(12)*, suggesting a surprising degree of interethnic variation. Within these variations will be many of the answers to ethnic differences in disease prevalence and presentation.

It is important to note that these broad ethnic categories do not mean that the SNP allele frequencies are uniform within these groups; geographically isolated populations within these ethnic groups will create further stratification. It has been suggested that this intraethnic variation may far outweigh the variation between the classic ethnic classifications *(13–15)*. However, the extent of intraethnic variability is controversial *(11,13,16–21)*. Recently it was shown that the Icelandic population, superficially a genetically homogeneous island, has a genetic substructure that requires consideration in Icelandic genetic-association studies *(22)*.

But what relevance does this inter- and intraethnic variation have for most association studies? An example of the impact of such variability among a geographically defined ethnic group is with the *lactase* (*LCT*) gene in which the SNP allele frequencies vary widely across Europe *(23)* and the world. It was subsequently shown that if you do not account for the European ancestry of

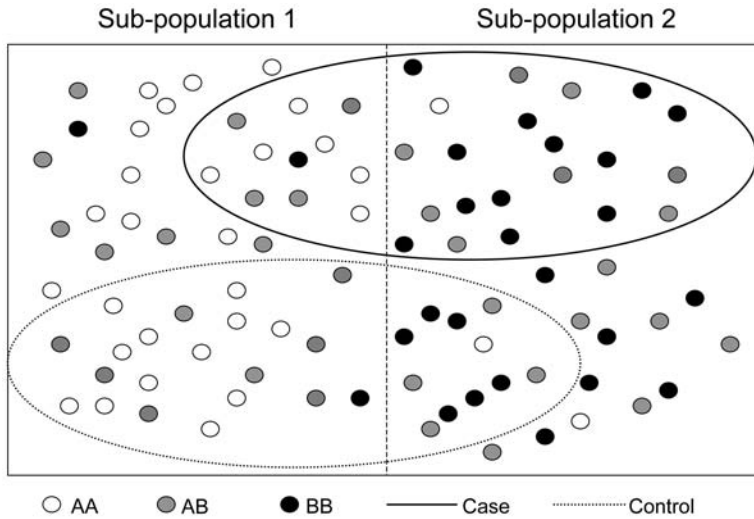



Fig. 1. False association at a single-nucleotide polymorphism (SNP) locus arising from population substructure in a case–control study. A schematic demonstrating the underlying problem of population substructure in a case–control study. The allelic distribution at this particular SNP is not the same in subpopulations 1 and 2, with the frequency of the A allele higher in subpopulation 1 and B higher in subpopulation 2. Owing to the fact that more cases are recruited from subpopulation 2, it would appear as if there is an association between the disease and this SNP. In reality, this apparent association is because of the difference in allele frequency between the two populations at this SNP.

origin there was an apparent association between a SNP in *LCT* and height that was owing solely to population stratification ***(24)***. This stratification was not detected by common methods—ancestry informative markers and the computer program *Structure* ***(20)***—and the authors suggest that new methods may be routinely required to detect underlying stratification in case–control samples. This false association arises because the population substructures differ between cases and controls (**Fig. 1**).

In epidemiological terms, this type of phenomenon is known as confounding. To avoid such spurious associations, cases and controls must be matched as closely as possible in terms of their ancestry of origin to minimize these erroneous associations.

One goal of the recently published HapMap Project was to identify all common variation in four ethnically diverse populations ***(25)***. And while the project was successful in identifying common variants in each of these populations and allele frequency differences between each population, it did not have sufficient sample size to uncover substructures within these populations (269 subjects'

DNA were analyzed). The data from the HapMap Project allowed for conclusions to be made about the samples studied; each ethnic sample represented a subpopulation within a larger ancestral ethnic group. For example, the African population studied was the Yoruba in Ibadan, Nigeria. As is pointed out by the study, this group is not representative of the general population of Africa as a whole or of smaller subpopulations outside of the Yoruba. The extent of intra-ethnic variation is not possible to capture from these limited samples, but rather allowed for the study of the patterns of allelic variation between ethnically diverse subgroups.

Usually, when we talk about genetic studies, we think about the nuclear genome. However, the mitochondrial genome is inherited separately in humans and should be considered. The mitochondrial DNA (mtDNA) is a circular DNA molecule that is about 16-kb long. It is inherited maternally, through the cytoplasm of the oocyte. Therefore, mutations, and not recombinations, are the primary force of genetic change in the mtDNA genome. This property has lead many researchers to use the mtDNA to study the demographics and history of human populations *(26)*. Mitochondrial DNA sequences can be classified into distinct lineage-specific haplogroups based on a small number of sequence features *(27)*. Furthermore, many disease-causing loci have been mapped to specific mutations in the mitochondrial genome *(28)*. The diseases caused by these mutations range from those that are classically associated with mtDNA mutations (e.g., Kearns–Sayre syndrome and Leber hereditary optic neuropathy) and common diseases for which mtDNA mutations account for only a small fraction of the disease-causing variants in the population (e.g., diabetes). Much remains to be learned about the role of mtDNA mutations in human disease, and this small but important genome should not be forgotten in association studies of human diseases.

## 3. Genotyping and Sequencing

In performing an association study, the first step is to determine the genotypes at the loci of interest in all of your sample individuals. For our purposes, we will assume that the markers being genotyped are SNPs. The number of genotypes to determine can quickly grow extremely large. For instance, genotyping 1000 individuals at 1000 SNPs yields 1 million genotypes. There are a large number of technologies on the market for genotyping large number of SNPs. Each of these technologies has their pros and cons, and they need to be weighed in terms of the experimental design and requirements. Factors to consider in evaluating these technologies include their throughput (how easy it is to genotype multiple SNPs and multiple individuals), cost, call rate (how often the assay gives you a genotype), accuracy (how often the genotype given is correct), and expected power for your study.

### 3.1. High-Throughput Genotyping Technologies

Genotyping technologies can be broken down into three classes, based on the number of SNPs to genotype: small-scale, medium-scale, and large-scale. Many of the small-scale genotyping methods are quite simple. They are good for genotyping a small number of SNPs in any number of individuals; large DNA collections can be efficiently assayed when the number of SNPs of interest is low. As the number of SNPs increases, this method becomes prohibitively time consuming and expensive. Medium-scale genotyping makes it straightforward to genotype up to several thousand SNPs simultaneously. This is most useful when you want to perform fine mapping in a region or regions of interest. Finally, large-scale genotyping allows the genotyping of 10,000 or more SNPs simultaneously, and are most useful for genome-wide scans and will be discussed in detail here.

For large scales (10,000 SNPs or more), there are a variety of microarray-based assays, from Affymetrix (Santa Clara, CA) and Illumina (San Diego, CA) that can be chosen. Unlike smaller scale genotyping methods, these are fixed-content systems. Using a variety of criteria, the manufacturer has selected which SNPs will be genotyped. Therefore, in these assays, the appropriateness of a given selection of SNPs, as well as the total SNP count, should be considered in evaluation.

Affymetrix makes three large-scale genotyping array products. One product genotypes 10,000 SNPs at a time on a single chip (10K). Because of its low genome-wide density and fixed-content nature, this array is best suited for genome-wide linkage scans rather than association studies, and is not discussed further here. The second product, with which we have the most experience, genotypes 100,000 SNPs at a time on two chips (100K). The most recent product genotypes 500,000 SNPs at a time on two chips (500K). Both these products allow for use of each chip in a genome-wide scan, thus providing two versions of 50K in the first and two versions of 250K in the second. At the time of writing, call rates for 100K approach 99%, whereas the 500K arrays are below 94% in our laboratory.

The first step of each Affymetrix assay consists of digestion with a specific restriction enzyme and amplification of short restriction fragments *(29,30)*. Each chip is distinguished by the restriction enzyme used and can be thought of as a separate assay. Thus, there is a 50K Xba chip, a 50K Hind chip, a 250K Nsp chip, and a 250K Sty chip, that can be mixed and matched as desired. On all of these chips, SNP selection is determined by the location of the small restriction fragments of the appropriate enzymes.

Illumina's genotyping system allows the manufacturer to select SNPs freely. The core of the assay is extension from a locus-specific or allele-specific primer attached to a bead *(31,32)*. Illumina has two genome-wide products. One ("Sentrix Human-1") contains 100,000 SNPs focused on polymorphisms

**Table 2**
**SNP Genome Coverage, Actual Number (Frequency), in Illumina 300K and Affymetrix 500K Genotyping Platforms**

| | Illumina 300K | Affymetrix 500K |
|---|---|---|
| Intragenic (between txStart and Stop) | 123,373 (0.389) | 184,912 (0.369) |
| Exon | 12,023 | 8513 |
| Intron | 111,350 | 176,399 |
| Within 10kB of txStart or txStop | 35,290 (0.111) | 52,208 (0.104) |
| Intergenic | 158,839 (0.500) | 263,706 (0.527) |

located in annotated genes. The other ("Sentrix HumanHap 300") contains 300,000 SNPs chosen to maximize coverage of the entire genome. This chip was designed using HapMap data for individuals from Utah whose ancestors are from northern and central Europe (CEU population *[25]*).

The genomic distribution of SNPs contained on the Affymetrix 500K and Illumina 300K chips differ and can be seen in **Table 2**. Although Affymetrix has a greater number of SNPs between transcriptional start and stop sites, the proportion of SNPs falling into each category is roughly the same (based on the build 35 annotation). Interestingly, Illumina, with the smaller number of total SNPs, has a greater number of SNPs occurring within an exon, presumably the benefit of their technology that allows for selection of SNPs by design rather than by enzyme cleavage.

### *3.2. Quality Assurance and Quality Control*

No matter the size of the experiment, it is imperative that you rigorously assess genotyping quality to make sure that your data are error free. A variety of errors could confound your data. Specific reactions or assays may fail. Samples may be misidentified. The technique may not be reproducible (precision error). The technique may give reproducibly poor results (accuracy error).

By implementing a variety of quality control processes into your genotyping, one may be confident of the data. Some of these checks are time- and resource intensive, and may not be feasible to do continuously, but we recommend doing all of them whenever beginning to use a new genotyping system in the lab and at reasonably spaced intervals thereafter.

The first problem one may encounter is the failure of a reaction or assay. This problem is most easily detected through lower than expected call rates—the fraction of genotyped SNPs for which the assay produces a genotype. Call rates can be looked at overall, on a per-SNP basis, or on a per-individual basis. In those assays (like those from Affymetrix) in which a single chip is used to

genotype many SNPs, call rates can be looked at on a per-chip basis as well as a per-individual basis. A low per-chip call rate (or other per-experiment call rate) suggests that something went wrong with a particular experiment. Seen once, a low per-experiment call rate suggests the experiment should simply be repeated. Repeated low per-experiment call rates suggest a problem in the assay itself, reagents employed, or in the performance of the assay. (The Affymetrix and Illumina systems require scrupulous adherence to protocol; modifications of methods frequently result in poor call rates.) A reproducibly low per-SNP call rate (a given SNP is infrequently called) suggests that the assay does not work for the SNP in question. A reproducibly low per-individual call rate (a given individual's DNA rarely gives a SNP call) suggests a problem with the DNA. After checking the quality of the DNA using your favorite method, it might be necessary to either get a new sample of DNA or remove the individual from the study.

Misidentification of samples is avoidable with proper laboratory controls. There are several controls that can allow you to catch these mistakes should they occur. These checks include checking the known gender against the gender determined from the genotypes and, in the case of the Affymetrix system, checking the redundant genotypes across chips. When starting to use a new or modified method, make sure that the results are both reproducible and accurate. Reproducibility can be checked by either running DNA samples twice or using reference DNA and comparing the results to reference genotype call. These methods are expensive and time consuming and although necessary cannot be done continuously.

Fortunately, assay accuracy can be accomplished on a continuing basis. The Hardy–Weinberg equilibrium ***(33)*** can be checked for possible genotyping errors. Equilibrium should be maintained in the normal control groups: assuming $p$ and $q$ are the frequencies of the major and minor alleles, respectively, then major allele homozygotes should occur with a frequency of $p^2$, minor allele homozygotes should occur with a frequency of $q^2$, and heterozygotes should have a frequency of $2pq$. Significant deviations from this expectation may indicate genotyping errors.

Another check can be performed if one genotypes related individuals from a pedigree (parents, children, grandparents, and so on). In these instances, one can check for Mendelian inconsistencies (i.e., genotypes that would be impossible for a child to have given the parents' genotypes). Such inconsistencies can be a sign of genotyping errors. Finally, the gold standard for SNP genotyping is bidirectional sequencing of the DNA. Perform this assay for several SNPs in several DNA samples and compare the genotypes with those obtained by the assay to be used on a large scale. The results should match.

### 3.3. Power

Before embarking on a genotyping project, one should estimate the power of a proposed genotyping strategy. Power is defined as the probability of detecting a statistically significant association given that such an association exists. Various factors determine the power of an association study, including the number of individuals in the study, the expected effect size, and the frequencies of the alleles at the loci to be genotyped. For genome-wide association studies (or any study where you expect the functional allele to be assayed indirectly through LD), power is also highly dependent on the pattern of LD between the genotyped markers and all of the variation in the population of interest.

Therefore, to compute the power of a genome-wide association study, it is necessary to take this LD into account. To do so exactly requires complete genotype data for common variation within the population of interest. Although such complete data does not exist, the genotype data from the HapMap Project *(25)* is a good approximation. Using this data, it is possible to compute the power of a genome-wide association study given the sample size, effect size, and a list of markers to be genotyped. The premise of this calculation is simple: the power of detecting an association at each polymorphism is calculated, assuming you are detecting it with the best marker from the set of genotyped markers. Then, assuming that each polymorphism is equally likely to be at the functional locus, the overall power is taken to be the average of the power at each locus *(34)*.

Using this approach, it is possible to compare the powers of the different whole-genome genotyping platforms in the different HapMap populations. The power of different genotyping platforms can be evaluated for different populations by examining the number of individuals required for 80% power. In the CEU population, the 300K Illumina platform requires fewer individuals than the 500K Affymetrix platform presumably because the Illumina optimized for this population. The Illumina platform has just as much power as the Affymetrix one in the CHB (Han Chinese in Beijing) and JPT (Japanese in Tokyo) populations, whereas the Affymetrix platform has more power in the YRI (Yoruba in Ibadan, Nigeria) population.

It is important to note that the number of individuals required for 80% power using only the Nsp or the Sty 250K Affymetrix chip is less than twice that needed for using the pair as a 500,000 SNP set. Therefore, more power is achieved by typing more samples on only one of the Affymetrix 250K chips than by typing half the number of samples on both of them *(34)*.

## 4. Statistical Analysis

The testing for significance in case–control association studies is fairly straightforward. In practice you are looking for an excess of one allele or geno-

**Table 3**
**A Simple Contingency Table (Counts) of Phenotype vs SNP Alleles**

| | | Allele | | |
|---|---|---|---|---|
| | | *A* | *B* | |
| Phenotype | *Case* | $N_{Case,A}$ | $N_{Case,B}$ | $N_{Case,.}$ |
| | *Control* | $N_{Control,A}$ | $N_{Control,B}$ | $N_{Control,.}$ |
| | | $N_{.,A}$ | $N_{.,B}$ | $N_{.,.}$ |

type, in your case population compared with the control population. This methodology is standard practice in epidemiological case–control studies in which you are examining your cases (HIV+) for excess exposure to a risk factor (i.v. drug abuse) compared with the controls. This association is easily tested using the standard $\chi^2$ statistic. Although a positive association does not prove causality it provides evidence and focus for functional investigation to demonstrate the causal relationship. Whole-genome screens test multiple (SNPs) hypotheses—one SNP or SNP combinations for each hypothesis—carrying out one test for each genotyped marker. This becomes a huge multiple testing problem, but methods to deal with this will be discussed later in this section.

### *4.1. Single SNP Analysis*

The simplest test of association is the allelic $\chi^2$ statistic where the number of A and/or B alleles for each individual is summed across cases and controls for a particular SNP marker, with each individual contributing two alleles for each SNP; this identifies meaningful contrasts between the cases and controls. The null hypothesis for the $\chi^2$ statistic is that there is no association between the rows (phenotype) and columns (SNP alleles) and a significant test statistic indicates an association between the phenotype and the SNP. The $2 \times 2$ table formed is simply (**Table 3**).

The expected counts (**Table 4**) are easily calculated from **Table 3**.

And the Pearson $\chi^2$ statistic (with 1df) becomes:

$$\chi^2 = \sum_{Phenotype} \sum_{Allele} \left( N_{Phen,Allele} - EN_{Phen,Allele} \right)^2 \div EN_{Phen,Allel} \quad (1)$$

To test the hypothesis that a particular genotype rather than an allele is associated with the phenotype you can use the genotypic $\chi^2$ statistic. This has the same layout as previously mentioned except that you are counting the specific genotype at a particular marker for each individual rather than the number of particular alleles the individual carries. This is a $2 \times 3$ table and this statistic will follow a $\chi^2$ distribution with two degrees of freedom. However, this test does

**Table 4**
**A Simple Contingency Table (Expected Counts) of Phenotype vs SNP Alleles**

| | | *Allele* | |
|---|---|---|---|
| | | *A* | *B* |
| Phenotype | *Case* | $(N_{Case,.}{*}N_{.,A})/N.,.$ | $(N_{Case,.}{*}N_{.,B})/N.,.$ |
| | *Control* | $(N_{Control,.}{*}N_{.,A})/N.,.$ | $(N_{Control,.}{*}N_{.,B})/N.,.$ |

not identify the risk genotypes that require a calculation of the genotypic odds ratio (OR; *see* **Subheading 4.2.**) to do so. The genotypic $\chi^2$ statistic will identify significant deviation from the expected distribution of genotypes at a particular SNP.

One can also use the likelihood ratio (LR) test, which can be found in all standard statistical textbooks. This is based on maximum likelihood estimation and the test always has one degree of freedom. Like the Pearson $\chi^2$ statistic, the null hypothesis of the LR $\chi^2$ statistic is no association between the phenotype and SNP alleles or genotypes and the alternative being that such an association exists. The log LR $\chi^2$ statistic is formed in the following manner (example given for allelic $\chi^2$ statistic) from the tables shown previously:

$$Log\ LR\chi^2 = \mathbf{2} * \sum_{Phenotype} \sum_{Allele} \boldsymbol{N}_{Phen,Allele} \ln\left(\frac{\boldsymbol{N}_{Phen,Allele}}{\boldsymbol{EN}_{Phen,Allele}}\right) \quad (2)$$

This statistic follows a $\chi^2$ distribution with 1d.f. and as the $N_{...}$ increases, the log LR $\chi^2 \sim$ Pearson $\chi^2$.

An alternative to low cell counts is to use the Fisher's exact test. This test assumes that the marginal numbers are fixed and randomizes the internal cell counts and the resulting statistic follows a hypergeometric distribution. If we consider **Table 3**, the null hypothesis we are testing is that there is no difference in the number of A and B alleles among cases and controls. The Fisher's exact test examines the number of A alleles among the case group ($N_{Case,A}$) and we compute the probabilities that $N_{Case,A} = n_{Case,A}$:

$$p(n_{Case,A}) = \frac{\binom{n_{Case,.}}{n_{Case,A}}\binom{n_{Control,.}}{n_{Control,A}}}{\binom{n_{.,.}}{n_{.,A}}} \quad (3)$$

We then use $N_{Case,A}$ as our test statistic and compute the probability from the null distribution *(35)*.

**Table 5**
**A Simple Contingency Table (Counts) of Phenotype vs SNP Genotypes**

| | | *Genotype* | |
|---|---|---|---|
| | | *AA* | *AB or BB* |
| Phenotype | *Case* | $N_{Case,AA}$ | $N_{Case,AB/BB}$ |
| | *Control* | $N_{Control,AA}$ | $N_{Control,AB/BB}$ |

### *4.2. Odds Ratio*

In the previous methods we have outlined both allelic and genotypic tests for association. Although the allelic tests indicate which allele is the risk allele, or conversely, which is the protective, the genotypic association tests tell you nothing about the underlying genetic model. In classic Mendelian genetics we think of dominant and recessive alleles. We can test these models for a particular SNP associated with the disease using the SNP genotypes and calculating the OR. Using the risk allele from the allelic $\chi^2$ statistic (higher frequency in cases than controls) we can test both dominant (individuals with one or two copies of the risk allele vs those with no copy) and the recessive (individuals with only two copies of the risk allele vs those with no or one copy). As an example, the odds ratio table for the recessive case (with A as the risk allele) is given in **Table 5**.

The resulting odds ratio is formed:

$$OR = \frac{N_{Case,AA*}\, N_{Control,AB/BB}}{N_{Control,AA}\, N_{Case,AB/BB}} \tag{4}$$

The significance of the odds ratio is tested by computing the 95% confidence interval around the OR value. Ninety-five percent confidence intervals around the OR (which must be >1) can be deemed significant at the $\alpha = 0.05$ level. The 95% CI is computed as:

$$95\%\,\mathrm{CI} = OR * \exp\left( \pm 1.96 * \sqrt{\frac{1}{N_{Case,AA}} + \frac{1}{N_{Case,AB/BB}} + \frac{1}{N_{Control,AA}} + \frac{1}{N_{Control,AB/BB}}} \right) \tag{5}$$

### *4.3. Multiple Testing*

No single issue in statistical genetics is more contentious than the issue of how to properly adjust for multiple testing. At the heart of the issue is that we are testing numerous hypotheses (one test for each marker genotyped) in the same dataset, which can be upward of 100,000–500,000 tests for one case–control study. For 100,000 markers, the expectation predicts 5000 markers will demonstrate a nominal *p*-value of ≤0.05 mostly or wholly from chance alone, making it difficult to distinguish the true associations from the much more frequent false

associations. One method to correct for this multiple testing is using Bonferroni correction. This method simply adjusts the cutoff value to declare significance at the $\alpha$ level by the number of tests performed by dividing $\alpha$ by the number of tests (significance threshold = $\alpha$/number of SNPs tested). The resulting value then is the threshold for declaring significance at the $\alpha$ significance level. This method for correcting for multiple tests has been criticized because it assumes independence of all markers and is overly conservative for tests of association in genetic-association studies because some of the closely spaced markers are correlated and the resulting lack of independence of the tests of association ***(36)***.

An alternative approach is to control the false discovery rate, which is defined as the expected number of false rejections divided by the number of rejections ***(37)***. It has been shown that this method is more powerful than the Bonferroni method and maintains a false-positive rate close to the nominal level ***(38)***. To perform the false discovery rate procedure for a number of SNPs ($n$) you simply order the $p$-values ($p_i$) from lowest to highest and assign the order value($i$, where $i = 1...n$). One then multiplies the global false-positive rate (0.05 in most cases) by $i$ and divides by $n$. This value is $q_i$. In the ordered list, SNPs, which have a $p_i$-value less than $q_i$ are considered significant at the prescribed global false-positive rate ***(37)***.

Another method is "step-down" permutation testing to compute the effective number of SNPs and empirical global $p$-values ***(39)***. In this approach, the case and control labels are shuffled a large number of times (for example, we used $10^6$ permutations in our AMD study). For each permutation, the $\chi^2$ test statistic is computed for each SNP. Then, the global $p$-value is computed for the real test statistic for each SNP. If the lowest $p$-value is less than 0.05, it is taken to be a real association, and that SNP along with the entire set of test statistics computed for its permutation, are removed from the dataset. The procedure is then repeated until the lowest $p$-value is more than 0.05. An effective number of SNPs can then be computed by dividing the empirical global $p$-value by the nominal $p$-value taken from the $\chi^2$ distribution for a given SNP; this number is always smaller than the number of tests used in the Bonferroni correction.

We performed the Westfall–Young permutation test and compared with the Bonferroni correction for the 116,000 null hypotheses in our AMD study. In this instance, we found that although stringent, Bonferroni indeed performed perfectly; the procedure eliminated thousands of associations with a nominal $p < 0.05$ none of which appear to have been erroneously discarded in our dataset (**Fig. 2**).

### 4.4. Haplotypes

The presence of haplotype blocks in the human genome creates other possibilities for association mapping. Haplotype blocks are regions of the chromosome in which only a few of the many possible haplotypes are observed. The

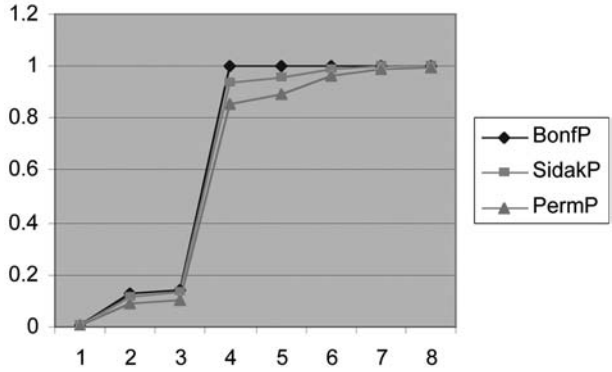



Fig. 2. *P*-values of three procedures for single-nucleotide polymorphisms (*x*-axis) ranked from small to large.

limited haplotype diversity found in these blocks allows for detecting association with an unobserved SNP that is found in one of these haplotypes simply by identifying which haplotype is present ***(40)***. One approach to performing such analyses involves exhaustively testing all possible haplotypes consisting of *n* adjacent SNPs, using permutation testing to determine statistical significance ***(41)***. This method has substantially better power than testing for association with individual SNPs alone, though in the genome-wide context it does require significant computing power.

Another approach is to define the edges of the haplotype block using the four-gamete test ***(42)***. In this test, for each pair of adjacent SNPs, the fraction of the four possible haplotypes that are observed, is tabulated. If all four possible haplotypes are observed, recombination is said to have taken place between the two SNPs and therefore a block boundary is located between the SNPs. Once the borders of haplotype blocks are defined, the haplotype frequencies need to be inferred and the haplotypes for specific individuals imputed based on the genotype data. A variety of methods can be used to do this. One method, SNPHAP, uses an expectation maximization algorithm to give a maximum likelihood estimate of haplotype frequencies (http://www-gene.cimr.cam.ac.uk/clayton/software/). Another, PHASE, uses a coalescent approach to determine haplotypes ***(43,44)***. A recent comparison of different haplotype inference algorithms is a useful starting point for determining what program is best for your specific situation ***(45)***.

Once the haplotypes in the block are determined, association between haplotype frequencies and disease status can then be calculated using a $\chi^2$ test on the appropriate $N \times 2$ contingency table. If it is necessary to reduce the degrees of freedom in the $\chi^2$ test, define different haplotypes to be either "risk" or "non-risk" and group the haplotypes by their risk level.

Phylogenetic analysis is powerful and useful, as it describes and maps the relationship between the haplotypes *(46)*. By building a phylogenetic tree of haplotypes, the evolutionary relationship between them can be visualized. In the simplest model, a single functional mutation arose somewhere on the tree. If the tree is divided at the point where this mutation arose, one would expect the haplotypes found on one side of the mutation in the tree to be at risk, and those found on the other side of the tree to be not at risk. The analysis of these trees depends on the complexity of the haplotypes and whether recombination or homoplasy (reoccurrence of the same mutation) is present.

### *4.5. Epistasis*

Often, a phenotype is not the simple sum of variations at different loci, but rather the interaction of variants at two or more loci in a more complex way. This phenomenon is called epistasis. When talking about association studies, we mean epistasis in the statistical sense, where the genotypes at two different loci nonadditively interact to alter the probability of a given phenotype. (Contrast this with how the term epistasis is typically used in model organism genetics, wherein it refers to ordering genes in a pathway by observing the phenotype of double mutants.) It has been shown that a trait can have relatively high heritability and yet have all of the genetic variance be from epistasis *(47)*. In such a case, simply looking at single locus effects would not reveal any associations.

There are several methods to look at epistasis when studying qualitative, dichotomous traits such as those found in a case–control association study. These methods typically look at the interaction between pairs of loci. One method is called multifactorial dimensionality reduction. It classifies each of the nine two-locus genotypes into different classes based on their risk level and evaluates the classification through cross-validation *(48)*. The method is computationally intensive, and therefore presently not well suited for the analysis of more than a few marker loci. Another approach uses logistic regression. Epistasis in quantitative traits has traditionally been analyzed by partitioning the epistatic variance into four orthogonal components: additive × additive, additive × dominant, dominant × additive, and dominant × dominant *(49)*. This technique has been extended to dichotomous traits by modeling the trait as having a value of either 0 or 1 and using logistic regression *(50,51)*.

We have developed a new method for finding epistasis that draws on both of these previous methods. Like the multifactorial dimensionality reduction, for a two-locus interaction we partition the nine possible genotypes into two or three genotype classes based on presumed risk level. However, the classification schemes are fixed and are inspired by Cockerham's partitioning of epistatic variance (*see* **Fig. 3**). For each of the four classification schemes, we count how many cases and controls are present at each risk level. Then, we can use a $3 \times 2$

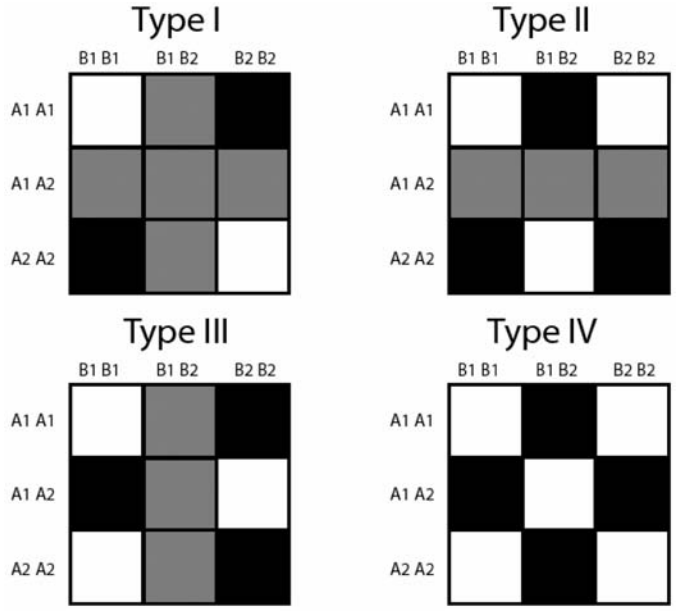



Fig. 3. Genotypic classes.

(types I–III interaction) or $2 \times 2$ (type-IV interaction) to compute a Pearson $\chi^2$ statistic and a *p*-value. In a genome-wide association setting, this can be done in an all-vs-all fashion (every SNP is compared with every other SNP) or one-vs-all fashion (a SNP that is known or suspected to be important is compared with all other SNPs). In any case, the *p*-values have to be corrected for multiple testing, not only for the number of interactions tested, but also an additional fourfold for the four different types of interactions for a given pair of SNPs.

## 5. Admixture Mapping

The presence of admixing presents an opportunity to discover genetic associations for diseases where prevalence is highly contrasted between the two mixed groups. It is necessary to measure the extent of admixture in exploiting this opportunity.

The first attempts to characterize admixture proportions in African Americans by means of genetic markers designated as "population-specific markers" dates back to the 1950s *(52)*. Recent effort in the International HapMap Project helps to increase the number of informative population-specific markers concerning the distributions of allele frequencies in the parental populations, Caucasian and African, to generate more precise estimates of the ancestral proportions of the admixed populations. For example, among the approx 3.5 million publicly available HapMap-genotyped SNPs there are 120,337 SNPs that show differences in

allele frequency greater than 50% between Africans (Yoruba tribe from Nigeria) and Europeans (CEPH, Utah residents with ancestry from northern and western Europe), and 814,272 SNPs of unique alleles are found in only one population. It is important in mapping disease genes to take into account the LD created when ethnic groups with different disease prevalence are mixed. For example, if the disease of interest has a higher prevalence among Europeans, the affected individuals tend to have a larger-than-average proportion of European ancestry. Thus, alleles that are common in Europeans are likely to be over represented among cases, leading to spurious associations if such a confounding factor is ignored. The genetic heterogeneity can mimic the signal of association; as the level of structure increases, the $\chi^2$ test for association becomes less conservative *(53)*.

Admixture mapping can be ideally applied if population 1 and population 2 carry a different allele at the disease locus. Whole-genome scanning under the admixture mapping strategy consists of scanning the genome and identifying the regions with an excess of population 1 ancestry in the cases vs the controls, assuming that population 1 carries the predisposition allele. The size of the LD blocks from different ancestors will depend on the number of generations because the populations were mixed.

There have long been analytical approaches to deal with genetic datasets from recently admixed populations. For example, the admixture proportions of the African-American and European-American populations were estimated by the weighted least squares *(54)* and gene-identity methods *(42)*. The expectation maximization algorithm has been employed to estimate haplotype frequencies and LD coefficients for pairs of loci *(55)*. Heterogeneity in allele frequencies of the parental populations can be tested by the likelihood ratio statistics with $\chi^2$ distribution for large sample sizes or by Monte Carlo data resampling under no association. The STRAT approach uses a set of unlinked markers to infer population structure, to estimate the ancestral proportions of sample individuals, and then to test for associations within subpopulations via the likelihood ratio test *(20)*.

Recently, marker maps have been developed to take advantage of the LD created in admixed populations *(56)* as well as a method to exploit these marker maps for disease gene mapping *(57)*. This method his been successful in mapping a candidate locus for multiple sclerosis susceptibility *(57)*, however, this result has not yet been replicated nor the functional mutation identified. One crucial issue with these admixture methods is that the underlying models used for adjustment are often inflexible and do not reflect the true admixture scenario *(58)*. Great care must be taken to properly adjust for the underlying admixture in a study to avoid over-correction for ethnicity specificity and false-negative findings *(9)*.

## 6. Bioinformatic Databases

Bioinformatic databases are another tool important for deciphering association studies. Although numerous bioinformatics databases on the human

genome and the variation present within the genome exist, we have found three databases to be especially helpful for our studies. The first of these databases is dbSNP (http://www.ncbi.nlm.nih.gov/SNP). This database contains information on known SNPs in the human genome. Each SNP has an accession number that is of the form "rs#" where # represents a number. Although some genotyping platforms have their own accession numbers for SNPs, all the genotyping methods we know about will give you an rs accession number for the SNP if available. If you look up a SNP by its accession number, you can get information on its genomic location, what gene it is located in, and what its putative function may be (UTR, synonymous coding, nonsynonymous coding, and so on). The two other databases we find useful are ENSEMBL (http://www.ensembl.org) and the UCSC genome browser (http://genome.ucsc.edu). These databases allow you to browse the human genome with different levels of annotations. You can visualize what SNPs are in a region, what genes are there, conservation between regions in other organisms, and expressed sequence tag (EST) expression data. Each database provides different yet overlapping information, and the data provided by each one is constantly growing. You can query both databases using a dbSNP rs accession number to visualize the genomic region surrounding the SNP. If you find association with a specific SNP, we recommend looking up the SNP in these databases to get the maximal information possible about the region.

There is one final caveat about these databases we want to mention. Care is essential when comparing chromosomal positions in that the same build of the human genome be used. In the process of finishing the human genome, several public releases, or "builds," were made. Each build has a slightly different sequence, and therefore the numbering of specific nucleotides in the chromosomes is different. It is imperative when using these databases that the same build is being used. There are three main builds of the human genome to be aware of: build 34 (July 2003), build 35 (May 2004), and build 36. The UCSC browser references the date of the build, whereas ENSEMBL references the build number. As of this writing, dbSNP does not reference the build but appears to be using build 35 coordinates. In dbSNP, use the coordinates for the "reference" group. There is another assembly ("Celera") based not on the public genome project but on the version sequenced by Celera and since made public. This assembly uses a different coordinate scheme and is not found in the other databases.

### *6.1. Growing Information Resources*

The number of available SNP markers has tripled to a total of 8 million SNPs (i.e., an average of one SNP/360 bp in the genome) and 1 million indel polymorphisms (based on dbSNP build 121 at http://www.ncbi.nlm.nih.gov/SNP).

1. Allele frequency data are available on approx 1.5 million SNPs distributed throughout the genome as a result of the International HapMap Project (http://www.hapmap.org) and of Perlegen, Inc. (http://www.perlegen.com). The first genome-wide panels of haplotype-tagging SNPs (tSNPs) for association studies are becoming available *(12)*.
2. A number of further SNP discovery studies are focused on genes: (1) the Seattle gene resequencing database (http://pga.gs.washington.edu/summary_stats.html); (2) the Innate Immunity Program (http://innateimmunity.net/); and (3) the Japanese JSNP site (http://snp.ims.u-tokyo.ac.jp/index.html).
3. From the ENCODE project (http://www.genome.gov/10005107) 10 0.5-Mb genomic regions selected are currently being sequenced in multiple individuals of Northern European, West African, and East Asian origin. Analysis of sequences from approx 48 individuals (96 chromosomes) resulted in the detection of 30–50% previously undiscovered SNPs.
4. Extensive sequencing has been carried out for chromosome 10 *(33)* and of individual haplotypes in the human leukocyte antigen region (HLA) *(59)*.
5. Approximately 450,000 SNPs have been rescreened for their allele frequency differences between African and European population samples *(56,57)*. And a mapping by admixture LD map for 3011 SNPs most differentiating two ethnic groups has been established (http://genepath.med.harvard.edu/~reich/).
6. A genome-wide association mapping for age-related macular degeneration, complete data, and analytical tools are soon to be available at http://variation.yale.edu.

## Acknowledgments

We gratefully acknowledge Dr. Richard Sackler for critically reading and editing the manuscript. This work was supported by the following funding sources: The Ellison Medical Foundation (JH), The Yale Pepper Center Fund (JH), an institutional award from the Howard Hughes Medical Institute to the Yale School of Medicine, and grants and fellowships from the National Institutes of Health (RK, JH).

## References

1 Cardon, L. R. and Bell, J. I. (2001) Association study designs for complex diseases. *Nat. Rev. Genet.* **2,** 91–99.
2 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517.
3 Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A., and Contopoulos-Ioannidis, D. G. (2001) Replication validity of genetic association studies. *Nat. Genet.* **29,** 306–309.
4 Risch, N. and Zhang, H. (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268,** 1584–1589.
5 Klein, R. J., Zeiss, C., Chew, E. Y., et al. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* **308,** 385–389.

6 Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33,** 228–237.

7 Kruglyak, L. and Nickerson, D. A. (2001) Variation is the spice of life. *Nat. Genet.* **27,** 234–236.

8 Hoh, J. and Ott, J. (2003) Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Rev. Genet.* **4,** 701–709.

9 Cordell, H. J. and Clayton, D. G. (2005) Genetic association studies. *Lancet* **366,** 1121–1131.

10 Palmer, L. J. and Cardon, L. R. (2005) Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* **366,** 1223–1234.

11 Romualdi, C., Balding, D., and Nasidze, E. S., et al. (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res.* **12,** 602–612.

12 Hinds, D. A., Stuve, L. L., Nilsen, G. B., et al. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* **307,** 1072–1079.

13 Lewontin, R. (1972) The apportionment of human diversity. In: *Evolutionary Biology*, (Dobzhansky, T., Hecht, M., and Steere, W., eds.), Appleton-Century-Crofts, New York, NY, pp. 381–398.

14 Cooper, R.S., Kaufman, J. S., and Ward, R. (2003) Race and genomics. *N. Engl. J. Med.* **348,** 1166–1170.

15 Haga, S. B. and Venter, J. C. (2003) Genetics. FDA races in wrong direction. *Science* **301,** 466.

16 Wilson, J. F., Weale, M. E., Smith, A. C., et al. (2001) Population genetic structure of variable drug response. *Nat. Genet.* **29,** 265–269.

17 Risch, N., Burchard, E., Ziv, E., and Tang, H. (2002) Categorization of humans in biomedical research: genes, race and disease. *Genome Biol.* **3,** comment 2007.

18 Burchard, E. G., Ziv, E., Coyle, N., et al. (2003) The importance of race and ethnic background in biomedical research and clinical practice. *N. Engl. J. Med.* **348,** 1170–1175.

19 Stephens, J. C., Schneider, J. A., Tanguay, D. A., et al. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293,** 489–493.

20 Rosenberg, N. A., Pritchard, J. K., Weber, J., et al. (2002) Genetic structure of human populations. *Science* **298,** 2381–2385.

21 Tang, H., Quertermous, T., Rodriguez, B., et al. (2005) Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am. J. Hum. Genet.* **76,** 268–275.

22 Helgason, A., Yngvadottir, B., Hrafnkelsson, B., Gulcher, J., and Stefansson, K. (2005) An Icelandic example of the impact of population structure on association studies. *Nat. Genet.* **37,** 90–95.

23 Bersaglieri, T., Saberti, P. C., Patterson, N., et al. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74,** 1111–1120.

24 Campbell, C. D., Ogburn, E. L., Lunetta, K. L., et al. (2005) Demonstrating stratification in a European American population. *Nat. Genet.* **37,** 868–872.

25 International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* **437,** 1299–1320.
26 Ingman, M., Kaessmann, H., Paabo, S., and Gyllensten, U. (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* **408,** 708–713.
27 Herrnstadt, C., Elson, J. L., Fahy, E., et al. (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. *Am. J. Hum. Genet.* **70,** 1152–1171.
28 Taylor, R. W. and Turnbull, D. M. (2005) Mitochondrial DNA mutations in human disease. *Nat. Rev. Genet.* **6,** 389–402.
29 Matsuzaki, H., Dong, S., Loi, H., et al. (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* **1,** 109–111.
30 Matsuzaki, H., Loi, H., Dong, S., et al. (2004) Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res.* **14,** 414–425.
31 Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G., and Chee, M. S. (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.* **37,** 549–554.
32 Steemers, F. J., Chang, W., Lee, G., Barker, D. L., Shen, R., and Gunderson, K. L. (2006) Whole-genome genotyping with the single-base extension assay. *Nat. Methods* **3,** 31–33.
33 Deloukas, P., Earthrowl, M. E., Grafham, D. V., et al. (2004) The DNA sequence and comparative analysis of human chromosome 10. *Nature* **429,** 375–381.
34 Klein, R. S. (Unpublished).
35 Rice, J. (1995) *Mathematical Statistics and Data Analysis. 2nd ed.,* Duxbury Press, Belmont, CA.
36 McIntyre, L. M., Martin, E. R., Simonsen, K. L., and Kaplan, N. L. (2000) Circumventing multiple testing: a multilocus Monte Carlo approach to testing for association. *Genet. Epidemiol.* **19,** 18–29.
37 Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.* **57,** 289–300.
38 Devlin, B., Roeder, K., and Wasserman, L. (2003) Analysis of multilocus models of association. *Genet. Epidemiol.* **25,** 36–47.
39 Westfall, P. and Young, S. (1989) pValue adjustments for multiple tests in multivariate binomial models. *J. Am. Stat. Assoc.* **84,** 780–786.
40 Gabriel, S. B., Schaffner, S. F., Nguyen, H., et al. (2002) The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229.
41 Lin, S., Chakravarti, A., and Cutler, D. J. (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat. Genet.* **36,** 1181–1188.
42 Wang, N., Akey, J. M., Zhang, K., Chakraborty, R., and Jin, L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* **71,** 1227–1234.
43 Stephens, M. and Donnelly, P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73,** 1162–1169.

44 Stephens, M., Smith, N. J., and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68,** 978–989.
45 Marchini, J., Cutler, D., Patterson, N., et al. (2006) A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* **78,** 437–450.
46 Templeton, A. R., Boerwinkle, E., and Sing, C. F. (1987) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics* **117,** 343–351.
47 Culverhouse, R., Suarez, B. K., Lin, J., and Reich, T. (2002) A perspective on epistasis: limits of models displaying no main effect. *Am. J. Hum. Genet.* **70,** 461–471.
48 Ritchie, M. D., Hahn, L. W., Roodi, N., et al. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **69,** 138–147.
49 Cockerham, C. C. (1954) An extension of the concept of partitioning hereditary variance for analysis of covariates among relatives when epistasis is present. *Genetics* **39,** 859–882.
50 Cordell, H. J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **11,** 2463–2468.
51 North, B. V., Curtis, D., and Sham, P. C. (2005) Application of logistic regression to case-control association studies involving two causative loci. *Hum. Hered.* **59,** 79–87.
52 Glass, B. and Li, C. C. (1953) The dynamics of racial intermixture; an analysis based on the American Negro. *Am. J. Hum. Genet.* **5,** 1–20.
53 Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004) The effects of human population structure on large genetic association studies. *Nat. Genet.* **36,** 512–517.
54 Long, J. C. (1991) The genetic structure of admixed populations. *Genetics* **127,** 417–428.
55 Long, J. C., Williams, R. C., and Urbanek, M. (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56,** 799–810.
56 Smith, M. W., Patterson, N., Lautenberger, J. A., et al. (2004) A high-density admixture map for disease gene discovery in african americans. *Am. J. Hum. Genet.* **74,** 1001–1013.
57 Patterson, N., Hattangadi, N., Lane, B., et al. (2004) Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74,** 979–1000.
58 Hoggart, C. J., Parra, E. J., Shriver, M. D., et al. (2003) Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet.* **72,** 1492–1504.
59 Stewart, C. A., Horton, R., Allcock, R. J., et al. (2004) Complete MHC haplotype sequencing for common disease gene mapping. *Genome Res.* **14,** 1176–1187.

# 8

# Linkage Disequilibrium Maps and Disease-Association Mapping

**Nikolas Maniatis**

**Summary**

Over the last few years, association mapping of disease genes has developed into one of the most dynamic research areas of human genetics. It focuses on identifying functional polymorphisms that predispose to complex diseases. Population-based approaches are concerned with exploiting linkage disequilibrium (LD) between single-nucleotide polymorphism (SNPs) and disease-predisposing loci. The utility of SNPs in association mapping is now well established and the interest in this field has been escalated by the discovery of millions of SNPs across the genome. This chapter reviews an association-mapping method that utilizes metric LD maps in LD units and employs a composite likelihood approach to combine information from all single SNP tests. It applies a model that incorporates a parameter for the location of the causal polymorphism. A proof-of-principle application of this method to a small region is given and its potential properties to large-scale datasets are discussed.

**Key Words:** Association mapping; linkage disequilibrium; LDU; single SNPs; CYP2D6.

## 1. Introduction

The central aim of association mapping is to identify genes that contribute to complex diseases. The first step identifies candidate regions in the genome that are associated with the disease of interest. Subsequently, association mapping focuses on finer localization of disease determinants and the ultimate identification of the causal variants. Association mapping superseded linkage-mapping approaches that were used with great success to locate major genes, which are relatively rare but have a large phenotypic effect. In contrast to linkage analyses, which permit comparatively low-resolution mapping with the available family resources, efforts to map genes of complex diseases are concerned with

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

exploiting linkage disequilibrium (LD) between markers and putative disease-predisposing loci, usually from population samples. LD analysis offers the prospect of fine-scale localization of genetic polymorphisms of medical importance, particularly when single-nucleotide polymorphisms (SNPs) are densely typed in a candidate region. This enthusiasm has been intensified by the discovery of millions of SNPs. Current estimates suggest that the human genome may have as many as 15 million such markers *(1)*. Among this vast number of polymorphisms only a very small number play a significant role in complex diseases. Therefore, the identification of genetic susceptibility factors for common disease remains a great challenge, but it is increasingly apparent that in order to enhance progress and assure success with association studies, detailed information on the underlying structure of LD is required. The aim of this chapter is to describe the theoretical framework of an association-mapping approach that utilizes LD maps, and to review its properties and potential applications to complex inheritance.

## 2. LD Maps

LD describes the tendency of alleles located close to each other on the same chromosome to be coinherited. It exists simply because small segments of ancestral haplotypes are transmitted unbroken over many generations. LD is present when recombination between alleles at a small distance is infrequent. There is much variation in the extent of LD across the genome and although recombination is the predominant cause of breakdown in LD, other evolutionary factors may also influence the LD patterns. LD plays a fundamental role in association mapping because it can provide high-resolution information to narrow a chromosomal region of interest and refine the location of the disease gene. There has been much emphasis on determining the haplotype-block structure of the human genome, but the utility of haplotype-block identification for disease gene localization still remains uncertain. Maniatis et al. *(2)* provided an alternative approach, which develops a metric map in LD units (LDUs) to describe the underlying pattern of LD by assigning an LDU for each marker SNP. These maps are analogous to linkage maps and discriminate blocks of conserved LD with additive distances. The methodology for their construction extends the seminal work of Collins and Morton *(3)* and Morton et al. *(4)* that describe a novel adaptation of the Malecot model *(5)* that predicts background levels of LD resulting from evolutionary history. LD declines exponentially with physical distance ($d$) in kilobases and the Malecot model describes the association rho ($\rho$) between any pair of SNPs as $\rho = (1-L)Me^{-\varepsilon d}+L$, where $M$ represents the initial value of LD before decay begins. It is the association at zero distance and represents an estimate of the association at the last major bottleneck. The proportionality parameter epsilon ($\varepsilon$) is the exponential decline

of association with physical distance $d$ and reflects the product of recombination $\theta$ and time $t$ in generations where recombination has taken place. However, the theory becomes more useful with $\varepsilon d$, simply because it is more accurately known than $\theta t$. LD declines until it reaches an asymptotic level $L$, which acts as a correction factor for spurious association. The parameters $\varepsilon$, $M$, and $L$ are not known but can be estimated iteratively using composite likelihood based on the observed pairwise marker-by-marker association $\hat{\rho}=D/Q(1-R)$. The minor allele frequencies $Q$, $R$, and the covariance $D$ are obtained from the $2 \times 2$ haplotype table. The columns and rows of the table are rearranged so $D$ is always positive, making $Q < 0.5$ the frequency of the putatively youngest allele. Regardless of any rearrangement of the $2 \times 2$ table, the $\chi^2_1$ is always the same. $M$ is the parameter with evolutionary interpretation and not just the association $\rho_0$ at 0 time $t$. It is estimated as ***(6)***:

$$M = (\rho_0 - L)e^{-(v+1/2Ne)t} / (1 - L)$$

where $Ne$ is the effective population size and $v$ is the linear pressure toward LD from migration and mutation ***(6)***.

The LDU map method estimates $\varepsilon$ in each ($i^{th}$) map interval by fitting the model only to the pairwise marker associations that are informative for that interval. A map interval in LDU is simply the product $\varepsilon_i d_i$ within a region having $\Sigma\varepsilon_i d_i$ LDU ***(2)***. **Figure 1A** shows the block-step structure of a 216-kb region of chromosome 6. Blocks of high LD are an uninterrupted sequence where LDU = 0 so SNPs within a block are completely associated with one another. Blocks are separated by areas of LD breakdown. These steps of reduced LD are defined as LDU > 0. Zhang et al. ***(7)*** have shown that there is a remarkable agreement between LDU steps (**Fig. 1A**) and recombination hot spots (**Fig. 1B**) presented by Jeffreys et al. ***(8)***, which confirmed the location of recombination hot spots by sperm typing. Sperm typing has two major limitations: it cannot determine female recombination, and it can only be applied to small regions because of its high cost and effort. LDU maps on the other hand can be easily constructed for the entire genome using the data provided by the HapMap Project ***(9)***. These fine-scale metric maps provide valuable information about the pattern of LD. Using whole-chromosome linkage maps at low resolution, Tapper et al. ***(9)*** have shown that more than 90% of the variation in LDU is explained by recombination. Low-resolution linkage maps are based on families with few meioses, whereas LDU maps reflect historical meiotic events. Although recombination dominates the LD structure, the great advantage of the LDU map method is that it models the decline of LD and not just recombination. The decline of LD is due to pressures other than recombination, such as mutation, selection, and long-range migration ***(4)***.

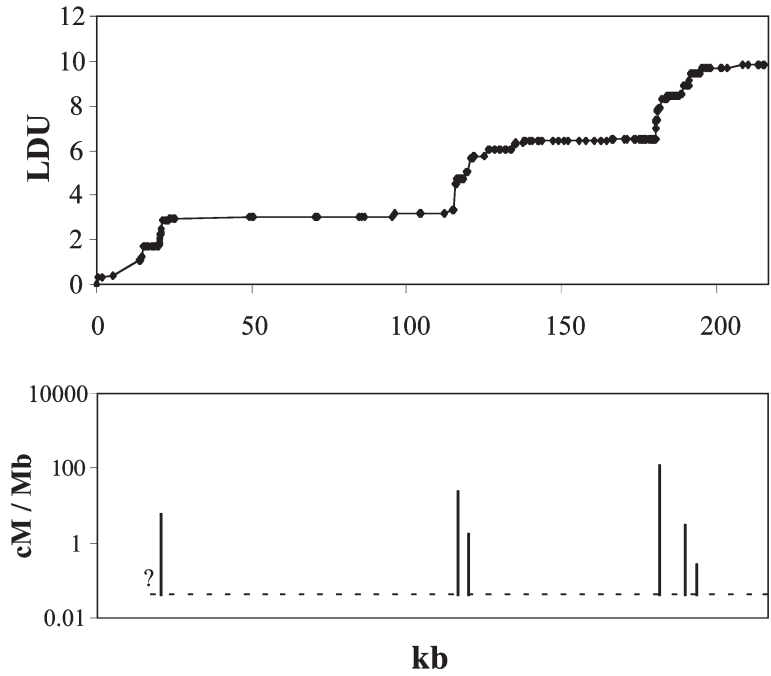



Fig. 1. Graph of a linkage-disequilibrium unit map **(A)** for a 216-kb segment of class II region of MHC *(7)* with corresponding hot spots **(B)** *(8)*.

## 3. Association Mapping Using LDU Maps

Following the development of LDU maps, their application to association mapping was examined by adapting the Malecot model and including a parameter for the location of the causal polymorphism *(10)*. This method exploits LD maps by assigning an LDU location for each marker. The method models association across single-marker tests within a composite likelihood framework. The properties of this method were first examined by simulating each SNP as causal from two existing real SNP datasets *(8,11)* that differed substantially in the number of markers and sample size *(10)*. By use of regression or correlations, false-negative indications of a disease locus (type-II error) were examined by treating each SNP as causal and predicting its location from the remaining markers. This simulation study showed that greater power is achieved when using an LDU map compared with a map in kilobases, especially in a densely typed region *(8)* that is characterized by intense recombination hot spots (**Fig. 1**). Composite likelihood has the advantage of considering all association SNP tests simultaneously and, as a result, avoids a heavy Bonferroni correction for multiple testing. On the other

**Table 1**
**Frequencies in a Random Sample: A 2 × 2 Table Between Affection Status and SNP Alleles**

| Affection status | | Genotypes | | |
|---|---|---|---|---|
| | | **AA** | **Aa** | **aa** |
| Affected | **1** | $n_{11}$ | $n_{12}$ | $n_{13}$ |
| Normal | **0** | $n_{21}$ | $n_{22}$ | $n_{23}$ |

| Affection status | | Disease-associated marker alleles | | |
|---|---|---|---|---|
| | | **A** | **a** | **Total** |
| **Affected** | count | **a**=$2n_{11}+n_{12}$ | **b** = $2n_{13}+n_{12}$ | ***f*** a+b |
| **Normal** | count | **c** = $2n_{21}+n_{22}$ | **d** = $2n_{23}+n_{22}$ | ***1−f*** c+d |
| Total | Frequencies | ***R*** | ***1-R*** | **1** |
| | count | a+c | b+d | n=a+b+c+d |

hand, it assumes independence among marker SNPs and hence raises questions about the reliability of significance tests because there is always autocorrelation in SNP datasets, especially for densely typed regions ***(10,12)***. This issue was addressed in the same simulation study ***(10)*** by investigating false-positive indications of a disease locus (type-I error). The analysis showed that the $\chi^2$ distribution of 1000 permutations yielded an acceptable goodness of fit even for a densely typed region (296 SNPs typed in a 216-kb region; **Fig. 1**). This demonstrates that composite likelihood works well, despite its assumption of independence among the marker SNPs. Devlin et al. ***(12)*** concluded that composite likelihood methods must be somewhat inefficient compared with a full likelihood model but the latter would be very difficult to specify without unrealistically stringent assumptions about population history.

### *3.1. Theoretical Framework*

Several factors determine the power to identify a candidate region for a gene contributing to a particular phenotype, and within that region to localize a causal polymorphism. The use of a metric with appropriate theory is essential. Ignoring the phenotype and using the pairwise marker-by-marker association ρ to create the LDU map, the metric ρ is adapted in order to obtain an association metric $\hat{z}$ from the 2 × 2 table between the affection status (0, 1 disease phenotype) and the two alleles of every marker SNP ***(13,14)***. **Table 1** shows the four counts (*a*, *b*, *c*, and *d*) in a 2 × 2 table under the simplest case of a random sample. The association metric is then estimated as:

$$\hat{z} = (ad-bc)/(a+b)(b+d), \text{ which is equal to:}$$

$$= D/f(1-R),$$

where $D$ is the covariance between affection status and the markers alleles, $f$ is the frequency of affected individuals in the population, and $R$ is the minor allele frequency. The columns of **Table 1B** can be interchanged so $ad-bc$ is always positive. The rows cannot be interchanged, however, because $f$ is the frequency of affected individuals in the sample. This is under the simplest case of a random sample but this metric can be extended to accommodate ascertainment in cases and controls using a correction factor based on population frequency ***(13)***. The number of single tests that performed is equal to the number of SNPs in the region under investigation. Subsequently, the association-mapping approach is applied, which is based on an extension of the Malecot model:

$$z = (1-L)Me^{-\varepsilon\,|(Si-S)|} + L$$

where $z$ is the expected association, $M$ is approx 1 if the disease alleles are monophyletic or less than 1 if there are multiple mutations at the disease locus. The objective of this method is to estimate $S$, which is the location of the disease gene in the marker map and $S_i$, the location of the ith marker, which can be expressed in kilobases or in LDU. The parameters in the model are estimated iteratively using composite likelihood ($\Lambda$) that combines information from all the single-marker tests. $\Lambda$ is estimated as: $\Sigma K_i (\hat{z}_i - z_i)^2$, where $\hat{z}_i$ and $z_i$ are the observed and expected association values, respectively, at the ith marker SNP, weighted for nominal information $K_i$. Therefore, every observed estimate of $\hat{z}_i$ has an amount of information, $K_i$, which is estimated as:

$$K = \chi^2_1 / \hat{z}^2 = n(a+b)(b+d)/(a+c)(c+d)$$

and the Pearson's $\chi^2_1$ for the $2 \times 2$ table is estimated as:

$$\chi^2_1 = K\hat{z}^2$$

$$= [n(a+b)(b+d)/(a+c)(c+d)]\,[(ad-bc)/(a+b)(b+d)]^2$$
$$= n(a+b)(b+d)\,(ad-bc)^2 / (a+b)^2(b+d)^2\,(a+c)(c+d)$$
$$= n(ad-bc)^2 / (a+b)(b+d)(a+c)(c+d)$$

Estimates of $\hat{z}$, $\chi^2_1$ and $K$ are under the null hypothesis because the expected counts $a$, $b$, $c$, and $d$ are not known prior to expectation by the Malecot model. The model can accommodate metrics other than $\hat{z}$, depending on the phenotype. The metric $z$, however, has been shown to outperform other metrics when the phenotype is the affection status ***(13)***. Even though the majority of phenotypes are presented as affected and normal, other metrics such as regression must be used for quantitative traits ***(15)***.

### 3.2. Modeling and Significance Testing

Various subhypotheses of the Malecot model could be used in order to test significance and the existence of a causal polymorphism. For example, the

baseline (model A) represents the null hypothesis of no association across the region. This implies that none of the parameters will be estimated and hence the intercept has $M$=0. Alternative models (model C and D) allow estimation of both parameters $M$ and $S$ but differing on parameter $L$. Model D iterates the parameter $L$, whereas model C fixes the value in $L$ with an estimate that is obtained from the mean deviation of al the information $K_i$. These two models have shown consistent results, but model C fits better than D in cases where the candidate regions are too short for the asymptote $L$ to be estimated accurately. Contrasting model A with any of these two alternative models (A–C and A–D contrasts) allows for significance testing for a disease determinant at location $S$. The significance for these two contrasts can be tested by the use of $\chi^2$ ***(13)***. For example, the A–C test, in large-sample theory, has an estimate of $\chi^2$ as:

$$\chi^2_2 = (\Lambda_A - \Lambda_C) / V_C,$$

where $V_C$ is the residual error variance of model C and is computed by dividing the weighted sum of squares $\Lambda_C$ with the degrees of freedom to give $V$=$\Lambda_C$/$m$. The degrees of freedom, $m$, equals the number of SNPs minus the number of parameters $k$ in the model (e.g., model C has two parameters and thus the A–C test has a $\chi^2_2$, whereas the A–D test has a $\chi^2_3$). However, an F test is more reliable than $\chi^2$ when the degrees of freedom are small (i.e., for small candidate regions, the number of marker tests is usually small). Therefore, significance can be computed by the use of the F test ***(14)*** with an F-value being estimated as the ratio of the between— models mean square to the error mean square ***(16)***, which is the error variance $V$. For example, for the A–C contrast the significance test is:

$$F(k,m) = [(\Lambda_A - \Lambda_C) / k] / V_C.$$

The F test for the A–D contrast can be computed in the same way using the corresponding sum of squares $\Lambda_D$ and error variance $V_D$. The 95% confidence interval (CI) for the estimated location $\hat{S}$ can be obtained as: $\hat{s} \pm t\ SE$, where $t$ is the tabulated value of Student's $t$-test for $m$ degrees of freedom. The empirical standard error of parameter $\hat{S}$ is $SE = \sigma_s\sqrt{V}$, where $\sigma_S$ is the nominal standard error of $\hat{S}$ estimated by quadratic approximation of the composite likelihood. The corresponding significance $\chi^2_i$ for every point $i$ with $S$ specified can also be estimated and hence estimates of $\Lambda_C(S)$ are obtained by fitting the Malecot model to specified values of $S$ in kilobases or LDU ***(14)***. Therefore, the lod surfaces ($2$=$\chi^2_i/2ln10$) *see also* **Fig. 3** for given values of $S$ can be obtained for both maps and models. These surfaces can also be used to estimate the lod support interval as an alternative to the CI. Plotting the likelihood surface against the LDU map may reveal different maxima other than the maximum likelihood estimate ($\hat{S}$), which is of great importance because disease gene mapping may identify several causal sites in a region.

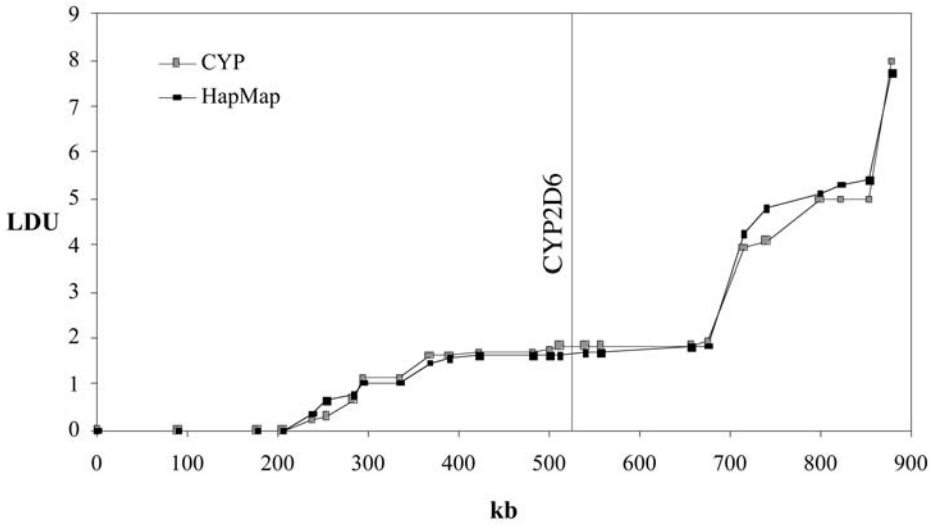



Fig. 2. The graph of the linkage disequilibrium unit (LDU) map for the *CYP2D6* region. The vertical line indicates the location of the locus at 525.3 kb.

### *3.3. A Proof-of-Principle Application*

Having established the superiority of LDU maps in simulation studies, these mapping procedures were subsequently applied to a 900-kb region flanking the *CYP2D6* gene that is associated with poor drug-metabolizing activity. This random sample was introduced as a test of association between the "poor-metabolizer" (0,1) phenotype and the 27 SNPs that flanked the gene (i.e., none of these markers were within the gene) ***(17)***. Previous analyses based on single SNP significance tests have shown that the region associated with the phenotype spans a 390-kb interval ***(17)***.

Using LDU locations for 27 SNPs flanking the *CYP2D6* gene on chromosome 22 (CYP LDU locations; **Fig. 2**), the most common functional polymorphism within the gene was predicted at a location only 14.9 kb away from its known true location, surrounded within a 95% CI of 172 kb ***(13)***. **Figure 3** shows the likelihood surface for the *CYP2D6* region, whereby the maximum likelihood estimation (peak) was found close to the true location of the *CYP2D6* locus (vertical line) despite being in the middle of a block (**Fig. 2**). Using an LDU map, analysis yielded a smaller CI and location error compared to the kilobase map. Simultaneous with this publication, data on 1.2 million SNPs in four populations were publicly released by the international HapMap Consortium, Phase I of the HapMap Project (http://www.hapmap.org/). The enormous body of data created by the HapMap Project enables the creation of high-resolution, population-specific LD maps, and their locations in LDU can be directly used in

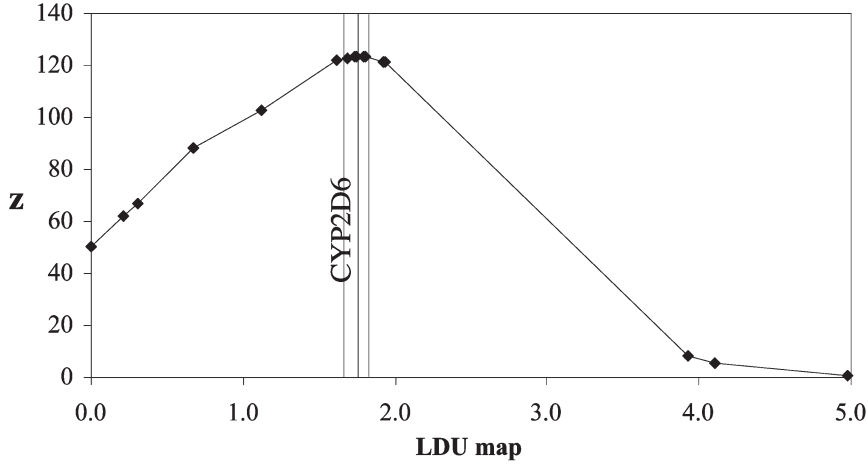



Fig. 3. Localization of *CYP2D6* locus (vertical line). The likelihood surface ($z = \chi^2_i / 2ln10$) plotted on the linkage-disequilibrium unit map and the 95% confidence interval (gray bars).

association-mapping studies to narrow a candidate region and increase the precision of localization. This led to our construction of whole-genome LDU maps ***(9)***. The performance of this mapping approach was further examined by comparing the high-density LDU map constructed from the HapMap data (the corresponding HapMap LDU locations for the 27 SNPs; **Fig. 2**) and the LDU locations obtained from *CYP2D6* (CYP LDU; **Fig. 2**). Such comparison is feasible not only for this example but for every association study simply because LDU maps are constructed on the basis of pairwise marker-by-marker associations where there is no phenotype involved and for only the unrelated healthy individuals or controls. Expressing the locations of the 27 SNPs in LDU from the HapMap LDU map analysis yielded an estimated location that is only 0.3 kb away from the gene despite the similarity of the two maps ***(14)***. This strongly supports the use of the high marker density HapMap-derived LDU map for association mapping.

### *3.4. Single SNPs vs Haplotypes*

So far we have only considered the properties of association-mapping methodology based on single SNPs. A popular belief is that haplotypes always provide greater power to detect disease genes when the SNPs tested are not functional but in LD with the causal locus. Interest in the analysis of haplotypes has increased as a result of the emphasis given by the International HapMap Project and other related initiatives ***(18)***. Using the mapping approach that is presented here, single SNPs and haplotypes were compared under the same

modeling procedures (i.e., using the Malecot model and composite likelihood). The simplest case of only two adjacent SNPs (2-SNP haplosets) was considered, e.g., SNP pairs 1,2 and 2,3 by assigning midpoint locations in kilobases and LDU. The statistical inference of haplotypes was conducted using the algorithm presented by Hill ***(19)*** for a pair of diallelic loci in a panmictic population. The association metric was obtained from the $2 \times 2$ table between affection status and most significant haplotype for two SNPs (e.g., AC vs Ac+aC+ac). In this study, it was shown that haplotypes can lead to poor estimates of localization compared with single SNPs in the analysis of the 27 SNPs flanking the *CYP2D6* gene. This analysis also showed that the poor localization of the gene was accompanied by great inflation of the significance level. No correction was used for selecting the most significant haplotype, and so this procedure may inflate the covariance between adjacent pairs that share the intervening marker. Nevertheless, this the simplest haploset of size 2, which is particularly useful because intervals do not overlap (intervals 1,2 and 2,3 do not overlap). Nevertheless, even this simplest case of haplotypes has a drawback. With the exception of the first and the last markers in the region, all SNPs are used twice, e.g., in haplosets 1–2, 2–3, 3–4, SNPs 2 and 3 are used twice. Such use of duplicated SNPs among haplosets can generate additional autocorrelation among SNPs, which can upwardly bias the significance test ***(14)***. The situation worsens with longer haplotypes simply because windows of three or more SNPs overlap (e.g., for haplosets 1–2–3 and 2–3–4, the interval for SNPs 2 and 3 is overlapped). Therefore, as the number of SNPs increases, the amount of autocorrelation in the data will also progressively increase.

Recent descriptions have focused on delimiting blocks of low haplotype diversity ***(20)***. This suggests that we can ignore overlapping windows and analyze haplotype blocks instead ***(21)***. This approach will generate additional problems using the procedures presented here because different haplosets may yield different numbers of significant haplotypes and, hence, variable degrees of freedom. Most importantly, longer haplotypes can be more ambiguous, especially in regions of high recombination because they require phase information from phase-unknown genotypes (diplotypes). Following the procedures presented here, haplotype analysis did not provide better estimates of the *CYP2D6* localization compared with single SNP tests, but the number of different ways to use haplotypes is large and alternative approaches ***(22)*** that account for autocorrelation may obtain more favorable results. In a comprehensive review of literature ***(18)***, it was shown that there are more than 40 published haplotyping methods.

Several authors have suggested that analysis of single SNPs loses power ***(22)***, especially for rare mutations ***(23)***. This is true, but depends on the approach used for single SNP analysis. Association mapping is possible

without an LDU map simply by selecting the most significant SNP. This is a very popular approach, and recently was used in a genome-wide scan on age-related macular degeneration ***(24)***. Single SNPs will not provide sufficient signal to narrow the region of interest, but this is also true when the most significant haplosets are selected. For example, using single SNPs, the region of significance around *CYP2D6* is 390 kb. Plotting the significance level on the kilobase map gives a pronounced "hole" at the *CYP2D6* locus because two distant SNPs on either side of the gene are highly significant, making the surface bimodal. Previous analysis using 5-SNP haplosets showed that the levels of significance were considerably higher than single SNPs, but with no further refinement of the support interval giving the same bimodal surface. Selection of significant SNPs avoids composite likelihood at the high cost of losing all information about other markers, and accepting a heavy correction that is unreasonable in a genome-wide scan ***(25)***. It further assumes that the functional SNPs have been included in the study. Although our multipoint approach is based on single SNPs, all association tests are considered simultaneously in a composite likelihood, which evades a heavy Bonferroni correction. Most importantly, the drawback of using a single SNP approach (selecting the most significant $\chi^2$ after Bonferroni correction) is that it does not consider that SNPs are in LD with one another. Using the mapping procedures presented here, the underlying LD structure is taken into account by mapping within an LDU map.

## 4. Discussion

It is increasingly apparent that in order to enhance progress and assure success with association studies, detailed information on the underlying structure of LD is required. The method presented here makes direct use of LDU maps, which aim to accurately characterize the fine-scale pattern of LD in the genome. This approach for disease gene association mapping applies a model with evolutionary theory, which incorporates a parameter for the location of the causal polymorphism. The method uses composite likelihood to combine information from all single-marker tests, which can be analyzed in random samples, but also in cases and controls using a correction factor based on the frequency of affected individuals in the general population ***(13)***. Data may be family-based where the affected offspring are the cases, with the nontransmitted alleles from the parents forming pseudocontrols. The pattern of LD of the studied region is taken into account by expressing the locations of marker SNPs in LDU. Analyses using these procedures have consistently shown the great utility of LDU maps in association mapping. Furthermore, the advantage of the proposed mapping approach is that it makes direct use of the HapMap data. The proof-of-principle application that is presented here provides evidence that strongly supports the use of the high marker density HapMap-derived LDU map for association

mapping. The ultimate goal of the HapMap Project is more than 3 million SNPs *(26)*. It is anticipated that the future HapMap releases and higher resolution LDU maps will enhance fine-mapping strategies. The HapMap Project has also been developed for four different populations, Yoruba, Japanese, Chinese, and CEPH (Centre d'Etude du Polymorphisme Humain) and, therefore, population-specific HapMap LDU maps offer versatility in association-mapping studies.

Whether the objective is a genome-wide association scan or a study on candidate regions, the HapMap LDU maps could be utilized directly and independently of the disease in question. Genome-wide association is a systematic strategy to obtain information on the association between SNPs and complex disease across the entire genome. Such large-scale analysis provides an unprecedented opportunity to identify genetic variants predisposing one to complex disease. This is an area of enormous interest and has attracted a great deal of attention worldwide. Although the association-mapping approach that is presented here has only been applied to small regions, it provides a substantial and solid foundation upon which we can build strategies for genome-wide association scans. Applications of composite likelihood can reduce the number of tests, while mapping within the high-resolution HapMap-derived LDU maps can greatly improve localization of causal polymorphisms. LDU maps are separated by several decades from the development of linkage, and their applications to high-resolution mapping of disease genes are just beginning.

## References


1 Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33,** 228–237.

2 Maniatis, N., Collins, A., Xu, C. F., et al. (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl. Acad. Sci. USA* **99,** 2228–2233.

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

4 Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P. Y., and Collins, A. (2001) The optimal measure of allelic association. *Proc. Natl. Acad. Sci. USA* **98,** 5217–5221.

5 Malecot, G. (1948) *Les Mathématiques de l'Hérédité*. Maison et Cie, Paris, France.

6 Morton, N. E. (2002) Applications and extensions of Malecot's work in human genetics. In: *Modern Developments in Theoretical Population Genetics,* (Slatkin, M. and Veuille, M., eds.), Oxford University Press, Oxford, UK.

7 Zhang, W., Collins, A., Maniatis, N., Tapper, W., and Morton, N. E. (2002) Properties of linkage disequilibrium (LD) maps. *Proc. Natl. Acad. Sci. USA* **99,** 17,004–17,007.

8 Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29,** 217–222.

9 Tapper, W., Collins, A., Gibson, J., Maniatis, N., Ennis, S., and Morton, N. E. (2005) A map of the human genome in linkage disequilibrium units. *Proc. Natl. Acad. Sci. USA* **102,** 11,835–11,839.

10 Maniatis, N., Collins, A., Gibson, J., Zhang, W., Tapper, W., and Morton, N. E. (2004) Positional cloning by linkage disequilibrium. *Am. J. Hum. Genet.* **74,** 846–855.

11 Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.* **29,** 229–232.

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

13 Maniatis, N., Morton, N. E., Gibson, J., Xu, C. F., Hosking, L. K., and Collins, A. (2005) The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. *Hum. Mol. Genet.* **14,** 145–153.

14 Maniatis, N., Collins, A., and Morton, N. (2006) Effects of single SNPs, haplotypes, and whole Genome LD Maps on accuracy of association mapping. Submitted.

15 Zhang, W., Maniatis, N., Rodriguez, S., et al. (2006) Refined association mapping of a causal site for a quantitative trait: weight in the *H19-IGF2-INS-TH* region. *Ann. Hum. Genet.*, in press.

16 Dobson, A. J. (1986) *An Introduction to Statistical Modelling*. Chapman and Hall, University Press, Cambridge, UK.

17 Hosking, L. K., Boyd, P. R., Xu, C. F., et al. (2002) Linkage disequilibrium mapping identifies a 390 kb region associated with CYP2D6 poor drug metabolising activity. *Pharmacogenomics J.* **2,** 165–175.

18 Salem, R. M., Wessel, J., and Schork, N. J. (2005) A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Hum. Genomics* **2,** 39–66.

19 Hill, W. G. (1974) Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33,** 229–239.

20 Gabriel, S. B., Schaffner, S. F., Nguyen, H., et al. (2002) The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229.

21 Clark, A. G. (2004) The role of haplotypes in candidate gene studies. *Genet. Epidemiol.* **27,** 321–333.

22 Morris, A. P. (2005) Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes. *Genet. Epidemiol.* **29,** 91–107.

23 Lin, S., Chakravarti, A., and Cutler, D. J. (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat. Genet.* **36,** 1181–1188.

24 Klein, R. J., Zeiss, C., Chew, E. Y., et al. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* **308,** 385–389.

25 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517.

26 International HapMap Consortium (2003) The International HapMap Project. *Nature* **426,** 789–796.

# 9

# Coalescent Methods for Fine-Scale Disease-Gene Mapping

**Andrew P. Morris**

**Summary**

Fine-scale mapping methods have been developed to localize functional polymorphisms within large candidate regions identified from previous linkage and/or association studies. Population-based association fine-mapping methods utilize linkage disequilibrium of alleles at high-density marker single-nucleotide polymorphisms with the functional polymorphism, generated as the result of shared ancestry of individuals within the population. Here, we review fine-mapping methods that model the shared ancestry of sampled chromosomes explicitly, using the coalescent process, resulting in greater accuracy and precision to localize functional polymorphisms than approaches that treat individuals as unrelated.

**Key Words:** Bayesian methods; coalescent process; fine-scale mapping; linkage disequilibrium; Markov chain Monte Carlo methods; population-based association studies.

## 1. Introduction

The traditional approach to mapping polymorphisms contributing to human diseases has been linkage analysis in pedigrees using well-defined maps of microsatellite markers to identify broad chromosomal regions containing functionally important genes. These chromosomal regions are typically of the order of several megabases in length and may contain many genes, making functional studies to identify the causal variant(s) too costly an undertaking to consider. As a result, we require additional fine-mapping studies to refine the location of disease genes within the candidate region. One such approach to fine mapping utilizes population association of disease with high-density single-nucleotide polymorphism (SNP) markers with samples of unrelated affected cases and unaffected controls. This approach is widely accepted to have the potential to map genetic polymorphisms contributing to complex traits, provided that the causal variants are not extremely rare *(1,2)*.

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

The key concept underlying any analysis of population-based association studies is linkage disequilibrium (LD), the nonrandom assortment of alleles among individuals generated as a result of their shared ancestry. Consider the simple case of a disease arising as a result of a relatively recent mutation at a functional polymorphism. At the instant the mutation occurs, it is carried on a single founding haplotype, and is in complete LD with alleles at any other marker SNP. Over subsequent generations, recombination will act to break down the haplotype carried by the founder chromosome, replacing ancestral-marker alleles with random genetic material from the background population, weakening LD with the mutation. However, with high-density maps of markers, the probability of recombination between the functional polymorphism and neighboring marker SNPs is small. Thus, the founder haplotype is expected to be preserved in the vicinity of the functional polymorphism on chromosomes carrying the causative mutation, with a mismatch of alleles occurring only as a result of rare marker mutation events.

Under this simple model of LD, an obvious strategy for fine mapping is to scan the candidate region for excess haplotype sharing among case chromosomes over controls. However, for complex diseases, we expect that genetic heterogeneity, epistasis, and environmental risk factors will affect the relative frequencies of case and control chromosomes carrying causative mutations at the functional polymorphism(s), introducing substantial noise in the relationship between phenotype and the underlying functional genotypes. Furthermore, excess haplotype sharing among cases can occur as a result of population structure not accounted for in the ascertainment process or subsequent analysis. The challenge for fine mapping with population-based association studies is to develop methodology that can efficiently detect LD resulting from the shared ancestry of chromosomes carrying the same causative mutations in a complex genetic setting, and can differentiate between it and marker SNP haplotype sharing resulting from demographic history and underlying structure.

In this chapter, we discuss how we can account for the shared ancestry of sampled chromosomes in fine-mapping studies, making use of standard models from population genetics. We compare a number of existing methods for fine mapping that allow for shared ancestry, and illustrate their application to localizing the ΔF508 mutation for cystic fibrosis. We discuss their advantages and limitations, and consider the future prospects for fine-mapping polymorphisms contributing to complex human disease.

## 2. Accounting for Shared Ancestry in Fine-Mapping Studies

Consider a population-based sample of unrelated cases and controls, with observed disease phenotypes, **y**, and phase-known haplotypes, **H**, at high-density marker SNPs throughout the candidate region. The ancestry of the sample of

chromosomes at a functional polymorphism, $x$, can best be represented by means of a bifurcating genealogical tree, an example of which is presented in **Fig. 1**. The genealogy is defined by the branching pattern, or *topology*, **T**, and the branch lengths, $\Psi$. The sampled chromosomes are represented by the leaves of the tree at the foot of the genealogy. Moving back in time involves climbing the tree through the internal nodes of the genealogy at which descendant chromosomes first share a common ancestor and the number of distinct ancestral lineages reduces by one. Ultimately, we reach the root of the tree at the top of the genealogy, corresponding to the most recent common ancestor (MRCA) of the entire sample of chromosomes at the functional polymorphism. **Figure 1** illustrates the position of two causative mutations at the functional polymorphism. Each chromosome descending from a branch of the genealogy on which the founding mutation event occurred will carry the causative mutation. Chromosomes carrying the *same* causative mutation are expected to share more recent common ancestry than a pair of chromosomes carrying the ancient wild-type allele, or a pair of chromosomes carrying *different* causative mutations.

Of course, individuals ascertained for population-based association are apparently unrelated, so the genealogy underlying the shared ancestry of their chromosomes at the functional polymorphism will not be known. A simple approximation to the bifurcating tree is the star genealogy for which every chromosome descends independently from the MRCA of the sample. Under this model, there is no correlation between pairs of chromosomes arising as a result of their *specific* shared ancestry. For example, a recombination event occurring in the shared ancestry of a pair of chromosomes must have occurred twice in the star genealogy, once in the descent of each chromosome from the MRCA. As a result, the star genealogy is optimistic about the variance of the estimated location of the functional polymorphism, $x$, because we effectively assume that we have more information about ancestral recombination and mutation events than we have in reality.

### *2.1. The Coalescent Process*

A common class of models for bifurcating genealogical trees is given by the *coalescent process* (*3*,*4*). Under this model, each topology is equally likely, with the leaves regarded as labeled to avoid combinatorial complication. Time is scaled to be measured in units of $N$ generations, where $N$ is the *effective* population size of chromosomes. For human populations, $N$ is generally assumed to be 10,000 individuals. The *standard* coalescent process is derived under the assumption that the leaves of the genealogical tree correspond to a random sample of chromosomes from a large random-mating population of constant size $N$, with no selection at the functional polymorphism. Under this model, McPeek and Strahs (*5*) calculated the expected time to the MRCA of a randomly

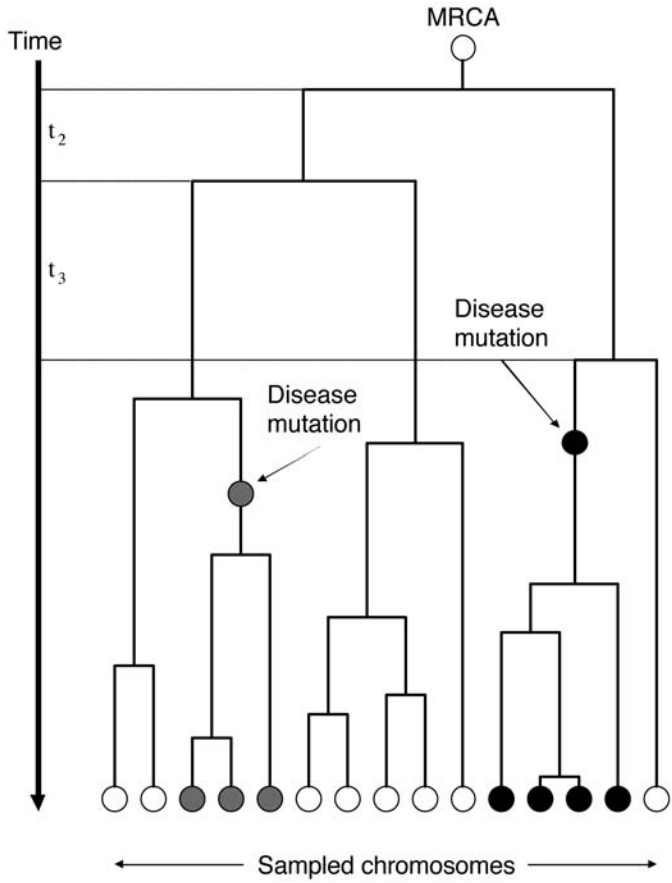



Fig. 1. Bifurcating genealogical tree representing the shared ancestry of 15 sampled chromosomes at the functional polymorphism. The most recent common ancestor of the sample carries the ancient wild-type allele at the functional polymorphism (indicated by the white circle). Mutations at the functional polymorphism change the form of the ancestral allele, indicated by the gray and black circles. The branch lengths of the genealogy are determined by the times, $\tau_k$, during which the genealogy has exactly $k$ lineages.

selected pair of chromosomes from the case–control sample, and thus the expected correlation in the length of the founder haplotype they share. This correlation can be used to inflate the variance of location estimates under the star genealogy, but still assumes that each pair of chromosomes shares the same ancestry at the functional polymorphism.

An alternative approach is to consider the space of possible genealogies consistent with the observed phenotype data and haplotypes across the candidate region within a Bayesian modeling framework. By Bayes' theorem, the posterior

density function of the location of the functional polymorphism, denoted $f(x|\mathbf{y}, \mathbf{H})$, can be expressed as

$$f\left(x|\mathbf{y},\mathbf{H}\right) \propto f\left(\mathbf{y},\mathbf{H}|x\right) f\left(x\right),$$

where $f(\mathbf{y},\mathbf{H}|x)$ denotes the *likelihood* of the observed data given the location of the functional polymorphism at $x$, and $f(x)$ denotes the prior density of $x$, often assumed to be uniform so that $f(x) \propto 1$. To allow for uncertainty in the underlying genealogy, $\mathsf{T} = \{\mathrm{T},\Psi\}$, we note that we can recover the marginal posterior distribution of the location of the functional polymorphism by integration,

$$f\left(x|\mathbf{y},\mathbf{H}\right) \propto \int_{\mathrm{T}} f\left(x,\mathrm{T}|\mathbf{y},\mathbf{H}\right) \partial\mathrm{T},$$

over tree space. Furthermore,

$$f\left(x,\mathrm{T}|y,\mathbf{H}\right) \propto f(y,\mathbf{H}|x,\mathrm{T})\, f(\mathrm{T}),$$

assuming the location of the functional polymorphism to be independent of the underlying genealogical tree, *a priori*.

Under the standard coalescent process, each topology is assumed equally likely. Then, the length of scaled time, $\tau_k$, during which the genealogical tree has exactly $k$ distinct lineages (**Fig. 1**), has an exponential distribution with rate parameter $\lambda_k = k(k-1)/2$, independently for each $k$, so that

$$f(\mathrm{T}) \propto \prod_{k\geq 2} \lambda_k \exp\left[-\lambda_k \tau_k\right]. \tag{1}$$

Generalizations of the standard coalescent process allow for exponential population growth, selection at the functional polymorphism, and population stratification. However, we cannot account for the over-representation of case chromosomes in the sample as a result of the biased ascertainment of affected individuals in the standard population-based association study design.

Nonrandom sampling affects the time-scale of genealogical trees, which can be accommodated by appropriate scaling of $N$, but will also distort the relative branch lengths ***(6,7)***. However, within the Bayesian paradigm, the standard coalescent process incorporates the principal effects of the shared ancestry of sampled chromosomes at the functional polymorphism, providing a relatively weak prior probability model that is readily overwhelmed by the observed data.

### *2.2. The Evolution of Marker SNP Haplotypes*

The transmission of genetic material through the genealogy depends on the occurrence of recombination and marker SNP mutation events in the candidate region. The marker haplotype carried by a node of the genealogy depends partly

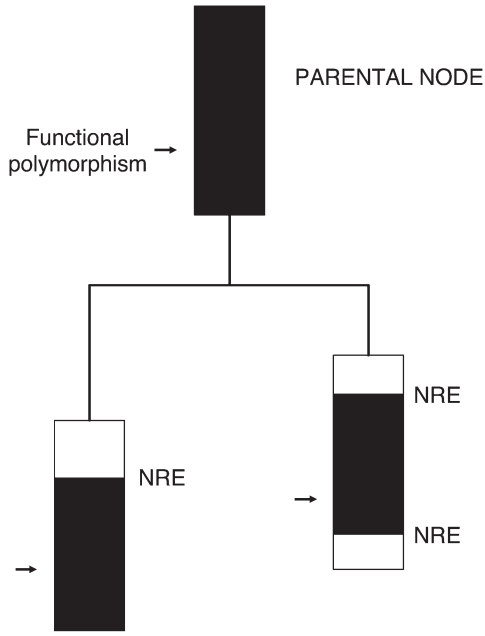



Fig. 2. Transmission of genetic material from a parental node of the genealogy to its two offspring nodes. The length of the parental haplotype preserved in the offspring nodes, indicated in black, is determined by the nearest recombination event on either side of the functional polymorphism along the connecting branches of the genealogy. When a recombination event occurs, the parental haplotype is replaced by random genetic material from the background population, indicated in white.

on that carried by its parental node, and the location of the nearest recombination event (NRE) on either side of the functional polymorphism (**Fig. 2**). The parental haplotype between the two NREs will then be preserved in the offspring node, with a mismatch of alleles occurring only as a result of marker SNP mutation. Of course, because the locations of the NREs are unknown, we must consider all possible intermarker SNP intervals as possible crossover breakpoints, according to appropriate models for the underlying recombination and mutation processes.

Assuming that recombination and mutation occur randomly, at uniform rates across the candidate region and through time, the distance to the NRE on either side of the functional polymorphism can be modeled by a Poisson process, and the number of mutations at a SNP marker within the preserved region can be modeled by a Poisson distribution. In either case, the mean of the process is then determined by the length of the branch connecting the two nodes in the genealogy.

Marker SNP mutation rates are generally accepted to be of the order of $10^{-8}$ per locus, per chromosome, per generation. However, crossover rates are known to vary throughout the genome, with increasing empirical evidence of highly localized hot spots of recombination. Nevertheless, methods for estimating fine-scale recombination rates from high-density SNP genotype data exist, and heterogeneity in the occurrence of crossovers in the candidate region could easily be incorporated in modeling the transmission of genetic material through the genealogy.

The haplotype extending beyond the NRE, on either side of the functional polymorphism, is generally assumed to have occurred as a result of recombination with a random chromosome from the background population. Of course, nonancestral genetic material may also share common ancestry at other loci across the candidate region, which cannot be modeled by the coalescent process at the functional polymorphism. The most appropriate model for the joint ancestry of polymorphisms across the entire region is the coalescent process with recombination ***(8–10)***, which greatly increases the complexity of the problem. By focusing on shared ancestry at the functional polymorphism, we ignore the shared ancestry of nonancestral haplotypes in favor of a simpler model dependent only on present-day population frequencies.

## 3. Modeling the Ancestry of Case Chromosomes

Consider again the simple case of a disease arising as a result of a relatively recent mutation at a functional polymorphism. Under this model, we expect the MRCA of a pair of chromosomes carrying the causative mutation to be more recent than that of a pair of chromosomes carrying the ancient wild-type allele. Rannala and Reeve ***(11)*** and Morris et al. ***(12)*** thus focus on the shared ancestry of case haplotypes, $\mathbf{H}_A$, and adopt a simpler model for the evolution of control haplotypes, $\mathbf{H}_U$, independent of the location of the functional variant. By ignoring the shared ancestry of control chromosomes, it follows that the marginal posterior distribution of $x$ is given by

$$f(x|\mathbf{y},\mathbf{H}) \propto \int_{\mathbb{T}} \int_{\mathbf{h}} f(\mathbf{H}_A|x,\mathbb{T},\mathbf{h}) f(\mathbf{H}_U|\mathbf{h}) f(\mathbb{T}) f(\mathbf{h}) \partial\mathbf{h}\, \partial\mathbb{T} \tag{2}$$

where $\mathbf{h}$ denotes relative haplotype frequencies in the background population, assumed independent of the underlying genealogy, *a priori*. Under this model, each case chromosome is assumed to carry the causative mutation at the functional polymorphism, whereas each control chromosome is assumed to carry the ancient wild-type allele. In this way, the disease model is fixed, and the causative mutation is assumed to have occurred on the MRCA of the sample of case chromosomes, reducing the complexity of the problem.

The most trivial model for $f(\mathbf{H}_U|\mathbf{h})$ assumes no LD between polymorphisms in the background population, so that the likelihood contribution of each control

haplotype is given by the product of the constituent population allele frequencies. Morris et al. *(12)* consider a more realistic model for background haplotype frequencies, incorporating LD between each pair of adjacent loci via a first-order Markov process *(13)*. This model could also be generalized to take account of longer range LD between nonadjacent SNPs, which may be important for high-density marker panels *(14)*. Appropriate prior distributions for **h** might assume uniform population allele frequencies at marker SNPs and uniform LD between adjacent loci.

The evolution of marker SNP haplotypes carried by case chromosomes can be modeled by the processes described in **Subheading 2.2.** Clearly, the likelihood $f(\mathbf{H}_A|x,\mathsf{T},\mathbf{h})$ will depend on the unobserved marker SNP haplotypes, **I**, carried by the internal nodes of the genealogy,

Within the Bayesian paradigm, internal-node haplotypes are treated as *augmented data* so that

$$f\left(\mathbf{H}_A|x,\mathsf{T},\mathbf{h}\right)=\sum_I f\left(\mathbf{H}_A,\mathbf{I}|x,\mathsf{T},\mathbf{h}\right) \tag{3}$$

Assuming the frequency of the causative mutation to be rare, nonancestral genetic material is modeled in the same way as haplotypes carried by control chromosomes, dependent only on **h**.

Thus, it follows that

$$f\left(\mathbf{H}_A,\mathbf{I}|x,T,\mathbf{h}\right)=\prod_b f\left(C_b|P_b,x,\Psi_b,\mathbf{h}\right) \tag{4}$$

where the product is over all branches of the genealogy, and $C_b$ and $P_b$, respectively, denote the offspring and parental nodes of branch $b$, having length $\Psi_b$ in scaled coalescent units of time.

To allow for over-representation of affected individuals in population-based association studies, Rannala and Reeve *(11)* employ the *intra-allelic* coalescent process as a prior model for the distribution of genealogical trees underlying the sample of case chromosomes *(15)*. However, this model requires specification of the age of the causative mutation, and may be unsuitable for rare variants *(7)*.

### 3.1. The Shattered Coalescent Model

A more important limitation of the standard coalescent process in modeling the ancestry of a sample of case chromosomes is the assumption that they descend from the same founding mutation event, represented by a single genealogical tree. Of course, not all case chromosomes will carry the causative mutation at the functional polymorphism, as a result of dominance, polygenic

effects, and environmental risk factors, and in this way are no different to control chromosomes. Furthermore, multiple causative mutations may occur at the functional polymorphism(s) within the disease gene.

Assuming that mutation events occur independently in the genealogy underlying the sample of chromosomes, we do not expect chromosomes carrying *different* mutations to share any more recent common ancestry than a pair of chromosomes carrying the wild-type allele.

To overcome this problem, Morris et al. ***(12)*** incorporate a *shattered* coalescent prior probability model for the genealogy underlying the sample of case chromosomes. A realization of this process is presented in **Fig. 3**, where the genealogy underlying a sample of case chromosomes is shattered by removing branches of the tree at random. *Sporadic* case chromosomes carrying the ancient wild-type allele at the functional polymorphism are represented by singleton leaf nodes. Disconnected subtrees correspond to independent causative mutations at the functional polymorphism (or more precisely completely linked functional polymorphisms). The transmission of genetic material through subtrees of the shattered genealogy can be modeled by the processes described in **Subheading 2.2.** However, by assuming that founder mutations and sporadic cases occur on random chromosomes from the background population, the haplotype carried by any node of the genealogy without a parent can then be modeled in the same way as that carried by a control chromosome.

Thus, the joint contribution of haplotypes carried by the observed case chromosomes and those carried by unobserved internal nodes of the genealogy (**Eq. 4**) generalizes to

$$f\left(\mathbf{H}_A,\mathbf{I}|x,\mathcal{T},\mathbf{h}\right)=\prod_b\left[z_b f\left(C_b|P_b,x,\Psi_b,\mathbf{h},N\right)+\left(1-z_b\right)f\left(C_b|\mathbf{h}\right)\right],$$

where $z_b$ takes the value 0 if branch $b$ is removed from the shattered genealogy, and 1 otherwise. As a result of the biased ascertainment of case chromosomes, the effective population size, $N$, is now treated as unknown, assumed uniformly distributed, *a priori*.

The prior probability that any branch is retained in the shattered genealogy is given by the *heterogeneity* parameter, $\rho$, where $\rho = 1$ corresponds to the standard coalescent process.

Low values of $\rho$ correspond to high levels of genetic heterogeneity at the functional polymorphism(s), with many singleton nodes and small subtrees in the genealogy.

Thus, the prior density function (**Eq. 1**), for the standard coalescent process generalizes to

$$f\left(\mathcal{T}|\rho\right)\propto\left[\prod_{k\geq 2}\lambda_k \boldsymbol{exp}\left(-\lambda_k\tau_k\right)\right]\left[\prod_b \rho^{z_b}\left(1-\rho\right)^{\left(1-z_b\right)}\right]$$

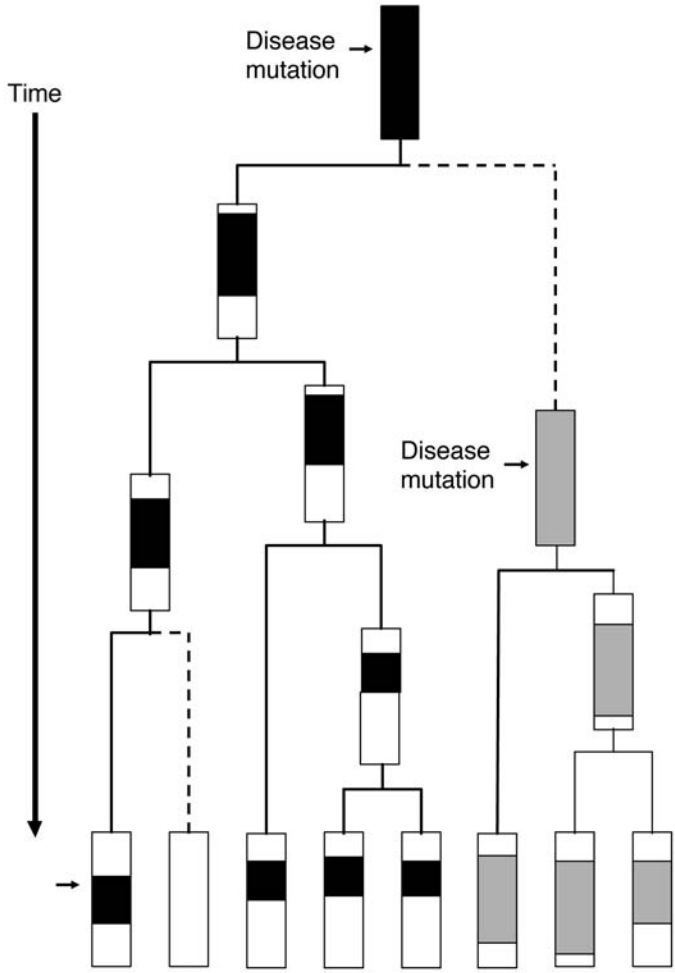



Fig. 3. Realization of the shattered coalescent process for a sample of eight case chromosomes. The branches indicated by dashed lines are removed in the shattered genealogy, creating two independent subtrees, indicated in black and gray, corresponding to independent causative mutation events at the functional polymorphism, and a singleton sporadic case chromosome, indicated in white.

### 3.2. Implementation

The parameter space, $S = \{x, \mathsf{T}, \mathbf{h}, \mathbf{I}, N\}$, is extremely complex, and hence the target posterior distribution, (**Eq. 2**), cannot be determined analytically. However, we can use Markov chain Monte Carlo (MCMC) methods to obtain an approximation to (**Eq. 2**) by sampling from the posterior density $f(S|\mathbf{y},\mathbf{H})$. At each step of a Metropolis–Hastings algorithm *(16,17)*, a candidate new value,

$s' \in S$, is proposed by making a *small* change to the current value, $s$. Changes might include an update of the haplotype carried by an internal node of the genealogy, adding or removing a branch of the shattered genealogy, or a shift in the location of the functional polymorphism. The proposed value, $s'$, is then accepted in place of $s$ with probability $f(s'|\mathbf{y},\mathbf{H})/f(s|\mathbf{y},\mathbf{H})$, otherwise the current value is retained. The algorithm can naturally deal with missing haplotype information as augmented data, updated in the MCMC algorithm in the same way as internal-node haplotypes.

The MCMC algorithm is run for an initial burn-in period to allow it to forget a randomly selected starting value in S. Convergence can be assessed using standard diagnostics ***(18)***. After convergence, each set of parameter values accepted, or retained, by the algorithm represents a random draw from the posterior distribution $f(S|\mathbf{y},\mathbf{H})$. Autocorrelation between draws is reduced by recording output at every $i$th iteration of the algorithm, for some suitably large value of $i$. The marginal posterior distribution of the location of the function variant is then approximated by the value of $x$ in each recorded output.

The MCMC algorithm also generates, at no extra computational cost, approximations to the posterior distribution of many other parameters in S that may be of interest. In particular, the posterior probability that any pair of case chromosomes carry the same causative mutation at the functional polymorphism can be approximated by the proportion of MCMC outputs in which they appear in the same subtree of the shattered genealogy. These probabilities can be used to construct a cladogram via standard hierarchical-clustering techniques ***(19)***, and can be used to represent genetic heterogeneity at the functional polymorphism from which clades of chromosomes carrying the same causative mutation can be identified.

## 4. Modeling the Ancestry of all Chromosomes

Rannala and Reeve ***(11)*** and Morris et al. ***(12)*** focus on modeling the shared ancestry only of case chromosomes in order to reduce the complexity of the underlying genealogy at the functional polymorphism. However, Zollner and Pritchard ***(20)*** consider the more ambitious task of taking account of the genealogy of the *entire* sample of chromosomes, which should extract substantially more information about shared ancestry in the candidate region. Furthermore, by including all chromosomes, we can directly model the disease risk of each causative mutation at the functional polymorphism, potentially allowing for additional environmental risk factors and polygenic effects, which will be extremely important for complex traits. By modeling risk in this way, we can also formally test for evidence of genetic variation within the candidate region that contributes to disease.

Begin by considering a fixed location, $x$, for the functional polymorphism, referred to as a *focal point*. Using the same arguments as presented in

**Subheading 2.**, it follows that the joint likelihood of observed phenotype and haplotype data, **y** and **H**, given location $x$, can be expressed as

$$f(\mathbf{y},\mathbf{H}|x) = \int_{\mathcal{T}} f(\mathbf{y},\mathbf{H}|x,\mathcal{T})f(\mathcal{T})\partial\mathcal{T}$$
$$= \int_{\mathcal{T}} f(\mathbf{y}|x,\mathcal{T})f(\mathbf{H}|x,\mathcal{T})f(\mathcal{T})\partial\mathcal{T},$$

where the integration is over tree space. The first term, $f(\mathbf{y}|x,\mathsf{T})$, corresponds to the probability of the observed phenotype data, given the underlying genealogy at location $x$. The second term, $f(\mathbf{H}|x,\mathsf{T})$, could be evaluated in the same way as in **Subheading 3.**, making use of **Eqs. 3** and **4**, but for the entire sample of chromosomes. Alternatively, it follows from Bayes' theorem that $f(\mathbf{H}|x,\mathsf{T})f(\mathsf{T}) = f(\mathsf{T}|\mathbf{H},x)$, so that

$$f(\mathbf{y},\mathbf{H}|x) \propto \int_{\mathcal{T}} f(\mathbf{y}|x,\mathcal{T})f(\mathcal{T}|\mathbf{H},x)\partial\mathcal{T}$$

To evaluate the likelihood, Zollner and Pritchard *(20)* suggest sampling $R$ genealogies from the posterior distribution, $f(\mathsf{T}|\mathbf{H},x)$, consistent with the observed haplotype data. Hence,

$$f(\mathbf{y},\mathbf{H}|x) \approx \frac{1}{R}\sum_{r} f(\mathbf{y}|x,\mathcal{T}_x^{(r)}),$$

where $\mathsf{T}_x^{(r)}$ denotes the topology and branch lengths of the $r$th sampled genealogy at focal point $x$.

To sample from the posterior distribution of genealogies consistent with the observed marker SNP haplotype data, a Metropolis–Hastings MCMC algorithm is developed, using the same techniques as Morris et al. *(12)*, described in **Subheading 3.2.** They incorporate a standard coalescent model for the tree topology and branch lengths, *a priori*, and model nonancestral genetic material by means of a first-order Markov process incorporating LD between adjacent marker SNPs. However, they focus on haplotype sharing from sampled chromosomes *backward* in time through the genealogy, rather than on the decay of ancestral genetic material forward in time. We would expect little of the founder haplotype carried by the MRCA of the sample of chromosomes to be preserved today, except in the region directly flanking the functional polymorphism. By moving backward through the genealogy, we can ignore ancestral material that is not present in the sample of chromosomes, reducing the complexity of the space of haplotypes carried by internal nodes of the genealogy, improving the efficiency of the MCMC algorithm.

### *4.1. Modeling Disease Phenotypes*

The key advantage of reconstructing the ancestry of the entire sample of chromosomes is that we can model the relationship between disease and the

underlying genotypes at the functional polymorphism directly. For a diallelic functional polymorphism, with alleles denoted *A* (ancestral wild-type) and *a* (causative mutation), this model can be parameterized in terms of three penetrances, one for each possible genotype. Assuming the effects on disease of each allele at the functional polymorphism to be independent, it is more convenient to reparameterize the model in terms probabilities, $\varphi_A$ and $\varphi_a$, that a chromosome carrying alleles *A* and *a*, respectively, come from an affected individual. Under this model, $\varphi_a/\varphi_A$ corresponds to the odds ratio of the causative mutation, relative to the wild-type allele at the functional polymorphism. The prior distribution of these penetrances is assumed to be uniform, without order, allowing the possibility that the wild-type allele is high-risk, whereas the causative mutation is protective.

Of course, the alleles carried by each chromosome at the functional polymorphism are not known in advance, but will depend on the position of causative mutation events in the underlying genealogy. Zollner and Pritchard ***(20)*** assume that causative mutations occur at a fixed rate $\upsilon/2$ per unit of coalescent time, independently on each branch, and that causative mutant alleles do not undergo further mutation. Thus, by considering all possible positions for causative mutations, $\mathsf{M} \in M$, in the underlying genealogy, $\mathsf{T}_x^{(r)}$, it follows that

$$f\left(\mathbf{y}|x,\mathcal{T}_x^{(r)}\right)=\sum_{M\in\mathcal{M}} f\left(M|x,\mathcal{T}_x^{(r)}\right)\prod_j f\left(y_j\middle|\varphi_A,\varphi_a,M_j\right),$$

where $M_j$ denotes the allele carried by the *j*th chromosome at the functional polymorphism for the set of causative mutations *M*. Because we exclude the possibility of back-mutation, this likelihood can be efficiently calculated using a peeling algorithm ***(21)***.

### *4.2. Fine Mapping and Significance Testing*

For fine mapping, Zollner and Pritchard ***(20)*** consider a dense set of focal points for the functional polymorphism, across the candidate region. By sampling *R* trees from the posterior distribution of genealogies consistent with the observed haplotype data for each focal point, independently, we can approximate the marginal posterior density, $f(x|\mathbf{y},\mathbf{H})$, at location *x* by computing

$$f\left(x|\mathbf{y},\mathbf{H}\right)=\frac{\sum_r f\left(\mathbf{y}|x,\mathcal{T}_x^{(r)}\right)}{\sum_{X\in x}\sum_r f\left(\mathbf{y}|X,\mathcal{T}_X^{(r)}\right)},$$

assuming each location to be equally likely, *a priori*. For significance testing, the likelihood

$$f\left(\mathbf{y}|x,\mathbf{H},\varphi_A,\varphi_a\right)\approx\frac{1}{R}\sum_r f\left(\mathbf{y}|x,\mathcal{T}_x^{(r)},\varphi_A,\varphi_a\right),$$

is maximized with respect to the penetrance parameters, and compared with the maximized null likelihood, $f(\mathbf{y}|\varphi_0)$, independent of location. The ratio of the maximized log likelihoods at focal point $x$ has an approximate $\chi^2$ distribution with one degree of freedom. To allow for multiple testing, empirical $p$-values are obtained by randomly permuting the case–control labels of pairs of chromosomes, and repeating the analysis for each permutation. Note that the sampling of trees from the posterior distribution of genealogies consistent with the observed marker haplotype data is *independent* of phenotype. Thus, we can make use of the same sets of trees for each location across all permutations, greatly reducing the burden of computation.

## 5. Example Application: Cystic Fibrosis

Cystic fibrosis (CF) is a well understood, fully penetrant-recessive disease most common in white populations with an incidence of approx 1 case per 2500 live births. Preliminary linkage analysis had suggested a 1.8-Mb candidate region for a single CF gene, on chromosome 7q31, between the MET locus and marker D7S426. More recently, a 3-bp deletion, ΔF508, has been identified within this region in the *CTFR* gene, at 885 kb from the MET locus. It is now well established that ΔF508 accounts for approx 66% of all chromosomal mutations in individuals with CF, with the remainder of cases due to many other, much rarer mutations in the same gene ***(22)***.

Kerem et al. ***(23)*** obtained marker haplotypes from 94 case chromosomes and 92 control chromosomes using 23 restriction fragment length polymorphism (RFLPs) in the candidate region. Of the case chromosomes, 62 have now been confirmed as carrying the ΔF508 mutation. Single-RFLP analyses revealed strong evidence of association with CF across the candidate region, particularly extending from 600 to 900 kb from the MET locus. A number of fine-mapping analysis methods have been applied to this dataset. In general, methods that do not explicitly model the genealogy underlying the sample of chromosomes are unable to localize the ΔF508 mutation, assigning too much confidence to their estimated locations (*see* **ref.** ***12*** for summary).

Using the shattered coalescent model, Morris et al. ***(12)*** obtain a median estimate of the location of ΔF508 at 851 kb from the MET locus, with a 95% credibility interval of 650–1003 kb, including the true location. Using a cladogram to represent genetic heterogeneity, as described in **Subheading 3.2.**, they were able to identify a cluster of 69 case chromosomes, including all 62 ΔF508 carriers. Zollner and Pritchard ***(20)*** estimate the location of the ΔF508 mutation at 867 kb from the MET locus, with a more precise 95% credibility interval of 814–920 kb, presumably reflecting the additional information gained by modeling the ancestry of all chromosomes.

## 6. Unphased Genotype Data

The methods described thus far assume that we are able to distinguish the pair of haplotypes carried by each sampled individual. However, with current SNP-genotyping technology, we cannot generally recover the required phase information without additional typing of parents or other family members. Within the Bayesian paradigm, we can treat the unobserved haplotypes as augmented data. Thus, the marginal distribution of the location of the functional polymorphism is given by

$$f(x|\mathbf{y},\mathbf{G}) \propto \sum_{\mathrm{H} \in \mathbf{G}} \int_{T} f(\mathbf{y},\mathrm{H}\,|x,T)\, f(T)\, \partial T \ ,$$

where the summation is over the space of haplotype configurations, H consistent with the observed unphased genotype data, **G**.

Morris et al. ***(24)*** consider this approach to allow for unphased genotype data in the shattered genealogy. Within the MCMC algorithm, haplotypes carried by cases and controls are updated by swapping the pair of alleles carried by an individual at a given marker SNP. Simulations of haplotypes and unphased genotype data suggest a minimal loss of information for fine-scale mapping resulting from unknown phase with this method. However, the additional layer of complexity in the MCMC algorithm carries a computational overhead of approx 50%, and may degrade mixing.

An alternative approach is to make use of statistical algorithms to reconstruct haplotypes from unphased genotype data, and to treat these haplotypes as if known in the subsequent fine-scale mapping analysis. Zollner and Pritchard ***(20)*** use PHASE ***(25,26)*** to impute marker SNP haplotypes by sampling from the posterior distribution of phase assignments consistent with the observed genotype in a pseudo-Bayesian MCMC framework. PHASE assumes a coalescent process with recombination to model expected patterns of haplotype diversity among populations, *a priori*. As a byproduct, the algorithm estimates fine-scale recombination rates across the candidate region, which are extremely useful in modeling the degradation of ancestral haplotypes through the genealogy at the functional polymorphism.

One potential disadvantage of this approach is that imputed haplotypes are *estimates*, and are unlikely to be known with certainty. Over small genetic distances, this uncertainty in phase assignment is likely to be minimal because of strong LD between SNP markers. However, over large candidate regions on the order of megabases, the reconstruction process is less robust, and there may be many possible haplotype configurations consistent with each unphased genotype. By ignoring uncertainty in the fine-mapping analysis of imputed haplotypes, Morris et al. ***(24)*** have demonstrated a noticeable bias, and over-confidence in estimates of the location of the functional polymorphism.

A possible solution to this problem would be to consider all haplotype configurations with a phase assignment probability above a prespecified threshold for each individual, and to use this as a prior distribution in the subsequent fine-mapping analysis. In this way, we still allow for uncertainty in phase assignment, but reduce the space of possible haplotype configurations consistent with each unphased genotype, lessening the computational burden of the MCMC algorithm.

## 7. Discussion

Fine-mapping methods that take account of the shared ancestry of sampled chromosomes in population-based association studies show great promise for localizing functional polymorphisms contributing to complex traits. By modeling shared ancestry, we can take account of genetic heterogeneity at the functional polymorphism, and allow for polygenic effects and environmental risk factors. However, Bayesian MCMC methods are computationally intensive, and may not be particularly well suited to the analysis of large-scale genetic association studies of thousands of individuals at many hundred marker SNPs in large candidate regions.

Future research could focus on developing faster methods for fine mapping, yet which still take account of shared ancestry. One promising approach is to cluster haplotypes directly, according to their *similarity*, without an explicit model of shared ancestry. These methods take advantage of the expectation that similar haplotypes in the region flanking the functional polymorphism are likely to share recent common ancestry, and thus are likely to share similar disease risk. Molitor et al. ***(27,28)*** measure similarity in terms of the length of shared haplotype between chromosomes on either side of a putative functional polymorphism. They cluster haplotypes according to a Bayesian partition model such that haplotypes within the same cluster are assigned the same risk of disease. They utilize Bayesian MCMC methods to approximate the posterior distribution of the location of the functional polymorphism, given observed disease phenotypes and phase-known haplotype data, although the method could also be extended to deal with unphased genotypes.

Overall, the prospects for fine-mapping complex disease polymorphisms via population-based association studies look extremely promising. With improvements in the efficiency of high-throughput genotyping technology, high-density marker panels are becoming increasingly realistic for the sample sizes we require for the modest genetic effects we expect for complex diseases. Furthermore, with the publication of the SNP map of the human genome ***(29,30)***, and the release of data from the International Haplotype Map (HapMap) Project ***(31)***, we have a much better understanding of the patterns of common genetic variation throughout the genome. As a result, there will continue to be an exciting

period of development of statistical methods needed to meet the challenges posed by fine mapping with this type of data.

## References


1 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex diseases. *Science* **273,** 1516–1517.

2 Zondervan, K. T. and Cardon, L. R. (2004) The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* **5,** 89–100.

3 Kingman, J. F. C. (1982) The coalescent. *Stoch. Proc. Appl.* **13,** 235–248.

4 Nordborg, M. (2001) Coalescent theory. In: *Handbook of Statistical Genetics* (Balding, D. J., Bishop, M., and Cannings, C., eds.), Wiley, Chichester, UK, pp. 179–212.

5 McPeek, M. S. and Strahs, A. (1999) Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am. J. Hum. Genet.* **65,** 858–875.

6 Slatkin, M. (1996) Gene genealogies within mutant allelic classes. *Genetics* **143,** 579–587.

7 Wiuf, C. and Donnelly, P. (1999) Conditional genealogies and the age of a neutral mutant. *Theor. Popul. Biol.* **56,** 183–201.

8 Hudson, R. R. (1983) Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23,** 183–201.

9 Griffiths, R. C. and Marjoram, P. (1996) Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3,** 479–502.

10 Griffiths, R. C. and Marjoram, P. (1997) An ancestral recombination graph. In: *Progress in Population Genetics and Human Evolution,* (Donnelly, P. and Tavare, S., eds.), Springer-Verlag, New York, NY, pp. 257–270.

11 Rannala, B. and Reeve, J. P. (2001) High-resolution multipoint linkage disequilibrium mapping in the context of a human genome sequence. *Am. J. Hum. Genet.* **69,** 159–178.

12 Morris, A. P., Whittaker, J. C., and Balding, D. J. (2002) Fine-scale mapping of disease loci via coalescent modelling of genealogies. *Am. J. Hum. Genet.* **70,** 686–707.

13 Liu, J. S., Sabatti, C., Teng, J., Keats, B. J. B., and Risch, N. (2001) Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.* **11,** 1716–1724.

14 Strahs, A. L. and McPeek, M. S. (2003) Multipoint fine-scale linkage disequilibrium mapping: the importance of modeling background LD. In: *IMS Lecture Notes Monograph Series, Vol. 40 Science and Statistics,* (Festschrift, A. and Goldstein, D. R., eds.), Institute of Mathematical Statistics, Beachwood, OH, pp. 343–366.

15 Slatkin, M. and Rannala, B. (1997) Estimating the age of alleles by the use of intra-allelic variability. *Am. J. Hum. Genet.* **60,** 447–458.

16 Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21,** 1087–1092.

17 Hastings, W. K. (1970) Monte-Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.
18 Gammerman, D. (1997) *Markov Chain Monte Carlo: Stochastic Simulation for BayesianInference*. Chapman and Hall, London, UK.
19 Hartigan, J. A. (1975) *Clustering Algorithms*. Wiley, New York, NY.
20 Zollner, S. and Pritchard, J. K. (2005) Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* **169,** 1071–1092.
21 Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol*. **17,** 368–376.
22 Bertranpetit, J. and Calafell, F. (1996) Genetic and geographical variability in cystic fibrosis: evolutionary considerations. In: *Variation in the Human Genome,* (Weiss, K., ed.), Wiley, Chichester, UK, pp. 97–114.
23 Kerem, B., Rommens, J. M., Buchanan, J. A., et al. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science* **245,** 1073–1080.
24 Morris, A. P., Whittaker, J. C., and Balding, D. J. (2002) Little loss of information due to unknown phase for fine-scale linkage disequilibrium mapping with single nucleotide polymorphism genotype data. *Am. J. Hum. Genet*. **74,** 945–953.
25 Stephens, M., Smith, N. J., and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet*. **68,** 978–989.
26 Stephens, M. and Donnelly, P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet*. **73,** 1162–1169.
27 Molitor, J., Marjoram, P., and Thomas, D. (2003) Application of Bayesian spatial statistical methods to the analysis of haplotype effects and gene mapping. *Genet. Epidemiol*. **25,** 95–105.
28 Molitor, J., Marjoram, P., and Thomas, D. (2003) Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am. J. Hum. Genet*. **73,** 1368–1384.
29 International Human Genome Sequence Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.
30 International SNP Map Working Group (2001) A map of the human genome sequence variation contains 1.42 million single nucleotide polymorphisms. *Nature* **409,** 860–921.
31 International HapMap Consortium (2003) The international HapMap project. *Nature* **426,** 789–795.

# 10

# Family-Based Linkage Disequilibrium Tests Using General Pedigrees

**Yin Yao Shugart, Lina Chen, Rui Li, and Terri Beaty**

**Summary**

Linkage disequilibrium (LD) mapping has been established as a promising approach to identifying disease genes. The presence of a disease gene located near a marker locus may cause LD between the marker and the disease loci. In LD mapping, we assume that some of the affected individuals may have a common ancestor carrying the mutation and that mutation carriers are likely to share alleles at the markers loci close to the disease gene.

This chapter reviews the concept of LD mapping and outlines the advantages and disadvantages of two LD mapping approaches capable of handling general pedigrees: the family-based association test (FBAT) and pseudomarker. In summary, the pseudomarker statistical approach and the FBAT approach are both expected to offer reasonable statistical power to detect genes underlying complex traits. However, when the pedigree structure is more complicated, or when the number of informative families is limited, the pseudomarker approach is anticipated to outperform FBAT.

**Key Words:** LD mapping; family-based LD tests; FBAT; pseudomarker.

## 1. Introduction

In the past few decades, the most commonly utilized statistical tools to approach genetic disorders in humans are linkage analysis and linkage disequilibrium (LD) studies using pedigrees. Linkage analysis has proven effective at initially mapping the location of disease genes ***(1)*** and LD mapping has been more often used as either a tool for candidate gene or for fine-mapping genes under a linkage peak ***(2–9)***, particularly for genes with lower penetrances ***(10)***. The idea of LD mapping is important because detecting LD between disease status and a marker implies either that the disease locus and a tested marker are closely linked, and are associated with a particular marker allele. In many ways it is appealing to search for causal genes by simply genotyping a sample of

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

well-characterized cases and a sample of unrelated healthy controls and then compare the allele frequencies between the two groups. If a strong statistical association between case–control status and the marker alleles is documented, it could be because one marker allele is directly causal, even if it may only be part of a complex pathogenetic pathway (i.e., the association is direct). Alternatively, the statistical association could be an indirect one where no marker allele is causal but is associated with an unobserved high-risk allele owing to genetic linkage. The assumptions we have made in using this case–control approach to search for genes is that association resulting from close linkage decays rather dramatically with genetic distance (recombination fraction). Therefore, if we find an association through distinctly different marker allele/genotypic frequencies between the case and the control groups, then we may be able to conclude that the marker is tightly linked to an unknown disease susceptibility locus. However, this cost effective and traditionally accepted study design may not be valid in the presence of population stratification, which can create confounding and lead to a spurious statistical association in the absence of linkage. If cases and controls are not well matched for their genetic background, the probability of false-positive results can increase *(3)*.

The transmission/disequilibrium test (TDT) was designed to test the presence of both linkage and LD between a marker and a disease locus *(3,11,12)*. One advantage of the TDT is appealing because it remains a valid test even in the presence of population substructure, which can potentially produce false-positive findings in a case–control study design. Further, Terwilliger and Ott stated that the haplotype relative risk (HRR) approach is not completely immune to the problem of population structure *(12)*. Despite the fact that some extensions to the TDT have been explored by various authors, we will focus on two interesting approaches that can be used to test for linkage in the presence of LD using general pedigrees: one is known as "pseudomarker analysis" *(13)* and the other is called the family-based association test (FBAT) *(14–18)*.

## 2. The Pseudomarker Approach

Terwilliger and Goring *(13)* elegantly illustrated the conceptual difference between linkage and LD in a series of papers where they emphasized the importance of using joint linkage and LD analysis to detect genes underlying complex traits such as cardiovascular diseases, mental disorders, and cancer. Terwilliger and Goring *(13)* also developed "parametric statistics," which are analogs of the original HRR analysis, TDT analysis, and joint linkage and LD analysis. The detailed statistics are given in **Table 1**. The notation is described next: $\theta$ is the recombination fraction between the unobserved disease locus and the observed genetic marker, $\delta$, represents LD between the disease locus and the marker loci, without implying a specific parametric model for LD. As illustrated

**Table 1**
**Outline of General Likelihood Tests for Linkage and/or Linkage Disequilibrium[a]**

| Test statistic | Application | Analogous model-free tests |
|---|---|---|
| $B = 2\ln\dfrac{\max\limits_{\theta,\delta} L(\theta,\delta)}{\max\limits_{\delta} L(\theta = 0.5,\delta)}$ | Test of linkage allowing for linkage disequilibrium | Transmission/ disequilibrium tests (TDT) |
| $C = 2\ln\dfrac{\max\limits_{\theta,\delta} L(\theta,\delta)}{\max\limits_{\theta} L(\theta,\delta = 0)}$ | Test of linkage disequilibrium allowing for linkage | Haplotype relative risk (HRR) tests, case–control tests |
| $D = 2\ln\dfrac{\max\limits_{\theta,\delta} L(\theta,\delta)}{L(\theta = 0.5,\delta = 0)}$ | Joint test of linkage and linkage disequilibrium | |

[a]Adapted from Terwilliger, J. D., personal communication.

in **Table 1**, *B* is analogous to the TDT, and the authors stated that, in a set of large families, this B test would outperform the conventional TDT because it uses more information about the genetic relationships between affected individuals in the same pedigree. On the hand, in their pseudomarker analysis on triads, *C* is analogous to the original HRR, and on families of larger sizes, information from the entire pedigree is used to improve estimates of haplotype frequencies in the founders, therefore leading to a more powerful test. Interestingly, this type of pseudomarker analysis can also be applied to case–control data using their statistical test, which is equivalent to the normal case–control analyses, and when performed on combinations of singletons, triads, and larger pedigrees, *C* provides a test of LD allowing for linkage. *D* tests for linkage and LD because in the null hypothesis neither exists and in the alternative hypothesis both are allowed to exist (**Table 1**). All these three tests have been implemented in a freeware program called "Pseudomarker" (http://www.helsinki.fi/~tsjuntun/pseudomarker). One major limitation of the pseudomarker software is that it cannot currently handle multiple markers. Although all these three statistics are conceptually intuitive, relatively few studies of real family data have used them; therefore, both simulated data and real data need to be analyzed using the pseudomarker analysis.

### 3. Family-Based Association Tests

Another test termed FBAT has been more commonly used to map complex disease genes because it provides the test for both linkage and association in family samples in the presence of population admixture. FBAT was first

introduced by Rabinowitz and Laird ***(14)*** and Laird et al. ***(15)*** as a two-stage method: stage 1 defines a test statistic based on a linear combination of offspring genotypes and traits (to be detailed later) and stages 2 computes the genotype data distribution under the null hypothesis of no linkage and no association, or of no association in the presence of linkage (http://www.biostat.harvard.edu/~fbat).

FBAT treats the offspring genotype as a random variable, conditioning on all observed traits and parental genotypes to avoid making assumptions about the parental allele frequencies, the distribution of the traits, as well as how the data is ascertained. In the presence of missing parental genotypes, FBAT conditions on the sufficient statistics. Therefore, the distribution of offspring genotypes is independent of the missing parent genotype and the uncertainty of the marker allele frequencies. A set of exhaustive tables were given by Rabinowitz and Larid ***(14)***, building on an algorithm for computing the conditional distribution of offspring genotypes for nuclear families in the presence of missing genotypes. **Table 2** was used by Rabinowitz and Larid ***(14)*** as an example to illustrate the conditional distributions when testing for linkage with one homozygous parent with genotype A1A1.

The first column indicated the observed configurations of marker alleles in offspring. For example, the notation {A1A2, A1A3} corresponds to a sibship with at least one offspring with genotype A1A2 and another with genotype A1A3 and without any other genotypes represented. The second column describes the conditional distribution. Given, for example, the offspring's marker alleles being A1A2 and A1A3, the conditional distribution can be simulated by randomly assigning A1A2 and A1A3 with equal probability independently to each offspring until the requirement that at least one assignment of AB and one assignment of AC is violated.

FBAT can be viewed as a generalized TDT test. The FBAT statistic can be expressed as:

$$S = \sum_{ij} T_{ij} \left[ X_{ij} - E(X_{ij}) \right],$$

where $i$ indexes the families, $j$ indexes offspring within family, $T_{ij}$ is a function of the $(i,j)$th offspring's phenotype, and $X_{ij}$ denotes the marker "score" for $(i,j)$th offspring. Using the distribution of offspring's genotypes, $E(X_{ij})$ and var$(X_{ij})$ can be computed conditional on parental genotypes assuming the null hypothesis is true. $T_{ij} = Y_{ij} - \mu$ where $\mu$ is a constant and $Y_{ij}$ is an indicator variable for the disease status.

The FBAT toolkit (http://www.biostat.harvard.edu/~fbat/default.html) recommends that, for the dichotomous trait, the form $T_{ij} = Y_{ij} - \mu$ can be used to record $Y_{ij}$ into $T_{ij}$. Here, $\mu$ is an offset value defined by the FBAT authors, which ranges from 0 to 1 to allow both affected and unaffected individuals to contribute

**Table 2**
**Conditional Distributions When Testing for Linkage With Only One Homozygous Parent With Genotype AA[a]**

| Children's marker alleles | Conditional distribution |
|---|---|
| 1. {*A1A1*} or {*A1A2*} | Observed data have conditional probability |
| 2. {*A1A1, A1A2*} | Randomly assign *A1A1* or *A1A2* with probability 1/2, 1/2, independently to each sib, discarding outcomes without at least one assignment of *A1A1* and one assignment of *A1A2* |
| 3. {*A1A2, A1A3*} | Randomly assign *A1A2* or *A1A3* with probability 1/2, 1/2, independently to each sib, discarding outcomes without at least one assignment of *AB* and one assignment of *AC* |

[a]Adapted from Terwilliger, J. D., personal communication.

to the test statistic. It was suggested that inclusion of unaffected siblings could improve power when the offset μ is set to be the disease prevalence ***(19)*** when disease is not rare. Furthermore, Laird (personal communication, 2005) observed that, for common diseases, power of FBAT for using trio with one additional unaffected offspring is more powerful when the offset is chosen to be greater than the disease prevalence. The authors further suggested that offset choices similar to the population prevalence be regarded as "a rule of thumb."

However, in reality, disease prevalence is not known precisely. The typical method is to estimate the offset μ with the proportion of the affected in the sample, although a valid estimate of the offset usually cannot be obtained from the sample when the ascertainment depends on $Y_{ij}$. An alternative strategy for the offset estimation was developed by Lunetta et al. ***(20)***. They derived that variance of $S$ is minimized under $H_0$ by setting $\mu = n_{Aff}/(n_{Aff} + n_{Unaff})$, which is a sample estimate of the prevalence. This approach is incorporated in FBAT and can be invoked with *–o* option for the FBAT command. Beside the dichotomous trait, FBAT also can be applied in analysis of measured phenotypes, including continuous, by setting μ to the mean of $Y_{ij}$ in whole sample. More simulation-based research is needed in this area to provide further guidelines for using appropriate offsets. Experienced users of FBAT may wish to know the definition for a "common disease." Should a disease with prevalence of 0.1 be considered common or rare?

Another advantage of FBAT is it can be used in different genetic models such as the additive model, dominant model, and recessive model, and in different sample designs, including nuclear families, sibships, and the single pedigree consisting of multiple nuclear families ***(15)***. In the analysis, FBAT will automatically break the large pedigree into separate sibships or unclear families ***(20)***. FBAT can also handle missing parental genotypes. In FBAT, the distribution of offspring genotype is assumed to be random but is conditionally dependent on the parental genotypes and all phenotypes. Therefore, the missing parental

genotypes can be configured by the appropriate offspring genotype distribution. This means, in FBAT, the definition and distribution evaluation of the test statistic under the null hypothesis are separately implemented, which is an essential feature of this program ***(15)***. Moreover, this kind of genotype configuration has also been extended to the haplotype analysis to fix the unknown parent's genotypes and/or phase-unknown offspring by using a set of weights assigned to the phased genotypes by Horvath et al. ***(21)***. For example, if $G_{ij}$ is a phase-known genotype for offspring $j$ in the $i$th family, $X(G_{ij})$ was defined as the genotype coding of its haplotypes with known phase. However, if $G_{ij}$ is an unphased genotype of offspring, $X(G_{ij})$ will be redefined as:

$$X(G_{ij}) = \sum\nolimits_k X(G_{ijk}) W_{G_{ijk}}$$

Using an empirical example, Horvath et al. ***(21)*** demonstrated the advantage of conducting haplotype-based association analysis. For example, Horvath et al. utilized the haplotype FBAT program to test for associations between asthma phenotypes and single-nucleotide polymorphisms (SNPs) in the β2 adrenergic receptor gene. When no single SNP gave significant association signals with asthma diagnosis or bronchodilator responsiveness, using a haplotype-based global test, Horvath et al. found a highly significant association with asthma diagnosis ($p$-value < 0.00005), as well as the measure of bronchodilator responsiveness ($p$-value = 0.016). Further, Lake and Laid ***(22)*** described a new test for gene–environment interaction (FBAT-I). In essence, the FBAT-I test is similar in spirit to what was proposed by Umbach and Weinberg ***(23)***. In both tests, the null hypothesis of no gene–environment interaction is tested using a likelihood ratio test and the degree of freedom is determined by the assumptions made regarding the penetrance function. FBAT-I stratifies the observed data by mating types.

In general, the FBAT worked well under a number of simulated conditions including different genetic models with reduced penetrances and various allele frequencies (Rui and Shugart, unpublished data) when the number of informative families is reasonable (minimum number of informative families was recommended to be 10 by FBAT authors). As discussed previously, to test for association in the presence of linkage, the distribution of the offspring genotype is determined by conditioning on a sufficient statistic when the parental genotyping is missing. This test has been implemented in the FBAT software by Horvath et al. ***(24)***. One can use the "-e" option to perform the test for association by considering the asymptotic distribution and the empirical variance. However, the assumption of asymptotic normality may be violated when the number of informative families becomes too small. The second potential problem is that the multiple testing needs to be addressed when multiple markers and multiple traits are studied. Very recently, Schneiter et al. ***(25)*** developed an exact FBAT for biallelic data using a network algorithm, providing a useful

alternative approach to the asymptotic test. The advantage of using an exact test is that it returns a *p*-value obtained from the true distribution of the test statistic, whereas an asymptotic test may lead to bias. Interestingly, the simulation results presented by Schneiter et al. ***(25)*** indicated that: (1) when testing for linkage and LD jointly, the exact and asymptotic procedure gave similar results for parent–offspring trios; (2) FBAT is more powerful in the families with two discordant sibs and without parents; and (3) the exact test is more prominently powerful only for diseases with a recessive inheritance model.

Overall, we would like to conclude with enthusiasm that the idea of using joint linkage and LD tests to map for complex traits is promising. In theory, the pseudomarker statistical approach and the FBAT approach are both expected to offer reasonable statistical power to detect genes underlying complex traits. When the pedigree structure is more complicated or when the number of informative pedigrees is limited, a pseudomarker approach will outperform FBAT. The pseudo-marker approaches may have more flexibility in terms of combing different type of datasets including sib-pairs, trios, general pedigrees, and case–control samples in one analysis and as indicated by Terwilliger and Goring ***(13)***, throwing away information on correlated genotypes among related individuals seems to reduce power to detect linkage and LD.

On the other hand, FBAT allows one to evaluate any test statistic that can be expressed as the sum of products between an arbitrary function of an offspring's genotype with an arbitrary function of the offspring's phenotype even if there are missing parental information ***(24)***. This feature makes it already a very popular approach.

In recent years, tremendous amounts of effort has been made to extend the earlier work by Laird and her colleagues. A more recent version of PBAT ***(18,26)*** can handle not only nuclear families with missing data, but also general pedigrees with missing genotypes. One difference between FBAT and PBAT is that: "FBAT" handles pedigrees by breaking each pedigree into all possible nuclear families, and evaluating their contribution to the test statistic independently, whereas "PBAT" conditions on the founder genotypes, or their sufficient statistics if they are missing, to obtain the joint distribution of all the offspring in the pedigree, as proposed by Rabinowitz and Laird ***(14)***. Furthermore, tools for family-based association studies (PBAT) can perform a variety of statistical tests including analysis of SNPs, haplotype analysis, quantitative traits, multivariate/longitudinal data, and even time-to-onset phenotypes. In addition, screening tools have been implemented in PBAT (V2.5) to allow the user to successfully address the multiple comparisons problem at a genome-wide level, even for 100,000 SNPs and more to meet the new challenge of the forthcoming genomic-association studies, although further guidelines are needed for the new users in terms of the advantages and disadvantages of using the screening tools implemented in PBAT.

We also would like to note that more computer softwares other than Pseudomarker and FBAT/PBAT also implemented statistical tests for detecting transmission distortion in either nuclear or general pedigrees *(27–29)*. Particularly, Cantor et al. *(29)* developed a new statistical model that can be used to estimate both recombination and LD parameters and perform likelihood ratio tests for linkage alone, joint linkage and association, and association in the presence of linkage. Moreover, this method allows users to specify a relatively realistic penetrance model with reduced penetrance and phenocopies at the disease locus of interest. This model has been implemented in a computer software Mendel release 5.7 *(30)*. It is of importance to use both simulated and empirical data sets to compare and apply different methods discussed (but not limited to) in this chapter.

## References


1 Ott, J. (1999) *Analysis of Human Genetics Linkage 3rd ed.*, Johns Hopkins University Press, Baltimore, MD.

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

3 Spielman, R. S., McGinnis, R. E., and Ewen, W. J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52,** 506–516.

4 Terwilliger, J. D. (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.

5 Xiong, M. and Guo, S. W. (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. *Am. J. Hum. Genet.* **60,** 1513–1531.

6 Lazzeroni, L. C. (1998) Linkage disequilibrium and gene mapping: an empirical least-squares approach. *Am. J. Hum. Genet.* **62,** 159–170.

7 Graham, J. and Thompson, E. A. (1998) Disequilibrium likelihoods for fine-scale mapping of a rare allele. *Am. J. Hum. Genet.* **63,** 1517–1530.

8 Chapman, N. H. and Wijsman, E. M. (1998) Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. *Am. J. Hum. Genet.* **63,** 1872–1885.

9 Xiong, M. and Jin, L. (2000) Combined linkage and linkage disequilibrium mapping for genome scans. *Genet. Epidemiol.* **19,** 211–234.

10 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517.

11 Ott, J. (1989) Statistical properties of the haplotype relative risk. *Genet. Epidemiol.* **6,** 127–130.

12 Terwilliger, J. D. and Ott, J. (1992) A haplotype-based "haplotype relative risk" approach to detection allelic associations. *Hum. Hered.* **42,** 337–346.

13 Terwilliger, J. D. and Goring, H. H. (2000) Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Human Biol.* **72,** 163–132.

14 Rabinowitz, D. and Laird, N. M. (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.* **50,** 211–223.

15 Laird, N. M, Horvath, S., and Xu, X. (2000) Implementing a unified approach to family-based tests of association. *Genet. Epidemiol.* **19,** S36–S42.

16 Lake, S. L., Blacker, D., and Laird, N. M. (2000) Family-based tests of association in the presence of linkage. *Am. J. Hum. Genet.* **67,** 1515–1525.

17 Lange, C., Silverman, E., Weiss, S., Xu, X., and Laird, N. M. (2002) A multivariate family-based test using generalized estimating equations: FBAT-GEE. *Biostatistics* **1,** 1–15.

18 Lange, C., DeMeo, D., Silverman, E. K., Weiss, S. T., and Laird, N. M. (2004) PBAT: tools for family-based association studies. *Am. J. Hum. Genet.* **74,** 367–369.

19 Whittaker, J. C. and Lewis, C. M. (1998) Power comparisons of the transmission/disequilibrium test and sib–transmission/disequilibrium-test statistics. *Am. J. Hum. Genet.* **65,** 578–580.

20 Lunetta, K. L, Faraone, S. V., Biederman, J., and Laird, N. M. (2000) Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *Am. J. Hum. Genet.* **66,** 605–614.

21 Horvath, S., Xu, X., Lake, S. L., Silverman, E. K., Weiss, S. T., and Laird, N. M. (2004) Family based tests for association haplotypes with general trait data: application to asthma genetics. *Genet. Epidemiol.* **26,** 61–69.

22 Lake, L. S. and Laid, N. M. (2003) Tests of gene-environment interaction for case-parent triads with general environmental exposures. *Annals Hum. Genet.* **68,** 55–64.

23 Umbach, D. M. and Weinberg, C. R. (2000) The use of case-parent triads to study joint effects of genotype and exposure. *Am. J. Hum. Genet.* **66,** 251–261.

24 Horvath, S., Xu, X., and Laird, N. M. (2001) The family based association test method: strategies for studying general genotype-phenotype associations. *Eur. J. Hum. Genet.* **9,** 301–309.

25 Schneiter, K., Laird, N., and Corcoran, C. (2005) Exact family-based association tests for biallelic data. *Genet. Epidemiol.* **29,** 185–194.

26 Steen, K. V. and Lange, C. (2005) PBAT: a comprehensive software package for genome-wide association analysis of complex family-based studies. *Hum. Genomics* **2,** 67–69.

27 Monks, S. A., Kaplan, N. L., and Weir, B. S. (1998) A comparative study of sibship tests of linkage and/or association. *Am. J. Hum. Genet.* **63,** 1507–1516.

28 Martin, E. R., Bass, M. P., Hauser, E. R., and Kaplan, N. L. (2003) Accounting for linkage in family based tests of association with missing parental genotypes. *Am. J. Hum. Genet.* **73,** 1016–1026.

29 Cantor, R. M., Chen, G. K., Pajukanta, P., and Lange, K. (2005) Association testing in a linked region using large pedigrees. *Am. J. Hum. Genet.* **76,** 538–542.

30 Lange, K., Cantor, R., Horvath, S., et al. (2001) Mendel version 4.0: a complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Am. J. Hum. Genet.* **69,** 504.

# 11

# Association Studies Using Familial Cases

## *An Efficient Strategy for Identifying Low-Penetrance Disease Alleles*

**Emily L. Webb and Richard S. Houlston**

**Summary**

Low-penetrance alleles are likely to contribute to inherited susceptibility to many complex traits. Such alleles will rarely generate multiple-case families and are therefore difficult or impossible to identify through genetic linkage analyses. The search for low-penetrance alleles has therefore centred on comparing the frequencies of specific alleles in cases and controls via an association study. With recent improvements in genotyping technology and cost, and the completion of the HapMap Project, the long-predicted era of whole-genome association studies is now upon us, with several large-scale studies underway. Such studies require the simultaneous performance of a large number of statistical tests, with the result that power to detect association is in short supply, particularly if the disease allele is rare. One strategy to increase the power of an association study is to enrich cases for genetic predisposition; for this purpose, studies based on familial cases have attracted considerable interest. Using cancer as an example of a complex trait, we show that this approach greatly increases the power to detect association under a range of modes of inheritance, relative risks, and allele frequencies, but is especially efficient for detection of rare alleles.

**Key Words:** Association; familial; risk; power.

## 1. Introduction

### *1.1. Inherited Susceptibility to the Common Cancers as an Example of the Genetic Basis of Inherited Predisposition to a Complex Trait*

For most of the common malignancies, such as breast and colorectal cancer, patients' first-degree relatives (parents, siblings, and children) suffer an approximately twofold increased risk for cancer at the same site *(1)*. Recent twin studies suggest that much of this familial aggregation results from inherited susceptibility *(2)*, especially in breast cancer *(3)*. Mutations in known genes cannot

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

account for most of the excess risk. In breast cancer, for example, mutations in known predisposition genes, including *BRCA1* and *BRCA2*, account for only approx 20% of the twofold excess in patients' relatives ***(4)***. The remaining familial risk could be a result of high-penetrance mutations in as yet unidentified genes, but a polygenic mechanism may provide a more plausible alternative explanation. Several important loci for common cancers were found by linkage analysis several years ago, but even quite striking multiple-case cancer families have failed to reveal significant linkage to novel loci in most recent studies. Under the polygenic model, a large number of alleles each conferring a small genotypic risk (perhaps of the order of 1.5–2.0) combine additively or multiplicatively to confer a range of susceptibilities in the population. More than 100 such variants may contribute to susceptibility ***(5)***. Individuals carrying few such alleles would be at reduced risk, whereas those with many might suffer a lifetime risk as high as 50% ***(6)***.

Alleles conferring relative risks of 2.0 or less will rarely cause multiple-case families and are difficult or impossible to identify through linkage ***(7)***. The search for low-penetrance alleles has therefore centered on consideration of the frequency of candidate alleles in unselected cases and controls, either by direct comparison of proportions or by using the transmission disequilibrium test or allied tests. This approach is satisfactory for the evaluation of common variants but has limited power if the carrier frequency of the deleterious allele is less than 5%. The power of association studies can, however, be greatly enhanced by utilizing genetically enriched cases.

### *1.2. Design Strategies for Association Studies*

With the completion of the HapMap Project (http://www.hapmap.org/) and recent improvements in genotyping technology and reduced cost, the first full-genome association studies have recently appeared in the literature ***(8,9)***. The HapMap Project resource allows selection of reference panels with 250,000–500,000 single-nucleotide polymorphisms (SNPs) that capture at least 80% of common variation in African populations, with 94% of common variation in Caucasian and Chinese populations.

The HapMap approach has some success at capturing rare variants (frequency <5%). However, an alternative approach to the use of a map of anonymous haplotypes is a sequence-based method. The rationale for this approach is provided by the fact that in Mendelian diseases for which the underlying mutation has been identified, the majority (59%) are missense, 22% are deletions, 10% are splice-site mutations, and 7% are insertions or duplications with only 1% in regulatory regions ***(10)***. In recognition of this fact, the sequence-based approach aims to empower studies by increasing the prior probability that a SNP has a functional impact. However, irrespective of the association study

philosophy implemented, the issues of statistical power are equivalent, and the methods described here are equally applicable to each situation.

### 1.3. Analysis Strategies for Association Studies

Current strategies for the analysis of the genotype data generated by a whole-genome association study generally encompass statistical testing carried on a marker-by-marker basis. Using such an approach (and any more complex approaches), the number of tests undertaken will be very large, leading to a multiple testing problem. If a level of 5% is used to define a significant result, then by definition the corresponding type-I error rate (i.e., the probability of an association being declared true when it is, in fact, false) is 5%. Consequently, if 500,000 SNPs are to be tested on a locus-by-locus basis, 25,000 SNPs will display nominally significant associations purely by chance. Among those results, it is expected that there will reside a subset of SNPs where the association is truly causal. The simplest response to the multiple testing issue (which is often over-stringent given the correlated nature of closely spaced SNPs) is to employ a Bonferroni correction, dividing the global significance level by the number of tests undertaken. This leads to a required individual test *p*-value in the range of $10^{-6}$ to $10^{-8}$ to declare a significant result. However, even if a less conservative multiple testing adjustment is applied, it is improbable that the threshold below which a *p*-value must fall to declare a significant association would be drastically increased. It is also unlikely that alternative analysis methods to the marker-by-marker approach will significantly increase this, although research into alternative approaches is currently underway. It is therefore imperative to consider the design of the association study at the outset to ensure that power to detect association is maximized at this stage.

A number of methods of generating a genetically enriched series of cases have been proposed, including utilizing cases with early onset disease and those with a family history of the disease ***(11–13)***. Although the observation that for many cancers the probability of being a carrier of a highly penetrant mutation is higher in early onset cases supports the rationale of using this as a strategy for case selection, it is not a universal finding. Furthermore, the rationale is predicated on the assumption that there will be a similar marked difference in allele frequency with respect to low-penetrance alleles. This is inherently uncertain. As family history is the strongest risk factor for most cancers it provides a far more robust method of generating a genetically enriched series of cases. Here, we consider the performance of using familial cases over a range of genetic modes of inheritance, allele frequencies, and genotypic relative risks.

## 2. The Magnitude of Risk Associated With Low-Penetrance Alleles

For a complex disease, the excess familial risk is unlikely to be fully explained by disease alleles with high genotype relative risks. In fact, results

from meta-analyses have shown that the expected degree of risk associated with polymorphisms/subpolymorphisms is in the range 1.5 to 2.0 ***(5)***. Furthermore, in order to satisfactorily account for multiple testing issues, *p*-values of the order of $10^{-6}$–$10^{-8}$ are required in order to declare a disease risk association with confidence. Coupled together, these two inferences imply that case–control studies with sample sizes of an order much larger than those previously undertaken will be required in order to detect association, and the situation is especially critical if the variant allele is rare. For example, suppose that a whole-genome scan with 100,000 SNPs is undertaken and a Bonferroni correction used to stringently correct for multiple testing. Then under a multiplicative model and assuming equal numbers of cases and controls, an allele with population frequency of 0.05, conferring a relative risk of 1.5, requires a sample size of approx 7000 unselected cases in order to yield 90% power to detect an association.

In many situations the cost of genotyping a large number of SNPs across the number of people required to detect an association is prohibitively large; one method for reducing cost is to use a multistage design, selecting the most significant markers from the first stage for further analysis in the later stages. The preeminent aim for the first stage of such an approach is simply to identify those markers, which may be associated with the disease state, whereas stages two and three are used to determine the magnitude of risk associated with each marker. Although they should not be used to estimate disease risk, the use of familial or bilateral cases has particular facility at the first stage of the process where the number of genotypes undertaken is limited by cost.

## 3. Increasing Power Through Analysis of Familial Cases

The potential of association studies of familial cases to detect rare susceptibility alleles conferring a relative risk of less than 2.0 is illustrated by the recent analysis of the 1100delC *CHEK2* mutation in breast cancer patients. This allele, which is carried by approx 1% of the population, confers a 1.7-fold increase in breast cancer risk ***(14)***. The prevalence was not significantly increased among unselected breast cancer cases (1.4%), but was greatly increased among familial cases not carrying *BRCA1* or *BRCA2* mutations (5.1%; $p < 10^{-7}$).

The substantive increase in the probability of a familial case being a carrier of a deleterious low-risk allele compared with an unselected case is illustrated in **Fig. 1**. This shows the allele frequency for an affected individual with one, two, or three affected relatives for a range of population frequencies.

The frequency of the disease allele among cases increases as the number of affected relatives increases, being approx 1.3-fold for cases with one affected relative, 1.5-fold for cases with two affected first-degree relatives, and 2-fold for cases with three affected first-degree relatives. For breast cancer, the disease

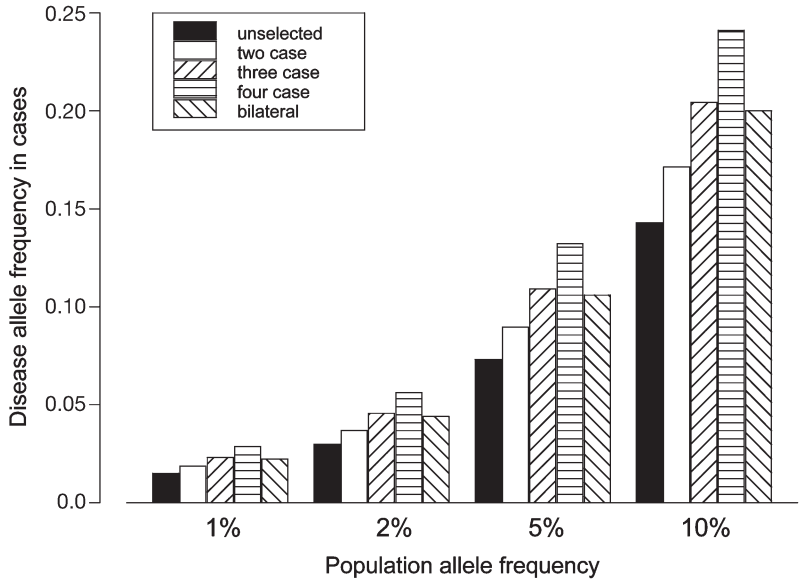



Fig. 1. Allele frequencies for an allele conferring a relative risk of 1.5 for an affected individual from different types of family under a range of population prevalences.

allele frequency for bilateral cases approximates to that of cases with two affected relatives.

## 4. Power Calculations

To evaluate the relative increase in power to identify a disease-causing allele using familial cases compared with unselected cases, we computed sample sizes required for an association study based on unselected cases; cases with an affected parent; cases with two affected first degree relatives; and bilateral cases, assuming an equal number of controls. Computations were carried out under multiplicative, dominant, recessive, and additive models over a range of allele frequencies and relative risks imposing a significance level of $1 \times 10^{-6}$ and power of 90%. The small significance level was chosen to reflect those that will be required for a whole-genome association study of 100,000 SNPs (applying a Bonferroni correction with a liberal 10% significance level). Power calculations were generated using the program FAMASSOC (available upon request) that makes use of the FASTLINK software (http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html) *(**15**,**16**)* to compute carrier probabilities for different types of pedigree. Based on these carrier probabilities, the power of an association study was calculated under the assumption that appropriate $\chi^2$ statistics were used to test for association, namely the allelic $\chi^2$ statistic was used for multiplicative and additive modes of inheritance, whereas dominant and recessive

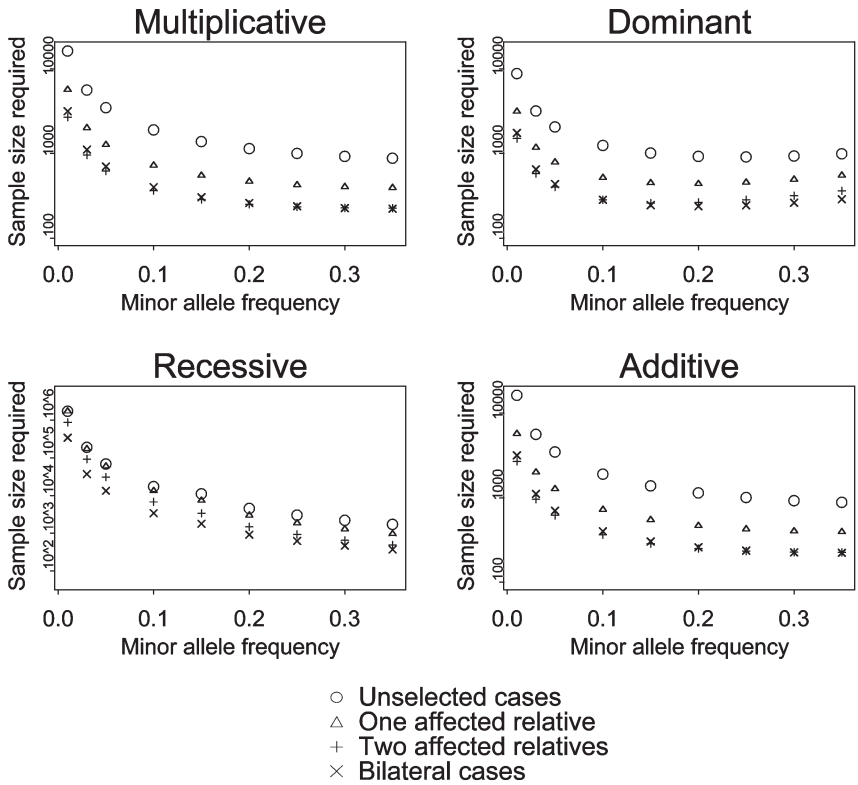



Fig. 2. Sample size required under multiplicative, dominant, recessive, and additive modes of inheritance to detect an allele conferring a relative risk of 1.8 in unselected cases, cases with one affected parent, cases with an affected parent and sibling, and bilateral cases.

$\chi^2$ statistics were used for the dominant and recessive modes of inheritance, respectively (**Fig. 2**).

For all modes of inheritance and allele frequencies, sample size requirements are significantly smaller for familial or bilateral cases than for unselected cases. The reduction in sample size required increases as the number of affected relatives increases. The reduction achieved by use of bilateral cases is comparable to that gained by use of cases with an affected parent and sibling, for all but the recessive mode of inheritance. The sample size required decreases as the disease allele becomes more common, although it remains fairly constant for allele frequencies more than 15%.

**Table 1** details the percentage decrease in sample size required to detect association under dominant and recessive models for cases with unaffected parent; cases with an affected parent and sibling; cases with two affected siblings; cases

**Table 1**
**Percentage Reduction in Sample Size Required to Detect a Disease Allele Conferring Relative Risk 2.0 Under Dominant and Recessive Modes of Inheritance, When Using Familial and Bilateral Cases Rather Than Unselected Cases**

| | | Dominant model | | | | Recessive model | | | |
|---|---|---|---|---|---|---|---|---|---|
| Family history | Allele frequency | 1% | 3% | 5% | 10% | 1% | 3% | 5% | 10% |
| Affected parent | | 65.5 | 64.1 | 62.5 | 58.9 | 10.1 | 8.5 | 13.5 | 23.0 |
| Affected parent, sibling | | 84.0 | 82.5 | 81.0 | 77.3 | 56.3 | 53.4 | 56.6 | 62.6 |
| Two affected siblings | | 84.0 | 82.6 | 81.2 | 77.8 | 75.5 | 70.8 | 71.7 | 73.5 |
| Affected parent, two siblings | | 91.3 | 89.9 | 88.4 | 89.1 | 75.5 | 73.2 | 75.3 | 79.2 |
| Bilateral | | 80.8 | 79.9 | 79.1 | 77.2 | 84.3 | 81.2 | 81.2 | 78.5 |

with an affected parent and two affected siblings; and bilateral cases, when compared with unselected cases. Computations were carried out assuming power of 90%, a relative risk of 2.0 and a significance level of $1 \times 10^{-6}$, assuming that a dominant $\chi^2$ test is used for dominantly inherited alleles and a recessive $\chi^2$ test used for recessively inherited alleles.

Sample size requirements are reduced across the range of population disease allele frequencies under both modes of inheritance. Percentage reductions are generally larger if the disease allele is dominantly inherited, with the exception of bilateral cases, which show approximately equal sample size reductions, irrespective of the mode of inheritance. The sample size reductions are greatest for rare alleles under the dominant model and are approximately uniform over allele frequencies under the recessive model. For example, in order to detect a dominantly inherited disease allele with frequency 1% in the general population, the use of cases with three affected relatives yields a 10-fold decrease in sample size requirements compared with using unselected cases.

The effects of family history on sample size were quantitatively the same irrespective of the magnitude of the relative risk specified. Sample size reductions under the multiplicative and recessive modes of inheritance adhered to a similar pattern shown by those computed for the dominant mode of inheritance (data not shown).

## 5. Concluding Remarks

We have shown that utilizing a set of familial cases and a series of unrelated controls greatly improves power to detect association. Moreover, ignoring family history as a covariate can actually impair the power of the study to detect

**Table 2**
**Expected Genotypes for 200 Bilateral Cases, 500 Unselected Cases, and All Cases With $\chi^2$ Statistics Generated to Test the Null Hypothesis of Equal Genotype Probabilities With Controls**

| Genotype | All cases | Bilaterals | Unselected | Controls |
|---|---|---|---|---|
| Homozygote mutation | 20.3 | 10.1 | 10.2 | 7 |
| Heterozygote | 186.5 | 64.0 | 122.5 | 126 |
| Homozygote wild-type | 493.2 | 125.9 | 367.3 | 567 |
| $\chi^2$ vs controls | 23.3 | 34.5 | 10.3 | |

association. For example, consider the situation where an allele with population frequency of 10% confers a genotypic relative risk of 1.5 under a multiplicative model and suppose that we attempt to detect this allele using a whole-genome association study of 100,000 markers with 700 cases—200 of whom are bilaterals—and 700 controls. **Table 2** shows the expected distribution of genotypes for all cases, bilateral cases, unselected cases and controls, along with $\chi^2$ statistics for the test of equal genotype numbers between cases and controls.

Considering all cases together and ignoring the fact that some are bilateral, the $\chi^2$ statistic attained from comparing allele frequencies in cases and controls is 23.3 yielding a $p$-value of $10^{-5}$, which, even if using a nonconservative correction is unlikely to be deemed significant. However, if the cases are stratified the $\chi^2$ statistic attained is 34.5, which with a corresponding $p$-value of $3 \times 10^{-8}$ would be declared highly significant even under the most conservative adjustments for multiple testing.

Although we have defined familial cases as those with affected first-degree relatives, the FAMASSOC program can be used to compute carrier probabilities for cases with more complex family histories. For example, among cases with one affected sibling and two affected first cousins, carrier probabilities and hence reductions in sample size lie between those observed for cases with two and three affected first-degree relatives. If a case has two affected first cousins but no affected first-degree relatives, the sample size reductions will still improve on those observed if using cases with one affected sibling. Hence, irrespective of the type of family history, the use of familial cases greatly empowers the detection of association. Although generally a greater number of cases in a family translates to a higher probability of a case being a carrier of a deleterious allele, a caveat to this, (depending on the nature of the interaction between high- and low-risk alleles), is that a failure to exclude the involvement of a high-penetrance allele in large multiple-case families may impact significantly on power ***(13)***.

The need to establish large cohorts for genome-wide association studies has received considerable attention recently, but existing family collections may

provide a more powerful resource to identify rare low-penetrance variants through association. We have discussed the value of familial cases for detecting low-penetrance cancer susceptibility alleles, but the strategy is likely to be equally pertinent in the analysis of other complex diseases.

## References


1 Houlston, R. S. and Peto, J. (2004) Genetics and the common cancers. In: *Genetic Predisposition to Cancer,* (Eeles, R. A. et al., eds.), Arnold, London, UK, pp. 235–247.

2 Lichtenstein, P., Holm, N. V., Verkasalo, P. K., et al. (2000) Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343,** 78–85.

3 Peto, J. and Mack, T. M. (2000) High constant incidence in twins and other relatives of women with breast cancer. *Nat. Genet.* **26,** 411–414.

4 The Anglian Breast Cancer Study Group (2000) Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. *Br. J. Cancer* **83,** 1301–1308.

5 Ponder, B. A. (2001) Cancer genetics. *Nature* **411,** 336–341.

6 Pharoah, P. D., Antoniou, A., Bobrow, M., Zimmern, R. L., Easton, D. F., and Ponder, B. A. (2002) Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.* **31,** 33–36.

7 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517.

8 Klein, R. J., Zeiss, C., Chew, E. Y., et al. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* **308,** 385–389.

9 Maraganore, D. M., de Andrade, M., Lasnick, T. G., et al. (2005) High-resolution whole-genome association study of Parkinson disease. *Am. J. Hum. Genet.* **77,** 685–693.

10 Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33,** 228–237.

11 Morton, N. E. and Collins, A. (1998) Tests and estimates of allelic association in complex inheritance. *Proc. Natl. Acad. Sci. USA* **95,** 11,389–11,393.

12 Houlston, R. S. and Peto, J. (2003) The future of association studies of common cancers. *Hum. Genet.* **112,** 434–435.

13 Antoniou, A. C. and Easton, D. F. (2003) Polygenic inheritance of breast cancer: Implications for design of association studies. *Genet. Epidemiol.* **25,** 190–202.

14 Meijers-Heijboer, H., van den Ouweland, A., Klign, J., et al. (2002) Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat. Genet.* **31,** 55–59.

15 Cottingham, R.W., Jr., Idury, R. M., and Schaffer, A. A. (1993) Faster sequential genetic linkage computations. *Am. J. Hum. Genet.* **53,** 252–263.

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

# 12

# Association Mapping Using Pooled DNA

**Hsin-Chou Yang and Cathy S. J. Fann**

**Summary**

The genetic dissection of complex disorders via genetic marker data has gained popularity in the postgenome era. Methods for typing genetic markers on human chromosomes continue to improve. Compared with the popular individual genotyping experiment, a pooled-DNA experiment (alleotyping experiment) is more cost effective when carrying out genetic typing. This chapter provides an overview of association mapping using pooled DNA and describes a five-stage study design including the preliminary calibration of peak intensities, estimation of allele frequency, single-locus association mapping, multilocus association mapping, and a confirmation study. Software and an analysis of authentic data are presented. The strengths and weaknesses of pooled-DNA analyses, as well as possible future applications for this method, are discussed.

**Key Words:** DNA pooling; preferential amplification; allele frequency; linkage disequilibrium; haplotype; SWEPT; microarray.

## 1. Introduction

The purpose of a pooled alleotyping experiment (PAE) is to provide a cost-saving and reliable alternative to an individual genotyping experiment (IGE) in a gene-mapping study. The low cost, reduced effort, high accuracy, and precision are critical aspects of the methods in large-scale association mapping studies. This is because large numbers of study individuals and genetic markers are required to detect disease susceptibility genes or quantitative trait loci that have polygenic effects for complex disorders. Briefly, the PAE involves mixing genomic DNAs from different individuals in the same tube and then extracting the integrated information that is condensed within the fluorescence intensities of different alleles within the DNA pool for follow-up statistical inferences and decision making. The cost of PAE is reduced because of the lower number of genetic typings that are carried out as compared with IGE; for example, 5 million

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

genotypings that are required to analyze 5000 single-nucleotide polymorphisms (SNPs) for 1000 individuals are reduced to 5000 allele typings for PAE with a pool size of 1000 (i.e., a 1000-fold decrease in typing reactions is achieved).

PAE has gained a lot of attention since the pioneering DNA-pooling study *(1)* was carried out, which used restriction fragment length polymorphisms to test linkage disequilibrium (LD) between HLA class II loci and insulin-dependent mellitus. A review of pooled-DNA developments during the past two decades shows its advancement in prominence and importance in association mapping. The growing list of methods for PAE includes (1) carrying out statistical tests, such as the single-locus test *(2,3)* and multilocus association test *(4,5)*; (2) analyzing trait attributes, both qualitative traits *(6)* and extreme quantitative traits (i.e., a selective typing study) *(3,7)*; (3) considering study designs, including a population-based strategy *(8,9)* and family-based strategy *(2,10–12)*; (4) investigating power and cost through a power analysis *(2,13)* or cost evaluation *(14,15)*; (5) studying typing error *(16,17)*; and (6) applications in identifying disease susceptibility genes *(18–24)*.

Other applications for PAE consist of polymorphism validation or identification *(25–30)*, analyzing genetic diversity *(31,32)*, and so on. This method can be used for genetic studies of humans as well as other species of animals *(33)* and plants *(31)*.

A successful PAE depends on accurate and precise estimation of allele frequencies of genetic markers. Different types of genetic markers have been used, e.g., restriction fragment length polymorphisms *(1,34)*, short tandem-repeat polymorphisms *(35,36)*, and SNPs *(37–42)*. The extended analysis consists of the estimation of the coefficient of LD and haplotype frequency *(43–46)*. Many platforms for alleotyping SNP markers have been established, such as MALDI-TOF mass spectrometry, denaturing high-performance liquid chromatography, quantitative single-strand conformation polymorphism analysis, kinetic PCR, pyrosequencing analysis, and so on. The performances of different platforms have been compared *(47–49)*. In addition, advanced microarray systems have been used *(50–52)*. For alleotyping with any of these different platforms, the manufacturer's instructions for suggested protocols should be referred to. As the biological techniques and statistical methodology that are associated with this analysis have improved, PAE has become a useful and popular tool for genetic association mapping studies.

In this chapter, we will introduce PAE association mapping through a five-stage framework, including the preliminary calibration of peak intensities, estimation of allele frequency, single-locus association mapping, multilocus association mapping and a confirmation study, which constitutes a complete pooled-DNA analysis.

## 2. Methods

Throughout this chapter, we focus on high-resolution association mapping using the most abundant SNP markers and a widely used case–control study design. We consider a disease association mapping study consisting of $n_{\text{case}}$ and $n_{\text{control}}$ individuals in groups $G_{\text{case}}$ and $G_{\text{control}}$, which may represent case and control groups or two groups with extremely high and low trait values, respectively. In total, $S$ diallelic SNPs having alleles $M$ and/or $m$ at each SNP are used to scan the chromosome region of interest. Let $p_M^{\text{case}}$ ($p_M^{\text{control}}$) denote the population allele frequency of allele $M$ in the group $G_{\text{case}}$ ($G_{\text{control}}$). The superscripts in $p_M^{\text{case}}$ and $p_M^{\text{control}}$ are omitted if the discussion is not necessary to specify groups.

### 2.1. Adjustment for Preferential Amplification

Accurate and precise allele frequency estimation is required for successful association tests and polymorphism validation/identification in a PAE. To achieve this, the estimating procedure for allele frequency must preclude the bias that is because of preferential amplification of different nucleotides. The unequal amplification may be caused by different efficiencies in nucleotide extension or in the ability to detect the fluorescence signal because of chemical features of the different nucleotides. These factors result in an estimation bias of allele frequency if no proper adjustment is made ***(9,53,54)***. The bias-causing mechanisms will be illustrated in **Subheading 2.2.** Hence, the first stage of a PAE is to perform an adjustment for preferential amplification, before allele frequency estimation is carried out in the next stage.

We define a parameter, $\kappa$, which is called the coefficient of preferential amplification (CPA), to quantify the enlargement or shrinkage of the amplification of a specific allele. The CPA is defined as a ratio of the population peak intensities of the dual alleles for a SNP, i.e., $\kappa = \mu_M / \mu_m$, where $\mu_M$ and $\mu_m$ are the population peak intensities of alleles $M$ and $m$, respectively. If $\kappa = 1$ then the two alleles are equally amplified. If $\kappa > 1$ then allele $M$ tends to be more greatly amplified relative to allele $m$. If $\kappa < 1$ then allele $M$ tends to be less well amplified relative to allele $m$.

In general, the parameter $\kappa$ is unknown and must be estimated. Heterozygous individuals can provide a standard for a 50:50 ratio for a pair of peak intensities of two different nucleotides and thus can be used for quantifying the preferential amplification. Let $\left\{\left(h_{M,1}^{\text{I}}, h_{m,1}^{\text{I}}\right), \left(h_{M,2}^{\text{I}}, h_{m,2}^{\text{I}}\right), \ldots, \left(h_{M,n_h}^{\text{I}}, h_{m,n_h}^{\text{I}}\right)\right\}$ denote the pairs of peak intensities of dual alleles of a SNP for $n_{\text{h}}$ heterozygous individuals based on IGE, and let $\left(\bar{h}_M^{\text{I}}, \bar{h}_m^{\text{I}}\right)$ denote the sample means of the pairs of peak intensities. The first CPA estimator, the arithmetic mean of the peak intensity ratios, was proposed ***(53)*** as follows:

$$\kappa_{\mathrm{H}} = \frac{1}{n_{\mathrm{h}}} \sum_{i=1}^{n_{\mathrm{h}}} \frac{h_{M,i}^{\mathrm{I}}}{h_{m,i}^{\mathrm{I}}} \tag{1}$$

The other two methods, bias-correction CPA and geometric mean of peak intensity ratios, were suggested *(9)* as follows:

$$\hat{\kappa}_{\mathrm{U}} = \hat{\kappa}_{\mathrm{H}} + \frac{n_{\mathrm{h}}}{n_{\mathrm{h}} - 1}\left[\frac{\bar{h}_M^{\mathrm{I}}}{\bar{h}_m^{\mathrm{I}}} - \hat{\kappa}_{\mathrm{H}}\right] \text{ and } \hat{\kappa}_{\mathrm{G}} = \sqrt[n_{\mathrm{h}}]{\prod_{i=1}^{n_{\mathrm{h}}} \frac{h_{M,i}^{\mathrm{I}}}{h_{m,i}^{\mathrm{I}}}} \tag{2}$$

A simulation study that compared the performance of the three CPA estimators showed that all three methods were similar as far as their ability to correct for estimation biases. The method in **Eq. 1.** produces a slightly higher mean squared error relative to the methods summarized in **Eq. 2**, however, using one of these latter methods is generally preferred *(9)*.

The processing of additional samples and the increased costs that are associated with typing additional heterozygous individuals are contrary to the spirit of a PAE. Two strategies have been suggested to overcome this difficulty. The first suggests small samples, e.g., about eight or slightly more heterozygous individuals *(53)* are sufficient to perform calibrations for most practical applications. Although these data can be used subsequently in an IGE confirmation study (the fifth stage), which will be discussed in **Subheading 2.5.**, this procedure slightly increases the typing costs for a PAE. Alternatively, a reference database can be established for preferential amplification. Using CPAs from a publicly available database helps avoid typing additional heterozygous individuals. Two public databases *(55,56)* are available at http://cogent.iop.kcl.ac.uk/rcorrection.cogx and http://www.ibms.sinica.edu.tw/%7Ecsjfann/first%20flow/database.htm, respectively. More systematic investigations about the impact of using CPAs derived from different experimental conditions or populations on allele frequency estimation should be carried out.

### 2.2. Estimation of Allele Frequency

In this stage, DNAs from individuals of the same group are mixed together. Based on peak intensities in a DNA pool, $(h_M^{\mathrm{P}}, h_m^{\mathrm{P}})$, allele frequency and the corresponding variance are estimated and used for the next stage, case–control association mapping. In an IGE, the frequencies of the dual alleles of a SNP are

$$p_M = N_M/(N_M + N_m) \text{ and } p_m = 1 - p_M \tag{3}$$

where $N_M$ and $N_m$ are the number of each of the two alleles in the study population. The frequency can be estimated by directly counting alleles in representative samples as follows:

$$\hat{p}_M = n_M/(n_M + n_m) \text{ and } \hat{p}_m = 1 - \hat{p}_M \tag{4}$$

where $n_M$ and $n_m$ are the number of each of the two alleles in the samples.

A PAE involves the mixing of genomic DNAs from different individuals to form a DNA pool. The allele-counting approach fails because alleles cannot be counted individually and frequency can be measured only collectively by peak intensities in a DNA pool. The readings of florescence intensities for alleles, calculated from the mass spectrum of primer extension products or fluorescent signals resulting from allele hybridization, are extracted from alleotyping experiments. The intensities reflect the accumulation amount of the alleles in a DNA pool. Hence, allele frequencies of a diallelic SNP are estimated indirectly by calculating the relative florescence intensities of the two alleles. Ideally, the ratio between the two allele frequencies is equal to the ratio between peak intensities of the two alleles in a SNP, i.e., $p_M : p_m = H_M : H_m$, satisfying the constraint $p_M + p_m = 1$, where $H_M$ ($H_m$) denotes the accumulated peak intensity of allele $M$ ($m$) in a DNA pool. The two equations result in the following allele frequency estimators:

$$\hat{p}_M = h_M^{\mathrm{P}} / (h_M^{\mathrm{P}} + h_m^{\mathrm{P}}) \text{ and } \hat{p}_m = 1 - \hat{p}_M \tag{5}$$

where $h_M^{\mathrm{P}}$ and $h_m^{\mathrm{P}}$ are the measured peak intensities in the DNA pool.

In practice, one often sees preferential amplification in a PAE as described in **Subheading 2.1.** This interference results in the relationship $p_M : p_m = H_M : \kappa H_m$, which makes estimated allele frequencies in **Eq. 5** biased if $\kappa \neq 1$. Conditional on $\kappa$, adjusted allele frequencies derived from the equation system of $p_M : p_m = H_M : \kappa H_m$ and $p_M + p_m = 1$ are as follows:

$$\hat{p}_{\mathrm{Adjust},M} = h_M^{\mathrm{P}} / (h_M^{\mathrm{P}} + \kappa h_m^{\mathrm{P}}) \text{ and } \hat{p}_{\mathrm{Adjust},m} = 1 - \hat{p}_{\mathrm{Adjust},M} \tag{6}$$

Let $R_H = H_M / H_m$ and $R_h = h_M^{\mathrm{P}} / h_m^{\mathrm{P}}$. A feasible PAE experiment satisfies $R_h / R_H \approx 1$. The bias of frequency estimation of allele $M$ without the CPA adjustment is $\left[\kappa \times (R_h / R_H) - 1\right] / \left[(R_h + 1) \times (1 + \kappa / R_H)\right]$. Therefore, if allele $M$ tends to be more amplified relative to allele $M$ (i.e., $\kappa > 1$), then ignorance of the adjustment for preferential amplification results in a positive bias (i.e., overestimation) of the frequency of allele $M$, and vice versa. The estimation bias leads to misleading conclusions of the follow-up statistical inferences. The final adjusted allele frequency can be estimated by plugging the estimate of $\kappa$ in **Eqs. 1** or **2** into **Eq. 6**. The bias of frequency estimation of allele $M$ with the CPA adjustment is $\left[\kappa \times (R_h / R_H - 1) - (\hat{\kappa} - \kappa)\right] / \left[(R_h + \hat{\kappa}) \times (1 + \kappa / R_H)\right]$. Therefore, the unbiased estimation of the CPA guarantees the allele frequency estimation is unbiased.

Some important sources of variation affect the performance of a PAE (***14***,***15***). The total variance can be partitioned into sampling variation ($\sigma_{\mathrm{S}}^2$), amplification variation ($\sigma_{\mathrm{A}}^2$) and residual variation ($\sigma_{\mathrm{R}}^2$), which results from a combination of other uncontrollable (or hard to control) sources of experimental variation (***8***,***9***) as follows:

$$V\left(\hat{p}_M^{\text{case}} - \hat{p}_M^{\text{control}}\right) = \sigma_{\text{S}}^2 + \sigma_{\text{A}}^2 + \sigma_{\text{R}}^2 \tag{7}$$

where

$$\sigma_{\text{S}}^2 = \frac{p_M^{\text{case}} p_m^{\text{case}}}{2n_{\text{case}}} + \frac{p_M^{\text{control}} p_m^{\text{control}}}{2n_{\text{control}}} \text{ and } \sigma_{\text{A}}^2 = V(\hat{\kappa}) \times \left[\frac{p_M^{\text{case}} p_m^{\text{case}} - p_M^{\text{control}} p_m^{\text{control}}}{\kappa}\right]^2.$$

The variance in **Eq. 7** can be estimated by plugging in the proper estimates of several parameters. The allele frequencies $p_M^{\text{case}}$ and $p_M^{\text{control}}$ can be estimated using **Eq. 6**; $\kappa$ can be estimated using **Eqs. 1** or **2**; $V(\hat{\kappa})$ can be estimated using a bootstrapping procedure ***(9)***; experimental variation $\sigma_{\text{R}}^2$ can be estimated based on a hierarchical design ***(14)*** or variance component analysis ***(15)***. Some estimated variation from different sources have been reported ***(14,15)***; however, the values vary depending on the laboratory and genotyping platform. Different decompositions of the total variance are possible. Therefore, establishing the standard for your own studies is recommended. The two terms $\sigma_{\text{A}}^2$ and $\sigma_{\text{R}}^2$ are the extra variances in a PAE with respect to an IGE. A good PAE should keep these two variances as low as possible by establishing the optimal standard operating procedure in order to retain an efficiency similar to that of an IGE.

### 2.3. Analysis of a Single-Locus Association Test

Before initializing a large-scale association study, the analyses of sample size and power are necessary. To determine the required sample size for a PAE, we can first calculate the required sample for an IGE under prespecified conditions including population allele frequencies, testing power, effect size, and test size. This is routinely done for IGE association studies, and many statistical packages have provided the necessary calculations (e.g., **ref. *57***). Next, we can determine the required sample size, $n_{\text{case}}$ and $n_{\text{control}}$, for a PAE based on the efficiency of the PAE relative to the IGE, i.e., the inverse proportion of variances for PAE and IGE allele frequency estimators ***(14)***. For example, if the numbers of patients and normal controls in a case–control IGE are equal to $n_{\text{IGE}}$, then the respective sample sizes in case and control groups in a PAE attaining the required testing power are as follows:

$$n_{\text{PAE}} = \left[\left(\omega / n_{\text{IGE}}\right) + 2\sigma_{\text{A}}^2 + 2\sigma_{\text{R}}^2\right]^{-1} \tag{8}$$

where

$$\omega = p_M^{\text{case}} p_m^{\text{case}} + p_M^{\text{control}} p_m^{\text{control}}.$$

For a single-locus association test, the frequency of allele *M* for a specific SNP between the case and control groups is compared to test for an association between a marker locus and a putative disease locus. The null hypothesis (i.e., $H_0$: $p_M^{\text{case}} = p_M^{\text{control}}$) specifies no association between marker and disease locus; the alternative hypothesis (i.e., $H_a$: $p_M^{\text{case}} \neq p_M^{\text{control}}$, $H_a$: $p_M^{\text{case}} < p_M^{\text{control}}$ or $H_a$: $p_M^{\text{case}} > p_M^{\text{control}}$)

indicates that the marker is associated with the disease locus. The test statistic for comparing two proportions of two independent populations is

$$Z = \frac{\hat{p}_M^{\text{case}} - \hat{p}_M^{\text{control}}}{\sqrt{\hat{V}\left(\hat{p}_M^{\text{case}} - \hat{p}_M^{\text{control}}\right)}} \tag{9}$$

where $\hat{V}(\hat{p}_M^{\text{case}} - \hat{p}_M^{\text{control}})$ is the estimator of variance in **Eq. 7**. The asymptotic distribution of the test statistic in **Eq. 9** follows a standard normal distribution. The statistical significance can be evaluated by comparing the test statistic $Z$ with a standard normal distribution or by comparing $Z^2$ with a $\chi^2$ distribution with one degree of freedom under a prespecified test size, $\alpha$. A large absolute value of the $Z$ score demonstrates the discrepancy in the allele frequencies between case and control groups and suggests that this SNP may be a candidate associated with the putative disease gene. This method provides a powerful tool for screening candidate loci for large-scale genetic studies. The candidates identified in this stage should be further studied in a follow-up confirmation study.

### 2.4. Analysis of a Multilocus Association Test

Haplotype-based and locus-scoring methods are two main approaches used for multilocus association tests in an IGE *(58)*. Haplotype-based methods have already been applied to PAEs; the estimations of the coefficient of LD and haplotype frequency in a PAE were first discussed in year 2002 *(43)*. In the next year, several methods based on expectation–maximization algorithms were proposed *(44–46)*. At present, only a few software packages are publicly available for these analyses. The review paper *(59)* summarized the available software including LDPooled *(44)*, EHP.R *(46)*, and Pools2 *(60)* for PAE haplotype analysis. The main restrictions of haplotype analysis in a PAE are the small pool size and the small number of markers, which restrict the development of a multilocus analysis in a PAE.

Locus-scoring methods were first applied to study DNA pools based on a sliding-window empirical $p$-value test (SWEPT) *(4)*. This method provides a convenient procedure (sliding window) for screening a large chromosomal region. For example, consider a study region that covers $S$ SNPs ordered according to the physical or genetic map and consider that $\{p_j,\ j = 1,\ldots,S\}$ denotes the $p$-value of a single-locus association test for the $S$ SNPs. The SWEPT procedure uses sliding windows, moving from the beginning to the end of the SNPs and covering $k$ SNPs ($k \leq S$) for each window to scan the entire chromosomal region of study. In each window, the SWEPT combines single-locus $p$-values by different transformation functions. For example, the multiplicative and minimum SWEPT statistics at the $i$th window ($i = 1,\ldots,\ S + 1 - k$) are defined as follows:

$$Z_{\times}(i,k) = \prod_{j=i}^{i+k-1} p_j^{w_{ij} \times I[p_j < \mu]} \text{ and } Z_{\wedge}(i,k) = \min_{j=i,\ldots,i+k-1} \left\{p_j\right\}$$

where $w_{ij}$ is a weight for the $j$th SNP in the $i$th window, $\mu$ is the threshold of $p$-value truncation and $I[E]$ is an indicator function of an event $E$, which takes a value of 1 if event $E$ holds and has a value of 0 otherwise. The significance of the SWEPT results can be evaluated by using a Monte Carlo procedure *(4)*. This testing procedure is not restricted as a result of small pool sizes and/or the number of SNPs analyzed as in the haplotype-based approach. A CPA adjustment has not been incorporated into conventional PAE multilocus association tests; however, this SWEPT method can easily incorporate preferential amplification by using an adjusted single-locus association test (e.g., the test statistic in **Eq. 9**). The SNPs identified via PAE association tests can be regarded as potential loci, but the results should be confirmed by an IGE confirmation study in the next stage.

### 2.5. Confirmation Study Via Individual Genotyping

In the final stage, an IGE study is performed to confirm the association results from the previous four stages. Samples are genotyped only for the significant SNPs identified by PAE single-locus and multilocus association tests. The number of these SNPs out of the original $S$ SNPs is reduced dramatically. Therefore, the cost is reduced greatly in a PAE relative to an IGE. All available statistical methods for the analysis of individual genotyping data can be used at this stage.

## 3. An Illustrative Example

A two-population dataset was analyzed using the software PDA (Pooled DNA Analyzer) *(4)*, which was developed for performing a complete analysis of pooled-DNA data, to illustrate the five-stage procedure introduced in this chapter. The PDA software along with data used in this example can be downloaded at http://www.ibms.sinica.edu.tw/%7Ecsjfann/first%20flow/pda.htm. Ten SNPs of interest within the human major histocompatibility complex were analyzed in this example. The summary results are shown in **Table 1**.

In the first stage, the estimated CPAs and the corresponding standard errors of these 10 SNPs were calculated based on heterozygous individuals selected from an additional sample of 95 individuals. In **Table 1**, the second column shows the number of heterozygous individuals out of 95 individuals at each SNP. The third column shows the estimated CPA ($\hat{\kappa}_U$) for each of the SNPs. For example, 38 heterozygous individuals out of 95 samples were used to estimate the CPA of SNP 1, and the corresponding standard error was 0.003. All estimated CPAs were greater than 1, and the maximum value attained was 2.64, suggesting a strong preferential amplification for the SNPs in this example. Hence, the adjustment of preferential amplification was necessary in this example. The corresponding standard errors of the estimated CPAs were small as shown in the fourth column in **Table 1**.

**Table 1**
**Results From an Example PAE Analysis (*see* Subheading 3.)**

| | | | | Population 1 | | Population 2 | | $Z$ | $Z_\times$ |
|---|---|---|---|---|---|---|---|---|---|
| SNP | $n_h$ | $\hat{\kappa}_U$ | $se(\hat{\kappa}_U)$ | $\hat{p}_M$ | $se(\hat{p}_M)$ | $\hat{p}_M$ | $se(\hat{p}_M)$ | $p$-value | $p$-value |
| 1 | 38 | 2.64 | 0.003 | 0.63 | 0.024 | 0.57 | 0.024 | 0.175 | 0.047 |
| 2 | 37 | 1.76 | 0.002 | 0.37 | 0.024 | 0.45 | 0.024 | 0.065 | 0.084 |
| 3 | 36 | 1.69 | 0.002 | 0.33 | 0.023 | 0.43 | 0.024 | 0.025 | 0.229 |
| 4 | 12 | 1.33 | 0.003 | 0.92 | 0.014 | 0.91 | 0.014 | 0.078 | 0.718 |
| 5 | 41 | 1.78 | 0.003 | 0.51 | 0.024 | 0.45 | 0.024 | 0.163 | 0.629 |
| 6 | 37 | 1.66 | 0.168 | 0.65 | 0.023 | 0.62 | 0.024 | 0.549 | 0.874 |
| 7 | 4 | 1.99 | 0.004 | 0.95 | 0.011 | 0.94 | 0.012 | 0.720 | – |
| 8 | 38 | 2.02 | 0.003 | 0.36 | 0.023 | 0.37 | 0.024 | 0.808 | – |
| 9 | 38 | 1.75 | 0.003 | 0.52 | 0.024 | 0.55 | 0.024 | 0.446 | – |
| 10 | 37 | 1.58 | 0.002 | 0.56 | 0.024 | 0.55 | 0.024 | 0.782 | – |

In the second stage, DNAs from 210 individuals from population 1 were mixed as the first pool, and DNAs from the other 210 individuals from population 2 were mixed as the second pool. Based on the estimated CPA and peak intensity in the DNA pools, the allele frequencies of the two populations were estimated and shown in the fifth and seventh columns, along with their standard errors in the sixth and eighth columns. Results show that most of the 10 SNPs have similar allele frequencies and standard errors between the two populations, except for the third SNP.

In the third stage, we assigned an experimental standard error of 0.02, which we estimated based on a variance component model ***(15)***. The single-locus association test was performed to test for a difference in allele frequencies between the two populations. The $p$-values of the 10 SNPs are shown in the ninth column of **Table 1**. Only the third SNP, with a $p$-value 0.025, was significant at a test size of 0.05.

In the fourth stage, we conducted a multilocus association test. No haplotype methods can handle such pooled-DNA data with a pool size of 210 individuals. We therefore performed the multilocus association test using the SWEPT method. A window size of 5, a multiplicative $p$-value statistic, and 10,000 Monte Carlo simulations were considered to calculate the empirical $p$-value. The empirical $p$-values of six sliding windows are shown in the tenth column of **Table 1**. The first window, which contains SNPs 1–5, reveals a signal of association.

SNPs 6–10 should not be genotyped individually in the fifth stage for a confirmation study because they were not identified by PAE association tests. We typed all 10 SNPs in order to illustrate the different conclusions that are likely

to occur when comparing PAE and IGE analyses. The exact *p*-values of IGE-based association tests for the 10 SNPs were 0.022, 0.005, 0.012, 0.686, 0.023, 0.944, 0.163, 0.447, 0.408, and 0.944, suggesting that the first, second, third, and fifth SNPs may have different genotypic distributions between the two populations. These results were compared with those from the previous four-stage PAE analyses. The PAE-based single-locus association test (the third stage) identified only the third SNP; the multilocus association test (the fourth stage) suggested that SNPs 1–5 should be investigated in the confirmation study. This example demonstrates the whole process of a complete PAE analysis. Pooled-DNA analysis reduces cost of the study because only SNPs 1–5 should be genotyped individually in the confirmation IGE study. The confirmation IGE not only locates truly important SNPs, but also corrects false-positive results from the PAE association mappings.

## 4. Discussion

As genetic typing methodologies improve, DNA pooling is encountering rigorous challenges from individual genotyping techniques that are becoming more and more cost effective. The enhancement of PAE by combining pooled-DNA techniques with modern genetic typing techniques, such as individualized Affymetrix GeneChip ***(61)***, to further reduce typing costs and efforts needed for genome-wide disease gene mapping is worth investigating. For example, the performance of a combination of PAE with the Affymetrix GeneChip Human Mapping 10K array was evaluated ***(62,63)***; the technique was applied to identify loci that confer addiction vulnerability ***(64)*** and mild mental impairment ***(65)***. A recent review summarized the use of microarray-based experimental methods for pooled DNA ***(66)***. Recently, Affymetrix has developed the GeneChip Human Mapping 100K set for genotyping 116,204 SNPs simultaneously based on a pair of oligonucleotide arrays using *Xba*I and *Hind*III restriction enzymes. Gene chips that cover more SNPs will soon be available, such as GeneChip Human Mapping 500K array set. Although the success of this new technique needs more verification, it points out a potential direction for PAE in future genetic studies.

Combining PAE with microarray-based genotyping does not require a completely revamped analytic flow. Compared with conventional mass spectroscopy methods, the Affymetrix GeneChip uses a microarray-based technique and measures peak intensities for fluorescent signals of allele hybridization to genotype SNPs ***(61,67)***. To increase the reliability of the genotyping calls, multiple probe quartets are used. Hence, the major difference in data formats between these two types of genotyping platforms is that for each SNP, only a pair of peak intensities is generated for an individual or a DNA pool with the conventional method, whereas there are multiple pairs for the microarray

method. Existing methods for PAE can be applied to microarray-based data with a data transformation ***(56,62,63)*** that properly reduces the multiple-pair data into single-pair data. Hence, the concept of the five-stage design introduced in this chapter is also applicable to microarray-based PAE data, and the PDA software can also provide the ability to analyze data on multiple intensities.

The advantage of a PAE is its cost savings and high-throughput analyses. The major weakness of PAE, however, is the information from individuals is "almost" lost, although part of this information can be extracted under some restricted assumptions, e.g., a very small pool size ***(60)***. This disadvantage results in difficulties when checking Hardy–Weinberg disequilibrium, carrying out a haplotype or regression-based analysis and so on. PAE may be useful only when used as a screening tool at this stage; however, we believe that as more biological techniques and sophisticated statistical methods are developed, they will overcome the present limitations of PAE and further increase the applicability of pooled-DNA analyses.

## References


1 Arnheim, N., Strange, C., and Erlich, H. (1985) Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: Studies of the *HLA* class II loci. *Proc. Natl. Acad. Sci. USA* **82,** 6970–6974.

2 Risch, N. and Teng, J. (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res.* **8,** 1273–1288.

3 Jawaid, A., Bader, J. S., Purcell, S., Cherny, S. S., and Sham, P. (2002) Optimal selection strategies for QTL mapping using pooled DNA samples. *Eur. J. Hum. Genet.* **10,** 125–132.

4 Yang, H. C., Pan, C. C., Lin, C. Y., and Fann, C. S. J. (2006) PDA: pooled DNA analyzer. *BMC Bioinformatics* **7,** 233.

5 Zeng, D. and Lin, D. Y. (2005) Estimating haplotype-disease associations with pooled genotype data. *Genet. Epidemiol.* **28,** 70–82.

6 Scott, D. A., Carmi, R., Elbedour, K., Yosefsberg, S., Stone, E. M., and Sheffield, V. C. (1996) An autosomal recessive nonsyndromic-hearing-loss locus identified by DNA pooling using two inbred Bedouin kindreds. *Am. J. Hum. Genet.* **59,** 385–391.

7 Darvasi, A. and Soller, M. (1994) Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics* **138,** 1365–1373.

8 Visscher, P. M. and Le Hellard, S. (2003) Simple method to analyze SNP-based association studies using DNA pools. *Genet. Epidemiol.* **24,** 291–296.

9 Yang, H. C., Pan, C. C., Lu, R. C. Y., and Fann, C. S. J. (2005) New adjustment factors and sample size calculation in a DNA-pooling experiment with preferential amplification. *Genetics* **169,** 399–410.

10 Bader, J. S. and Sham, P. (2002) Family-based association tests for quantitative traits using pooled DNA. *Eur. J. Hum. Genet.* **10,** 870–878.
11 Lee, W. C. (2005) A DNA pooling strategy for family-based association studies. *Cancer Epidem. Biomar.* **14,** 958–962.
12 Zou, G. and Zhao, H. (2005) Family-based association tests for different family structures using pooled DNA. *Ann. Hum. Genet.* **69,** 429–442.
13 Baro, J. Á., Carleos, C., Corral, N., López, T., and Cañón, J. (2001) Power analysis of QTL detection in half-sib families using selective DNA pooling. *Genet. Sel. Evol.* **33,** 231–247.
14 Barratt, B. J., Payne, F., Rance, H. E., Nutland, S., Todd, J. A., and Clayton, D. G. (2002) Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann. Hum. Genet.* **66,** 393–405.
15 Downes, K., Barratt, B. J., Akan, P., et al. (2004) SNP allele frequency estimation in DNA pools and variance components analysis. *BioTech* **36,** 840–845.
16 Zou, G. and Zhao, H. (2004) The impacts of errors in individual genotyping and DNA pooling on association studies. *Genet. Epidemiol.* **26,** 1–10.
17 Quade S. R. E., Elston, R. C., and Goddard, K. A. B. (2005) Estimating haplotype frequencies in pooled DNA samples when there is genotyping error. *BMC Genet.* **6,** 25.
18 Barcellos, L. F., Klitz, W., Field, L. L., et al. (1997) Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am. J. Hum. Genet.* **61,** 734–747.
19 Bansal, A., van den Boom, D., Kammerer, S., et al. (2002) Association testing by DNA pooling: an effective initial screen. *Proc. Natl. Acad. Sci. USA* **99,** 16,871–16,874.
20 Mohlke, K. L., Erdos, M. R., Scott, L. J., et al. (2002) High-throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools. *Proc. Natl. Acad. Sci. USA* **99,** 16,928–16,933.
21 Williams, N. M., Spurlock, G., Norton, N., et al. (2002) Mutation screening and LD mapping in the VCFS deleted region of chromosome 22q11 in schizophrenia using a novel DNA pooling approach. *Mol. Psychiatr.* **7,** 1092–1100.
22 Herbon, N., Werner, M., Braig, C., et al. (2003) High-resolution SNP scan of chromosome 6p21 in pooled samples from patients with complex diseases. *Genomics* **81,** 510–518.
23 Hinds, D. A., Seymour, A. B., Durham, L. K., et al. (2004) Application of pooled genotyping to scan candidate regions for association with HDL cholesterol levels. *Hum. Genomics* **1,** 421–434.
24 Johnson, M. P. and Griffiths, L. R. (2005) A genetic analysis of serotonergic biosynthetic and metabolic enzymes in migraine using a DNA pooling approach. *J. Hum. Genet.* **50,** 607–610.
25 Wolford, J. K., Blunt, D., Ballecer, C., and Prochazka, M. (2000) High-throughput SNP detection by using DNA pooling and denaturing high performance liquid chromatography (DHPLC). *Hum. Genet.* **107,** 483–487.
26 Buetow, K. H., Edmonson, M., MacDonald, R., 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.

27 Shubitowski, D. M., Venta, P. J., Douglass, C. L., Zhou, R. X., and Ewart, S. L. (2001) Polymorphism identification within 50 equine gene-specific sequence tagged sites. *Anim. Genet.* **32,** 78–88.

28 Pan, X. and Weissman, S. M. (2002) An approach for global scanning of single nucleotide variations. *Proc. Natl. Acad. Sci. USA* **99,** 9346–9351.

29 Nelson, M. R., Marnellos, G., Kammerer, S., et al. (2004) Large-scale validation of single nucleotide polymorphisms in gene regions. *Genome Res.* **14,** 1664–1668.

30 Yang, H. C., Lin, C. H., Hung, S. I., and Fann, C. S. J. (2006) Polymorphism validation using DNA pools prior to conducting large-scale genetic studies. *Ann. Hum. Genet.* **70,** 350–359.

31 Dubreuil, P., Rebourg, C., Merlino, M., and Charcosset, A. (1999) Evaluation of a DNA pooled-sampling strategy for estimating the RFLP diversity of maize populations. *Plant Mol. Biol. Rep.* **17,** 123–138.

32 Hillel, J., Groenen, M. A. M., Tixier-Boichard, M., et al. (2003) Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. *Genet. Sel. Evol.* **35,** 533–557.

33 Gonda, M. G., Arias, J. A., Shook, G. E., and Kirkatrick, B. W. (2004) Identification of an ovulation rate QTL in cattle on BAT14 using selective DNA pooling and interval mapping. *Anim. Genet.* **35,** 298–304.

34 Kraft, T., Fridlund, B., Hjerdin, A., Sall, T., Tuvesson, S., and Hallden, C. (1997) Estimating genetic variation in cultivated and wild beets using pools of individuals. *Genome* **40,** 527–533.

35 Shaw, S. H., Carrasquillo, M. M., Kashuk, C., Puffenberger, E. G., and Chakravarti, A. (1998) Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res.* **8,** 111–123.

36 Schnack, H. G., Bakker, S. C., van't Slot, R., et al. (2004) Accurate determination of microsatellite allele frequencies in pooled DNA samples. *Eur. J. Hum. Genet.* **12,** 925–934.

37 Breen, G. Harold, D., Ralston, S., Shaw, D., and St. Clair, D. (2000) Determining SNP allele frequencies in DNA pools. *Biotechniques* **28,** 464–466.

38 Germer, S., Holland, M. J., and Higuchi, R. (2000) High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Res.* **10,** 258–266.

39 Sasaki, T., Tahira, T., Suzuki, A., et al. (2001) Precise estimation of allele frequencies of single-nucleotide polymorphisms by a quantitative SSCP analysis of pooled DNA. *Am. J. Hum. Genet.* **68,** 214–218.

40 Norton, N., Williams, N. M., Williams, H. J., et al. (2002) Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Hum. Genet.* **110,** 471–478.

41 Werner, M., Sych, M., Herbon, N., Illig, T., König, I. R., and Wjst, M. (2002) Large-scale determination of SNP allele frequencies in DNA pools using MALDI-TOF mass spectrometry. *Hum. Mutat.* **20,** 57–64.

42 Lavebratt, C., Sengul, S., Jansson, M., and Schalling, M. (2004) Pyrosequencing™-based SNP allele frequency estimation in DNA pools. *Hum. Mutat.* **23,** 92–97.
43 Pfeiffer, R. M., Rutter, J. L., Gail, M. H., Struewing, J., and Gastwirth, J. L. (2002) Efficiency of DNA pooling to estimate joint allele frequencies and measure linkage disequilibrium. *Genet. Epidemiol.* **22,** 94–102.
44 Ito, T., Chiku, S., Inoue, E., et al. (2003) Estimation of haplotype frequencies, linkage disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. *Am. J. Hum. Genet.* **72,** 384–398.
45 Wang, S., Kidd, K. K., and Zhao, H. (2003) On the use of DNA pooling to estimate haplotype frequency. *Genet. Epidemiol.* **24,** 74–82.
46 Yang, Y., Zhang, J., Hoh, J., et al. (2003) Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. *Proc. Natl. Acad. Sci. USA* **100,** 7225–7230.
47 Le Hellard, S., Ballereau, S. J., Visscher, P. M., et al. (2002) SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Res.* **30,** e74.
48 Shifman, S., Pisanté-Shalom, A., Yakir, B., and Darvasi, A. (2002) Quantitative technologies for allele frequency estimation of SNPs in DNA pools. *Mol. Cell. Probe.* **16,** 429–434.
49 Sham, P., Bader, J. S., Craig, I., O'Donovan, M., and Owen, M. (2002) DNA pooling: A tool for large-scale association studies. *Nat. Rev. Genet.* **3,** 862–871.
50 Uhl, G. R., Liu, Q. R., Walther, D., Hess, J., and Naiman, D. (2001) Polysubstance abuse-vulnerability genes: genome scans for association, using 1,004 subjects and 1,494 single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **69,** 1290–1300.
51 Lindroos, K., Sigurdsson, S., Johansson, K., Rönnblom, L., and Syvänen, A. C. (2002) Multiplex SNP genotyping in pooled DNA samples by a four-colour microarray system. *Nucleic Acids Res.* **30,** e70.
52 Ye, B. C., Zuo, P., Yi, B., and Li, S. (2004) Estimation of relative allele frequencies of single-nucleotide polymorphisms in different populations by microarray hybridization of pooled DNA. *Anal. Biochem.* **333,** 72–78.
53 Hoogendoorn, B., Norton, N., Kirov, G., et al. (2000) Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Hum. Genet.* **107,** 488–493.
54 Moskvina, V., Norton, N., Williams, N., Holmans, P., Owen, M., and O'Donovan, M. (2005) Streamlined analysis of pooled genotype data in SNP-based association studies. *Genet. Epidemiol.* **28,** 273–282.
55 Simpson, C. L., Knight, J., Butcher, L. M., et al. (2005) A central resource for accurate allele frequency estimation from pooled DNA genotyped on DNA microarrays. *Nucleic Acids Res.* **33,** e25.
56 Yang, H. C., Liang, Y. J., Huang, M. C., et al. (2005) A genome-wide study of preferential amplification/hybridization in microarray-based pooled DNA experiments. *Nucleic Acids Res.* DOI: 10, 1093/nar/gkl 4c.
57 Purcell, S., Cherny, S. S., and Sham, P. C. (2003) Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19,** 149–150.

58 Seaman, S. R. and Müller-Myhsok, B. (2005) Rapid simulation of p values for product methods and multiple-testing adjustment in association studies. *Am. J. Hum. Genet.* **76,** 399–408.

59 Salem, R. M., Wessel, J., and Schork, N. J. (2005) A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Hum. Genomics* **2,** 39–66.

60 Hoh, J., Matsuda, F., Peng, X., Markovic, D., Lathrop, M. G., and Ott, J. (2003) SNP haplotyping tagging from DNA pools of two individuals. *BMC Bioinformatics* **4,** 14.

61 Matsuzaki, H., Dong, S., Loi, H., et al. (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* **1,** 109–111.

62 Butcher, L. M., Meaburn, E., Liu, L., et al. (2004) Genotyping pooled DNA on microarrays: a systematic genome screen of thousands of SNPs in large samples to detect QTLs for complex traits. *Behav. Genet.* **34,** 549–555.

63 Meaburn, E., Butcher, L. M., Liu, L., et al. (2005) Genotyping DNA pools on microarrays: tackling the QTL problem of large samples and large numbers of SNPs. *BMC Genomics* **6,** 52.

64 Liu, Q. R., Drgon, T., Walther, D., et al. (2005) Pooled association genome scanning: validation and use to identify addiction vulnerability loci in two samples. *Proc. Natl. Acad. Sci. USA* **102,** 11,864–11,869.

65 Butcher, L. M., Meaburn, E., Knight, J., et al. (2005) SNPs, microarrays and pooled DNA: identification of four loci associated with mild mental impairment in a sample of 6000 children. *Hum. Mol. Genet.* **14,** 1315–1325.

66 Craig, I., Meaburn, E., Butcher, L., Hill, L., and Plomin, R. (2005) Single-nucleotide polymorphism genotyping in DNA pools. In: *Pharmacogenomics: Methods and Protocols,* (Innocenti, F., ed.), Humana, Totowa, NJ, pp. 147–164.

67 Liu, W. M., Di, X., Yang, G., et al. (2003) Algorithms for large-scale genotyping microarrays. *Bioinformatics* **19,** 2397–2403.

# 13

# Selecting Single-Nucleotide Polymorphisms for Association Studies With SNPbrowser™ Software

**Francisco M. De La Vega**


## Summary

The design of genetic association studies using single-nucleotide polymorphisms (SNPs) requires the selection of subsets of the variants providing high statistical power at a reasonable cost. SNPs must be selected to maximize the probability that a causative mutation is in linkage disequilibrium (LD) with at least one marker genotyped in the study. The HapMap Project performed a genome-wide survey of genetic variation with over 3 million SNPs typed in four populations, providing a rich resource to inform the design of association studies. A number of strategies have been proposed for the selection of SNPs based on observed LD, including construction of metric LD maps and the selection of haplotype-tagging SNPs. Power calculations are important at the study design stage to ensure successful results. Integrating these methods and annotations can be challenging: the algorithms required to implement these methods are complex to deploy, and all the necessary data and annotations are deposited in disparate databases. Here, we review the typical workflows for the selection of markers for association studies utilizing the SNPbrowser™ software, a freely available, stand-alone application that incorporates the HapMap database together with gene and SNP annotations. Selected SNPs are screened for their conversion potential to genotyping platforms, expediting the set up of genetic studies with an increased probability of success.


**Key Words:** SNP selection; tagging SNPs; association study; statistical power; genotyping.

## 1. Introduction

One problem researchers face when designing and executing human genetic studies with single-nucleotide polymorphisms (SNPs) is the difficult task of selecting the most suitable set of the variants for the goal at hand in a cost-effective manner. This task is time consuming and overwhelming because of the millions

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

of SNPs currently listed on the public databases and the fact that relevant information is often distributed among multiple repositories.

There are two strategies for conducting association studies: (1) indirect association, which uses surrogate SNP markers with the expectation that one of them is linked to the causative mutation through linkage disequilibrium (LD); and (2) direct association, in which putatively causative SNPs are directly tested. Indirect association studies require that SNPs be selected to maximize the probability that significant LD exists between the unknown causative mutation and at least one of the markers genotyped in the study. Empirical studies have shown that LD can typically extend for tens and even hundreds of kilobases *(1)*, but its distribution varies widely along the genome *(2,3)*. Block-like regions with extensive LD (so-called haplotype blocks) are found interspersed with regions of medium and low LD *(4)*, the latter probably reflecting the distribution of recombination hot spots *(2)*. The International HapMap Consortium recently completed the genotyping of millions of SNPs across four major populations, with the aim of creating a resource of validated SNPs representative of the common sequence variation and describing the fine patterns of LD along the human chromosomes *(3)*. The hope is that this information would enable the optimal selection of markers for cost effective and powerful association studies.

**Figure 1** shows the typical workflow for selecting SNP markers for association-mapping studies for a particular candidate region or gene list (steps 1–3) followed by procurement and validation of reagents (steps 5–68), genotyping of study samples (step 9), data analysis (step 10), and reassessment of additional genotyping for replication (step 11). With the availability of the HapMap Consortium data *(3)* the validation of SNPs and reagents (steps 7–8) has become less necessary. However given that the researcher may be using a different genotyping platform to the mix used by the HapMap Project, good assay performance for the selected SNPs is not guaranteed. Furthermore, when coverage of rare or putative functional variation not captured by the HapMap Project is desired, these validation steps may still be needed.

The goal of selecting the correct SNP coverage is to provide the statistical power required to detect association. A large number of methods have appeared recently in the literature for selecting optimal subsets of SNPs, also referred to as tag-SNPs *(5–16)*. These differ mostly in two major aspects: the quality or correlation measure used to define tagging, i.e., measures to assess how well a set of SNPs captures the variance observed in the original dataset, and the algorithm used for the minimization of the final number of tagging SNPs. This large number of methods creates another dilemma for the researcher: which method for tagging SNP selection should one use? For more detail of methods for optimal subsets of SNPs the reader is referred to recent reviews in the subject *(22)*. On the other hand, LD is not the only important factor when designing an

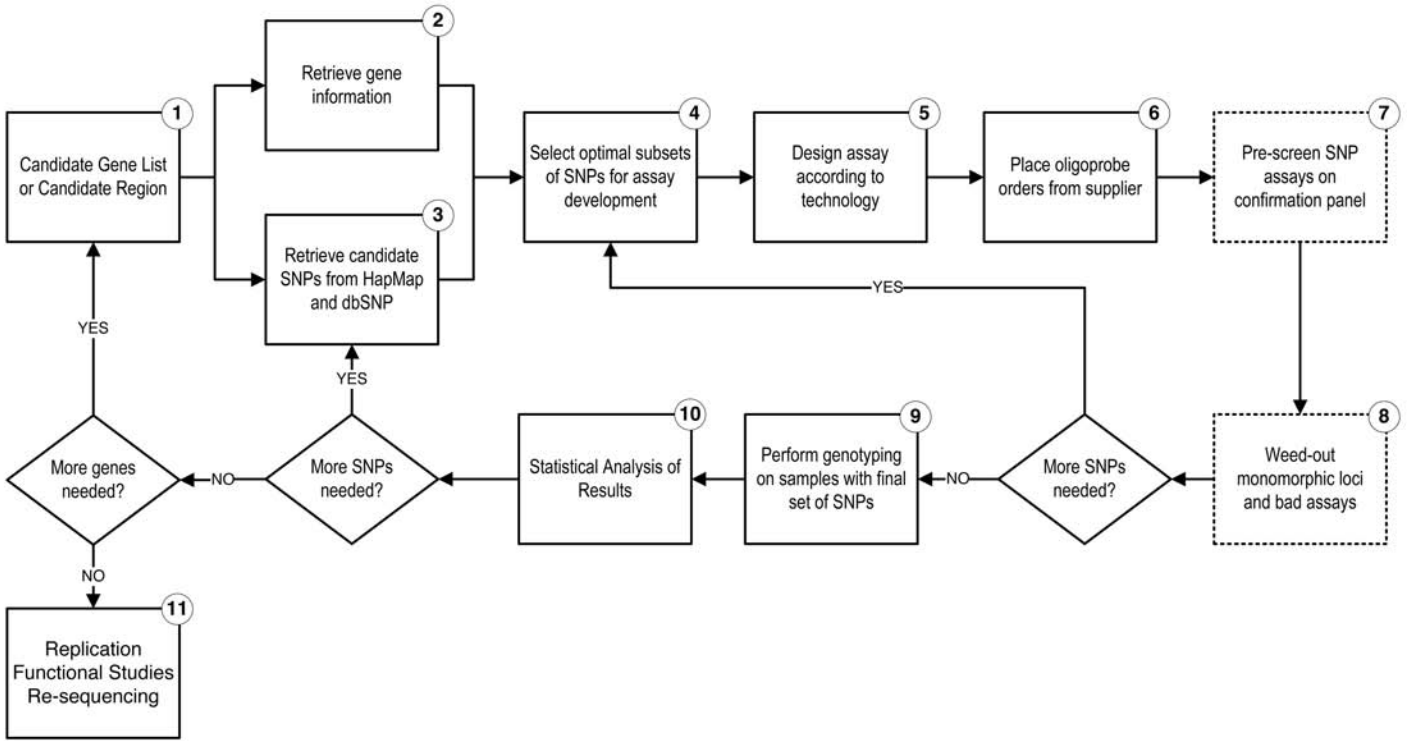



Fig. 1. Typical workflow for the selection of single-nucleotide polymorphism markers and genotyping reagents for association-mapping studies starting from a candidate region or list of candidate genes.

appropriately powered association study. Estimating the statistical power and the required sample size (i.e., the number of cases and controls) required to carry out a successful study is critical, as meta-analyses suggest that many irreproducible associations reported in the recent literature are partly caused by improperly powered studies *(17)*.

When selecting SNPs for a study, integrating all the criteria described previously can be challenging, even with the current availability of larger number of validated SNPs and empirical LD data. In particular, the algorithms required to analyze LD, develop metric LD maps, select haplotype-tagging SNPs, and estimate power are rather specialized. In addition, the necessary SNP annotations (e.g., allele frequency, double-hit status, suitability for a genotyping platform) are deposited in heterogeneous data sources. Furthermore, once a set of SNPs is selected, researchers lack a rapid way to obtain reliable, predictable assays for multiple SNPs that work together under the same experimental conditions. To reduce these barriers, we developed the SNPbrowser™ software (Applied Biosystems, Foster City, CA), a freely available tool providing an intuitive interface to search a stand-alone, embedded database that contains detailed information on millions of validated SNPs. Included in this SNP collection are more than 3 million genome-wide distributed SNPs that the International HapMap Project genotyped *(3)*, as well as 160,000 intragenic SNPs previously validated by us in four populations using TaqMan® SNP genotyping assays *(2,18)*. The depth of SNP and genomic information in the database together with the swift visual interface and embedded selection algorithms provides researchers greater flexibility when designing associations studies with an increased probability of success.

## 2. Methods

### 2.1. SNP Genotype and Annotation Data

We obtained the SNP annotation and genotypes from the public release 19 of the HapMap Consortium ignoring the SNPs with failed assays or those reported as monomorphic, and the data from the children on the CEPH and Yoruba trios *(3)*. Only SNPs having a unique mapping location on the NCBI b35 assembly and a minor allele frequency (MAF) of greater than 5% were considered for further analysis. We previously genotyped DNA samples from 45 African-Americans, 46 Caucasians, 45 Chinese, and 45 Japanese, all unrelated individuals *(2,18)*. Over 160,000 TaqMan SNP genotyping assays were used to genotype these samples. Gene annotation including HUGO names, exon and intron boundaries of all reported (RefSeq NM) and predicted (XM) transcripts were obtained from NCBI Entrez. Transcripts were coalesced into "supertranscript" constructs with boundaries delimited by the coordinates of the first and last base transcribed.

### 2.2. SNP Screening for Genotyping Assay Development

All the SNPs in our database were passed through the high-throughput design pipelines for both TaqMan SNP genotyping assays, and the SNPlex™ genotyping system *(19)*. SNPs that passed the design rules of either platform and are thus candidates for the development of good assays were flagged and subsequent analyses were performed separately for each subset. In the case of TaqMan, assay designs (primers and probes) were uploaded into our TaqMan predesigned database for immediate commercial availability. In the case of the SNPlex system, because it is a multiplexed assay format, we perform a pre-screen for the "single-plex" part of the pipeline; when users submit an actual design request, a few SNPs may still be lost at the final design because of multiplexing rules *(19)*.

### 2.3. Analysis of LD

We constructed metric maps scaled to the strength of LD that can guide the selection of SNPs for association studies. LD units (LDUs) define a metric coordinate system where locations are additive and distances are proportional to the allelic association between markers *(20)*. The LDMAP software v0.9 (available at http://cedar.genetics.soton.ac.uk/public_html/helpld.html) was applied separately to each chromosome and population to construct the corresponding LDU maps. Haplotype blocks were estimated dynamically by LDUs as user-defined intervals with a very small distance in this coordinate system (the default value is 0.3, which returns similar blocks to previous methods *[4]*), or by a rule-based algorithm which uses the D' confidence interval *(4)*, optimized through a dynamic programming algorithm *(21)*.

### 2.4. Selection of Minimum Informative Subsets of SNPs

We utilized three algorithms *(22)* to select minimum informative subsets of SNPs or tag-SNPs: (1) simple genotype correlation between samples (allowing for one item of missing data); (2) pairwise $r^2$ *(7)*; and (3) haplotype $R^2$ *(13)*. We embedded all of these algorithms within the software tool which allows "on-the-fly" selection of tagging SNPs where the user has a greater choice of parameters that could be tailored to the characteristics of the regions of interest. The algorithm implementation is extremely efficient and permits the selection of tag-SNPs for even the largest chromosome in a matter of seconds.

### 2.5. Power Calculations for Case–Control Studies

We calculated power for a fixed sample size of cases and controls on a per gene basis. For each gene, power is calculated using a haplotype-based test, for each of the common haplotypes in the window, considering the empirically

observed average LD of that region. Using a multiplicative genetic model with a relative risk ratio of 3 and a prevalence of 1.5%, power is calculated for each haplotype and a frequency-weighted average is provided as the summary. This is repeated separately for each population, for three settings of sample sizes of cases and controls (250/250, 500/500, 1000/1000), and assuming a disease allele frequency of either 10 or 20%. The resulting estimated power is visualized using a color scale ranging from 0.5 to 1.0 displayed as a background to each gene region *(23)*.

### 2.6. Downloading, Installing, and Updating SNPbrowser

SNPbrowser software was developed using Microsoft® Visual C++, and currently is available only as a native Windows application requiring a system with 512 Mb of RAM. However, the software can be readily used with the MacOS platform with Parallels Desktop (Parallels Inc., Renton, WA), a commercially available emulation environment. The latest version of SNPbrowser is always freely available for download at http://www.allsnps.com/snpbrowser/. Once installed, the software checks for updated versions either automatically or manually.

## 3. Results

### 3.1. SNPbrowser Software-Embedded Database

When SNPbrowser is launched, the user has the option to select which reference database they want to utilize: either the maps and data derived from the HapMap Project *(3)*, or the gene-centric maps obtained by Applied Biosystems by typing 160,000 SNPs in four populations *(2,24)*. After the user selection, SNP and gene annotations, their physical coordinates on the NCBI build 35 assembly, genotypes on the corresponding reference populations, and the results of a series of LD analysis, tagging SNP, and power calculations pipelines performed offline are loaded from a set of binary files distributed with the application into a highly compressed and indexed embedded database maintained in memory. The SNPbrowser database also includes a set of metric LD maps *(20)*, which are empirically derived from the patterns of allelic association observed on the hundreds of millions of genotypes analyzed, and provide information on how to best position SNPs across the genes or regions of interest in a study *(24)*.

### 3.2. Visualization and Query Tools

The SNPbrowser main interface is a visualization panel consisting of a chromosome map viewer representing the location in the physical map of SNPs, and their relationship to annotated human genes and exons. Researchers studying a particular gene or a set of genes can easily pan and zoom to the region of the genome of interest. The users can type directly the gene HUGO name (or aliases) into the search box and press the search button. Instantaneously, the

visualization is focused on the gene of interest and intron/exon structure becomes readily apparent, as is the haplotype-block structure and the location of SNPs along the chromosomal axis (**Fig. 2**).

Vertical blue lines represent SNPs validated in population panels, either by the HapMap Project or Applied Biosystems and for which genotypes are available, whereas grey lines represent SNPs corresponding to over 5 million putative SNP deposited at public databases. The user can select between by clicking the bottom tabs to display the SNPs, which can be developed as assays for the SNPlex genotyping system or as TaqMan SNP genotyping assays ***(19)***. By clicking the "All SNPs" tab, the union of both sets is displayed, including all HapMap-validated SNPs that could not be converted to either TaqMan or SNPlex platforms.

The vertical lines representing SNPs connect to their locations on the LDU coordinates shown on the bottom horizontal axis, in many cases coalescing together into a single position when LD is extensive (i.e., a haplotype block ***[4]***). By clicking and dragging with the mouse any interval can quickly be measured in both base pairs or interpolated LDUs (*see* distance box upper left, **Fig. 2**).

Finally, overall statistical power of the full SNP map, estimated per gene for a preselected genetic model, assumed disease allele frequency, and sample size ***(23)***, is shown color coded within the intronic regions (scale is visible at the upper-right corner).

Searches can be performed with a variety of terms, including gene name, RefSeq transcript ID, NCBI ID, SNP ID, assembly basepair range, or linkage mapping set microsatellite marker set intervals. For most of these identifiers, batch searches are also allowed. Because the SNPbrowser database is loaded into RAM memory, searches are almost instantaneous, which is an advantage over web-based tools. The batch search feature allows users to quickly search genes in big candidate lists and to explore interactively the results of various selection scenarios.

### *3.3. Selection of Evenly Spaced Markers*

SNPbrowser software provides a number of SNP selection "wizards" where researchers can define a region and select SNPs at a given density, based on either LDU or kilobase distances. When selecting SNPs by spacing, the wizards also allow researchers to prioritize the SNPs that are included in the set based on criteria such as MAF and type of SNP. For example, with a few clicks researchers can configure the software to include only SNPs with a MAF of more than 10% in any population and for which a validated SNP assay is available (**Fig. 3A**).

Another typical use case for study design is the candidate region study, where the researchers already performed a linkage or genome-wide association

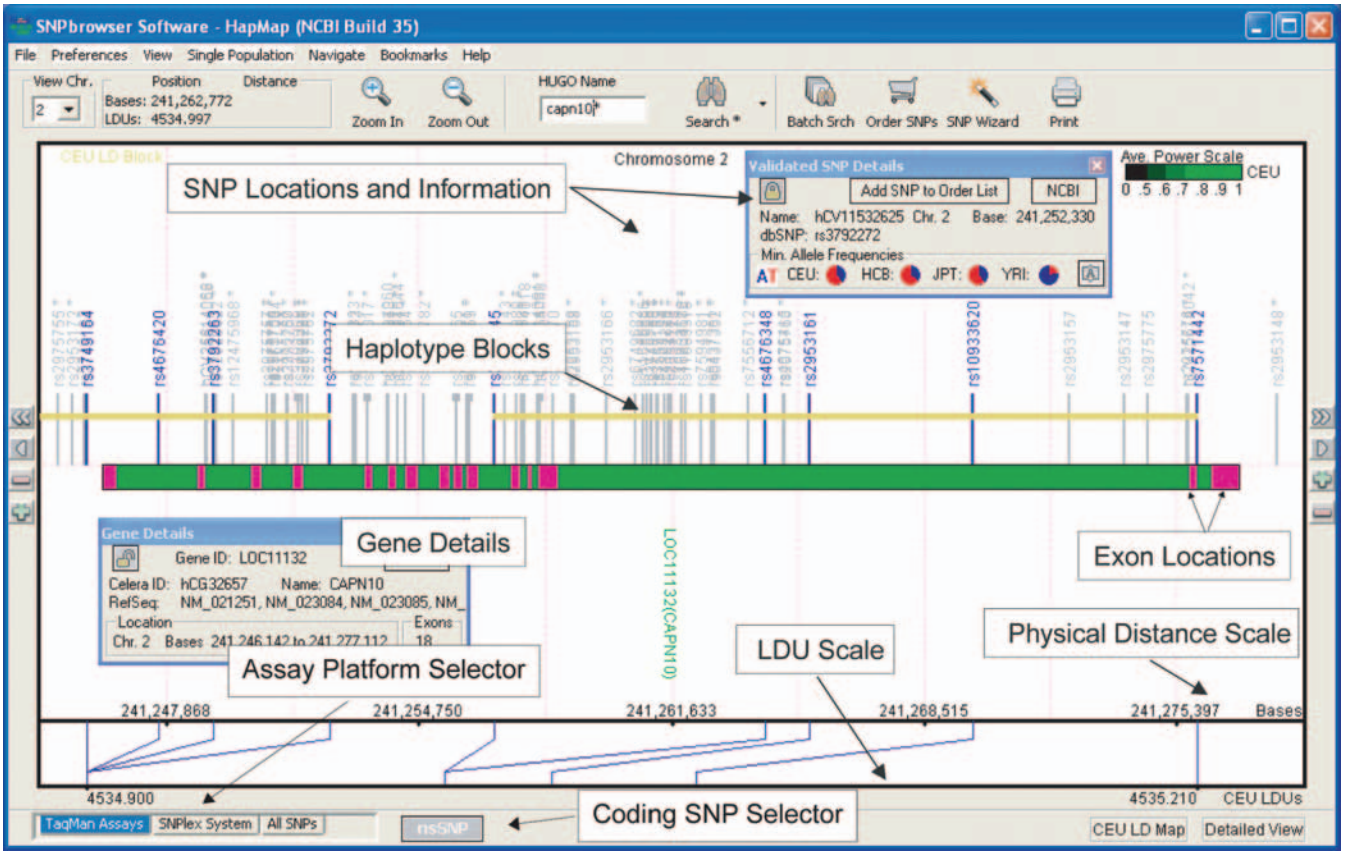


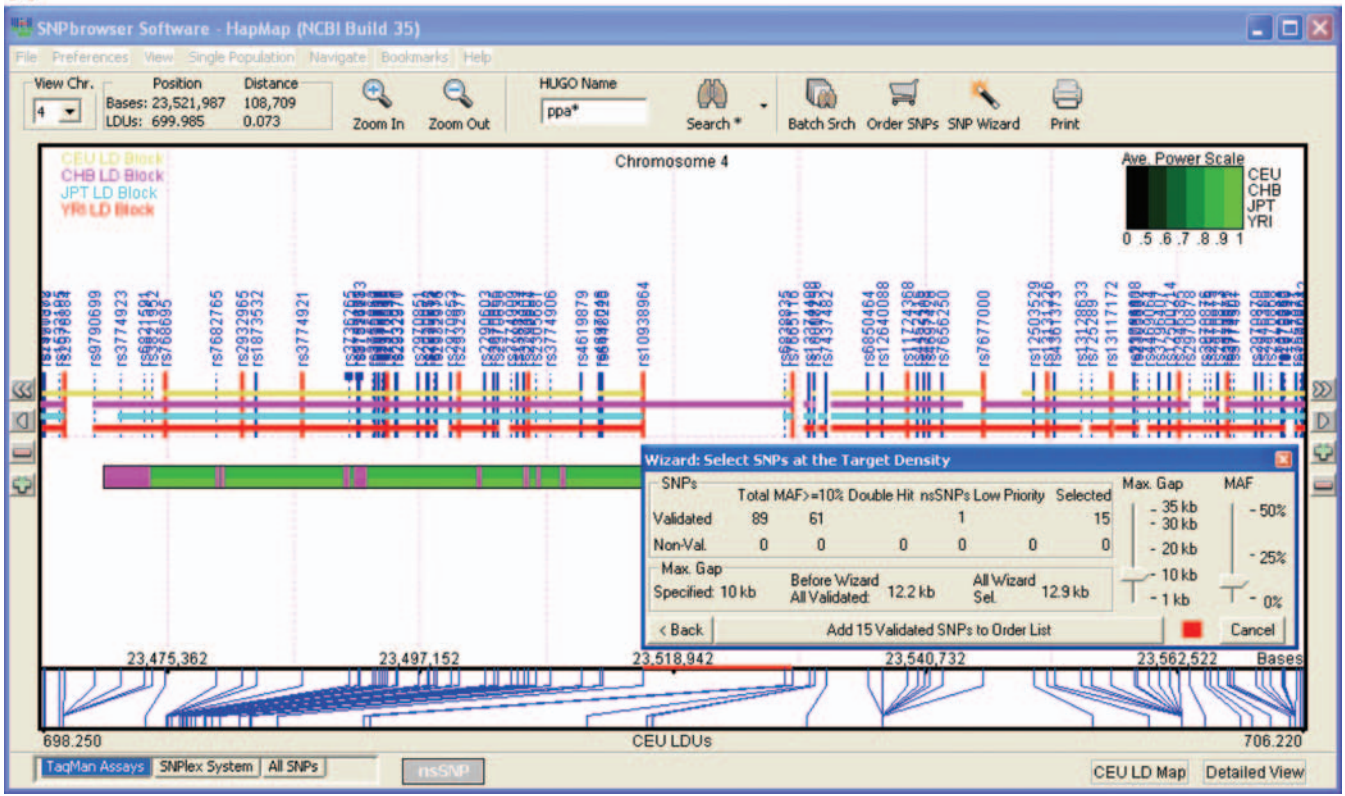

Fig. 2. The SNPbrowser software allows visualization of extensive gene and genomic information, including the physical and linkage disequilibrium maps, intron/exon structure, the locations and allele frequencies of single-nucleotide polymorphisms, and putative haplotype blocks in four different populations.

study and the goal is to perform fine mapping of an implicated chromosomal region to find the disease gene. For example, choosing a region and searching for validated SNPs spaced a desired length in kilobases across the region, the SNPbrowser software identifies appropriate SNPs and indicates if it is possible to achieve the desired spacing across the entire region (**Fig. 3A**). If validated SNPs had not been available, a red indicator bar would replace the green indicator bar in the bottom right-hand corner of the read-out window. The slider in the wizard allows researchers to modify the spacing or MAF parameters to quickly visualize the level of coverage that is possible in the region given their other requirements. Alternatively, SNPs can be selected to try to achieve an even spacing of a given number of LDUs on the metric LD map by simply going back and changing the density parameters (**Fig. 3B**). If extensive LD is present in the region, the wizard will select at least one SNP, although there is might be intervals where the target LDU distance is not achievable (red bar), suggesting the presence of recombination hot spots. If desirable, the user can select the wizard to fill this "gap" with as many nonvalidated SNPs (gray vertical lines) as required. All this process can be carried out in seconds. Because the selection process is carried out selecting a particular genotyping platform (selected by the platform tabs), additional time is saved by not having to go back and refill gaps created by SNPs that cannot be converted to a given assay format. Furthermore, the SNPbrowser wizard can also take into account SNPs for which the user already developed assays and fill gaps around them.

### 3.4. Selection of Tagging SNPs

Because SNPs that show high LD result in chromosomal segments in which a limited number of haplotypes are found in a population (i.e., a haplotype block ***[4]***), it is possible to select a small subset of SNPs that distinguish, or "tag," the common haplotypes previously found in a gene or region. This eliminates a large number of SNPs from the study that would only provide redundant information. In principle, this reduction in markers brings down the cost and time necessary to conduct a study retaining good statistical power ***(23)***. Although in some instances capturing haplotype diversity could be the desired goal, the selection of nonredundant SNP subsets using other LD metrics that have fewer assumptions, e.g., pairwise $r^2$ ***(7)*** and simple genotype correlations, could be preferable in other cases. All calculations are carried out by the software "on-the-fly" and almost instantaneously for most regions (up to several seconds for whole chromosomes), by directly accessing the genotype database maintained in memory, and without needing to download data or maintain an internet connection during use. This allows the easy comparison of the effect of different parameters and filters in the final selection of SNPs. Finally, the Tagging wizard can also export the values of the relationship between tag and tagged SNPs with the selected metric

A


SNPbrowser Software - HapMap (NCBI Build 35)
File Preferences View Single Population Navigate Bookmarks Help
View Chr. 4
Position Bases: 23,521,987 LDUs: 699.985
Distance 108,709 0.073
Zoom In
Zoom Out
HUGO Name ppa*
Search *
Batch Srch
Order SNPs
SNP Wizard
Print
CEU LD Block
CHB LD Block
JPT LD Block
YRI LD Block
Chromosome 4
Ave. Power Scale
CEU
CHB
JPT
YRI
0 .5 .6 .7 .8 .9 1
Wizard: Select SNPs at the Target Density
SNPs
Total MAF>=10% Double Hit nsSNPs Low Priority Selected
Validated 89 61 1 15
Non-Val. 0 0 0 0 0 0
Max. Gap
Specified: 10 kb
Before Wizard All Validated: 12.2 kb
All Wizard Sel. 12.9 kb
Max. Gap
35 kb
30 kb
20 kb
10 kb
1 kb
MAF
50%
25%
0%
< Back
Add 15 Validated SNPs to Order List
Cancel
23,475,362
23,497,152
23,518,942
23,540,732
23,562,522
Bases
698.250
CEU LDUs
706.220
TaqMan Assays
SNPlex System
All SNPs
nsSNP
CEU LD Map
Detailed View

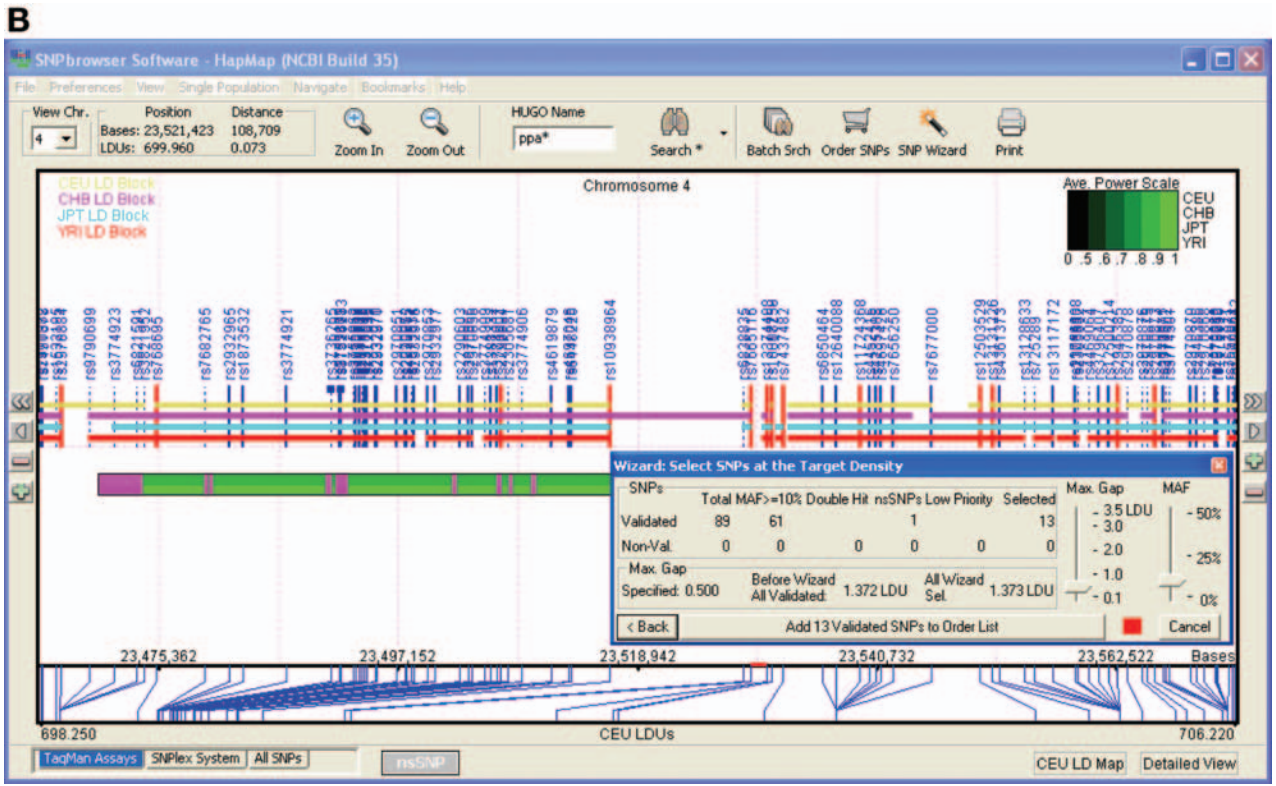



Fig. 3. The single-nucleotide polymorphism (SNP) selection wizard in density mode allows researchers to find SNPs at even linkage disequilibrium units (LDU) or kilobase intervals. **(A)** Validated SNPs in the CEU population with minor allele frequencies of at least 10% allowing spacing no larger than 20 kb. **(B)** The same region requiring SNP selection with spacing of 0.5 LDU. The selected SNPs are shown as vertical red lines.

for further use during analysis. Optionally, the correlation between tagSNPs and other HapMap-validated SNPs that cannot be converted to assays can also be exported for further consideration during analysis.

### 3.5. Selecting Coding SNPs

SNPbrowser software also makes it easy to include putatively functional coding SNPs (cSNPs) in association studies. SNPs that result in nonsynonymous codon changes and consequently, amino-acid substitutions (or premature stop codons) in the gene's protein product that can potentially affect its function *(25)*, also referred to as nonsynonymous cSNPs (nsSNPs). By simply clicking on the "nsSNP" button only these types of variants are visualized. If cSNPs are the study focus, it is possible to limit the search at two points. First, the Density Wizard includes a checkbox to make selecting cSNPs the search priority. Second, the "shopping basket" has one-click functionality that will add only the cSNPs to the cart.

### 3.6. Implementing the Study

Selected SNPs can be added to a working list of markers (or "shopping basket") by either simply clicking on the results bar of the SNP wizards, manually adding individual markers with the right-click option, or by invoking the "shopping basket" window and adding markers from the current view in many forms. There are two separate shopping baskets, one for each TaqMan and SNPlex platforms. In the case of TaqMan, assay availability and previous performance validation is indicated. The contents of each basket can be exported and saved for use in a future session. Finally, once the researcher has identified the ideal set of SNPs for an association study, genotyping reagents can easily be obtained. The user can also export the list of SNPs from the shopping basket to a text file, including a number of the annotations maintained in the software internal database (*see* **Fig. 4**).

## 4. Discussion

Because there is no single SNP selection approach that can serve all the requirements of different types of studies, the SNPbrowser software offers researchers a choice of methods for picking markers suited to a wide range of objectives and disease characteristics. Two basic paradigms for selecting SNP markers are supported: (1) selection of evenly spaced markers on the physical or metric LD maps *(20)*, and (2) selection of nonredundant subsets of haplotype "tagging" SNPs *(22)*. Furthermore, the tool immediately indicates the SNPs for which genotyping assays are viable and available from commercial sources. This means that researchers can get promptly started in their study after identifying an optimal SNP set. Previously, identifying the most efficient and highly

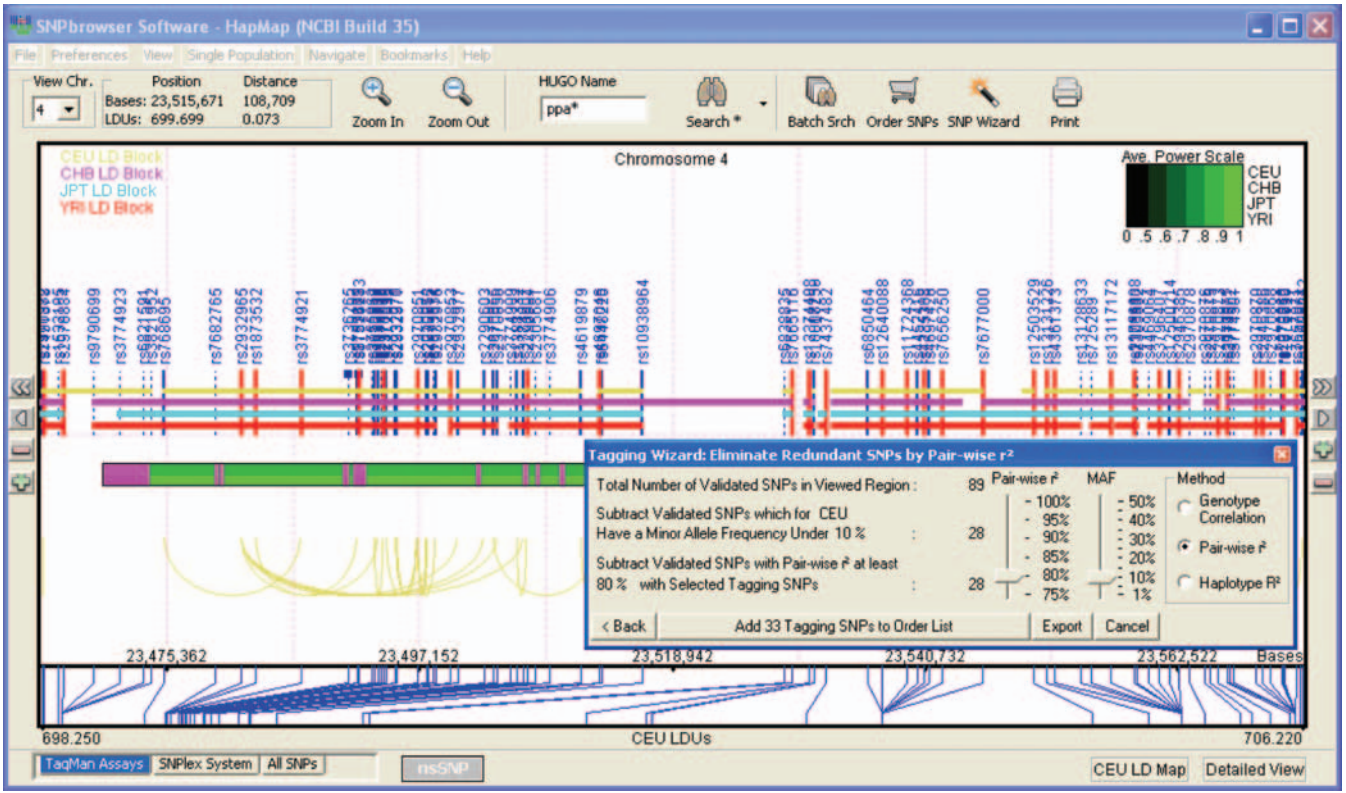



Fig. 4. The single-nucleotide polymorphim (SNP)-tagging wizard allows the selection of minimum subsets of SNPs (shown in red) retaining a threshold value of the selected quality metric, in this case pairwise $r^2$. Yellow arcs show the relationship between tag and tagged SNPs, which can also be exported as a file with the numerical values of the metrics.

informative SNP set for a multimegabase region (e.g., a candidate region from a previous linkage study performed with microsatellites) was extremely time-consuming. With the SNPbrowser wizards it only takes a few seconds, for example, to get a list of evenly spaced, highly-informative SNPs across the region of interest either on the physical (kilobase) or metric LD (LDU) maps. A metric LD map, expressed in LDUs, calculated by the LDMAP software *(20)*, places SNPs on a coordinate system where distances between SNPs are additive and directly related to the degree of LD between them. For example, SNPs in perfect LD (completely correlated) have zero distance between them, whereas SNPs with no significant correlation are separated by over three LDUs in this map. Analogous to the genetic map expressed in centimorgans commonly used for selecting markers for linkage studies in families, the LD map can be used to efficiently position markers for population-based disease association studies *(24)*.

Normally, the HapMap database would be preferred because of the depth of coverage, but often the AB maps could be useful, for example, if the study involves African-Americans (the HapMap Project did not genotype samples of this population). Although it is always preferable to utilize validated SNPs when designing genetic studies, there may be circumstances when it would be desirable to include SNPs present in the public databases but that have not been validated, e.g., by the HapMap Project. SNPbrowser allows displaying the complete SNP complement for the visible region that can be converted to commercially available genotyping assays, whether validated or not, making it easy to select additional SNPs that can be used to fill gaps left by the validation projects.

Sometimes nsSNPs are included because they are referenced in the literature, and other times adding nsSNPs to the study may increase its power because in some instances an nsSNP can be indeed a causative variant for the phenotype under investigation. It is important to note that noncoding SNPs, such as those in regulatory regions or splice junctions, can also influence the trait of interest and thus cannot be completely ignored, but these are difficult to identify or predict. Further, if their penetrance is high, cSNPs may not occur in sufficient frequency in the population to be informative in a study with a typical sample size. Ultimately, most researchers find that it is most productive to include a mix of nsSNPs and surrogate marker SNPs with high MAFs.

In summary, SNPbrowser is a free tool that allows researchers to easily select SNPs for genetic association or other types of studies involving human SNPs. Its main advantages include ease of use, swift interaction and searches, informative visualization, intuitive wizards that automate the most common selection workflows, no need to be online to access the data, completeness in terms of data, and selection algorithms enabling rapid experimental cycles by considering an assay platform conversion potential from the beginning. The software

also includes extensive online help describing in detail additional features and facilities that owing to length limitations cannot be discussed in this chapter. The extensive and detailed information available through the SNPbrowser software solves many of the major challenges that researchers face when designing human-association studies, including visualizing complete genomic information in their region or gene of interest, leveraging the extensive reference genotype datasets becoming available from the HapMap Project, identifying the best set of SNPs for their studies, and easily obtaining reliable assays that correspond to those SNPs.

## Acknowledgments

I am very grateful to Hadar Isaac (Imagenix Corp, Los Altos, CA), who carried out all programming, algorithm development, and software design for the SNP-browser software, Charles Scafe (Applied Biosystems), who provided the data and performed computations, Andrew Collins (University of Southampton, Southampton, UK), for providing the LDMAP software, Bjarni Halldórsson and Ross Lippert (formerly at AB), who provided tagging SNP selection pipeline and haplotype-phasing code, and Derek Gordon (Rutgers University, Piscataway, NJ), who provided power calculation methods. I also acknowledge the valuable support and feedback provided by Pius Brzoska, Joanna Curlee, Dennis Gilbert, Toinette Hartshorne, Fiona Hyland, Michael Rhodes, Katherine Rogers, Leila Smith, Eugene Spier, Rob Tarbox, and Trevor Woodage (all from AB) during the development of SNPbrowser.

## References

1 Reich, D. E., Cargill, M., Bolk, S., et al. (2001) Linkage disequilibrium in the human genome. *Nature* **411,** 199–204.

2 De La Vega, F. M., Isaac, H., Collins, A., et al. (2005) The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. *Genome Res*. **15,** 454–462.

3 Consortium, T. I. H. (2005) A haplotype map of the human genome. *Nature* **437,** 1299–1320.

4 Gabriel, S. B., Schaffner, S. F., Nguyen, H., et al. (2002) The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229.

5 Avi-Itzhak, H. I., Su, X., and De La Vega, F. M. (2003) Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity. In: *Pacific Symposium on Biocomputing*, (Altman, R. B., et al., eds.), World Scientific Press, Lihue, Hawaii, pp. 466–477.

6 Byng, M. C., Whittaker, J. C., Cuthbert, A. P., Mathew, C. G., and Lewis, C. M. (2003) SNP subset selection for genetic association studies. *Ann. Hum. Genet*. **67,** 543–556.

7 Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi,Q., Kruglyak, L., and Nickerson, D. A. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms

for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74,** 106–120.

8 Hampe, J., Schreiber, S., and Krawczak, M. (2003) Entropy-based SNP selection for genetic association studies. *Hum. Genet.* **114,** 36–43.

9 Johnson, G. C., Esposito, L., Barratt, B. J., et al. (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29,** 233–237.

10 Ke, X. and Cardon, L. R. (2003) Efficient selective screening of haplotype tag SNPs. *Bioinformatics* **19,** 287–288.

11 Sebastiani, P., Lazarus, R., Weiss, S. T., Kunkel, L. M., Kohane, I. S., and Ramoni, M. F. (2003) Minimal haplotype tagging. *Proc. Natl. Acad. Sci. USA* **100,** 9900–9905.

12 Stram, D. O., Haiman, C. A., Hischhorn, J. N., et al. (2003) Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.* **55,** 27–36.

13 Weale, M. E., Depondt, C., Macdonald, S. J., et al. (2003) Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.* **73,** 551–565.

14 Zhang, K., Sun, F., Waterman, M. S., and Chen, T. (2003) Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. *Am. J. Hum. Genet.* **73,** 63–73.

15 Horne, B. D. and Camp, N. J. (2004) Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genet. Epidemiol.* **26,** 11–21.

16 Hu, X., Schrodi, S. J., Ross, D. A., and Cargill, M. (2004) Selecting tagging SNPs for association studies using power calculations from genotype data. *Hum. Hered.* **57,** 156–70.

17 Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., and Hirschhorn, J. N. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* **33,** 177–182.

18 De La Vega, F. M., Dailey, D., Ziegle, J., Williams, J., Madden, D., and Gilbert, D. A. (2002) New generation pharmacogenomic tools: a SNP linkage disequilibrium map, validated SNP assay resource, and high-throughput instrumentation system for large-scale genetic studies. *Biotechniques* **Suppl,** 48–50.

19 De la Vega, F. M., Lazaruk, K. D., Rhodes, M. D., and Wenz, M. H. (2005) Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System. *Mutat. Res.* **573,** 111–135.

20 Maniatis, N., Collins, A., Xu, C. F., et al. (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl. Acad. Sci. USA* **99,** 2228–2233.

21 Schwartz, R., Halldorsson, B. V., Bafna, V., Clark, A. G., and Istrail, S. (2003) Robustness of inference of haplotype block structure. *J. Comput. Biol.* **10,** 13–19.

22 Halldorsson, B. V., Istrail, S., and De La Vega, F. M. (2004) Optimal selection of SNP markers for disease association studies. *Hum. Hered.* **58,** 190–202.

23 De La Vega, F. M., Gordon, D., Su, X., et al. (2005) Gene-centric power and sample size calculations for genetic case/control studies using empirical genotype data from dense SNP maps. *Hum. Hered.* **60,** 46–60.

24 Collins, A., Lau, W., and De La Vega, F. M. (2004) Mapping genes for common diseases: the case for genetic (LD) maps. *Hum. Hered.* **58,** 2–9.

25 Thomas, P. D. and Kejariwal, A. (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc. Natl. Acad. Sci. USA* **101,** 15,398–15,403.

# 14

# Avoiding False Discoveries in Association Studies

**Chiara Sabatti**

**Summary**

We consider the problem of controlling false discoveries in association studies. We assume that the design of the study is adequate so that the "false discoveries" are potentially only because of random chance, not to confounding or other flaws. Under this premise, we review the statistical framework for hypothesis testing and correction for multiple comparisons. We consider in detail the currently accepted strategies in linkage analysis. We then examine the underlying similarities and differences between linkage and association studies and document some of the most recent methodological developments for association mapping.

**Key Words:** Type-I error; global error; FWER; FDR; power; permutations.

## 1. False Discoveries

The expression "false discovery" has an easy scientific interpretation: in the context of gene mapping, we have a false discovery when we erroneously conclude that a genomic region harbors a susceptibility gene for a disease or a gene contributing to a quantitative trait. Clearly, one wants to avoid such false discoveries, mainly because they induce the scientific community to focus on the wrong leads, diverting time and resources from more useful avenues. Additionally, the people directly involved in the study ultimately do not benefit from false discoveries; they often end up spending a large portion of their career unsuccessfully trying to finalize the discovery and their reputation in the community is at stake. In the short run, however, the "result-oriented" mode of research financing we live in, all too often makes it appealing to publish "discoveries" that are not quite such. This is a serious problem, and the recent history is full of examples where a novel technology or investigation approach has been initially applied with the all too liberal interpretation of what constitutes

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

a discovery. After a rather short while, the scientific community inevitably agrees that it is necessary to "raise the bar" for significance and a more stringent system to protect against false discoveries is put into place. Association studies, despite having been discussed in the literature for at least a decade *(1)* are still in their infancy, as only in 2005 the genotyping technology that will make them economically feasible has become available. Hence, the scientific community involved in such studies has not yet reached that mature stage where a well-established protocol to avoid excessive false discoveries is in place. In this chapter, we will define the problem and describe the available instruments to address it. After refreshing some statistical concepts and terminology, we will review the strategies for controlling false discoveries that have proven useful for linkage mapping. Although the conceptual framework for controlling false discoveries in linkage and association studies is quite similar, there are specific characteristics of association mapping that are worth considering. After having underlined these, we will review some of the approaches most recently proposed for the control of false discoveries in association studies.

## 2. Statistical Tests of Hypotheses and Error Types

The scientifically evocative term "discovery" corresponds, in statistical terminology, to a rejection of the null hypothesis. Although undoubtedly more prosaic, "rejection" relates more accurately to the actual practice in scientific research; rather than proving any connection between one genomic region and a disease, for example, we are simply able to deem as inadequate the default hypothesis that there is no connection between the two. There are a number of very good reasons why scientific discovery relies on testing null hypotheses, which need not be discussed here *(2,3)*. Given that this is the framework, it is useful to recall the statistical approach to hypothesis testing in some detail. Given a null hypothesis $H_0$, and a decision rule that tells when to reject it or not, there are four possible situations, exemplified in the table below:

| | Not reject | Reject |
|---|---|---|
| $H_0$ true | – | Type-1 error |
| $H_0$ false | Type-2 error | – |

Ideally one would like to minimize both error types, but, for any given procedure, reducing one increases the other. Two solutions to this impasse are common practice. Stressing that the null hypothesis, $H_0$, represents the status quo—which should be questioned only when strictly necessary—frequentist procedures fix an acceptable probability for type-1 error (called "level" of the test, and often indicated with $\alpha$) and then search, among all tests that guarantee such a level, the one that minimizes the type-2 error (or maximizes power).

The term "*p*-value" is used to indicate the minimum level of the test such that the observed data would justify rejection of the null hypothesis. Bayesian procedures, instead, minimize an average of the probabilities of the two errors, where the average is obtained using a posterior distribution on the veridicity of $H_0$. In many cases, the shape of the region of data values that would result in the rejection of the null hypothesis is identical in Bayesian and frequentist procedures, and the Bayesian approach can be seen as a rational recipe to define what is an appropriate level for the test. Indeed, the choice of an $\alpha$ value can be rather disconcerting. Introductory statistics textbooks will often suggest a value of 0.05 (one type-1 error accepted every 20 tests) or 0.01 (one type-1 error accepted every 100 tests), but an appropriate choice of the level really depends on the "cost" for the scientist to erroneously reject the null hypothesis. Consider the costs associated with the four possible cases described before:

| | Not reject | Reject |
|---|---|---|
| $H_0$ true | – | $C_1$ |
| $H_0$ false | $C_2$ | – |

Then the Bayesian approach suggests making the decision that minimizes the posterior expected costs, that is, one would reject $H_0$ when $C_1 Pr(H_0|Data) < C_2 Pr(H_0 false|Data)$, that is, when $Pr(H_0|Data) < C_2/(C_1+C_2)$, or indicating with $\pi_0$ the prior probability of $H_0$ and with $f(Data|H_0)$ the probability of the data under the hypothesis $H_0$, $H_0$ is rejected if

$$\frac{f(\text{Data}\,|\bar{H}_0)}{f(\text{Data}\,|H_0)} > \frac{C_2\pi_0}{C_1(1-\pi_0)} \tag{1}$$

The left-hand side of **Eq. 1** is the likelihood ratio, used also in frequentists tests: in that context, the cut-off value is determined as the $(1 - \alpha)$ quantile of the distribution of the likelihood ratio under the null hypothesis, where $\alpha$ is the desired level of the test. Clearly then, specifying an $\alpha$ level can be put in correspondence with a specific choice of costs and prior distributions of the hypotheses. We will see how this analogy has been useful in deciding appropriate significance levels for the gene-mapping problem.

When multiple tests are conducted at the same time, other issues arise in the choice of an appropriate significance cutoff. One quickly realizes that there is another level of complexity, by looking at the estimated number of false-positives. With a level $\alpha = 0.05$, we expect to make a false rejection every 20 tests. This means that if we actually carry out 1000 tests, we expect to erroneously reject 50 null hypotheses. This is easily too high a number of false discoveries for scientific practice. In order to look at the problem with more generality, let us

introduce some notation. Let $m$ be the total number of tests, with $T_1, \ldots, T_m$ a set of test statistics for testing the hypotheses $\{H_1, \ldots, H_m\}$. Let $H_0$ be the hypothesis that corresponds to each of the $H_i$ being true $H_0 = \bigcap_{i=1}^{m} H_i$. When we conduct tests of these hypotheses, often we are trying to answer two types of questions: (1) can $H_0$ be rejected? (2) If $H_0$ is rejected, which of the $H_i$ should be rejected? The statistical technique used to answer the first question is called a global test, although the one addressing the second is called a multiple test procedure. Both (1) and (2) are relevant for gene mapping. Suppose we are conducting a genome screen, that is, we are testing linkage/association between a disease and a number of markers supposed to cover the entire genome. Question (1) depends on the genetic base of the disease: can we, on the basis of the data collected, exclude the hypothesis that the disease has no genetic basis (at least among the genomic regions reasonably covered in our investigation)? Question (2) pertains to the identification of genomic regions that are involved with the disease.

The case of a global test is conceptually easily dealt with. One can use the same criteria that we described previously for one test, and just apply them to the test of the global null $H_0$. The catch here is that, although it is typically relatively easy to identify the rejection region corresponding to a given level for a test based on one statistic $T_i$ only, identifying the appropriate rejection region for $H_0$ (which needs to involve all the $T_1, \ldots, T_m$) may be difficult.

Answering the question presented by multiple comparisons, requires the definition of a new error measure. Let us consider the set of $m$ tested hypotheses. With respect to the true nature of the hypotheses and the tests results, they can be organized in the following table.

| | No. non rejections | No. rejections | |
|---|---|---|---|
| # true null | $U$ | $V$ | $m_0$ |
| # false null | $T$ | $S$ | $m_1$ |
| | $m - R$ | $R$ | $m$ |

Given this setting, we want to define a measure of global error that describes how well we do overall and then devise methodologies to control such error. A notion of global error that has been used for a long time in the statistics and genetics literature is the family-wise error rate (FWER), defined as the probability of committing at least one false rejection: $Pr(V > 0)$. This is quite a stringent criterion, which is typically appropriate when the costs of following up a wrong lead are really high (which is the case, for example, in pharmaceutical industries). Although there are a variety of other definitions of global errors, the other one that is nowadays also widely used is the false discovery rate (FDR), introduced in the mid-1990s by Benjamini and Hotchberg ***(4)***. The FDR is the expected fraction of mistakes among the rejected hypotheses $E(Q)$:

$$Q = \begin{cases} \dfrac{V}{V+S} & \text{if } V+S>0 \\ 0 & \text{otherwise} \end{cases}$$

Defined as an error rate, the solution to the multiple testing problem depends on finding a procedure that will control it. The well-known Bonferroni procedure, that suggests testing each hypothesis at the level $\alpha/m$ controls FWER at the level $\alpha$. When one expects that more than one hypothesis is false, stepwise procedures increase the power of detection. Such procedures are applied to a set of ordered $p$-values $p_{(1)} < p_{(2)} < \ldots < p_{(m)}$. One such procedure that controls FWER is the following ***(5)***:

**(Holm)** Start with $i = 1$. If $p_{(i)} > \alpha/(m - i + 1)$ accept $H_{(i)}, \ldots, H_{(m)}$ and stop. Otherwise, reject $H_{(i)}$ and continue.

As can be easily seen, the first cut-off value of this rule is the same as the one suggested by Bonferroni. It is only following rejections that the criterion becomes less stringent. It is also clear that it will take a sizeable amount of rejections for the cut-off value to substantially decrease. Another issue with the Bonferroni procedure is that it does not take into account the dependency between tests. Indeed, if we were to conduct $m$ identical tests to control FWER at the level $\alpha$ one would need to simply use level $\alpha$ for each single test. In general, the presence of dependence between the test statistics makes it unnecessary to opt for such a strong correction as Bonferroni. The appropriate cutoff can be sometimes determined analytically when the joint distribution of the test statistics is known. Otherwise, resampling procedures can be often successfully used (*see* the interesting book by Westfall and Young, ***[5]***). Examples of both these strategies can be found in mapping literature and we will come back to them in subsequent paragraphs. The following procedure ***(4)*** controls the FDR:

**(BH)** Proceed from $i = m$ to $i = m$-1 etcetera, until, for the first time, $p_{(i)} \leq i\alpha/(p_0 m)$. Where $p_0 = m_0/m$. If this quantity is unknown, set it to 1. Denote that $i$ by $k$ and reject all $H_{(i)}$ with $i = 1, \ldots, k$.

This step-down rule was proposed by Benjamini and Hochberg in 1995. At the time the authors proved that it controlled FDR for independent tests. Subsequent work by Benjamini and Yekutieli ***(6)***, Genovese and Wasserman ***(7)***, and Storey and Tibshirani ***(8)*** showed that this is true also for a quite general class of dependent distributions. Benjamini and Yekutieli ***(6)*** also proposed another procedure that compares $p_{(i)}$ with $i\alpha/(m\sum_{j=1}^{m} 1/j)$; this is guaranteed to strongly control FDR for any type of dependence, even if it may lead to a significant loss of power. Before concluding this overview of the statistical methodologies for control of false discoveries, we introduce the notion of adjusted $p$-value. These correspond to one specified notion of global error and indicate,

for each hypothesis, the minimal level of such error necessary to reject that hypothesis. These are a quite useful statistics to report as they enable other future researchers to make informed decision on which effects should be considered significant. The two described step-wise procedures lead also to the definition of corresponding adjusted $p$-values $\tilde{p}_{(i)}$:

$$\textbf{(Holm)}\tilde{p}_i^H = \max_{k=1,\ldots,i}\{\min((m-i+1)p_{(k)}, 1)\}$$
$$\textbf{(BH)}\tilde{p}_i^{BH} = \min_{k=i,\ldots,m}\{\min(m/kp_{(k)}, 1)\}.$$

## 3. The Case of Linkage Mapping

In order to understand the parameters of our problem, and what impact the statistical methodologies described previously have in gene mapping, it is useful to review the approaches to establish significance cut-off values in linkage analysis for monogenic diseases, which has been well studied. It is particularly interesting to realize how the criteria to assess significance have evolved with the genotyping technology and the resulting types of datasets available for mapping. The first linkage studies were conducted in a context where very few markers were available and genotyping represented a substantial cost. With this background, LOD score greater than three became an accepted cut-off for significance. With a LOD score, one indicates the logarithm base 10 of the likelihood ratio introduced in **Eq. 1.** With the definition of a genetic hypothesis in mind, one has

$$\text{LOD} = \log_{10} \frac{Pr(\text{Data} \mid \text{Linkage})}{Pr(\text{Data} \mid \text{non Linkage})} \tag{2}$$

Considering the distribution of LOD scores under the null hypothesis of no linkage, a cutoff of 3 corresponds roughly to a significance level of $\alpha = 0.0001$. Why was such a high threshold adopted?

Initially, Morton *(9)*, with the idea of minimizing the costs of family collection and genotyping, proposed a sequential procedure for sampling and analyzing pedigrees until the evidence in favor of linkage with a marker was met. To guarantee against biases introduced by the sequential sampling procedure, he proved that such evidence was to be considered sufficient only when the LOD score reached three. Subsequently, Morton and others *(9,10)* used Bayesian arguments to show how, even without adopting a sequential procedure, it was necessary to obtain such strong evidence to establish linkage. The crux of the argument was that, given the availability of only a few markers, there was a very small prior probability that one of these markers was actually linked to the disease gene of interest. Based on genome length and the distance between two loci over which one could detect linkage, one can show that the prior probability of linkage between a given disease locus and a random location in the genome is 0.02. Going back to the formulation of the Bayesian test given in **Eq. 1**, if our

cost structure is such that we want to reject the null hypothesis of no linkage only when the posterior probability is less than 0.05 and we have a prior probability on $H_0$ of 0.02, we find that the likelihood ratio has to be, approximately, greater than 1000, corresponding to a LOD score of 3.

In its original formulation, then, the stringent threshold for the LOD score was dictated by the necessity of accounting for too little searching, both in the number of pedigrees and markers investigated. As the number of markers available for analysis increased substantially, the perspective changed. Because, in current genome screens, a considerable number of markers are typed across the genome, the prior probability that at least one of them is linked to the disease is high. However, conducting multiple tests makes it necessary to worry about multiple comparisons. Even if the disease was not genetic and all the null hypotheses of no linkage to each of the considered markers were true, the simple fact of looking at many markers increases the probability of finding one that shows a significant pattern. To address this issue, two simplifying hypotheses on the marker structure have been made, corresponding to the idea of a "sparse map" and "dense map" ***(11)***. In the first case, markers are assumed to be independent, in the second case, markers are assumed to cover the entire genome so that the LOD scores observed are actually continuous processes. In both of these frameworks, LOD-score values of 3–3.5 appear to produce good evidence for linkage. In the sparse-map assumption, consider a genome screen with 400 markers, which is a fairly typical one. Then one can apply a Bonferroni correction and find that to have an overall level of significance of 0.05, one needs an individual $p$-value lower than 0.001, corresponding to a LOD score of 3.3. This approximation is not valid as we further increase the number of markers, as the assumption of independence between the markers becomes increasingly unrealistic and would lead to an unnecessary loss of power. In this context, the dense-map approximation is useful as the tests for linkage can be shown to follow an Ornstein–Uhlenbeck process and extreme probabilities from this process can used to define the level of significance ***(12)***, and obtain results comparable to the 3.3 cut-off value.

A further issue that has been introduced over time in the discussion of the appropriate cut-off values for a mapping study is the increased interest in complex diseases, where more than one locus is expected to be involved. In the context of linkage studies ***(11,13)***, comparisons have been made of the performances of marginal search (focusing on one locus at the time), and simultaneous and conditional search that are carried out under the explicit assumption that more than one locus should be involved. From a practical standpoint, the conclusion of these studies has been that even if conditional and simultaneous searches are potentially more powerful, they require such high levels of correction for multiple comparisons that they are often not worth pursuing. Certainly marginal search plays the leading role in practical applications.

In parallel to the results described previously that rely on approximations of the joint distributions of test statistics with a continuous process, permutation and resampling methodologies have been used to determine the empirically adjusted *p*-value corresponding to different tests. The work on quantitative trait locus (QTLs) ***(14)*** is perhaps the first such contribution, but methodology along these lines has been developed and applied to binary traits as disease status also.

Besides all the theoretical approaches described previously, it is worth mentioning that the genetics community now has a substantial experience with linkage studies and it has clearly been proven that the stringent cutoffs described are necessary to avoid false-positives and lead to reliable results in the case of Mendelian diseases. It is interesting to note that the crucial parameter does not seem to be the number of test actually conducted, but rather the total number of independent tests that one could potentially carry out. It is now time to consider how similar considerations apply to the case of association mapping.

## 4. Association Mapping

Moving now to the consideration of association mapping, there are a few points that one can make quite easily that underscore similarity/differences with the context of linkage maps.

### *4.1. Marker Availability and Genotyping Costs*

At the time of writing, at the end of 2005, a number of companies have made commercially available technologies that allow genotyping of hundreds of thousands of single-nucleotide polymorphisms (SNPs) of known genomic position at reasonable cost. A number of genome-wide association studies are ongoing, taking advantage of this genotyping technology. Despite the availability of a large number of markers, however, not all the association studies are in the category of genome-wide screens. Indeed, by and large, the majority of the studies carried out thus far have been based on the candidate gene approach, where only a very limited subset of the genome is investigated for association with the trait under study. Hence, the panorama of gene-mapping studies still include both the paradigms delineated in the previous section, that is, the case where the domain of investigation is so limited that one should really consider the probability of the susceptibility gene being contained in the analyzed region as being quite small; and the case where one, instead, needs to worry about multiple testing, given the large number of markers analyzed.

### *4.2. Candidate Genes and Prior Probabilities*

For any given disorder, the list of candidate genes is compiled on the basis of current knowledge. It typically includes genes previously implicated on related phenotypes and genes of known function involved in biological

processes that the investigators judge as related to the disorder. Despite all "good intentions," the quality and comprehensiveness of such lists vary considerably, and quantifying the probability that they are inclusive of the real susceptibility gene is often not possible. On the one hand, the criteria to identify a gene "previously implicated on related phenotypes" are quite loose. The strength of the noted linkage or association signal is often far from conclusive, so it is impossible to really distinguish the result from a false-positive, and the definition of "related" phenotype can be stretched to a degree that is hard to justify. On the other hand, one has to acknowledge that the most immediate goal of gene mapping is to unravel novel biological pathways determining diseases, so that guessing which processes are likely to be faulty in patients cannot be considered in any way exhaustive, even when based on state-of-the-art knowledge. Given these limitations, the generic candidate gene study is much closer than researchers usually admit to the context of the few available markers for which the LOD score cutoff of three was established. In other words, the prior probability that the true susceptibility gene is included in the list should be considered very small and appropriately high significance thresholds are therefore required. The common practice of simply correcting for multiple comparisons on the basis of the tests actually conducted is not acceptable. Clearly there may be cases where the candidate gene list is very carefully constructed and very strong. Think, for example, of all the genes identified in a region that showed an unquestionable linkage signal for precisely the phenotype under study. It is, however, a burden of the researcher to show this to be the case, and the simple creation of a sensible candidate gene list does not *de facto* satisfy these requirements (for a detailed discussion of this issue, *see* **refs. *15*** and ***16***).

### *4.3. Sparse or Continuous Map*

In order to maximize the power to detect association with a disease, genome-wide screens are planned and conducted using a considerable number of markers. So between the continuous and the sparse-map assumptions, it seems that the first is definitely preferred. However, a couple of remarks are in order, which bring us closer to the sparse-map situation. (1) The maximal distance between two markers that guarantees that the association tests at the two locations are dependent is much smaller than the corresponding distance for linkage tests. Indeed, the current fine-scale genotyping platforms are developed not with redundancy in mind, but with the goal of providing minimal coverage. Moreover, the relationship between distance and dependence between association tests is less consistent than the one with linkage tests, reflecting the fact that linkage disequilibrium (LD) is not in perfect correlation with the recombination fraction. (2) Despite the fact that there will be some dependency between tests at nearby markers, it has to be emphasized that this dependency is going

to be very local, and that, by and large, there will be independence between the vast majority of tests conducted. Both of these observations suggest that, even if the effective number of independent test to be used in correction for multiple comparisons may not be quite as large as the number of tests actually carried out, it is likely not to be much smaller in terms of order of magnitude (*see*, for example, the correction for multiple comparisons suggested in **ref. *1***).

### *4.4. Structure of Dependency Between Tests*

As mentioned in the previous paragraph, we do not have a model for dependency between association tests with markers at a given distance. Nor does it look like one will be available soon. Because LD depends on population history, allele frequency, and so on, on top of recombination, obtaining a precise model of the dependence between tests created by nearby markers is quite challenging. This results in the impossibility of coming up with an approximation of the O–U process that may be used with success to control global error. Mounting evidence on empirical levels of LD is being gathered, however, and it is possible that we may be able to use this effectively to achieve this goal in the near future.

### *4.5. Complex Diseases*

While devising methodologies for controlling false discoveries in association studies, it is important to bear in mind the characteristics of the phenotypes that we are trying to map. Typically, one resorts to association mapping not for simple/Mendelian/monogenic disorders, but for complex ones, where we expect multiple locations in the genome to be involved. Translated into the multiple test context, this means that we do expect more than one hypothesis to be false. It is then important, for whatever criteria of global error we choose to control, to select a procedure that maximizes the power of detection of the second, third, and so on most significant loci.

### *4.6. Follow-Up Costs*

Another consequence of the advances in technology and the abundance of genomic information gathered thus far is the decrease in follow-up costs for suggestive locations. With the human genome entirely sequenced and increasingly annotated, a vast amount of gene expression results are publicly available and the costs of resequencing are considerably reduced; therefore, at least some initial steps of follow-up can be currently carried out at manageable costs. This should reflect on the significance rules one chooses for identifying suggestive locations.

### *4.7. Association After Linkage*

A final point worth noting is that association studies are generally carried out after a number of attempts to map the phenotype of interest with linkage

methods with not completely satisfactory results. On the one hand, this observation immediately qualifies the problems one tackles with association as difficult. On the other hand, it underscores the fact that some information on the gene localization is available to researchers that embark on association studies. Although they may have not been successful in localizing all the underlying genes, linkage studies may have excluded regions of the genome or may have identified potential localization. This available prior information should be taken into account when devising association studies and when evaluating the significance of their results.

## 5. New Directions in the Control of False Discoveries for Association Mapping

In this last section, we will review some of the most recent contributions to the literature that specifically address the questions we have outlined so far.

### *5.1. Choosing an Appropriate Global Error*

It is becoming clearer that genome-wide studies have to be considered precisely as screening tools, rather than experiments that will immediately lead to the identification of the disease gene. Because of the complex nature of the disease, we expect that more than one locus in the genome may be implicated, so we are effectively interested in multiple hits. In this context, one wants to be able to follow all the good leads that may result in the identification of a disease locus. Although too many wrong clues are to be avoided as costly, what really matters is the proportion of these over the total number of clues that are warranted for further investigation.

With this background, a global error measure like the FDR is particularly interesting because of its increased power in circumstances where more than one of the null hypotheses is false. The step-down procedure that we have described in the previous section controls the FDR if the test statistics $T_1, \ldots, T_m$ are independent, as assumed in the sparse-map approximation. However, the data collected so far on the levels of background LD suggest that this is not a realistic hypothesis in the case of tests of association between a disease and a set of finely spaced markers. There are two implications from the departure from independence. The first is that the step-down procedure may not control the FDR for tests with generic dependence. The second is a consideration similar to that already discussed for the Bonferroni procedure: the cut-off value of $\alpha/m$ for $p_{(1)}$ may be excessively conservative if the tests are positively associated. We deal here with the first problem and refer to the next section for a discussion of the second.

Recent work of Benjamini and Yekutieli shows that under some forms of dependence (positive regression dependency on each one from a subset [PRDS]), the procedure described in **(BH)** controls FDR. If the test statistic

under association satisfies this requirement, then we can use the presented step-down procedure and be reassured that it will control the overall FDR. Technically the definition of PRDS is as follows. The set $D$ is called increasing if $x \in D$ and $y \geq x$ imply that $y \in D$. The random variables $X_1, \ldots, X_m$ are PRDS on $I_0$ if, for any increasing set $D$, and for each $i \in I_0$, $P(X_1, \ldots, X_m \in D \mid X_i = x)$ is nondecreasing in $x$. Benjamini and Yekutieli were able to prove that the procedure illustrated for independent tests also controls the FDR at a level $m_0 / m\alpha$, where $m_0$ is the number of false null hypotheses if the joint distribution of the test statistics is PRDS on the subset of test statistics corresponding to the true null hypothesis. The definition of PRDS may seem rather arcane. However, one illustration with reference to linkage should serve to clarify the nature of this hypothesis and illustrate its adaptability to the mapping context. Benjamini and Yekutieli show that PRDS translates in the following requirement for multivariate normal tests statistics. Consider $X \sim N(\mu, \Sigma)$, a vector of test statistics, each testing the hypothesis $H_i$ that $\mu_i = 0$ against the alternative

$\mu_i > 0$, for $i = 1, \ldots, m$. For $i \in I_0$, the true set of null hypotheses, $\mu_i = 0$; otherwise $\mu_i > 0$. If for each $i \in I_0$, and for each $j \neq i$, $\sigma_{ij} \geq 0$, then the distribution of $X$ is *PRDS* over $I_0$. If we now consider the Gaussian models for genetic linkage analysis proposed by Feingold et al. ***(12)***, it is easy to see that they satisfy this condition. Consider the first of the models proposed in the article, for grandparent–grandchild pairs. If we restrict our attention to a finite subset of genome locations we get a multivariate Gaussian. The mean values of the test statistics at each unlinked location is zero and it is positive for linked loci. The covariance between the two test statistics is non-negative and a function of the recombination fraction across loci. Because the covariances are non-negative, we can conclude that the tests are PRDS on $I_0$ and hence the cut-off values defined by the BH procedure are actually guaranteed to control the FDR, even when we relax the independence assumption.

It should be clear by now that PRDS is a property that is likely to hold also for LD test statistics in the sense that if two markers are in LD and one happens to show random association to the disease, the other marker in LD would have increased chances of showing association higher than a given threshold. Because we do not have a general model for the dependency between tests of association, it is difficult to translate this intuitive idea in a precise general statement.

This author and colleagues ***(17)*** considered a series of simulation frameworks concluding that PRDS is likely to hold in the context of association testing. We refer to that article ***(17)*** for a detailed illustration of the power gains attainable by using FDR-controlling procedures rather than FWER-controlling ones in the context of gene mapping. The use of FDR in the context of QTL mapping had been argued originally by Weller et al. ***(18)*** and more recently in **ref.** ***19***; in association studies its applications have been explored in **refs.** ***8***, ***20***, and ***21***.

### 5.2. Dealing With Dependence: Empirical Studies and Permutation Methods

The paper of Storey and Tibshirani, while pointing the attention of the community to the possibility of using an FDR-controlling procedure, also familiarizes the reader with ways of estimating the empirical FDR level using resampling procedures. In general, resampling and permutation-based approaches have become increasingly popular in dealing with genomic investigation where the necessity of controlling for multiple comparisons coexists with the lack of knowledge about the joint distribution of the multiple test statistics. A clear description of many such methods is offered in **ref. *22***. The main challenge presented by resampling approaches is the computational requirements.

If a *p*-value for each hypothesis can be obtained analytically, it is rather simple to set up a resampling procedure to obtain FEWR-corrected *p*-values. Enumerating all permutations, or sampling a number of them (depending on the size of the dataset) or using bootstrap resampling to better tailor the specific hypothesis tested, one creates a collection β of fictional datasets generated under the complete null hypothesis (or the true combination of true and false nulls, *see* **ref. *23***) and uses them to estimate adjusted *p*-values.

Problems arise when one cannot rely on analytical approximation of the *p*-value and when permutations are needed to estimate them. This is often the case in associations studies when, to maximize power, one uses "exact methods" or resorts to permutations to approximate the null distribution of complex statistics. With these problems in mind, Westfall and Young ***(5)*** describe two rounds of permutations, one to get the *p*-values and one to obtain the global *p*-values. However, if the number of hypotheses evaluated is considerable, this strategy quickly becomes excessively time consuming to be realistically performed.

One way of reducing the number of permutations to be evaluated consists of storing the matrix of permutations, and directly estimating global *p*-values proceeding gene by gene. This translates, however, into a memory burden that can be decreased with smart storing rules. One such procedure is suggested in **ref. *22***.

In order to overcome the computational costs incurred with permutations, some analytical approximations have been recently proposed and used with some success. Lin ***(24)*** suggests an approximation of the tests statistics that isolates the "data contribution" in a component that is kept fixed, while another component is sampled from a standard distribution in order to simulate values of the test statistics under the null hypothesis. A contribution similar in spirit is **ref. *25***. Nyholt ***(26)*** suggests using spectral decomposition of the correlation matrix between a set of SNPs to identify the "effective number" of independent tests to be used in multiple comparison corrections. A limitation of this approach is that the dependence between SNPs does not translate unequivocally in dependence between association tests. However, it is clear that knowing

accurately the levels of LD between SNPs is a fundamental ingredient toward the modeling of dependence between association tests. In this regard, the large number of genotyping studies now underway will be particularly useful in accumulating information on LD between standard sets of markers and systematically across the genome. For progress in this direction *see*, for example, **refs.** ***27–30***.

### *5.3. Using Prior Information*

We have already pointed out how often some prior information is available in the context of association studies. The most common way in which results of prior studies have been used is in the selection of candidate genes. Although this is a quite natural approach, it has often been implemented in a rather unsatisfactory way. In particular, the accumulated evidence is generally not as strong as to conclude that one of the genes in the candidate list is surely going to be the susceptibility one—and yet this is the approach that informs the corrections for multiple comparisons commonly performed.

There are ways to use prior evidence that more accurately reflect the strength of the available information. On the one hand, if one intends really to focus on candidate genes, it would be useful to quantify precisely the prior chance that each of the genes is a susceptibility one. This would add transparency to the research, in that every reader would be able to see which lines of evidence were used and how strong they are. It is possible that it will be very difficult to quantify in a generally acceptable way what is the prior probability of association for every gene in the candidate list. This is, however, a strong argument in favor of not including it in the list.

On the other hand, one can use prior results to inform evaluation of significance of genome-wide investigations. The inability to really zoom in on an appropriately defined gene list does not equate with a total lack of prior information. Roeder et al. ***(31)*** have recently proposed the use of weighting schemes in the correction for multiple comparisons of genome-wide association studies. They consider, in particular, the case of FDR control, but weighting procedures can be applied also when the goal is to control FWER, for example. The idea behind a weighting scheme is that not all the tested hypotheses have the same "importance" or "chance of being true" because they are not "exchangeable." When we define a global error as we described previously we treat all hypotheses in the same way; however, there is no need to stick to this uniform viewpoint. One can define a global error that gives different weights to the different hypotheses: hypotheses that have higher weights are compared with a less stringent threshold than others. The contribution in **ref.** ***31*** exemplifies quite clearly how linkage traces can be used to define weights, giving better chances of being rejected to a hypothesis that corresponds to genomic locations implicated in previous studies.

### 5.4. Multiple Stages and Independent Confirmation

A discussion of control of false discovery in association mapping would not be complete without a reference to replications and multiple stage procedures. Replication of results in an independent sample is often considered the gold standard for confirmed significance. Replication is, however, very difficult in that samples often are collected with slightly different ascertainment criteria, have slightly different ethnic backgrounds, and so on. Often, the lack of replication is not considered a declaration of "false-positive." On the contrary, motivated by the fact that sample differences are known to influence the strength of the association, researchers often end up considering as replication very weak results (like association with the disease, but on a different allele with respect to the one originally implicated, even when the studied polymorphism is supposed to be causal).

In this respect, replication studies are often not as definitive as one would wish, and it is worth emphasizing that they should be conducted with precise rules and protocols ***(32)***.

Given the seriousness of the problem of multiple comparison and the costs associated with genotyping, researchers have proposed using a two-stage design, where only part of the sample is genotyped with the entire marker set and the second portion of the sample is used for replication ***(33)***. The optimal strategy depends on the costs of genotyping, and so readers have to be careful in their design conclusion.

An interesting development along these lines is a screening and replication design using trios proposed by C. Lange and coauthors ***(34)***. In this work, the parental genotypes are used to conduct an initial screening and the offspring genotypes are used for the replication stage, with a transmission/disequilibrium test. The authors report encouraging results in terms of power and control of false-positives.

Association studies present a formidable challenge in terms of avoiding false discoveries, without missing all the true ones. Although there is still room for much improvement in our methods to address such challenges, the directions outlined in this last section suggest that the community is well aware of the difficulties and is moving in the right direction.

## References

1 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517.
2 Popper, K. (1959) *The Logic of Scientific Discovery*. Routledge, London and New York.
3 Fisher, R. A. (1935) *The Design of Experiments*. Oliver Boyd, Edinburgh, UK.
4 Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statist. Soc. B* **57**, 289–300.
5 Westfall, P. and Young, S. (1993) *Resampling-Based Multiple Testing*. Whiley, New York, NY.

6 Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under independence. *The Annals of Statistics* **29,** 1165–1188.
7 Genovese, C. R. and Wasserman, L. (2002) Operating characteristics and extensions of the false discovery rate procedure. *J. Royal Statist. Soc. B* **64,** 499–518.
8 Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. USA* **100,** 9440–9445.
9 Morton, N. (1955) Sequential tests for the detection of linkage. *Am. J. Human. Genet.* **7,** 277–318.
10 Elston, R. C. and Lange, K. (1975) The prior probability of autosomal linkage. *Ann Hum Genet.* **38,** 341–350.
11 Lander, E. and Botstein, D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121,** 185–190.
12 Feingold E., Brown, P. O., and Siegmund, D. (1993) Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am. J. Hum. Genet.* **53,** 234–251.
13 Dupuis, J., Brown, P. O., and Siegmund, D. O. (1995) Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. *Genetics* **140,** 843–856.
14 Churchill, G. and Deorge, R. (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138,** 963–971.
15 Freimer, N. B. and Sabatti, C. (2005) Guidelines for association studies in Human Molecular Genetics. *Hum. Mol. Genet.* **14,** 2481–2483.
16 Freimer, N. B. and Sabatti, C. (2004) The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. *Nat. Genet.* **36,** 1045–1051.
17 Sabatti, C., Service, S., and Freimer, N. (2003) False discovery rates in linkage and association linkage genome screens for complex disorders. *Genetics* **164,** 829–833.
18 Weller, J. I., Song, J. Z., Heyen, D. W., Lewin, H. A., and Ron, M. (1998) A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* **150,** 1699–1706.
19 Benjamini, Y. and Yekutieli, D. (2005) Quantitative trait loci analysis using the false discovery rate. *Genetics* **171,** 783–790.
20 Devlin, B., Roeder, K., and Wasserman, L. (2003) False discovery or missed discovery? *Heredity* **91,** 537–538.
21 Devlin, B., Roeder, K., and Wasserman, L. (2003) Analysis of multilocus models of association. *Genet. Epidemiol.* **25,** 36–47.
22 Ge, Y., Dudoit, S., and Speed, T. (2003) Resampling-based multiple testing for microarray data-analysis. *Test* **12,** 1–77.
23 Dudoit, S., van der Laan, M. J., and Pollard, K. S. (2004). Multiple testing. Part I. Single-step procedures for control of general Type I error rates. *Stat. Appl. Genet. Mol. Biol.* **3,** Article 13.
24 Lin, D. Y. (2005) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* **21,** 781–787.

25 Seaman, S. R. and Muller-Myhsok, B. (2005) Rapid simulation of p values for product methods and multiple-testing adjustment in association studies. *Am. J. Hum. Genet.* **76,** 399–408.

26 Nyholt, D. R. (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am. J. Hum. Genet.* **74,** 765–769.

27 Dawson, E., Abecasis, G. R., Bumpstead, S., et al. (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418,** 544–548.

28 Maniatis, N., Collins, A., Xu, C. F., et al. (2002) The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis. *Proc. Natl. Acad. Sci.* **99,** 2228–2233.

29 Evans, D. M. and Cardon, L. R. (2005) A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am. J. Hum. Genet.* **76,** 681–687.

30 De La Vega, F. M. Isaac, H., Collins, A., et al. (2005) The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. *Genome Res.* **15**, 454–462.

31 Roeder, K, Bacanu, S., Wasserman, L., and Devlin, B. (2006) Using linkage genome scans to improve power of association genome scans. *Amer. J. Hum. Genet.*, in press.

32 Patterson, M. and Cardon, L. (2005) Replication publication. *PLoS Biol* **3,** e327.

33 Satagopan, J. M. and Elston, R. C. (2003) Optimal two-stage genotyping in population-based association studies. *Genet. Epidemiol.* **25,** 149–157.

34 Van Steen, K., McQueen, M. B., Herbert, A., et al. (2005) Genomic screening and replication using the same data set in family-based association testing. *Nat. Genetics* **37,** 683–691.

# 15

# Gene Mapping in Asthma-Related Traits

**Tarja Laitinen**

**Summary**

In asthma, as in many other common multifactorial diseases, the identification of the susceptibility genes has been challenging because consistent results at the genome-wide significance level have been scarce. So far, genome-wide scans have been reported in 17 study populations. By means of genome-wide linkage and hierarchical association analysis, six positional candidate genes (*ADAM33*, *PHF11*, *DPP10*, *GPR154*, *HLA-G*, and *CYFIP2*) for asthma-related traits have been cloned. The interactions of the proteins encoded by these genes and the biological relevance of these signaling pathways in the development of asthma are still poorly understood. Also, the disease mechanisms resulting from the genetic variance in the genes identified remain largely unknown. Although this information is gradually accumulating, we can examine the statistical robustness of each genetic finding in combination with the limited data available on the functional properties of the corresponding proteins to estimate the strengths and weaknesses in the chains of evidence.

**Key Words:** Asthma; atopy; genome scan; positional cloning; positional candidate gene; linkage disequilibrium.

## 1. Introduction

Asthma and other immunoglobulin E (IgE)-mediated allergic diseases have become the most common chronic diseases (prevalence 5–15%) among children and adolescents of all industrialized countries. Owing to environmental changes, the reported prevalence rates of asthma have shown a steady increase for several decades. The results from the most recent epidemiological studies in Europe, however, suggest the rising trend in allergic sensitization among adolescents might come to an end ***(1,2)***.

Based on numerous epidemiological studies, the genetic component in the development of asthma is consistent. In the families in which one of the offspring have already developed asthma, the risk of another sibling to become

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

affected is three to five times higher than in the general population. The genetic component is at the same level as in hypertension, obesity, type 2 diabetes, and alchoholism, but at a lower level than in diseases such as inflammatory bowel disease or type 1 diabetes *(3)*. Nationwide twin studies in young age cohorts in Nordic countries showed the heritability of 60–80% *(4–7)*, which is higher than expected based on the sibling risk alone.

During the 1990s, the efforts of identifying genes causing the increased susceptibility of asthma-related traits started from biologically interesting candidate genes and gene families organized as clusters in the genome *(8)*. A recent review of 500 papers listed altogether 25 genes that have been associated with asthma-related traits in six or more populations *(9)*. An additional 54 genes have been associated in two to five populations. One obvious reason for failures has been underpowered study designs. Association studies offer a potentially powerful approach to identify genetic variants that influence susceptibility to common diseases, but the studies are prone to false-positive results that might be seen as inconsistencies of reproducibility of results in different study populations *(10)*. It has been estimated that a quarter of published studies are reports of false-positive. Publication bias most likely makes the proportion even greater. Asthma-susceptibility alleles appear to have relatively high population frequencies with greatly reduced penetrances because of the influence of other genes, the environment, or stochastic events.

In addition to study sizes, the genetic informativeness of publications has also improved because of the increased knowledge on the physical orientation of the genes, genetic markers, and haplotype structures. In the earliest studies, the genetic analyses were applied using a linkage approach with microsatellite markers. Microsatellite markers are highly polymorphic, located usually outside coding sequences, and are rarely considered to be functional. In these studies the physical distances and the order of the markers were based on approximations from several genetic maps, often somewhat ambiguous, harming multipoint linkage analyses. Since the year 2003, the complete human DNA sequence has been available, the human high-throughput sequencing projects have increased the density of single-nucleotide polymorphism (SNP) maps and enhanced the development of the technologies for SNP genotyping. Furthermore, international collaborations, such as the HapMap Project, have released valuable information on the human haplotype structures across the genome and the catalog of SNPs that can recognize these conserved haplotype blocks that are 20–60 kb in size. These blocks may contain a large number of SNPs, but far fewer SNPs are needed to identify the haplotypes within a block.

## 2. Genome Scans in Asthma-Related Traits

Genome scans are performed in complex traits to identify novel genes and signaling pathways involved in the development of a disease. The basic principals

of gene mapping and linkage analysis have been adopted from the gene-mapping studies in Mendelian diseases in which this approach has successfully identified the mutated gene causing the disease. In complex traits, however, the mode of inheritance is unknown. This led to the development of new, nonparametric multipoint analysis methods, such as GENEHUNTER. In these methods, linkage information is extracted only from the affected family members and the performance of nonparametric likelihood of genetic linkage (LOD) scores are roughly comparable to the LOD score analysis under the correct model of inheritance ***(11)***.

Several genome-wide scans with an increasing number of study families have been already published in 17 study populations of different ethnic origin ***(12–42)*** (**Table 1**). Suggestive or weak linkage has been reported in all chromosomes and very few of the reported loci have reached the genome-wide significance level. The most significant linkage and/or association has been reported for 2pter ***(31,42)***, 2q14 ***(12)***, 5q33 ***(36)***, 6p21 ***(20,27)***, 7p15-14 ***(12,30)***, 13q14 ***(12)***, 14q24 ***(33)***, and 20p13 ***(32)***. These findings have led to cloning of six candidate genes: *PHF11* on 13q14 ***(43)***, *PPD10* on 2q14 ***(44)***, *GPR154* on 7p15-14 ***(45)***, *HLA-G* on 6p21 ***(46)***, and *CYFIP2* on 5q33 ***(47)*** (discussed in detail in **Subheadings 5.1.–5.6.**).

The scans represent a very heterogeneous group of studies differentiated from each other in their study designs. First, the recruitment criteria for the participating families differed. Study subjects were selected into the study either according certain clinical criteria or as a cohort representing a random population sample, the latter being more informative for quantitative traits rather than dichotomized (affected vs unaffected) traits. Quatitative phenotypes such as bronchial hyperreactivity, total serum IgE level, blood eosinophil account, specific sensitivity to airway allergens measured either by skin-prick testing or by specific IgE levels have been frequently used. The clinical inclusion criteria are based on different clinical conditions and the diagnosis is based on a different type of information such as patient-reported questionnaire data, diagnostic measurements at a single cross-sectional time-point, or retrospective medical history. Both the size of the cohorts and the age range of the probands vary. Second, the ethnic background, population history (inbred vs outbred populations), and family structures (large pedigrees vs affected sibpairs) of the study cohorts differed. Because of differences in recruitment methods, the analytical methods also differed. Compared to genome scans performed in the diseases such as diabetes type 1 or psoriasis in which linkage to the major histocompatibility complex on chromosome 6p is found constantly across publications, the genome-wide scan data in asthma suggest that there is no single locus for asthma with such a dramatic effect on morbidity. For the six previously mentioned loci, the signal of linkage is strong and most likely clinically relevant. For the loci that have shown only weak linkage, but were identified by multiple

**Table 1**
**Published Genome Scans in Asthma and Atopy-Related Traits**

| Study | Study population primary/replication | Study design | N of study subjects | Recruitment criteria | Studied phenotype | Best reported linkages |
|---|---|---|---|---|---|---|
| Daniel et al. 1996 *(12)* | Australian /British | affected sib pairs | 80 families, 364 individuals | population based cohorts | 4 atopy related traits | 4q, 6, 7, 11q, 13q, 16 |
| GSCS 1997 *(13)*, Xu et al. 2001 *(14)*, Mathias et al. 2001 *(15)*, Huang et al. 2003 *(16)*, Blumenthal et al. 2004 *(17)*, and 2006 *(18)* | Different ethnicities -African American -Caucasian -Hispanic American | affected sib pairs | 287 families[b], 1931 individuals | asthma | asthma, BHR, IgE, sensitivity to HDM and seasonal allergens | 13q34, 20p12, 21q21, 11 |
| Ober et al. 1998 *(19)*, and 2000 *(20)* | Hutterite | extended pedigree | 1 family, 693 individuals | 15 generation pedigree | BHR and self-reported asthma | 5p, 5q, 8p, 14q, 16q, 11 |
| Wjst et al. 1999 *(21)*, Altmuller 2005 *(22)* | German | affected sib pairs | 201 families, 465 individuals | asthma | asthma and 6 related phenotypes | 1p33, 7p11.2, 11pter, 12q15 |
| Dizier et al. 2000 *(23)*, Bonzigon et al. 2004 *(24)* | French | affected sib pairs | 295 families, 1317 individuals | asthma | 8 asthma related traits | 6q14, 12p13, 17q22-q24, 21q21 |
| Xu et al. 2000 *(25)*, 2002 *(26)*, Koppelman et al 2002 *(27)*, Meyers et al. 2005 *(28)*, Postma et al. 2005 *(29)* | Dutch | affected family members | 200 families, 1174 individuals | asthma | specific and total IgE, eos count, sensitivity to HDM | 5q31-33, 6q25, 7q, 9p1, 12q1 |
| Laitinen et al. 2001 *(30)* | Finnish/Canadian | affected family members | 86 families, 443 individuals | chronic asthma | asthma, serum total IgE | 7p15, 4q |
| Xu et al. 2001 *(31)* | Chinese | affected sib pairs | 533 families, 2551 individuals | asthma | 9 asthma related traits | 2pter |

| | | | | | | |
|---|---|---|---|---|---|---|
| Eerdewegh et al. 2002 *(32)* | US/British | affected sib pairs | 460 families, 920 individuals | asthma | asthma and BHR | 20p13 |
| Hakonarson et al. 2002 *(33)* | Icelander | extended pedigrees | 175 families, 596 individuals | diagnosed asthma | asthma | 14q24 |
| Haagerup et al. 2001 *(34)* and 2002 *(35)* | Danish | affected sib pairs | 100 families, 424 individuals | clinical allergies | Allergic rhinitis and 3 other atopy related traits | 1p36, 3q21-q22, 4q24, 5q31, 6p24-p22 |
| Yokouchi et al. 2000 *(36)* and 2002 *(37)* | Japanese | affected sib pairs | 48 families, 188 individuals | childhood atopic asthma | Rhinitis, total and specific IgE levels | 1p36, 2p13, 3p24.1, 4q13, 5q33, 9q34, 12q24 |
| Wang et al. 2005 *(38)* | Taiwanese | singletons[a] | 190 individuals | asthma | Allergic and nonallergic asthma | 5q31, 6q15, 6p24, 9p22 , 8q11, 19p12 |
| Kurz et al. 2005 *(39)* | multiethnic European | affected sib pairs | 82 families, 366 individuals | asthmatic children | sensitivity to HDM | 2p12, 3q21, 16q21 |
| Ferreira et al. 2005 *(40)* | Australian | affected twin pairs | 202 families, 591 individuals | asthmatic twin proband | HDM sensitivity and 6 other asthma related traits | 20q13, 12q24 |
| Bu et al. 2006 *(41)* | Swedish | affected relative pairs | 250 families | subcohort of AD | allergic rhinoconjuctivitis | 3q13, 4q34-q35, 18q12 |
| Pillai et al. 2006 *(42)* | Caucasian | affected sib pairs | 364 families | asthma | asthma, atopic asthma, BHR | 2p, 4p |

References for the same or extended study cohort using different study design are given in the same cell of the table. The number of the study subjects and best-reported linkage loci are given according to the latest publications.

[a]Genome-wide linkage disequilibrium with 763 autosomal STR markers.

studies, the conclusions should be cautious as the interpretations of the results can change.

The limitations of linkage analysis in complex human diseases resulting from phenocopies and diagnostic inconsistencies have become very obvious. This has shifted emphasis away from linkage analysis using microsatellite markers toward SNP genotyping and analytical strategies based on allele- and haplotype-association analyses. Compared with microsatellite markers, SNPs are less polymorphic, but are more stable and robust to the genotype, and they may be found in the coding regions of genes. Changes in the analysis methods have also shifted the focus from family-based studies to extremely large, well-defined population based case–control cohorts to observe ancient, shared haplotype blocks carrying disease-causing SNPs and, therefore, were found more frequently among the patients than controls.

## 3. Challenges Proving That the Identified Gene is a True Susceptibility Gene for Asthma

In gene-mapping studies of monogenetic diseases, the causative relationship between the mutated gene and the disease has been rather easy to prove when the mutation is found only among affected individuals and not among the healthy family members. In multigenic disease, this is rarely the case and the picture gets much more blurred. The results at their best find a robust association of genetic variants found more frequently among affected than unaffected. Therefore replications of the gene-mapping results in independent datasets are crucial. Because of strong linkage disequilibrium (LD) between neighboring polymorphisms the causative variant remains often unknown. The apparent susceptibility alleles may be able to produce a fully functional protein. However, the protein structure can have changed by one or two amino acids. If these amino acids are in critical regions of the molecule (e.g., at the binding site causing altered binding) the change can have a dramatic effect on the function of the molecule and the development of the disease is understandable. Still, even these kinds of changes are rare in the gene variants associated with complex diseases. There are more and more reports of noncoding SNPs that are reported to be causative. Intronic SNPs obviously do not change the structure of the encoded protein, but they are known to change the expression level and splicing of genes. The number of human genes was less than expected and it is likely that the complexity of human genetics is determined not only by polymorphisms in the genes, but also by how genes are expressed leading to transcriptional and translational modifications. Therefore, the proportions of different splice variants produced can be crucial in the development of the disease. Thus, the gene variation seen in complex traits is more like a quantitative (functioning less effectively) effect than dichotomized effect (functioning vs not functioning).

This also fits well with the low penetrance of the susceptibility alleles, late onset of the disease, and the strong impact of environmental factors.

## 4. Identified Positional Candidate Genes

Genome-wide screens, combined with fine mapping of the loci that showed initial linkage have so far revealed six interesting positional candidates for asthma-related traits. In these projects a so-called hierarchical genotyping method is usually applied. The linkage peak is systematically saturated with tens of markers and marker maps are made more dense in a stepwise manner by following the strongest signal of allele and haplotype association.

### *4.1.* ADAM33 *on Chromosome 20p13*

The first positional candidate for asthma was reported on chromosome 20p13 ***(32)***. A standard microsatellite marker genome scan implicated a linkage peak for the asthma phenotype combined with bronchial hyperreactivity (LOD 3.9) in 362 UK and 98 US families. A genomic region of about 2.5 Mb was then considered for gene content, resulting in the listing of 40 genes, and a subsequent genetic association study using cases from positive linkage families and population controls was performed. A 185-kb segment was implicated, in which over 20 SNPs showed nominally significant allelic associations in either the United Kingdom, United States, or both samples, supported also by significant transmission disequilibrium test results. Most SNPs localized within the 3′ half of the disintegrin and metalloproteinase-33 gene (*ADAM33*). *ADAM33* is a membrane-anchored metaloprotease with diverse functions ***(48)***. *ADAM33* is mainly expressed in muscles of every type and fibroblasts, lymph nodes, thymus, and liver, but not in leukocytes, bronchial epithelium, or bone marrow. Therefore, it has been suggested that *ADAM33* might have a role in bronchial contractibility or bronchial remodeling. The 3′ untranslated region of *ADAM33* and domains downstream of the catalytic domain regulate *ADAM33* protein maturation and, thus, potential activity. *ADAM33* undergoes complex alternative splicing, and an isoform has an active catalytic domain with a known crystal structure and is processed to the cell surface ***(49)***.

Currently there are nine replications published and summarized in **Table 2** ***(50–58)***. Some of them report modest association with different markers in the region and results as such are confusing. However, meta-analysis across the replication datasets (N = 2721) showed that the intronic SNP ST+7 was significantly associated to asthma both in case–control and transmission disequilibrium test (TDT) study designs ***(57)***. The reported odds ratio remain relatively small (OR 1.4, $p = 0.0013$), but is robust, and of importance because the susceptibility allele is extremely common in the population (85% among patients vs 79% among control population). The authors estimate that genetic variation in *ADAM33* would

**Table 2**
**Replication Studies for *ADAM33***

| Study | No. of independent cases | Population | No. of SNPs studied | Associated SNPs | Associated phenotype |
|---|---|---|---|---|---|
| Howard et al. 2003 *(50)* | 219 | US White | 8 | none | |
| | 160 | African American | | none | |
| | 153 | Dutch | | ST+7, and V4 | asthma |
| | 112 | US Hispanic | | none | |
| Lind et al. 2003 *(51)* | 190 | Mexico | 6 | none | |
| | 183 | Puerto Rico | | none | |
| Raby et al. 2004 *(52)* | 436 | US White | 17 | 16-marker haplotype | childhood asthma |
| | 66 | African American | | none | |
| | 47 | US Hispanic | | T1, and T+1 | childhood asthma |
| Lee et al. 2004 *(53)* | 326 | Korean | 5 | none | |
| Jongepier et al. 2004 *(54)* | 152 | Dutch | 8 | S2 | excess decline in $FEV_1$ in asthma |
| Werner et al. 2004 *(55)* | 91 | German | 15 | ST+5, ST+7 | asthma |
| Simpson et al. 2004 *(56)* | 470 children | UK | 17 | F+1, S1, ST+5, and V4<br>F+1, M+1, T1, and T2 | increased airway resistance<br>excess decline in $FEV_1$ in early childhood |
| Blakey et al. 2005 *(57)* | 348 | Icelander | 13 | none | asthma |
| | 60 families | UK | | none | asthma |
| van Diemen et al. 2005 *(58)* | 1390 | Dutch | 8 | S_1 , S_2, and Q–1<br>F+1, S_1, S_2, and T_2 | excess decline in $FEV_1$<br>chronic obstructive pulmonary disease |

account for 50,000 excess asthma cases in the United Kingdom. These studies have also demonstrated that very careful replication studies in large population samples are needed before any conclusions can be made. This especially concerns genetic findings in which the functions of the identified gene are poorly understood.

### *4.2.* PHF11 *on Chromosome 13q14*

The first genome-wide scan performed among Australian families and replicated among UK families reported six loci ***(12)***. Based on simulations the authors showed that it is highly unlikely that all the reported loci are false-positive. Later, based on their fine-mapping results, a significant association to atopy-related quantitative traits with one these loci, 13q14, have been published in 230 Australian families and replicated in 150 Australian families with consistent results ***(43)***. Linkage identified a genomic region of 7.5 cM that was saturated with microsatellite markers. The strongest signal of association was followed up by creating a SNP map of a 620-kb region around the best marker. The strongest haplotype associations identified a gene named plant homeodomain (PHD) finger protein 11 (*PHF11*). The gene showed alternative splicing and contained two plant homeodomain zinc fingers suggesting that it regulates transcription. The replicated association extended for approx 100 kb and the best associated SNPs were located in the two introns and in the 3′-coding region of the gene. Because rearrangements in PHD domains in other members of the gene family and small deletions of 13q14 has been reported in B-cell leukemia, the authors speculate that *PHF11* might have a role in the regulation of B-cell clonal expansion. A variant utilizing an alternative exon 1 was expressed in immune-related tissues and cells. Variants utilizing two alternative exons following exon 5, both of which contain premature stop codons, were detected only in lung and unactivated peripheral blood leukocytes. Although the precise functions of *PHF11* are not known, the conserved domains that suggest a role in chromatin remodeling or transcriptional regulation. A family-based association study across the neighboring *SETDB2* and *PHF11* genes has also identified two SNPs in the *PHF11* gene significantly associated with childhood atopic dermatitis in an Australian cohort further supporting the importance of the loci ***(59)***.

### *4.3.* DPP10 *on Chromosome 2q14*

In the original genome-wide scan by Daniel et al. ***(12)*** the chromosomal region 2p14 showed suggestive linkage. The homologous region has been suggested to be linked to bronchial hyperreactivity in mouse genome screens. Fine mapping of the region in 244 families showed significant association close to the D2S308 microsatellite marker and excluding the interleukin-1 gene cluster as a candidate region ***(44)***. Based on high-density mapping they identified

significant associations in two LD blocks. The critical region located in the initial exons of a new member of serine proteases, named dipeptidyl peptidase 10 (*DPP10*). *DPP10* is a 796-amino acid protein contains a transmembrane domain, 10 N-glycosylation sites, and several conserved amino acids found in the six domains characteristic of members of the peptidase, lipase, esterase, epoxide hydrolase, or serine hydrolase superfamily ***(60)***. However, *DPP10* lacks the active-site serine, which is substituted with a glycine residue. *DPP10* has been shown to modulate the Kv4 subfamily of voltage-gated potassium channels that are of importance in regulating neuronal firing frequencies and in the modulation of incoming signals in dendrites ***(61)***. No coding polymorphism in *DPP10* were detected, but the gene shows several N-terminal slice variants some of which are membrane bound and some are cytosolic.

### *4.4.* GPR154 *on Chromosome 7p14.3*

Contrary to the previous studies describing a positional candidate gene, the Finnish genome scan was performed among adult asthma patients and the families were recruited from a particular area that was inhabited substantially later than the Southern and coastal areas of Finland ***(30)***. The probands represent a younger subisolate of the Finns (age of 20–25 generations) and 10% of the regional patients who have chronic asthma and a need for daily medication. Both linkage and association results were replicated in another founder population of the Saguenay-Lac-St-Jean of North-Eastern Quebec ***(45)***. The genetic homogeneity has been shown to be of tremendous help in gene-mapping studies of rare Mendelian diseases, but in common multifactorial disorders its role is not well established. Nongenetic factors such as a more uniform environment than ethnically mixed populations may be of greater importance.

In the original genome scan only one locus on chromosome 7p15-14 reached the genome-wide significance based on simulations (NPL 3.9, nominal *p*-value 0.0001, and empirical genome-wide *p*-value 0.035). The same locus has also been reported among Australian families ($p = 0.0003$) ***(62)***. Using the hierarchical gene-mapping strategy, a genomic region of 133 kb was identified showing robust haplotype association and strong LD between markers among both the Finnish and Canadian study populations ***(45)***.

Within the region a total of 152 SNPs and deletion–insertion polymorphisms were discovered. The polymorphisms formed seven common (frequency >2%) haplotypes. The same haplotypes with varying frequencies were discovered in several other Caucasian populations ***(63,64)*** (**Table 3**). Haplotypes 4, 5, and 7 are associated with high IgE level among the Finnish patients and haplotype 2 with asthma among the Canadian families. Subsequently the results have been replicated in four large European cohorts using the same set of haplotype-tagging SNPs (**Table 4**). The variants can be categorized by risk (H2, H4, H5,

**Table 3**
**The Frequencies of the Common *GPR154* Haplotypes Differ Between the Caucasian Populations That May Have an Effect on the Informativeness of Different Association Studies**

| | Haplotype frequency | | | | | |
|---|---|---|---|---|---|---|
| *GPR154* haplotypes | Finnish | Canadian | Swedish | North European | German | Italian (*Malerba et al., unpublished*) |
| H1 | 33 | 22 | 31 | 31 | 31 | 28 |
| H2 | 13 | 21 | 22 | 21 | 21 | 21 |
| H3 | 19 | 18 | 11 | 14 | 24 | 15 |
| H4 | 11 | 17 | 7 | 9 | 9 | 5 |
| H5 | 9 | 8 | 6 | 6 | 6 | 7 |
| H6 | 5 | 9 | 8 | 11 | – | 12 |
| H7 | 5 | 6 | 6 | 6 | 6 | 7 |
| H8 | – | – | – | – | 3 | 2 |

H6, and H7) and non-risk/protective haplotypes (H1 and H3). The haplotypes in both groups are closely related and reflect different evolutionary origin in the phylogenetic analysis of the haplotypes ***(45,63)***. The observed risk for a single haplotype varies between studies, but H1 and H3 are either neutral or protective for the disease risk in all cohorts. In atopic dermatitis no *GPR154* association has been found ***(65,66)***.

The genomic region harbored two genes: G-protein coupled receptor 154 (*GPR154*) also known in the literature as *GPRA* (alias *PGR14*, *VRR1*, and *NPSR*) and asthma-associated alternatively spliced gene 1 (*AAA1*). The genes are overlapping and encoded to opposite directions, but do not share any exons. *AAA1* has numerous alternative splice variants with the longest comprising only 74 potential amino acids. Based on several lines of evidence, *AAA1* may not be a protein-coding gene. In vitro translation failed to yield a stable polypeptide and transiently transfected cells did not produce recombinant protein, with the polyclonal antibodies detecting the antigen but no proteins in Western blot or immunohistochemistry ***(45)***.

*GPR154*, on the other hand, was shown to be an interesting candidate for asthma susceptibility. *GPR154* is a cell membrane-expressed receptor with seven transmembrane domains and shares significant sequence identity with human vasopressin and oxytocin and *Drosophila CG611* receptors. Initially *GPR154* was cloned from retinal tissue. More recently it has been shown that the receptor is also expressed in smooth muscle cells of internal organs, all mucosal surfaces, and skin. It is also expressed by multiple types of inflammatory cells

**Table 4**
**Statistically Significant Haplotype Associations Reported for *GPR154* Among the European Populations**

| Study cohort | N of study subjects | Common GPRA haplotypes | | | | | | | Characteristics of the study cohort |
|---|---|---|---|---|---|---|---|---|---|
| | | H1 | H2 | H3 | H4 | H5 | H6 | H7 | |
| Finnish (Laitinen et al. 2004) ***(45)*** | 220 large pedigrees and trios | | | | Risk Serum total IgE | Risk Serum total IgE | | Risk clinical asthma | Recruited through an adult asthma patient, studied phenotypes clinical asthma and total IgE |
| Canadian (Laitinen et al. 2004) ***(45)*** | 193 nuclear families | | Risk clinical asthma | | | | | | Recruited through an asthmatic child, studied phenotypes clinical asthma and total IgE |
| Swedish (Melen et al. 2005) ***(63)*** | 800 children at the age of 4 yr | | | | | | | | Population based cohort, physician diagnosed current asthma and allergic sensitization |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| North European (Melen et al. 2005) ***(63)*** | 3113 children aged 5 to 13 yr | Protective allergic asthma Allergic sensitization | | Protective allergic rhinoconjunctivitis | | Risk clinical asthma allergic sensitization | Risk allergic sensitization | | Population based cohort, physician diagnosed asthma, allergic sensitization, and allergic rhinoconjunctivitis |
| German (Kormann et al. 2005) ***(64)*** | 1872 children aged 9–11 yr | Protective clinical asthma | | | Risk asthma+ BHR | | | | Subpopulation from a very large population based cohort, studied phenotypes physician diagnosed asthma, BHR, and total IgE |
| Italian (Malerba et al. unpublished data) | 211 families | Protective clinical asthma | Risk clinical asthma | | Risk clinical asthma | Risk clinical asthma | Risk clinical asthma | Risk clinical asthma | Recruited through an allergic child, phenotypes studied: asthma, BHR, total IgE, SPT, atopy |

Based on the results the common haplotypes H1–H7 can be divided into haplotypes that increase (H2, H4–H7) or decrease (H1 and H3) the risk of asthma-related phenotypes.

such as T-cells, monocytes/macrophages, and eosinophils in peripheral blood cells and in human sputum (Pulkkinen et al., unpublished information). Based on proteomic screenings, a high-affinity endogenous agonist for *GPR154* has been described and named neuropeptide S because of its high expression in the brain and the physiological responses related to locomotor activity and anxiolytic-like effects in mice ***(67,68)***. This novel linear 20-residue peptide is able to activate the receptor, which increases both intracellular cAMP and $Ca^{2+}$ levels ***(69)***. Based on *in situ* hybridizations neuropeptide S is co-expressed also by mucosal epithelial cells in bronchi ***(70)***.

The main two transcripts (A and B) have alternative 3′ exons encoding proteins of 371 and 377 amino acids, respectively. The N-terminus and the extracellular domains form the ligand-binding pocket, the C-temini of the isotypes differ suggesting that the isotypes may have different signaling properties. Based on immunohistology, the expression profiles of the isotypes also differ in the tissues relevant for asthma. GPR154-B is expressed in virtually all human epithelial tissues lining outer and internal body surfaces including skin, digestive tract, and bronchial epithelium. GPR154-A has additional prominent expression in submucosal smooth muscle cells. In bronchial biopsies from asthma patients, the expression of GPR154-B was upregulated in the bronchial smooth muscle cells in contrast to the negative finding in control samples ***(45)***. Lately the same staining pattern has been found also in newborn children suffering from respiratory distress syndroma and bronchopulmonary dysplasia ***(71)***.

In the associated region most of identified polymorphisms were noncoding. Several splice variants of the *GPR154* have been identified and for all of them the exon rearrangements occurred within the associated region. Most splice variants lose the 7TM structure and failed to be expressed on the cell membrane ***(70)***. This may suggest that an unfavorable pattern of produced splice variants may increase the risk of asthma. One missense SNP (Asn107Ile) has been identified from the binding pocket of the receptor ***(45)***. The latest association study has been published among the Chinese asthma patients ***(72)***. Compared with the European populations, the observed haplotype pattern was less polymorphic identifying only three common haplotypes. The strongest allele association was observed to Asn107Ile. Based on sequencing results among the Finnish patients, Asn107Ile also tags the two groups of risk and non-risk/protective haplotypes among the European study cohorts. In addition, a recent in vitro study has shown that Asn107Ile results in a gain-of-function characterized by an increase in agonist potency in the Ile107 residue-carrying receptor variant ***(67)***. Therefore, the finding among the Chinese patients not only further supports the genetic hypothesis of the categorization of GPR154 risk and non-risk/protective haplotypes, but also suggests direct involvement of this missense polymorphism in the ligand-binding capabilities of the receptor as a potential disease-

causing mechanism. G protein-coupled receptors are well-known drug targets for small molecule compounds, which makes *GPR154* a promising candidate for drug development and novel therapeutic intervention in asthma-related traits.

### *4.5.* HLA-G *on Chromosome 6p21*

Four Caucasian datasets (a total of 867 families) previously showing linkage of varying degrees to the 6p21 region in initial genome scans were pooled and a 10- to 20-kb SNP map was created for a 1-Mb region within the linkage peak ***(46)***. Based on summarized genotyping results and immuohistochemical staining the authors concluded that *HLA-G* was the susceptibility gene. *HLA-G* has more limited tissue expression and polymorphism than the genes encoding the classical human leukocyte antigens. It has an import immunoregulatory role in promoting maternal tolerance of the allogeneic fetus.

### *4.6.* CYFIP2 *on Chromosome 5q33*

The T helper type 2 cytokines, primarily interleukins (IL)-4, -5, and -13, control the major components that characterize an asthmatic immune response. The genes are located as a tight cluster on the long arm of chromosome 5q31, which has been one of the most intensively studied genomic regions in asthma-related traits. Most of the studies have used a candidate gene approach of which IL-13 shows the most consistent association results ***(73)***. IL-13 encodes a 132 amino acid immunoregulatory cytokine produced primarily by activated T helper type 2 cells. In particular the nonsynonymous SNP causing an amino acid change from arginine to glutamine (named in the literature as R110Q, R129Q, R144Q, or R130Q), has shown association by several studies among patients with different ethnic backgrounds and different atopic conditions or quantitative traits ***(74–82)***. R130Q is of particular interest because it obviously changes the functional properties of the protein ***(83)***.

Noguchi et al. ***(47)*** have studied the region using the hierarchical gene-mapping approach. Mutation screening the 9.4-Mb region and association analysis using a TDT test of 105 polymorphisms in 155 families with asthma revealed six polymorphisms in cytoplasmic fragile X mental retardation protein interacting protein 2 gene (*CYFIP2*) were associated significantly with the development of asthma (OR 5.9, $p = 0.000075$).

### *4.7.* PTGDR *on Chromosome 14q22.1 and* NOD1 *on Chromosome 7p14*

The history behind becoming candidate genes for asthma is similar to the prostanoid DP receptor (*PTGDR*) and the nucleotide-binding oligomerization domain (*NOD1*) ***(84,85)***. Both genes have been selected for haplotype analysis

because they are located within previously described linkage peaks (14q and 7p, respectively) and both are well-known immune response-modifying genes. Prostaglandin D2 is known to be an abundant prostanoid that is capable of inducing the key features of an acute asthma phenotype. The four marker *PTGDR* haplotypes that associated with the low transcriptional efficiency of the receptor in vitro also protected against asthma in ethnical populations studied (US white OR 0.55, 95% CI 0.38–0.80 and US black OR 0.32, 95% CI 0.12–0.89).

*NOD1* represents the pattern recognition receptors that identify intracellular microbial products. An association study of 12 SNPs showed that an insertion polymorphism in exon nine accounted for 7% of the variation in total IgE serum level in two studied family cohorts (OR 6.3; 95% CI 1.4–28.3).

## 5. Conclusion

Asthma is a complex chronic inflammatory disorder. In molecular genetic terms the clinical outcome of asthma is a group of overlapping diseases rather than one disease. These pathological processes involve a variety of molecular signaling pathways—possibly also at the individual, but definitely at the population level. Genetic variation in these signaling pathways influences on the interactions between proteins as well as with environmental variables. From this information we will gain insight into the genes involved in producing that subject's allergic and asthmatic phenotype, understand the natural history of that patient's disease, and predict responses to pharmacological agents. That allows us to tailor a more specific treatment procedure for each patient, which hopefully also reduces the overall cost of health care and drug development related to asthma.

## References

1 Braun-Fahrländer, C., Gassner, M., Grize, L., et al. (2004) No further increase in asthma, hay fever and atopic sensitisation in adolescence living in Switzerland. *Eur. Respir. J.* **23,** 407–413.

2 Verlato, G., Corsico, A., Villani, S., et al. (2003) Is the prevalence of adult asthma and allergic rhinitis still increasing? Results of an Italian study. *J. Allergy Clin. Immunol.* **111,** 1232–1238.

3 Vyse, T. J. and Todd, J. A. (1996) Genetic analysis of autoimmune disease. *Cell* **85,** 311–318.

4 Lichtenstein, P. and Svartengren, M. (1997) Genes, environments, and sex: factors of importance in atopic diseases in 7–9-year-old Swedish twins. *Allergy* **52,** 1079–1086.

5 Laitinen, T., Räsänen, M., Kaprio, J., Koskenvuo, M., and Laitinen, L. A. (1998) Importance of genetic factors in adolescent asthma: a population-based twin-family study. *Am. J. Respir. Crit. Care Med.* **157,** 1073–1078.

6 Skadhauge, L. R., Christensen, K., Kyvik, K. O., and Sigsgaard, T. (1999) Genetic and environmental influence on asthma: a population-based study of 11,688 Danish twin pairs. *Eur. Respir. J.* **13,** 8–14.

7 Clarke, J. R., Jenkins, M. A., Hopper, J. L., et al. (2000) Evidence for genetic associations between asthma, atopy, and bronchial hyperresponsiveness: a study of 8- to 18-yr-old twins. *Am. J. Respir. Crit. Care Med.* **162,** 2188–2193.

8 Marsh, D. G., Neely, J. D., Breazeale, D. R., et al. (1994) Linkage analysis of IL4 and other chromosome 5q31.1 markers and total serum immunoglobulin E concentrations. *Science* **264,** 1152–1156.

9 Ober, C. and Hoffjan, S. (2006) Asthma genetics 2006: the long and winding road to gene discovery. *Genes Immun.* **7,** 95–100.

10 Ioannidis, J. P. (2005) Why most published research findings are false. *PLoS Med.* **2,** e124.

11 Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58,** 1347–1363.

12 Daniels, S. E., Bhattacharrya, S., James, A., et al. (1996) A genome-wide search for quantitative trait loci underlying asthma. *Nature* **383,** 247–250.

13 The Collaborative Study on the Genetics of Asthma (CSGA) (1997) A genome-wide search for asthma susceptibility loci in ethnically diverse populations. *Nat. Genet.* **15,** 389–392.

14 Xu, J., Meyers, D. A., Ober, C., et al. (2001) Genomewide screen and identification of gene-gene interactions for asthma-susceptibility loci in three U.S. populations: collaborative study on the genetics of asthma. *Am. J. Hum. Genet.* **68,** 1437–1446.

15 Mathias, R. A., Freidhoff, L. R., Blumenthal, M. N., et al. (2001) Genome-wide linkage analyses of total serum IgE using variance components analysis in asthmatic families. *Genet. Epidemiol.* **20,** 340–355.

16 Huang, S. K., Mathias, R. A., Ehrlich, E., et al. (2003) Evidence for asthma susceptibility genes on chromosome 11 in an African-American population. *Hum. Genet.* **113,** 71–75.

17 Blumenthal, M. N., Langefeld, C. D., Beaty, T. H., et al. (2004) A genome-wide search for allergic response (atopy) genes in three ethnic groups: Collaborative Study on the Genetics of Asthma. *Hum. Genet.* **114,** 157–164.

18 Blumenthal, M. N., Langefeld, C. D., Barnes, K. C., et al. (2006) A genome-wide search for quantitative trait loci contributing to variation in seasonal pollen reactivity. *J. Allergy Clin. Immunol.* **117,** 79–85.

19 Ober, C., Cox, N. J., Abney, M., et al. (1998) Genome-wide search for asthma susceptibility loci in a founder population. The Collaborative Study on the Genetics of Asthma. *Hum. Mol. Genet.* **7,** 1393–1398.

20 Ober, C., Tsalenko, A., Parry, R., and Cox, N. J. (2000) A second-generation genomewide screen for asthma-susceptibility alleles in a founder population. *Am. J. Hum. Genet.* **67,** 1154–1162.

21 Wjst, M., Fischer, G., Immervoll, T., et. al. (1999) A genome-wide search for linkage to asthma. German Asthma Genetics Group. *Genomics* **58,** 1–8.

22 Altmuller, J., Seidel, C., Lee, Y. A., et al. (2005) Phenotypic and genetic heterogeneity in a genome-wide linkage study of asthma families. *BMC Pulm. Med.* **5,** 1–10.
23 Dizier, M. H., Besse-Schmittler, C., Guilloud-Bataille, M., et al. (2000) Genome screen for asthma and related phenotypes in the French EGEA study. *Am. J. Respir. Crit. Care Med.* **162,** 1812–1818.
24 Bouzigon, E., Dizier, M. H., Krahenbuhl, C., et al. (2004) Clustering patterns of LOD scores for asthma-related phenotypes revealed by a genome-wide screen in 295 French EGEA families. *Hum. Mol. Genet.* **13,** 3103–3113.
25 Xu, J., Postma, D. S., Howard, T. D., et al. (2000) Major genes regulating total serum immunoglobulin E levels in families with asthma. *Am. J. Hum. Genet.* **67,** 1163–1173.
26 Xu, J., Bleecker, E. R., Jongepier, H., et al. (2002) Major recessive gene(s) with considerable residual polygenic effect regulating adult height: confirmation of genomewide scan results for chromosomes 6, 9, and 12. *Am. J. Hum. Genet.* **71,** 646–650.
27 Koppelman, G. H., Stine, O. C., Xu, J., et al. (2002) Genome-wide search for atopy susceptibility genes in Dutch families with asthma. *J. Allergy Clin. Immunol.* **109,** 498–506.
28 Meyers, D. A., Postma, D. S., Stine, O. C., et al. (2005) Genome screen for asthma and bronchial hyperresponsiveness: interactions with passive smoke exposure. *J. Allergy Clin. Immunol.* **115,** 1169–1175.
29 Postma, D. S., Meyers, D. A., Jongepier, H., Howard, T. D., Koppelman, G. H., and Bleecker, E. R. (2005) Genomewide screen for pulmonary function in 200 families ascertained for asthma. *Am. J. Respir. Crit. Care Med.* **172,** 446–452.
30 Laitinen, T., Daly, M. J., Rioux, J. D., et al. (2001) A susceptibility locus for asthma-related traits on chromosome 7 revealed by genome-wide scan in a founder population. *Nat. Genet.* **28,** 87–91.
31 Xu, X., Fang, Z., Wang, B., et al. (2001) A genomewide search for quantitative-trait loci underlying asthma. *Am. J. Hum. Genet.* **69,** 1271–1277.
32 Van Eerdewegh, P., Little, R. D., Dupuis, J., et al. (2002) Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* **418,** 426–430.
33 Hakonarson, H., Bjornsdottir, U. S., Halapi, E., et al. (2002) A major susceptibility gene for asthma maps to chromosome 14q24. *Am. J. Hum. Genet.* **71,** 483–491.
34 Haagerup, A., Bjerke, T., Schoitz, P. O., Binderup, H. G., Dahl, R., and Kruse, T. A. (2001) Allergic rhinitis—a total genome-scan for susceptibility genes suggests a locus on chromosome 4q24-q27. *Eur. J. Hum. Genet.* **9,** 945–952.
35 Haagerup, A., Bjerke, T., Schiotz, P. O., Binderup, H. G., Dahl, R., and Kruse, T. A. (2002) Asthma and atopy—a total genome scan for susceptibility genes. *Allergy* **57,** 680–686.
36 Yokouchi, Y., Nukaga, Y., Shibasaki, M., et al. (2000) Significant evidence for linkage of mite-sensitive childhood asthma to chromosome 5q31-q33 near the interleukin 12 B locus by a genome-wide search in Japanese families. *Genomics* **66,** 152–160.

37 Yokouchi, Y., Shibasaki, M., Noguchi, E., et al. (2002) A genome-wide linkage analysis of orchard grass-sensitive childhood seasonal allergic rhinitis in Japanese families. *Genes Immun*. **3,** 9–13.

38 Wang, J. Y., Lin, C. G., Bey, M. S., et al. (2005) Discovery of genetic difference between asthmatic children with high IgE level and normal IgE level by whole genome linkage disequilibrium mapping using 763 autosomal STR markers. *J. Hum. Genet*. **50,** 249–258.

39 Kurz, T., Altmueller, J., Strauch, K., et al. (2005) A genome-wide screen on the genetics of atopy in a multiethnic European population reveals a major atopy locus on chromosome 3q21.3. *Allergy* **60,** 192–199.

40 Ferreira, M. A., O'Gorman, L., Le Souef, P., et al. (2005) Robust estimation of experimentwise P values applied to a genome scan of multiple asthma traits identifies a new region of significant linkage on chromosome 20q13. *Am. J. Hum. Genet*. **77,** 1075–1085.

41 Bu, L. M., Bradley, M., Soderhall, C., Wahlgren, C.F., Kockum, I., and Nordenskjold, M. (2006) Genome-wide linkage analysis of allergic rhinoconjunctivitis in a Swedish population. *Clin. Exp. Allergy* **36,** 204–210.

42 Pillai, S. G., Chiano, M. N., White, N. J., et al. (2006) A genome-wide search for linkage to asthma phenotypes in the genetics of asthma international network families: evidence for a major susceptibility locus on chromosome 2p. *Eur. J. Hum. Genet*. **14,** 307–316.

43 Zhang, Y., Leaves, N. I., Anderson, G. G., et al. (2003) Positional cloning of a quantitative trait locus on chromosome 13q14 that influences immunoglobulin E levels and asthma. *Nat. Genet*. **34,** 181–186.

44 Allen, M., Heinzmann, A., Noguchi, E., et al. (2003) Positional cloning of a novel gene influencing asthma from chromosome 2q14. *Nat. Genet*. **35,** 258–263.

45 Laitinen, T., Polvi, A., Rydman, P., et al. (2004) Characterization of a common susceptibility locus for asthma-related traits. *Science* **304,** 300–304.

46 Nicolae, D., Cox, N. J., Lester, L. A., et al. (2005) Fine mapping and positional candidate studies identify HLA-G as an asthma susceptibility gene on chromosome 6p21. *Am. J. Hum. Genet*. **76,** 349–357.

47 Noguchi, E., Yokouchi, Y., Zhang, J., et al. (2005).Positional identification of an asthma susceptibility gene on human chromosome 5q33. *Am. J. Respir. Crit. Care Med*. **172,** 183–188.

48 Yoshinaka, T., Nishii, K., Yamada, K., et al. (2002) Identification and characterization of novel mouse and human ADAM33s with potential metalloprotease activity. *Gene* **282,** 227–236.

49 Powell, R. M., Wicks, J., Holloway, J. W., Holgate, S. T., and Donna, D. E. (2004) The splicing and fate of ADAM33 transcripts in primary human airways fibroblasts. *Am. J. Respir. Cell Mol. Biol*. **31,** 13–21.

50 Howard, T. D., Postma, D. S., Jongepier, H., et al. (2003) Association of a disintegrin and metalloprotease 33 (ADAM33) gene with asthma in ethnically diverse populations. *J. Allergy Clin. Immunol*. **112,** 717–722.

51 Lind, D. L., Choudhry, S., Ung, N., et al. (2003) ADAM33 is not associated with asthma in Puerto Rican or Mexican populations. *Am. J. Respir. Crit. Care Med.* **168,** 1312–1316.
52 Raby, B. A., Silverman, E. K., Kwiatkowski, D. J., Lange, C., Lazarus, R., and Weiss, S. T. (2004) ADAM33 polymorphisms and phenotype associations in childhood asthma. *J. Allergy Clin. Immunol.* **113,** 1071–1080.
53 Lee, J. H., Park, H. S., Park, S. W., et al. (2004) ADAM33 polymorphism: association with bronchial hyper-responsiveness in Korean asthmatics. *Clin. Exp. Allergy* **34,** 860–865.
54 Jongepier, H., Boezen, H. M., Dijkstra, A., et al. (2004) Polymorphisms of the ADAM33 gene are associated with accelerated lung function decline in asthma. *Clin. Exp. Allergy* **34,** 757–760.
55 Werner, M., Herbon, N., Gohlke, H., et al. (2004) Asthma is associated with single-nucleotide polymorphisms in ADAM33. *Clin. Exp. Allergy* **34,** 26–31.
56 Simpson, A., Maniatis, N., Jury, F., et al. (2005) Polymorphisms in a disintegrin and metalloprotease 33 (ADAM33) predict impaired early-life lung function. *Am. J. Respir. Crit. Care Med.* **172,** 55–60.
57 Blakey, J., Halapi, E., Bjornsdottir, U. S., et al. (2005) Contribution of ADAM33 polymorphisms to the population risk of asthma. *Thorax* **60,** 274–276.
58 van Diemen, C. C., Postma, D. S., Vonk, J. M., Bruinenberg, M., Schouten, J. P., and Boezen, H. M. (2005) A disintegrin and metalloprotease 33 polymorphisms and lung function decline in the general population. *Am. J. Respir. Crit. Care Med.* **172,** 329–333.
59 Jang, N., Stewart, G., and Jones, G. (2005) Polymorphisms within the PHF11 gene at chromosome 13q14 are associated with childhood atopic dermatitis. *Genes Immun.* **6,** 262–264.
60 Qi, S. Y., Riviere, P. J., Trojnar, J., Junien, J. L., and Akinsanya, K. O. (2003) Cloning and characterization of dipeptidyl peptidase 10, a new member of an emerging subgroup of serine proteases. *Biochem. J.* **373,** 179–189.
61 Zagha, E., Ozaita, A., Chang, S. Y., et al. (2005) DPP10 modulates Kv4-mediated A-type potassium channels. *J. Biol. Chem.* **280,** 18,853–18,861.
62 Leaves, N. I., Bhattacharyya, S., Wiltshire, S., and Cookson, W. O. (2002) A detailed genetic map of the chromosome 7 bronchial hyper-responsiveness locus. *Eur. J. Hum. Genet.* **10,** 177–182.
63 Melen, E., Bruce, S., Doekes, G., et al. (2005) Haplotypes of G protein-coupled receptor 154 are associated with childhood allergy and asthma. *Am. J. Respir. Crit. Care Med.* **171,** 1089–1095.
64 Kormann, M. S., Carr, D., Klopp, N., et al. (2005) G-Protein-coupled receptor polymorphisms are associated with asthma in a large German population. *Am. J. Respir. Crit. Care Med.* **171,** 1358–1362.
65 Veal, C. D., Reynolds, N. J., Meggitt, S. J., et al. (2005) Absence of association between asthma and high serum immunoglobulin E associated GPRA haplotypes and adult atopic dermatitis. *J. Invest. Dermatol.* **125,** 399–401.

66 Soderhall, C., Marenholz, I., Nickel, R., et al. (2005) Lack of association of the G protein-coupled receptor for asthma susceptibility gene with atopic dermatitis. *J. Allergy Clin. Immunol.* **116,** 220–221.
67 Reinscheid, R. K., Xu, Y. L., Okamura, N., et al. (2005) Pharmacological characterization of human and murine neuropeptide s receptor variants. *J. Pharmacol. Exp. Ther.* **315,** 1338–1345.
68 Reinscheid, R. K. and Xu, Y. L. (2005) Neuropeptide S and its receptor: a newly deorphanized G protein-coupled receptor system. *Neuroscientist* **11,** 532–538.
69 Gupte, J., Cutler, G., Chen, J. L., and Tian, H. (2004) Elucidation of signaling properties of vasopressin receptor-related receptor 1 by using the chimeric receptor approach. *Proc. Natl. Acad. Sci. USA* **101,** 1508–1513.
70 Vendelin, J., Pulkkinen, V., Rehn, M., et al. (2005) Characterization of GPRA, a novel G protein-coupled receptor related to asthma. *Am. J. Respir. Cell Mol. Biol.* **33,** 262–270.
71 Pulkkinen, V., Haataja, R., Hannelius, U., et. al. (2006) G protein-coupled receptor for asthma susceptibility associates with respiratory distress syndrome. *Ann. Med.* **38,** 357–366.
72 Feng, Y., Hong, X., Wang, L., et al. (2006) G protein-coupled receptor 154 gene polymorphism is associated with airway hyperresponsiveness to methacholine in a Chinese population. *J. Allergy Clin. Immunol.* **117,** 612–617.
73 Elias, J. A., Zheng, T., Lee, C. G., et al. (2003) Transgenic modeling of interleukin-13 in the lung. *Chest* **123,** 339S.
74 Liu, X., Nickel, R., Beyer, K., et al. (2000) An IL13 coding region variant is associated with a high total serum IgE level and atopic dermatitis in the German multicenter atopy study (MAS-90). *J. Allergy Clin. Immunol.* **106,** 167–170.
75 Graves, P. E., Kabesch, M., Halonen, M., et al. (2000) A cluster of seven tightly linked polymorphisms in the IL-13 gene is associated with total serum IgE levels in three populations of white children. *J. Allergy Clin. Immunol.* **105,** 506–513.
76 Heinzmann, A., Mao, X. Q., Akaiwa, M., et al. (2000) Genetic variants of IL-13 signalling and human asthma and atopy. *Hum. Mol. Genet.* **9,** 549–559.
77 Kauppi, P., Lindblad-Toh, K., Sevon, P., et al. (2001) A second-generation association study of the 5q31 cytokine gene cluster and the interleukin-4 receptor in asthma. *Genomics* **77,** 35–42.
78 Tsunemi, Y., Saeki, H., Nakamura, K., et al. (2002) Interleukin-13 gene polymorphism G4257A is associated with atopic dermatitis in Japanese patients. *J. Dermatol. Sci.* **30,** 100–107.
79 Wang, M., Xing, Z. M., Lu, C., et al. (2003) A common IL-13 Arg130Gln single nucleotide polymorphism among Chinese atopy patients with allergic rhinitis. *Hum. Genet.* **113,** 387–390.
80 DeMeo, D. L., Lange, C., Silverman, E. K., et al. (2002) Univariate and multivariate family-based association analysis of the IL-13 ARG130GLN polymorphism in the Childhood Asthma Management Program. *Genet. Epidemiol.* **23,** 335–348.

81 He, J. Q., Chan-Yeung, M., Becker, A. B., et al. (2003) Genetic variants of the IL13 and IL4 genes and atopic diseases in at-risk children. *Genes Immun*. **4,** 385–389.
82 Heinzmann, A., Jerkic, S. P., Ganter, K., et al. (2003) Association study of the IL13 variant Arg110Gln in atopic diseases and juvenile idiopathic arthritis. *J. Allergy Clin. Immunol*. **112,** 735–739.
83 Vladich, F. D., Brazille, S. M., Stern, D., Peck, M. L., Ghittoni, R., and Vercelli, D. (2005) IL-13 R130Q, a common variant associated with allergy and asthma, enhances effector mechanisms essential for human allergic inflammation. *J. Clin. Invest*. **115,** 747–754.
84 Oguma, T., Palmer, L. J., Birben, E., Sonna, L. A., Asano, K., and Lilly, C. M. (2004) Role of prostanoid DP receptor variants in susceptibility to asthma. *N. Engl. J. Med*. **351,** 1752–1763.
85 Hysi, P., Kabesch, M., Moffatt, M. F., et al. (2005) NOD1 variation, immunoglobulin E and asthma. *Hum. Mol. Genet*. **14,** 935–941.

# 16

# Identifying Susceptibility Variants for Type 2 Diabetes

**Eleftheria Zeggini and Mark I. McCarthy**

**Summary**

The etiology of type 2 diabetes (T2D) is complex and remains poorly understood. Differences in individual susceptibility to this condition reflect the action of multiple variants, each of which confers a modest effect, and their interactions with a variety of environmental exposures. Several complementary approaches to the identification of the etiological variants have been adopted, though, for all, association analyses provide the final common pathway. The genes and/or chromosomal regions studied have been selected on the basis of their presumed biological relevance to diabetes, known involvement in monogenic forms, or animal models of the condition and/or signals arising from whole-genome linkage scans. These association studies have featured a wide variety of designs and analytical approaches, but reliable biological insights have been few, largely because of difficulties in obtaining reproducible findings. However, in recent years, several examples of robustly replicated associations have emerged, largely as a result of an emphasis on the need for improved power and more appropriate analysis and interpretation. New strategies for the large-scale identification of T2D susceptibility variants are now becoming possible, including the prospect of genuine genome-wide association scans, but caution in their design, analysis, and interpretation remains essential.

**Key Words:** Type 2 diabetes; linkage disequilibrium; association; candidate gene; genetic analysis; genome-wide; study design; replication.

## 1. Type 2 Diabetes as a Complex Trait

Type 2 diabetes (T2D) is a serious disease with a rapidly increasing prevalence. It represents a growing public health issue for societies across the globe *(1)*. Although recent secular trends in diabetes prevalence clearly reflect the pervasive effects of adverse environmental exposures (notably the global move toward lifestyles characterized by less exercise and more food), this occurs on the background of substantial individual differences in the inherited

From: *Methods in Molecular Biology, vol. 376: Linkage Disequilibrium and Association Mapping: Analysis and Applications* Edited by: A. R. Collins © Humana Press Inc., Totowa, NJ

predisposition to disease. Evidence for this genetic component has come from a wide range of classical approaches including studies of twins, migrant and admixed populations, and of familial aggregation *(2)*. All indications are that this genetic component is (with the exception of a small proportion of families with monogenic or syndromic forms of diabetes *[3]*) owing to the combined action of many variants, each of modest effect.

As with most complex traits, efforts to identify the specific variants influencing T2D susceptibility have proceeded relatively slowly, at least until recently *(4)*. Indeed, it can be argued that most of the several thousand published studies of T2D genetics have generated more in the way of heat than light. The inconsistency of association findings has been a major problem, with exciting reports of positive signals typically followed by a series of follow-up studies that fail to confirm the original findings *(5–7)*. Of course, this is by no means unique to T2D. In retrospect, it is all too easy to see that such a pattern is entirely in line with expectation given that most studies in this area have suffered from a "deadly triad" of methodological inadequacies: low power (inadequate sample size and/or inappropriate marker selection), low prior odds (i.e., selection of candidates with low absolute probabilities of involvement in disease pathogenesis), and use of overly liberal thresholds for declaring significance (including a reluctance to account properly for the inflated type-1 error associated with the testing of multiple hypotheses). Such a combination is guaranteed to result in most positive (i.e., nominally significant) associations being false *(8)*.

Over the past 5 yr, in the face of a growing understanding of these failings, T2D research has made considerable progress in the development and application of more robust strategies for disease gene identification *(7)*. A key feature has been the move toward large collaborative endeavours, in the realization that the scale of the challenge requires the pooling of disparate expertise and access to large-scale clinical resources *(7)*.

In this chapter, we reinforce some of these themes, using selected examples of the application of linkage disequilibrium (LD)-based mapping approaches to the identification of T2D-susceptibility genes.

## 2. An Overview of Strategies for T2D-Susceptibility Gene Identification

As exhaustive genotyping of human genome sequence variation is not yet feasible, almost all studies to date have focused on selected candidate genes or chromosomal regions. Selection relies on evidence felt to indicate that the gene or region of interest is particularly likely to harbor variants influencing diabetes susceptibility. Such evidence can come from a variety of sources: a prior genome-wide linkage scan in families segregating T2D, for example; or known involvement in the causation of monogenic forms of diabetes; or, simply,

a perceived match between the known (or presumed) function of the gene in question, and the known (or presumed) pathophysiology of the disease *(7)*.

## 2.1. Candidate Gene Studies

### 2.1.1. Limitations of Candidate Gene Studies

The published literature contains reports of T2D-association studies involving several hundred different candidate genes. Very few of these studies have been adequately powered, and even fewer have attempted to survey variation across the gene in any systematic or comprehensive fashion. It is not altogether surprising, therefore, that relatively few true susceptibility variants have emerged from this approach *(9)*.

One of the weakest aspects of the candidate gene approach has been the process of selecting the genes in the first place. Because we know so little about the causal mechanisms (indeed, this is one of the chief justifications for the genetics approach in the first place), the selection of candidates with strong prior odds for involvement in diabetes pathogenesis is, at best, an imprecise art *(7)*. Empirically, the most powerful strategy to date (though the number of successes is too small to generalize too far) has been to select candidates for which there is incontrovertible evidence that perturbation of function alters glucose homeostasis in man. Such insights may come through "experiments of nature," that is, rare mutations that result in extreme phenotypes of either hyperglycemia (such as monogenic and syndromic forms of diabetes) or hypoglycemia (hyperinsulinemia of infancy) *(3)*. An alternative route to success derives from an understanding of the mechanisms through which antidiabetic drugs operate. In each case, this prior evidence identifies genes for which a causal link to diabetes has been established rather than merely postulated. Besides, such studies indicate that there is limited capacity for redundant or compensatory mechanisms to substitute for changes in expression or function of the particular gene product.

### 2.1.2. Successes From Monogenic Disease Genes

Several of the genes for which rare mutations have been implicated in the causation of monogenic and syndromic forms of diabetes *(3)* have also been shown to contain commoner variants, with less severe effects on gene function, which influence susceptibility to multifactorial T2D or related traits. These include the gene encoding hepatocyte nuclear factor-4 α (*HNF4A*), known to be one of the causes of maturity onset diabetes of the young (MODY) *(3)*; common variants around the P2 promoter of this gene have been shown to be associated with T2D in recent studies of Ashkenazi, Finnish, and UK subjects, among others *(10–12)*. Similarly, common variants upstream of the glucokinase

gene (mutations in which cause another subtype of MODY) have been shown to have a modest effect on fasting glucose levels ***(13)***.

### 2.1.3. Successes From Drug Targets

Genes encoding known targets of antidiabetic drugs have also proved fertile territory. Peroxisome proliferator-activated receptor gamma (PPARG) is the target of the thiazolidinedione drugs and known to play a crucial role in adipocyte differentiation and function. Over the past few years, a coding polymorphism in the *PPARG* gene (P12A) has emerged as the "poster-child" of T2D genetics ***(14)***, with extensive replication across many different studies. Remarkably, the risk allele (Pro) is also the commoner allele (~80%), so that, although the relative risk is modest (~1.2), the population-attributable risk is high (>20%) ***(14)***.

A second example of "reverse therapeutics" lies in the membrane $K_{ATP}$ channel expressed in the pancreatic β-cell. This channel, which plays a critical role in glucose-stimulated insulin secretion (and hence in glucose homeostasis) is made up of two components ***(15)***. One, the sulfonylurea receptor, is the target of a group of drugs (sulfonylureas) used to restore failing insulin secretion in T2D, and is encoded by the gene *ABCC8*. The other is the inwardly rectifying potassium channel, Kir6.2, transcribed from the gene *KCNJ11*. Though they have no homology, these two genes map to adjacent positions on chromosome 11.

The chronology of association studies on these particular genes is illustrative. The initial focus on a pair of *ABCC8* polymorphisms (in exons 16 and 18) generated a checkered pattern of positive and negative association ***(16–21)***. Many of these studies were small, and the evidence contradictory. There was little logic to the selection of these particular variants (out of the many thousands mapping to the region). When, subsequently, attention turned to *KCNJ11*, most studies focused on a single coding variant, E23K. The first small-scale studies of E23K failed to reveal any evidence of association ***(14,21,22)***. However, as sample sizes increased and with meta-analyses of published data ***(24–30)***, a consistent pattern of association (between the K allele and T2D) emerged. As with P12A in *PPARG,* the relative risk is modest (between 1.15 and 1.2) ***(25)*** which, of course, helps to explain why the association was missed in studies involving a few hundred case–control pairs.

But is E23K actually the etiological variant, and does it alone explain all of the susceptibility attributable to the *KCNJ11*/*ABCC8* region? At this stage, it is too early to be sure. Though there are in vitro data to suggest that the polymorphism influences channel function ***(31)***, not all groups have been able to detect such effects ***(22)***. Efforts to define the etiological variant are also frustrated by extensive LD across the region (although this does not reach as far as the exon

16 and 18 variants in *ABCC8*). For example, in Europeans at least, the E23K variant is in very tight LD with a nonsynonymous coding variant in exon 33 of *ABCC8* (A1369S) ***(27)***, and the statistical evidence does not allow the etiological credentials of these two to be distinguished.

The *ABCC8*/*KCNJ11* story highlights the importance of designing well-powered case–control studies and the usefulness of meta-analyses in deconvoluting conflicting evidence for association in the literature. Additionally, it illustrates the need for exhaustive assessments of regional variation, even after robust association signals have been obtained. Until such a comprehensive study of the *ABCC8*/*KCNJ11* region has been performed, it is simply not possible to be certain that these genes do not contain additional common susceptibility variants. It is entirely possible that such variants have a far more dramatic effect on diabetes risk at the population level than those already examined.

### *2.1.4. Issues of Marker Selection in Candidate Gene Studies*

Until recently, the high costs of genotyping, combined with an incomplete catalog of human genome sequence variation, have restricted most candidate gene studies in T2D to a small set of variants within each gene. The early successes in *PPARG* and *KCNJ11* were, in some measure, because of the fact that the association signals involved nonsynonymous coding variants. The strong presumption that such variants are likely to have functional repercussions made them obvious targets for study.

However, such selective approaches have obvious limitations ***(7)***. In recent years, enhanced appreciation of the patterns and distribution of human genome sequence variation provided by the International Haplotype Map Project ***(32)***, has made it possible to tackle candidate genes (and regions, *see* **Subheading 2.2.**) in more systematic and comprehensive fashion. Not only it is possible to choose an increasingly complete set of variants within a given gene, it is also possible to use information on local LD to select a subset of those variants to maximize the efficiency of the genotyping effort ***(33)***. As is well known, such a "tagging" approach is predicated on the assumption that a significant proportion of etiological variation is attributable to common variants of modest effect. If so, such variants will be captured by the subset of genotyped tags, whether or not they are directly typed. In contrast, the capacity of common variant tags to detect rare variants contributing to susceptibility is, under most circumstances, poor ***(34)***. The relative merits of these two alternative views of complex trait susceptibility have been hotly contested in recent years ***(35)***. Novel approaches to deep resequencing will be required to allow researchers to recover the full spectrum of allelic diversity, and gain insights into how etiological variation is distributed across that spectrum.

### 2.2. Candidate Region Studies

#### 2.2.1. Replicated Linkage Signals as a Source of Candidate Regions

The classical complement to the candidate gene approach to gene identification has lain in positional-cloning strategies. Typically, genome-wide linkage data (from man or rodent models) has been used to define intervals felt likely to contain one or more susceptibility variants. To date, over 30 linkage scans for T2D have been completed *(36)*, each with its own particular configuration of study population, sample type, marker set, and analytical strategy. Failure to identify consistent replication across these studies certainly does indicate that there is no single overwhelming locus for T2D susceptibility (akin to human leukocyte antigen in type 1 diabetes). However, given that most of these studies were underpowered to detect genes of more modest effect, the implicit heterogeneity in study design has impacted on the clarity with which one can use such data to obtain a clear view of the susceptibility landscape of T2D.

Notwithstanding these limitations, several chromosomal regions have emerged sufficiently often from T2D genome-wide scans to excite interest in further analysis. Typically, such regions are large (between 10 and 40 Mb) with somewhat nebulous boundaries, reflecting the poor localization capacity of linkage for genes of modest effect *(37)*. Until recently, the size of these regions has been a major limitation and few groups have had the resources to tackle them in systematic fashion, preferring instead to pick off biologically attractive positional candidates one-by-one.

#### 2.2.2. Successes From the Positional Candidate Approach

This approach has met with some success. Within the region of replicated linkage on chromosome 3q, there is compelling evidence that variants within the genes encoding adiponectin (*ADIPOQ)* and PSARL (a mitochondrial rhomboid protease) contribute to susceptibility to diabetes *(38–40)*, though it is not clear that they explain the linkage signal.

On chromosome 20 *(41–46)*, *HNF4A* was already an obvious positional candidate given its causative role in MODY (*see* **Subheading 2.1.2.**), and a role for variation in and around the P2 promoter has been demonstrated in several recent studies *(10–12)*.

A second gene within this region has recently been implicated in T2D susceptibility. *PTPN1* codes for the ubiquitously expressed protein tyrosine phosphatase 1B (PTP1B). PTP1B is known to be responsible for the dephosphorylation of phosphotyrosine residues within the activated insulin receptor, and PTP1b-deficient mice show altered insulin sensitivity and propensity to diet-induced obesity *(47,48)*. Early association studies at this gene typed nonoverlapping sets of single-nucleotide polymorphisms (SNPs) and generated

conflicting results ***(49–52)***. In the case of *PTPN1*, the usual recommendations for more extensive evaluation of variation across the region in larger sample sets have only added to the confusion. Two groups ***(53,54)*** found that *PTPN1* resides in a region of low haplotype diversity and reported significant associations of several noncoding SNPs with T2D (odds ratios ~1.3), though without identifying any single etiological variant. In contrast, Florez et al. ***(55)***, in the largest case–control study of *PTPN1* variation to date, failed to identify any association with disease whatsoever. The list of possible explanations for these discrepant findings is long, but includes differences in genetic background (i.e., effects of modifier genes) between the study populations. The *PTPN1* example illustrates the problem of conflicting, variably powered, incomplete variation surveys in the genetic association literature and, furthermore, highlights the potential contribution of gene–gene interactions to the etiological architecture of complex traits.

Examples such as this were the motivation for the recent recommendation that the gene should provide the unit of replication for genetic studies; in other words, the ability to align different association studies onto the same haplotypic scaffold represents a powerful tool for synthesizing results from studies that use nonoverlapping sets of markers ***(56)***.

### 2.2.3. Indirect LD Mapping Across Regions of Interest

The positional candidate strategy suffers from the generic limitations of any approach that is dependent on the prior assessment of the biological plausibility that a given gene is involved in disease pathogenesis, with all the imprecision that implies. Advances in genotyping technology have made a more systematic approach to such regions of interest possible, through large-scale indirect LD mapping.

As with candidate gene studies, the selection of the variants to be typed in such a study can be approached in several ways. Some groups have adopted a gene-centric strategy, investing a disproportionate amount of their genotyping effort into the protein-coding sequence. However, increasing evidence for a role played by conserved nongenic regions ***(57)*** and the incomplete characterization of the human genome gene content have highlighted the advantages of selection strategies that are less reliant on our current, rather rudimentary, ability to map function onto the genome sequence. The HapMap Project has made it possible to use the patterns of LD within such regions to define parsimonious SNP sets that provide efficient capture of common variation ***(32)***. Combinatorial strategies, ensuring common variation capture while enriching maps with potentially functional SNPs in candidate genes or conserved regions, are also beginning to emerge.

It is also worth emphasizing that the parameters of indirect LD mapping within a region of interest are likely to differ significantly from those that might

apply in a genome-wide association scan. Given that the linkage signal may reflect only a single etiological variant, researchers are likely to want to capture as much variation within the region as is possible. This may extend not only to more dense SNP sets, but also to more strenuous efforts to capture less common variants (those in the range of 2 to 5%). Linkage analysis is best powered for the detection of rare, penetrant alleles, and it is likely that the etiological variants within linkage signals are disproportionately drawn from that part of the allelic frequency spectrum. On the other hand, the opportunity to draw cases from multiplex families enriched for the linkage signal offers a considerable boost to power, because such ascertainment can appreciably increase the susceptibility allele frequency within those cases ***(58)***.

The earliest example of an indirect mapping approach was that applied to a region of chromosome 2q that had emerged from a genome-wide linkage scan for T2D conducted in Mexican Americans ***(59)***. Positional cloning efforts within this region ultimately led the investigators to *CAPN10,* a gene encoding a ubiquitously expressed protease, calpain 10. Within this gene, a haplotype consisting of three intronic SNPs was found to be associated with T2D both in the original Mexican American samples as well as a Finnish population ***(60)***. Notably, variation within this gene appeared to account for almost all the evidence for linkage on chromosome 2q.

Subsequent association studies across a range of populations produced a mixture of confirmatory ***(61–65)*** and conflicting ***(60,66–74)*** results. However, in a recent large-scale meta-analysis that incorporated all available (published and unpublished data), Weedon and colleagues ***(75)*** did find evidence (albeit at a relatively undramatic significance value) for a modest effect (odds ratio ~1.2) on T2D risk for SNP44, a variant located in a *CAPN10* enhancer element. SNP44 is perfectly correlated with a missense-coding polymorphism (T504A) and two further *CAPN10* SNPs in the 5′ UTR of the gene. Unequivocal identification of the disease-causing variant has, therefore, not yet been achieved. Studies of calpain biology, prompted by these findings, have suggested that calpain 10 may be involved in insulin secretion ***(76,77)***.

Similar efforts, boosted by recent advances in genome informatics and genotyping capacity, are now underway in several other regions of replicated linkage, including chromosomes 1q and 10q. Chromosome 1q constitutes one of the best replicated regions of T2D linkage, with positive signals in several human ***(78–84)***, as well as rodent studies ***(85–88)***. The 25-Mb interval defined by these linkage signals is the subject of a high-density, large-scale indirect LD mapping effort by an international consortium. Several features of this study are worth mentioning. First, case–control and family-based association analyses are being conducted in eight different populations in parallel, providing opportunities for internal replication as well as the potential to make use of differences in the LD

structure between them to assist with fine-mapping. Second, most of the cases included in the study come from families included in the original linkage studies. This ensures several levels of enrichment for disease-susceptibility alleles (for familial T2D, as well as linkage to 1q) that should significantly augment the power of the study, as well as allowing tests of the extent to which association signals can explain the original linkage findings *(58)*. Third, the selection of markers has been based on exhaustive coverage of common variation, combined with the addition of all known nonsynonymous-coding SNPs and variants of potential functional importance. All of these maneuvers should enhance the prospects for success, and indeed several replicated association signals have emerged *(89)*. Nonetheless, such studies remain major undertakings, typically calling for between 30 and 100 million genotypes.

### 3. Toward Genome-Wide Association

The landscape of genetic association studies for T2D (and other complex traits) is set to be transformed over the coming year, with the advent of truly genome-wide association scans. Access to array-based reagents that allow massively parallel genotyping (between 250,000 and 1 million SNPs per assay), combined with the insights into the haplotypic architecture of human populations provided by the HapMap, provide the basis for a global view of the association architecture of T2D. As a result of phase II HapMap, it is now possible (for some populations at least) to select a set of 250,000 "tag" SNPs capable of capturing a high proportion (over 80%) of common variants across the genome. These advances promise to render arguments about the relative merits of different approaches (positional vs candidate) somewhat redundant.

The pace of technical development in this area has largely outstripped equivalent analytical advances, and there is an urgent need for resolution of a variety of design, analysis, and interpretation issues *(90–95)* if we are to maximize the reliability of the insights forthcoming from these high-profile (and expensive) studies. Principal among them will be the requirement to develop experimental and analytical strategies that preserve power in the face of the enormous multiple-testing issue that is implicit in such a design. Replication across multiple scans is likely to play a major role here. It is also worth emphasizing that the overall success of such genome-wide association experiments (that is, the proportion of the total set of etiological variants recovered) is likely to depend on the, as yet unknown, relative contributions of common and rare variation to disease susceptibility.

Several genome-wide association scans for T2D are now underway, with a collective sample size of over 10,000 case–control samples (preliminary results can be expected during 2006). Although it is easy to become enthused by the potential offered by such a massive increase in genotyping capacity, it is

important not to underestimate the prestigious informatics, statistical, and logistical challenges which such studies will pose.

## 4. Conclusion

T2D genetics is a fast-evolving field. Progress in genome- and system-wide characterization through transcriptomics, metabolomics, and proteomics will undoubtedly inform researchers in their quest for better-supported candidate genes and biological pathways. Well-planned, powerful indirect LD-based approaches to disease gene identification offer new prospects for target detection. However, many aspects of optimal association study design, analysis, and interpretation remain unresolved.

## References

1 Zimmet, P., Alberti, K. G. M. M., and Shaw, J. (2001) Global and societal implications of the diabetes epidemic. *Nature* **414,** 782–787.
2 Gloyn, A. L. and McCarthy, M. I. (2001) The genetics of type 2 diabetes. *Best Pract. Res. Clin. Endocrinol. Metab.* **15,** 293–308.
3 Stride, A. and Hattersley, A. T. (2002) Different genes, different diabetes: lessons from maturity-onset diabetes of the young. *Ann. Med.* **23,** 207–216.
4 McCarthy, M. I. (2004) Progress in defining the molecular basis of type 2 diabetes through susceptibility gene identification. *Hum. Mol. Genet.* **13,** R33–R41.
5 Ioannidis, J. P. A., Ntzani, E. E., Trikalinos, T. A., and Contopoulos-Ioannidis, J. G. (2001) Replication validity of genetic association studies. *Nat. Genet.* **29,** 306–309.
6 Ioannidis, J. P. A., Trikalinos, T. A, and Ntzani, E. E., Contopoulos-Ioannidis, J. G. (2003) Genetic associations in large versus small studies: an empirical assessment. *Lancet* **361,** 567–571.
7 Hattersley, A. T. and McCarthy, M. I. (2005) A question of standards: what makes a good genetic association study? *Lancet* **366,** 1315–1323.
8 Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., and Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl. Cancer Inst.* **96,** 434–442.
9 Glazier, A. M., Nadeau, J. H., and Aitman, T. J. (2002) Finding genes that underlie complex traits. *Science* **298,** 2345–2349.
10 Love-Gregory, L., Wasson, J., Ma, J., et al. (2004) A common polymorphism in the upstream promoter region of the hepatocyte nuclear factor-4 gene on chromosome 20q is associated with type 2 diabetes and appears to contribute to the evidence for linkage in an Ashkenazi Jewish population. *Diabetes* **53,** 1134–1140.
11 Silander, K., Mohlke, K. L., Scott, L. J., et al. (2004) Genetic variation near the hepatocyte nuclear factor-4 gene predicts susceptibility to type 2 diabetes. *Diabetes* **53,** 1141–1149.
12 Weedon, M. N., Owen, K. R., Shields, B., et al. (2004) Common variants of the hepatocyte nuclear factor-4alpha P2 promoter are associated with type 2 diabetes in the UK population. *Diabetes* **53,** 3002–3006.

13 Weedon, M. N., Frayling, T. M., Shields, B., et al. (2005) Genetic regulation of birth weight and fasting glucose by a common polymorphism in the islet cell promoter of the glucokinase gene. *Diabetes* **54,** 576–581.
14 Altshuler, D., Hirschhorn, J. N., Klannemark, M., et al. (2000) The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* **26,** 76–80.
15 Gloyn, A. L., Pearson, E. R., Antcliff, J. F., et al. (2004) Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. *N. Engl. J. Med.* **350,** 1838–1849.
16 Inoue, H., Ferrer, J., Welling, C. M., et al. (1996) Sequence variants in the sulfonylurea receptor (SUR) gene are associated with NIDDM in Caucasians. *Diabetes* **45,** 825–831.
17 't Hart, L. M., De Knijff, P., Dekker, J. M., et al. (1999) Variants in the sulphonylurea receptor gene: association of the exon 16 –3t variant with type II diabetes mellitus in Dutch Caucasians. *Diabetologia* **42,** 617–620.
18 Rissanen, J., Markkanen, A., Kärkkäinen, P., et al. (2000) Sulfonylurea receptor 1 gene variants are associated with gestational diabetes and type 2 diabetes but not with altered secretion of insulin. *Diabetes Care* **23,** 70–73.
19 Meirhaeghe, A., Helbecque, N., Cottel, D., et al. (2001) Impact of sulfonylurea receptor 1 genetic variability on non-insulin-dependent diabetes mellitus prevalence and treatment: a population study. *Am. J. Med. Genet.* **101,** 4–8.
20 Hani, E. H., Clement, K., Velho, G., et al. (1997) Genetic studies of the sulfonylurea receptor gene locus in NIDDM and in morbid obesity among French Caucasians. *Diabetes* **46,** 688–694.
21 Hansen, T., Echwald, S. M., Hansen, L., et al. (1998) Decreased tolbutamide-stimulated insulin secretion in healthy subjects with sequence variants in the high-affinity sulfonylurea receptor gene. *Diabetes* **47,** 598-605.
22 Sakura, H., Wat, N., Horton, V., Millns, H., Turner, R. C., and Ashcroft, F. M. (1996) Sequence variations in the human Kir6.2 gene, a subunit of the beta-cell ATP-sensitive K-channel: no association with NIDDM in white Caucasian subjects or evidence of abnormal function when expressed in vitro. *Diabetologia* **39,** 1233–1236.
23 Inoue, H., Ferrer, J., Warren-Perry, M., et al. (1997) Sequence variants in the pancreatic islet b-cell inwardly rectifying K+ channel Kir6.2 (Bir) gene: identification and lack of role in caucasian patients with NIDDM. *Diabetes* **46,** 502–507.
24 Hani, E. H., Boutin, P., Durand, E., et al. (1998) Missense mutations in the pancreatic islet beta cell inwardly rectifying $K^+$channel gene (*KIR6.2/BIR*): a meta-analysis suggests a role in the polygenic basis of type II diabetes mellitus in Caucasians. *Diabetologia* **41,** 1511–1515.
25 Gloyn, A. L., Weedon M. N., Owen, K. R., et al. (2003) Large scale association studies of variants in genes encoding the pancreatic beta-cell K-ATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with increased risk of type 2 diabetes. *Diabetes* **52**, 568–572.

26 Barroso, I., Luan, J., Middelberg, R. P., et al. (2003) Candidate gene association study in type 2 diabetes indicates a role for genes involved in beta-cell function as well as insulin action. *PLoS Biol*. **1,** E20.
27 Florez, J. C., Burtt, N., de Bakker, P. I., et al. (2004) Haplotype structure and genotype-phenotype correlations of the sulfonylurea receptor and the islet ATP-sensitive potassium channel gene region. *Diabetes* **53,** 1360–1368.
28 Nielsen, E. D., Hansen, L., Carstensen, B., et al. (2003). The E23K variant of Kir6.2 associates with impaired post-OGTT serum insulin response and increased risk of type 2 diabetes. *Diabetes* **52,** 573–577.
29 Love-Gregory, L., Wasson, K., Lin, J., Skolnick, G., Suarez, B., and Permutt, M. A. (2002) An E23K single nucleotide polymorphism in the islet ATP-sensitive potassium channel gene (Kir6.2) contributes as much to the risk of type II diabetes in Caucasians as the PPARγ Pro12Ala variant. *Diabetologia* **45,** 136–137.
30 van Dam, R. M., Hoebee, B., Seidell, J. C., Schaap, M. M., de Bruin, T. W., and Feskens, E. J. (2005) Common variants in the ATP-sensitive K+ channel genes KCNJ11 (Kir6.2) and ABCC8 (SUR1) in relation to glucose intolerance: population-based studies and meta-analyses. *Diabet Med*. **22,** 590–598.
31 Schwanstecher, C., Neugebauer, B., Schulz, M., and Schwanstecher, M. (2002) The common single nucleotide polymorphism E23K in KIR6.2 sensitizes pancreatic β-cell ATP-sensitive potassium channels toward activation through nucleoside diphosphonates. *Diabetes* **51,** S363–S367.
32 The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature* **437,** 1299–1320.
33 Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet*. **74,** 106–120.
34 Zeggini, E., Rayner, W., Morris, A., et al. (2005) HapMap sample size and tagging SNP performance: an evaluation in large-scale empirical and simulated data sets. *Nat Genet*. **37,** 1320–1322.
35 Palmer, L. J. and Cardon, L. R. (2005) Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* **366,** 1223–1234.
36 McCarthy, M. I. (2003) Growing evidence for diabetes susceptibility genes from genome scan data. *Curr. Diab. Rep*. **3,** 159–167.
37 Roberts, S. B., MacLean, C. J., Neale, M. C., Eaves, L. J., and Kendler, L. S. (1999) Replication of linkage studies of complex traits: an examination of variance in location estimates. *Am. J. Hum. Genet*. **65,** 876–884.
38 Gibson, F. and Froguel, P. (2004) Genetics of the *APM1* locus and its contribution to type 2 diabetes susceptibility in French Caucasians. *Diabetes* **53,** 2977–2983.
39 Vasseur, F., Helbecque, N., Dina, C., et al. (2002) Single nucleotide polymorphism haplotypes in the both proximal promoter and exon 3 of the APM1 gene modulate adipocyte-secreted adiponectin hormone levels and contribute to the genetic risk for type 2 diabetes in French Caucasians. *Hum. Mol. Genet*. **11,** 2607–2614.

40 Walder, K., Kerr-Bayles, L., Civitarese, A., et al. (2005) The mitochondrial rhomboid protease PSARL is a new candidate gene for type 2 diabetes. *Diabetologia* **48,** 459–468.

41 Bowden, D. W., Sale, M., Howard, T. D., et al. (1997) Linkage of genetic markers on human chromosomes 20 and 12 to NIDDM in Caucasian sib pairs with a history of diabetic nephropathy. *Diabetes* **46,** 882–886.

42 Ji, L., Malecki, M., Warram, J. H., Yang, Y., Rich, S. S., and Krolewski, A. S. (1997) New susceptibility locus for NIDDM is localized to human chromosome 20q. *Diabetes* **46,** 876–881.

43 Klupa, T., Malecki, M. T., Pezzolesi, M., et al. (2000) Further evidence for a susceptibility locus for type 2 diabetes on chromosome 20q13.1-q13.2. *Diabetes* **49,** 2212–2216.

44 Permutt, M. A., Wasson, J. C., Suarez, B. K., et al. (2001) A genome scan for type 2 diabetes susceptibility loci in a genetically isolated population. *Diabetes* **50,** 681–685.

45 Ghosh, S., Watanabe, R. M., Hauser, E. R., et al. (1999) Type 2 diabetes: evidence for linkage on chromosome 20 in 716 Finnish affected sib pairs. *Proc. Natl. Acad. Sci. USA* **96,** 2198–2203.

46 Zouali, H., Hani, E. H., Philippi, A., et al. (1997) A susceptibility locus for early-onset non-insulin dependent (type 2) diabetes mellitus maps to chromosome 20q, proximal to the phosphoenolpyruvate carboxykinase gene. *Hum. Mol. Genet.* **6,** 1401–1408.

47 Elchebly, M., Payette, P., Michaliszyn, E., et al. (1999) Increased insulin sensitivity and obesity resistance in mice lacking the protein tyrosine phosphatase-1B gene. *Science* **283,** 1544–1548.

48 Klaman, L. D., Boss, O., Peroni, O. D., et al. (2000) Increased energy expenditure, decreased adiposity, and tissue-specific insulin sensitivity in protein-tyrosine phosphatase 1B-deficient mice. *Mol. Cell Biol.* **20,** 5479–5489.

49 Echwald, S. M., Bach, H., Vestergaard, H., et al. (2002) A P387L variant in protein tyrosine phosphatase-1B (PTP-1B) is associated with type 2 diabetes and impaired serine phosphorylation of PTP-1B in vitro. *Diabetes* **51,** 1–6.

50 Weng, J., Yan, J., Huang, Z., Sui, Y., and Xiu, L. (2003) Missense mutation of Pro387Leu in protein tyrosine phosphatase-1B (PTP-1B) is not associated with type 2 diabetes in a Chinese Han population. *Diabetes Care* **26,** 2957.

51 Santaniemi, M., Ukkola, O., and Kesaniemi, Y. A. (2004) Tyrosine phosphatase 1B and leptin receptor genes and their interaction in type 2 diabetes. *J. Intern. Med.* **256,** 48–55.

52 Dahlman, I., Wahrenberg, H., Persson, L., and Arner, P. (2004) No association of reported functional protein tyrosine phosphatase 1B 3′ UTR gene polymorphism with features of the metabolic syndrome in a Swedish population. *J. Intern. Med.* **255,** 694–695.

53 Bento, J. L., Palmer, N. D., Mychaleckyj, J. C., et al. (2004) Association of protein tyrosine phosphatase 1B gene polymorphisms with type 2 diabetes. *Diabetes* **53,** 3007–3012.

54 Palmer, N. D., Bento, J. L., Mychaleckyj, J. C., et al. (2004) Association of protein tyrosine phosphatase 1B gene polymorphisms with measures of glucose homeostasis in Hispanic Americans: the insulin resistance atherosclerosis study (IRAS) family study. *Diabetes* **53,** 3013–3019.
55 Florez, J. C., Agapakis, C. M., Burtt, N. P., et al. (2005) Association testing of the protein tyrosine phosphatase 1B gene (PTPN1) with type 2 diabetes in 7,883 people. *Diabetes* **54,** 1884–1891.
56 Neale, B. M. and Sham, P. C. (2004) The future of association studies: gene-based analysis and replication. *Am. J. Hum. Genet.* **75,** 353–362.
57 Dermitzakis, E. T., Reymond, A., and Antonarakis, S. E. (2005) Conserved nongenic sequences - an unexpected feature of mammalian genomes. *Nat. Rev. Genet.* **6,** 151–157.
58 Fingerlin, T. L., Boehnke, M., and Abecasis, G. R. (2004) Increasing the power and efficiency of disease-marker case-control association studies through use of allele-sharing information. *Am. J. Hum. Genet.* **74,** 432–443.
59 Hanis, C. L., Boerwinkle, E., Chakraborty, R., et al. (1996) A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nat. Genet.* **13,** 161–171.
60 Horikawa, Y., Oda, N., Cox, N. J., et al. (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat. Genet.* **26,** 163–175.
61 Baier, L. J., Permana, P. A., Yang, X., et al. (2000) A calpain-10 gene polymorphism is associated with reduced muscle mRNA levels and insulin resistance. *J. Clin. Invest.* **106,** R69–R73.
62 Cassell, P. G., Jackson, A. E., North, B. V., et al. (2002) Haplotype combinations of calpain 10 gene polymorphisms associate with increased risk of impaired glucose tolerance and type 2 diabetes in South Indians. *Diabetes* **51,** 1622–1628.
63 Garant, M. J., Kao, W. H., Brancati, F., et al. (2002) Atherosclerosis Risk in Communities Study. SNP43 of CAPN10 and the risk of type 2 diabetes in African-Americans: the Atherosclerosis Risk in Communities Study. *Diabetes* **51,** 231–237.
64 Malecki, M. T., Moczulski, D. K., Klupa, T., et al. (2002) Homozygous combination of calpain 10 gene haplotypes is associated with type 2 diabetes mellitus in a Polish population. *Eur. J. Endocrinol.* **146,** 695–699.
65 Orho-Melander, M., Klannemark, M., Svensson, M. K., Ridderstrale, M., Lindgren, C. M., and Groop, L. (2002) Variants in the calpain-10 gene predispose to insulin resistance and elevated free fatty acid levels. *Diabetes* **51,** 2658–2664.
66 Evans, J. C., Frayling, T. M., Cassell, P. G., et al. (2001) Studies of association between the gene for calpain-10 and type 2 diabetes mellitus in the United Kingdom. *Am. J. Hum. Genet.* **69,** 544–552.
67 Fingerlin, T. E., Erdos, M. R., Watanabe, R. M., et al. (2002) Variation in three single nucleotide polymorphisms in the calpain-10 gene not associated with type 2 diabetes in a large Finnish cohort. *Diabetes* **51,** 1644–1648.
68 Hegele, R. A., Harris, S. B., Zinman, B., Hanley, A. J., and Cao, H. (2001) Absence of association of type 2 diabetes with CAPN10 and PC-1 polymorphisms in Oji-Cree. *Diabetes Care* **24,** 1498–1499.

69 Tsai, H. J., Sun, G., Weeks, D. E., et al. (2001) Type 2 diabetes and three calpain-10 gene polymorphisms in Samoans: no evidence of association. *Am. J. Hum. Genet.* **69,** 1236–1244.
70 Xiang, K., Fang, Q., Zheng, T., et al. (2001) The impact of calpain-10 gene combined-SNP variation on type 2 diabetes mellitus and its related metabolic traits. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi*. **18,** 426–430.
71 Daimon, M., Oizumi, T., Saitoh, T., et al. (2002) Calpain 10 gene polymorphisms are related, not to type 2 diabetes, but to increased serum cholesterol in Japanese. *Diabetes Res. Clin. Pract.* **56,** 147–152.
72 Elbein, S. C., Chu, W., Ren, Q., et al. (2002) Role of calpain-10 gene variants in familial type 2 diabetes in Caucasians. *J. Clin. Endocrinol. Metab.* **87,** 650–654.
73 Rasmussen, S. K., Urhammer, S. A., Berglund, L., et al. (2002) Variants within the calpain-10 gene on chromosome 2q37 (NIDDM1) and relationships to type 2 diabetes, insulin resistance, and impaired acute insulin secretion among Scandinavian Caucasians. *Diabetes* **51,** 3561–3567.
74 Horikawa, Y., Oda, N., Yu, L., et al. (2003) Genetic variations in calpain-10 gene are not a major factor in the occurrence of type 2 diabetes in Japanese. *J. Clin. Endocrinol. Metab.* **88,** 244–247.
75 Weedon, M. N., Schwarz, P. E. H., Horikawa, Y., et al. (2003) Meta-analysis confirms a role for calpain-10 variation in type 2 diabetes susceptibility. *Am. J. Hum. Genet.* **73,** 1208–1212.
76 Zhou, Y. P., Sreenan,S., Pan, C. Y., et al. (2003) A 48-hour exposure of pancreatic islets to calpain inhibitors impairs mitochondrial fuel metabolism and the exocytosis of insulin. *Metabolism* **52,** 528–534.
77 Parnaud, G., Hammar, E., Rouiller, D. G., and Bosco, D. (2005) Inhibition of calpain blocks pancreatic beta-cell spreading and insulin secretion. *Am. J. Physiol. Endocrinol. Metab.* **289,** E313–E321.
78 Wiltshire, S., Hattersley, A. T., Hitman, G. A., et al. (2001) A genome-wide scan for loci predisposing to type 2 diabetes in a UK population (The Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. *Am. J. Hum. Genet.* **69,** 553–569.
79 Hanson, R. L., Ehm, M. G., Pettitt, D. J., et al. (1998) An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. *Am. J. Hum. Genet.* **63,** 1124–1132.
80 Elbein, S. C., Hoffman, M. D., Teng, K., Leppert, M. F., and Hasstedt, S. J. (1999) A genome-wide search for type 2 diabetes susceptibility genes in Utah Caucasians. *Diabetes* **48,** 1175–1182.
81 Hsueh, W. C., St. Jean, P. L., Mitchell, B. D., et al. (2003) Genome-wide and fine-mapping linkage studies of type 2 diabetes and glucose traits in the Old Order Amish: evidence for a new diabetes locus on chromosome 14q11 and confirmation of a locus on chromosome 1q21-q24. *Diabetes* **52,** 550–557.
82 Vionnet, N., Hani, E. H., Dupont, S., et al. (2000) Genomewide search for type 2 diabetes-susceptibility genes in French Whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent

replication of a type 2-diabetes locus on chromosome 1q21-q24. *Am. J. Hum. Genet.* **67,** 1470–1480.

83 Xiang, K., Wang, Y., Zheng, R., et al. (2004) Genome-wide search for type 2 diabetes/impaired glucose homeostasis susceptibility genes in the Chinese. Significant linkage to chromosome 6q21-q23 and chromosome 1q21-q24. *Diabetes* **53,** 228–234.

84 Ng, M. C. Y., So, W.-Y., Cox, N. J., et al. (2004) Genome-wide scan for type 2 diabetes loci in Hong Kong Chinese and confirmation of a susceptibility locus on chromosome 1q21-q25. *Diabetes* **53,** 1609–1613.

85 Gauguier, D., Froguel, P., Parent, V., et al. (1996) Chromosomal mapping of genetic loci associated with non-insulin dependent diabetes in the GK rat. *Nat. Genet.* **12,** 38–43.

86 Wallace, K. J., Wallis, R. H., Collins, S. C., et al. (2004). Quantitative trait locus dissection in congenic strains of the Goto-Kakizaki rat identifies a region conserved with diabetes loci in human chromosome 1q. *Physiol. Genomics* **19,** 1–10.

87 Galli, J., Fakhrai-Rad, H., Kamel, A., Marcus, C., Norgren, S., and Luthman, H. (1999) Pathophysiological and genetic characterization of the major diabetes locus in GK rats. *Diabetes* **48,** 2463–2470.

88 Masuyama, T., Fuse, M., Yokoi, N., et al. (2003) Genetic analysis for diabetes in a new rat model of nonobese type 2 diabetes, spontaneously diabetic torii rat. *Biochem. Biophys. Res. Commun.* **304,** 196–206.

89 McCarthy, M. I., Zeggini, E., Rayner, W., et al. (2005) International Type 2 diabetes 1q consortium. Combined analysis of 4500 single nucleotide polymorphisms from chromosome 1q21-25 in samples from eight linked populations reveals shared type 2 diabetes susceptibility variants. *Am. J. Hum. Genet.* **77,** 88.

90 Wang, W. Y., Barratt, B. J., Clayton, D. G., and Todd, J. A. (2005) Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* **6,** 109–118.

91 Hirschhorn, J. N. and Daly, M. J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6,** 95–108.

92 Farrall, M. and Morris, A. P. (2005) Gearing up for genome-wide gene-association studies. *Hum. Mol. Genet.* **14,** R157–R162.

93 Carlson, C. S., Eberle, M. A., Kruglyak, L., and Nickerson, D. A. (2004) Mapping complex disease loci in whole-genome association studies. *Nature* **429,** 446–452.

94 Newton-Cheh, C. and Hirschhorn, J. N. (2005) Genetic association studies of complex traits: design and analysis issues. *Mutat. Res.* **573,** 54–69.

95 Lawrence, R. W., Evans, D. M., and Cardon, L. R. (2005) Prospects and pitfalls in whole genome association studies *Philos. Trans. R Soc. Lond. B Biol. Sci.* **360,** 1589–1595.