J Hum Genet (2006) 51:269–277
DOI 10.1007/s10038-006-0366-9

MINIREVIEW

Newton E. Morton

Fifty years of genetic epidemiology, with special reference to Japan

Received: 13 December 2005 / Accepted: 18 December 2005 / Published online: 15 February 2006

Abstract Genetic epidemiology deals with etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations. It took its first steps before its recognition as a discipline, and did not reach its present scope until the Human Genome Project succeeded. The intimate relationship between genetics and epidemiology was discussed by Neel and Schull (1954), just a year after Watson and Crick reported the DNA double helix, and 2 years before human cytogenetics and the Japan Society of Human Genetics were founded. It is convenient to divide the next half-century into three phases. The first of these (1956–1979) was before DNA polymorphisms were typed, and so the focus was on segregation and linkage of major genes, cytogenetics, population studies, and biochemical genetics. The next phase (1980–2001) progressively identified DNA polymorphisms and their application to complex inheritance. The last phase began with a reliable sequence of the human genome (2002), followed by exploration of genomic diversity. Linkage continues to be useful to study recombination and to map major genes, but association mapping gives much greater resolution and enables studies of complex inheritance. The generation now entering human genetics will have collaborative opportunities undreamed of a few years ago, without the independence that led to great advances during the past half-century.

Keywords Genetic epidemiology · Segregation · Linkage · Association mapping · Human genome · Genetic diversity · Linkage disequilibrium · Genetic loads

N.E. Morton
Human Genetics Division, Southampton General Hospital,
School of Medicine, University of Southampton,
Duthie Building (MP 808), SO16 6YD Southampton, UK
E-mail: nem@soton.ac.uk
Tel.: +44-2380-796536
Fax: +44-2380-794264

Introduction

The Japan Society of Human Genetics (JSHG) was established at a time when the science it represented was growing so explosively that Carlson (2004) signalised the death of classical genetics. Compared to their predecessors, geneticists in 1956 were more distrustful of eugenics and more committed to cytogenetics, biochemistry, and medical genetics, but not yet prepared for the genomic revolution that the double helix would ultimately bring (Yanase 1997). Population genetics was beginning to split into two major branches, one concerned with events that took place in the past under poorly known forces of systematic pressure and chance (evolutionary genetics), the other dealing with etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations (genetic epidemiology). Here, I shall deal with the latter, divided into four periods: before the JSHG, before DNA polymorphisms, before the human genome, and after genome diversity.

Before the JSHG (1950–1955)

Historians of science may well identify the paper by Watson and Crick (1953) demonstrating DNA structure as the most significant advance in the years immediately preceding formation of the JSHG, but its full effect was not felt for half a century. In its place one benchmark was identification of the first autosomal linkage in man (Mohr 1951), later revised by recognition that “recessive” Le is a marker for the Se locus (Mohr 1954). In the interval, Goodall et al. (1953) discovered linkage between elliptocytosis and the Rh blood group, and, simultaneously, Fujii and Tsuchitori reported five Japanese families with elliptocytosis. I was in the Atomic Bomb Casualty Commission (ABCC) at the time of their discovery and had the pleasure of analysing the blood group data by the standard method at that time (Fujii et al. 1955). James

Renwick, who was serving in the British Medical Corps stationed in Hiroshima, developed an intense interest in linkage analysis, which he applied in many distinguished papers on his return to England. Even at that early stage after a devastating war, Japan was influencing genetic developments in other countries. In this respect it contrasts sharply with Germany and Communist countries, where genetics was virtually destroyed by antisemitism and Lysenkoism, respectively.

The strength of Japanese genetics was illustrated dramatically by the link that developed with a young, frail, and brilliant geneticist named Motoo Kimura, then not widely known even in his own country. His brief reports in the National Institute of Genetics, Mishima, showed a remarkable command of the higher mathematics used in evolutionary genetics by Wright and Fisher. In turn, he was anxious to understand recent work by Gustave Malecot (1948), whose classical paper I took to Japan. I enjoyed translating the French to Kimura but hesitated over certain equations. He always found them obvious, and so we went successfully through the whole work. I wrote to James Crow in Madison, Wisconsin about this extraordinary experience. After a purgatorial year in Iowa, Kimura came to Wisconsin and we shared an office until we finished our doctorates with Crow. We enjoyed stimulating discussions with him, Sewall Wright, and each other in both branches of population genetics. Kimura's unique ability was quickly recognised internationally, and for many years he and Crow took every opportunity to work together in Madison and Mishima.

Of course, these academic pleasures were secondary to the main task of the Genetics Program of the ABCC, based on fear that animal experiments might underestimate genetic risks in humans. At the same time, it was recognised that most mutations could not be detected in the first generation by methods, phenotypes, and dosimetry then available. The Genetics Program produced no unpleasant surprises, but one puzzle remains: the initial analysis (Neel et al. 1953) showed a barely significant deficiency of sons from irradiated mothers (consistent with X-linked mutation), but this diminished as time went on, and in the last analysis was reversed as an effort was made to improve anamnestic dosimetry (Schull et al. 1966). Most experimental studies did not include sex ratio, and the few that did gave inconsistent estimates suggestive of germinal selection (Morton 1997). It was a great tribute to Neel and Schull that they steered the program through these complexities and still had boundless energy for other research with their many Japanese colleagues.

Before DNA polymorphisms (1956–1979)

The founding year of the JSHG saw the beginning of human cytogenetics (Tjio and Levan 1956) with new techniques that for the first time revealed the normal human karyotype of 46 chromosomes. From this came a shower of discoveries, including aberrations in chromosome number and rearrangements that provided critical evidence for gene localization. Simultaneously, the elegant but rather unreliable methods for linkage detection were replaced by lods based on exact probability ratios that were more powerful and easier to apply to large pedigrees (Morton 1955). The first success was with elliptocytosis, which was shown to map to more than one locus (Morton 1956). The need for a conservative lod ($Z > 3$) was demonstrated by a probability argument that now gives the false discovery rate (FDR) for other applications including genome scans and microarrays (Storey and Tibshirani 2003). It is indispensable for determining the likelihood that an apparently significant result in a preliminary sample is a type I error.

Several other factors contributed to the success of linkage analysis. First, the growth of medical genetics provided a number of rare diseases inherited as recessives or dominants with high penetrance that could be demonstrated by segregation analysis with appropriate allowance for ascertainment bias (Morton 1959). Good estimates of these parameters made for simple linkage analysis (Table 1). Secondly, the relatively small number of blood groups was exponentially augmented by enzymes and other proteins whose chromosomal locations could be estimated by family studies, somatic-cell hybridisation, DNA/RNA annealing, or gene-dosage effects. The growing linkage map led to annual conferences, beginning in 1973. LIPED (Ott 1976) implemented an algorithm of Elston and Stewart (1971) to give the first widely used computer program for pedigrees. Inevitably, linkage analysis was extended from major genes to complex inheritance in which oligogenes combine independently with polygenes and environmental factors (Fig. 1). Identity by descent for single markers was the initial method (Suarez et al. 1978). It subsequently had some successes, but many failures. Unaffected relatives provided little information, and ascertainment of affected relatives was arbitrary. Affectedness in two relatives could be caused by different genetic mechanisms. The advances in molecular genetics that would provide autozygosity mapping of recessive disease (Smith 1953) and unravel complex inheritance had not yet been made.

Table 1 Classes of causal alleles

Allelic class	Usual frequency	Penetrance	Segregation analysis	Linkage analysis
Major gene	Rare	High	+	++
Oligogene	Common	Low	–	+
Polygene	Common	Very low	–	–

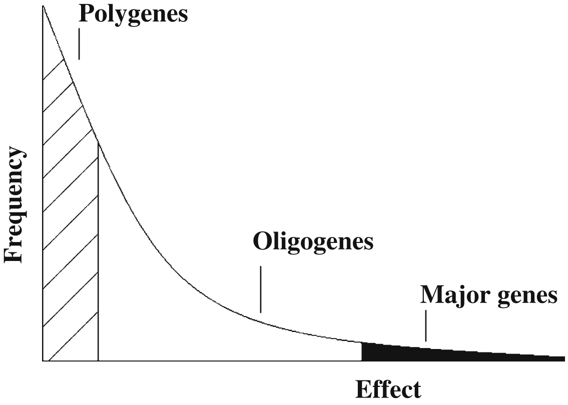

Fig. 1 Summed causal gene frequency by mean effect (Morton 1998)

Although linkage analysis was stimulated by discovery of polymorphisms through starch gel electrophoresis (Smithies 1955), some population geneticists continued to suppose that all polymorphisms are maintained by balanced selection. Support for that hypothesis came from three malaria-dependent polymorphisms, beginning with evidence on sicklemia (Allison 1964). Lewontin (1964) devised mathematical models for balanced polymorphisms and introduced the D' metric for linkage disequilibrium (LD, also called allelic association) between a pair of diallelic markers and a random sample. D' takes four values, of which two are negative and the others lie between 0 and 1. As defined, there is no probability interpretation, evolutionary theory, or extension to non-random samples or an affection dichotomy. Hill and Robertson (1968) introduced the r^2 metric as a simpler alternative to D', since $\chi^2 = r^2 n$ for a 2×2 table of n observations, but with the same limitations as D'. Sved (1971) developed a probability theory for the conditional probability of autozygosity at one diallelic locus, given autozygosity at another, and mistakenly took that probability as r^2. For many years these and other measures of LD had no significant application, but they were the forerunners of later developments that sprang from advances in molecular biology.

Pending that union, other subjects flourished in this early period besides cytogenetics and linkage. One line of research was into effects of consanguineous marriages, which were declining in Japan but still relatively frequent by world-wide standards (Imaizumi 1986). Therefore, much of the evidence came from Japan. There was lively discussion among population geneticists about estimation of a genetic load through inbreeding. Were social class and other non-genetic factors homogeneous among degrees of consanguinity? If genetic, could the effects be explained by heterozygote advantage rather than deleterious mutation? The available evidence was debated, but in subsequent years large numbers of deleterious mutations, much larger numbers of quasi-neutral mutations, but very few examples of heterozygote advantage have been demonstrated (Crow 1987).

Deleterious mutations were known from research on *Drosophila* and the mouse to have a large recessive component, but clinical and epidemiological branches of human genetics were in their infancy. The urgency of mutation research was apparent in efforts to assess hazards of ionising radiation and other mutagenic agents. Even if a “doubling dose” were estimated from another species, how could its human effect be anticipated without knowing our spontaneous mutation rate? Children of consanguineous marriages provide one avenue, which was exploited by a paper that estimated the deleterious mutation rate for quasi-recessive genes as 0.06–0.15 per gamete per generation (Morton et al. 1956). This corresponds to 6–15×10^{-6} such deleterious mutations per contributory locus, assuming 10^4 of the latter. With all the advances during the last half century, we have no better estimate today.

Neel and Schull (1953) were the first to perceive that epidemiology had failed to provide approaches to genetics. It was equally true that population genetics regarded the environmental variables dear to epidemiology as noise. Genetic analysis was handicapped by little knowledge about few genes, while epidemiology suffered from lack of interest in rare genes and the inability to make accurate measures of relevant environment. The first book on genetic epidemiology was published much later (Morton and Chung 1978), and the young science was slow to master DNA structure. Before its journal was founded in 1984 it had fewer practitioners than evolutionary genetics, and a society was not founded until 1992. Inevitably the special problems of a new discipline took first place. Initially several earlier problems were pursued. Among them were estimates of heritability, beginning with twin studies and culminating with path analysis of genetic and cultural variables (Rao et al. 1979). One rationale was to justify attempts to map and ultimately characterise genes contributing to complex inheritance. If the trait under study had negligible heritability, why bother to make genetic studies? From a Marxist position, Lewontin (1975) argued that zero heritability of socially important traits like intelligence cannot be excluded on present evidence. Some effort was devoted to testing this hypothesis, with the conclusion that heritability may well be less than conventional estimates, but no model yet invented is consistent with zero heritability (Rao et al. 1982). Analysis could be pushed further. William Shockley, inventor of the transistor, Nobel Laureate, and member of the National Academy of Sciences, in his last years argued that African descent is inextricably associated with a 15% decline in the intelligence quotient. Analysis of the data refuted that hypothesis (Rao et al. 1977). Given sufficient evidence, genetic epidemiology can resolve controversy generated by prejudice from either political extreme. However, it was a relief to turn from these social issues to deeper scientific questions.

Evolutionary genetics was quicker to exploit DNA structure, thanks to a lucky mistake. Haldane (1957) developed his theory on the cost of natural selection from industrial melanism, where pollution favoured dark moths over the paler wild type that became more vulnerable to predation by birds. As population size plummeted, increase of the dark mutants became a boon, not a burden. At first, Kimura (1960) accepted Haldane's assumption that selection of the fittest in a declining population imposes some kind of cost, and so he began to explore models of neutral mutation. As supporting evidence accumulated, he abandoned Haldane's cost (Kimura 1968) and for the rest of his life developed his neutral theory to its greatest extent.

At the end of this period genetic epidemiology was ready to become molecular. Many monogenic errors of metabolism had been discovered in Japan and elsewhere, a number of which had been placed on the linkage map. Then a decisive event occurred that greatly accelerated these developments: Solomon and Bodmer (1979) proposed construction of a linkage map incorporating restriction fragment length polymorphisms (RFLPs). The flower planted by Watson and Crick was beginning to bloom.

Before the human genome (1980–2001)

Transition to analytic genomics was gradual, and did not become the dominant trend until the beginning of this century. At the start of this period the main interest of genetic epidemiology was multifactorial inheritance, with emphasis on family studies including concordance in twins and recurrence risks in family members, leading to an estimate of heritability (Morton 1982; King et al. 1984). Segregation analysis was applied in two different ways. The Mendelian approach closely followed experience with major genes, giving careful attention to ascertainment and distinction from polygenes. When evidence for a major gene was obtained, estimates of its parameters were used in linkage studies. An alternative approach ignored ascertainment and tested non-Mendelian inheritance as a surrogate for environmental factors. Among the latter, maternal effects were sometimes identified. As major loci were detected, genetic analysis increasingly used affected sib pairs to test models of dominance, recessivity, and linkage. While giving greater simplicity and generality, this approach assumed effect, gene frequency and ascertainment into a "nonparametric" component that pointed in an interesting direction but could not clearly delimit a genetic component, even in the most favourable HLA system. Environmental factors were examined by path analysis, observation of migrants, and geographic distribution. Association of disease risk with a small number of genetic markers, especially in the HLA system, became increasingly popular. Use of the polymerase chain reaction by Mullis and Faloona (1987) led to the discovery of microsatellites and other DNA polymorphisms. Progress in DNA sequencing raised the possibility that the whole human genome could be sequenced.

By the end of this period linkage analysis had reached its apogee (Pawlowitski et al. 1997; Rao and Province 2001). Multiple marker loci were being used, but at low resolution and without reliable allowance for interference or sex differences in recombination. Many regional assignments of disease genes failed to narrow their localisation enough to identify them. An early success of association mapping was by inspection of two conventional measures for a 2×2 table of affection status × allelic pair in chromosomes from cystic fibrosis patients (Kerem et al. 1989). These data were analysed by Terwilliger (1995) using conditional probabilities under the assumption of independence. His 13.8 support interval included part of the CFTR gene but not the main causal marker ΔF508. Better results were obtained by Collins and Morton (1998), who placed ΔF508 within 50 kb of its physical location using an evolutionary model provided by Malecot (1948) and including composite likelihood to allow for autocorrelation. Subsequent papers proved the appropriateness of the Malecot model (Morton et al. 2001) and compared assignments from the physical and linkage maps, both at that time subject to considerable error (Lonjou et al. 1998). Most of the markers used then were RFLPs that cannot be identified in later maps, and their locations in the linkage map are subject to substantial error. Even with these handicaps, haemochromatosis (HFE) provided an early example of the gain in power that is possible with a linkage map. Linkage of HFE to HLA-A was demonstrated by Simon et al. (1976) 20 years before the location, and therefore identity, of HFE was determined (Feder et al. 1996). The obstacle was that the distance between HLA-A and HFE is 4.6 Mb on the physical map but only 0.75 cM on the linkage map, leading investigators in many countries to inch toward their goal.

In the middle of these developments Risch and Merikangas (1996) proposed that single nucleotide polymorphisms (SNPs) at sufficient density could be used to localise disease genes. The process they advocated, called association mapping or LD mapping, has been described by a genetic theory based on the Malecot model of evolution from a population bottleneck, measured in number of generations, and also by coalescent theory for a most recent common ancestor (MRCA) with time measured in effective population size, assumed constant. Use and abuse of these two models has become a topic for the current period, which began when the Human Genome Project provided a representative physical map of our genome. Its tangled early history was described by Cook-Deegan (1991), but was not finished until a decade later. Almost immediately, various efforts, including the HapMap Project, undertook to provide evidence on SNP diversity with the goal of using that information to localise disease genes and investigate other aspects of the diversity revealed by genetic polymorphism. Whereas the Human Genome Project produced a single sequence, the HapMap Project undertook to map a substantial part of sequence diversity.

After genome diversity (2002+)

Up to this point I have relied on the masterly summary of human genetics by Yanase (1997), adding only what is unique to genetic epidemiology. However, the combination of an (almost) complete representative sequence with substantial evidence on its variation marks a turning pointing in genetics, even though it is restricted to about 3 million of the 20 million or so SNPs, is strongly biased against SNPs with frequencies less than 0.05, and does not include diversity in genome structure (deletions, insertions, and nucleotide rearrangements).

To enter this new world, we must remember that its vocabulary does not agree with classical genetics. Single nucleotide polymorphisms are not necessarily polymorphic, and may include minor allele frequencies of less than the 0.01 that conventionally separates idiomorphs from polymorphs. Most SNPs are diallelic, but some are more complex and their minor allele frequencies may be pooled. Most, but not all, diallelic markers in common use are SNPs. A marker that contributes to disease risk is causal, whereas an associated marker is usually not causal, but merely predictive. Association mapping uses predictive markers to localise causal markers and thereby the genes to which they belong. Unlike the linkage methods that preceded association mapping, family studies are optional and may be reserved for rare genes with high penetrance (major genes). Freedom from genotyping parents is especially valuable for diseases of late onset.

Anticipating completion of the euchromatic sequence of the human genome (International Human Genome Sequencing Consortium 2004) led immediately to the realisation that maps in linkage disequilibrium units (LDU) would be far more useful for association mapping and related problems than linkage maps, whatever their resolution might be (Maniatis et al. 2002). LD reflects effective time in generations since the last bottleneck, typically more than 1,000 generations ago (Tapper et al. 2005). Therefore sex-specific recombination is averaged for autosomes, although interpolation from an LD map into a sex-specific linkage map can greatly improve the resolution and utility of the latter. The comment in a study of patterns of recombination that "averaging cM values from males and females creates a statistic with no biological meaning and dubious utility" (Lynn et al. 2004) ignores LD, evolution, and association mapping. Small recombination rates are more serious for linkage maps than for LD maps, while inbreeding, selection, and chromosome rearrangement act more powerfully on LD maps, creating differences within chromosomes and among populations. An LD map is not merely a stretched linkage map, and the latter is more than a compressed LD map.

Division of the genome into candidate regions is tentative, whether based on linkage at low resolution or LD at greater resolution. There are three specific reasons why linkage maps are unsatisfactory for analysis of LD. First, the number of independent regions increases with number of generations (Lander and Kruglyak 1995). Secondly, the linkage map exploits the multilocus-feasible Haldane function that does not allow for interference, which is important for linkage but negligible for LD (Tapper et al. 2005). Finally, different populations have different LD maps, but have not been shown by linkage evidence to have different linkage maps. Likelihood for multiple markers is routine for linkage, but considerable effort with LD has been diverted to most significant single SNPs and haplotypes (Thomas et al. 2005). Composite likelihood for multiple markers gives smaller and more reliable confidence intervals.

Since this approach is so recent (Table 2), it is not generally understood. A comprehensive derivation is unfortunately scattered in the cited literature, but a brief summary of results is given here. The LDU distance between the i^{th} pair of adjacent markers is $\varepsilon_I \mathrm{d}_I$, where d_I is the physical distance, conventionally expressed in kilobases with precision given by a 3-digit decimal. The association probability ρ between two markers with LDU distance $\sum \varepsilon_I \mathrm{d}_I$ is predicted under the Malecot model as $\rho_I = (1 - L)M\mathrm{e}^{-\sum \varepsilon_I d_I} + L$. This model makes several predictions: (1) the effective population size N_{e} and systematic pressure v due to mutation and long-range migration act on M, but not on the expected value of LDU or, to an appreciable extent, on L; (2) sample size acts on the expected value of L, but not on the expected value of M; (3) the sample size acts on the expected value of LDU only if some included value $\varepsilon_I d_I$ is so large as to be indeterminate and therefore given a nominal value of 3 (such an interval is called a "hole", the frequency of which declines as marker density increases); (4) the expected value of LDU is $\mathrm{E}(\sum \varepsilon_I d_I) = t\sum \mathrm{w}_I$, where t is the number of generations since a major bottleneck and w_I is the length of the Ith interval on the sex-average linkage map in Morgans. The latter prediction is limited by low resolution of the linkage map (making small values of w_I unreliable) and by inaccurate modelling of interference (making $\sum \mathrm{w}_I$ unreliable). On the contrary, the LD map is unaffected by interference, since recombination rates of order w_I occur in different generations and therefore without interference.

Table 2 History of linkage disequilibrium (LD) maps and association mapping. *LDU* LD units, *SNP* single nucleotide polymorphism

Years	Association mapping	kb map	LDU map	SNPs	Theory
< 1998	±	0	0	0	0
1998–2001	+	±	0	±	±
> 2001	++	+	+	+	+

Two other properties of the Malecot model should be noted. First, the observed value of ρ in a random sample (but not otherwise) is equal to the largest value of D' (Lewontin 1964). Unlike D', ρ has an evolutionary theory that predicts from linkage to LD over t generations. Secondly, the LD map is not dependent on the linkage map and is therefore free to reflect evolutionary factors such as selection, changes in DNA sequence, variation in t, and population differences.

Given a physical, LD, or other map that has the definitive property of additive distance, the parameters to be estimated are M, L, ε, and S, the location of a causal marker. Pedigree data for a major locus allow determination of alleles shared by affected relatives. The 2×2 table of affection status × allele gives an estimate of ρ, with suitable allowance for non-random sampling. Complex inheritance introduces genes with low penetrance that prevents classification of marker alleles as affected or normal. Fortunately, the transformation $z = \gamma\rho$ allows diplotypes to be scored assuming additivity. This has a theoretical basis in MacLauren's theorem when penetrance is low. Even if a causal marker deviates from additivity, the deviation for a predictive marker not completely associated is initially less and diminishes with recombination over time.

Quantitative traits must be examined by regression (b), or less efficiently by correlation (r). Let ψ represent the metric used, whether ρ, z, b, r, or some less appropriate one (Fig. 2). Composite likelihood is $\mathrm{lk} = \varepsilon^{-\Lambda/2}$, where $\Lambda = \sum K_\psi(\hat{\psi} - \psi)^2, \hat{\psi}$ denotes the observed estimate, ψ its expected value, and K_ψ the information under the null hypothesis that $\psi = 0$. In large sample theory the predicted value of L is a simple function of K_ψ, and $\chi_1^2 = K_\psi\hat{\psi}^{2\cdot}$.Composite likelihood is maximised by minimising Λ, giving an estimate of error variance for each hypothesis about the model.

The first LD maps were for small regions (Maniatis et al. 2002), progressively extended to whole genomes (Tapper et al. 2005). They accurately reflect blocks and steps (Fig. 3). LD maps have proven generally to be more powerful than kilobase maps for association mapping (Maniatis et al. 2005). Haplotypes have also been used, but they have many arbitrary variants (reflecting length, choice of markers, inference from diplotypes, imputation of untyped markers, and other factors) and have not performed as well as composite likelihood for SNPs (N. Maniatis et al., MS submitted). Several other methods have been proposed for association mapping, but so far they have seldom been used and when used have not given better localisation or stronger evidence than the Malecot model with a map in LDU. After 3 years of development, map construction is rapid and accurate. Theory and practice for allelic association in small regions is stable, and interest has shifted to genome scans, a misnomer since a large region, chromosome, or set of chromosomes is analysed in the same way. Whatever the terminology or method, such a scan is only stage 1 in a multistage design, the later stages being concerned with smaller regions. The number of markers in a dense scan is extremely large compared with the number of causal sites likely to be detected even in a large sample, whether or not supported by functional tests. This is a challenge to the FDR, which cannot be reliably calculated when the probability of the null hypothesis approaches 1 (Storey and Tibshirani 2003; Dalmasso et al. 2004). In its absence alternative corrections are being developed to determine the real significance corresponding to nominal significance in a large number of tests.

For the stage 1 analysis of large regions, the most promising approach exploits composite likelihood to divide each chromosome into small regions, a few of which flank the most significant locations. We conjecture that the flanking interval need not exceed 10 LDU (5 LDU on each side), since greater distances provide virtually no information. However, until high resolution is attained this criterion must be relaxed to ensure enough markers (30+) to estimate the error variance well. Within this design there are several variants to be explored. Later stages (Table 3) use tested methods based

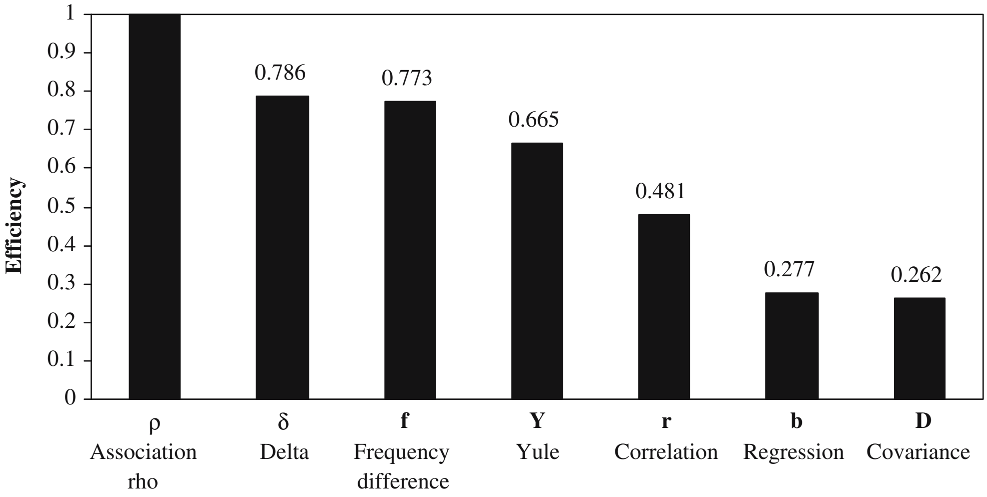

Fig. 2 Efficiency relative to association ρ for kilobase map (Collins et al. 1996)

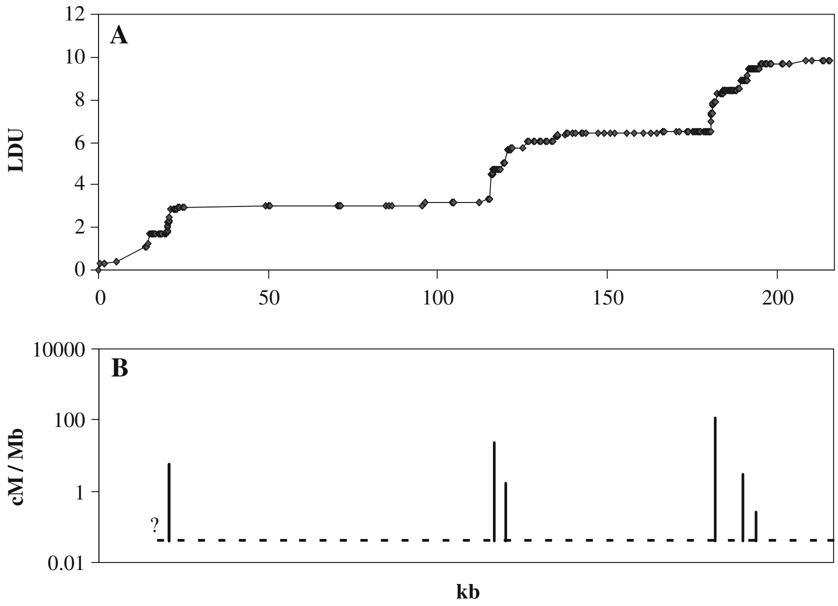

Fig. 3 Graph of the linkage disequilibrium units (LDU) map for 216 kb segment of MHC-II region A (Zhang et al. 2002) compared with recombination hotspots B presented by Jeffreys et al. (2001)

on composite likelihood for small regions, and ultimately functional tests (Zhang et al. 2004). Advances in marker typing can make ordered analysis of independent samples highly efficient, testing the most significant regions first. These and other approaches must be examined to optimise association mapping in a genome scan.

The database created by the HapMap Project has recently been described, together with some early results (International HapMap Consortium 2005). The primary goal is to facilitate "comprehensive, genome-wide association studies". However, methods to accomplish this have not yet been detailed nor assessed. Some effort is being made to select tag SNPs that will minimise the number of tests without substantial loss of power, which by definition is dependent on the method of analysis (Zhang et al. 2004). Haplotype analysis was an original goal, now less popular because of the many problems it presents (N. Maniatis et al., MS submitted).

The basis for selecting tag SNPs is controversial, with allele frequency, level of association, non-synonymous substitution or not, location in a spliceosomal intron in a block or step, and many other factors competing for attention. Unless such issues are settled or abandoned, how can useful haplotypes be selected? A more interesting issue is whether increasing the number of markers in a scan increases nominal power enough to compensate for conjectured increase in the criterion for the type I error. Otherwise stated, is a 500,000 SNP microarray a better investment than a 100,000 SNP model? In this confusion, coalescent and other alternatives to LD maps and the Malecot model have been intimated, but are not yet implemented. When they become clear, they must be assessed.

We are fortunate to live in interesting times, with Japan playing a seminal role. The Japanese Millennium Genome Project described by Haga et al. (2002) included 174,269 SNPs and 16,293 insertion/deletion polymorphisms selected from 154 Mb in DNA from 24 Japanese individuals. Although only 5% of the genome, it covered at least 13,758 genes. This prototype database was designed to identify genes associated with susceptibility to common diseases or to therapeutic drugs. Using the simplest stage 1 method (selection of most significant polymorphisms), coupled with stage 2 confirmation in an independent sample, there has been a notable success (Ozaki et al. 2002). The young BioBank Japan Project has wisely limited its questions to genes that affect disease and drug response. The first result is evidence for several genes that may indicate the appropriate warfarin dose for individual patients and provide safer management of anti-coagulant therapy (Nakamura 2005; T. Mushiroda et al., MS submitted). Britain and the United States are anxious to emulate that project, with the added objective of identifying interaction between unspecified loci and poorly measured environments. The United States BioBank, envisaged but not yet launched, would be wise to follow Japan both in obtaining its sample from hospital inpatients and outpatients rather

Table 3 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 4 Today's choices

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

than from general practices, and in emphasis on genetic rather than environmental factors, since the latter are inaccurately measured and much less promising than pharmacogenetics for personalised therapy. BioBank Japan currently follows 47 diseases and an unspecified number of therapeutic drugs. The British BioBank is equivocal about the diseases it will study intensively, but it will contain a large number of individuals not expressing any of them. There is general concern about the ambiguity and sluggish pace of British BioBank (Watson and Cyranoski 2005). Operating in secrecy under a succession of chief executives and without research results, the value for Britain can be inferred only from reduced support for independent research and consequent emigration of young scientists. Cancer and heart disease are popular with both BioBanks, but cognitive, behavioural, and sensory disease like macular degeneration are underrepresented, even though they reduce the quality of life for many patients as severely (and with present treatment as irrevocably) as fatal diseases.

All these ventures raise interesting problems, of which association mapping is among the first (Table 4). The structure and content of a multistage analysis will determine its type I and type II errors, and therefore its success. The future of genetic epidemiology rests on its value for association mapping, not on its untested potential to perform miracles with gene–environment interactions. The affair of genetic epidemiology with complex inheritance outside genomics has run its course, whereas the marriage with analytic genomics has been consummated and shows every sign of permanence.

References

Allison AC (1964) Polymorphism and natural selection in human populations. Cold Spring Harbour Symp Quant Biol 24:137–149

Carlson EA (2004) Mendel's legacy. The origin of classical genetics. Cold Spring Harbor Laboratory Press, Long Island

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

Collins A, Frezal J, Teague J, Morton NE (1996) A metric map of humans: 23,500 loci in 850 bands. Proc Natl Acad Sci USA 93:14771–14775

Cook-Deegan RM (1991) The human genome project: the formation of federal policies in the United States, 1986–1990. In: Hanna KE (ed) Biomedical politics. National Academy Press, Washington, DC

Crow JF (1987) Muller, Dobzhansky, and overdominance. J Hist Biol 20:351–380

Dalmasso C, Broet P, Moreau T (2004) A simple procedure for estimating the false discovery rate. Bioinformatics 21:660–668

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542

Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, Domingo R Jr, Ellis MC, Fullan A et al (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet 13:399–408

Fujii T, Moloney WC, Morton NE (1955) Data on linkage of ovalocytosis and blood groups. Am J Hum Genet 7:72–75

Goodall HB, Hendry DWW, Lawler SD, Stephen SA (1953) Data on linkage in man: elliptocytosis and blood groups II. Family 3. Ann Eugen 17:272–278

Haga H, Yamoda R, Ohnishi Y, Nakamura Y, Tanaka T (2002) Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190562 genetic variations in the human genome. J Hum Genet 47:605–610

Haldane JBS (1957) The cost of natural selection. J Genet 55: 511–524

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

Imaizumi Y (1986) A recent survey of consanguineous marriages in Japan. Clin Genet 30:230–233

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

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

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

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui L-C (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080

Kimura M (1960) Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load. J Genet 57: 21–34

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624–626

King M-C, Ree GM, Spinner NB, Thomson G, Wrensch MR (1984) Genetic epidemiology. Annu Rev Public Health 5:1–52

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:244–247

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

Lewontin RC (1975) Genetic aspects of intelligence. Annu Rev Genet 9:387–405

Lonjou C, Collins A, Ajioka RS, Jorde LB, Kushner JP, Morton NE (1998) Allelic association under map error and recombinational heterogeneity: a tale of two sites. Proc Natl Acad Sci USA 95:11366–11370

Lynn A, Ashley T, Hassold T (2004) Variation in human meiotic recombination. Annu Rev Genomics Hum Genet 5:317–349

Malecot G (1948) Les mathématiques de l'hérédité. Masson, Paris

Maniatis N, Collins A, Xu C-F, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke K, Morton NE (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc Natl Acad Sci USA 99:2228–2233

Maniatis N, Morton NE, Gibson J, Xu C-F, Hosking LK, Collins A (2005) The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. Hum Mol Genet 14:145–153

Mohr J (1951) Estimation of linkage between the Lutheran and the Lewis blood groups. Acta Pathol Microbiol Scand 29:339–344

Mohr J (1954) A study of linkage in man. Munksgaard, Copenhagen

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

Morton NE (1956) The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood group. Am J Hum Genet 8:80–96

Morton NE (1959) Genetic tests under incomplete ascertainment. Am J Hum Genet 11:1–16

Morton NE (1982) Outline of genetic epidemiology. Karger, Basle

Morton NE (1998) Significance tests in complex inheritance. Am J Hum Genet 62:690–697

Morton NE (1997) Genetic epidemiology. Ann Hum Genet 61:1–13

Morton NE, Chung CS (1978) Genetic epidemiology. Academic Press, New York

Morton NE, Crow JF, Muller HJ (1956) An estimate of the mutational damage in man from data on consanguineous marriages. Proc Natl Acad Sci USA 42:855–865

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

Mullis KB, Faloona FA (1987) Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol 155:335–350

Nakamura Y (2005) Scientists create Japan's largest BioBank for genetic studies of 47 common diseases (Interview). Affymetrix Microarray Bull 1:16–19

Neel JV, Schull WJ (1954) Human heredity. University of Chicago Press, Chicago

Neel JV, Morton NE, Schull WJ, McDonald DJ, Kodani M, Takeshima K, Anderson RC, Wood J, Brewer K, Wright S, Yamazaki J, Suzuki M, Kitamura S (1953) The effects of exposure of parents to the atomic bombs on the first generation offspring in Hiroshima and Nagasaki: (preliminary report). Jpn J Genet 28:211–218

Ott J (1976) A computer program for linkage analysis of general human pedigrees. Am J Hum Genet 28:528–529

Ozaki K, Ohnishi Y, Iida A, Akihiko S, Yamada R, Tsunoda T, Sato H, Hori M, Nakamura Y, Tanaka T (2002) Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat Genet 32:650–654

Pawlowitski I-H, Edwards JH, Thompson EA (1997) Genetic mapping of disease genes. Academic, San Diego

Rao DC, Province MA (2001) Genetic dissection of complex traits. Academic, San Diego

Rao DC, Morton NE, Elston RC, Yee S (1977) Causal analysis of academic performance. Behav Genet 7:147–159

Rao DC, Morton NE, Cloninger CR (1979 Path analysis under generalized assortative mating. I Theory Genet Res 33:175–188

Rao DC, Morton NE, Lalouel JM, Lew R (1982) Path analysis under generalized assortative mating. II American IQ. Genet Res 39:187–198

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

Schull WJ, Neel JV, Hashizumi A (1966) Some further observations on the sex ratio among infants born to survivors of the atomic bombings of Hiroshima and Nagasaki. Am J Hum Genet 18:328–338

Simon M, Bourel M, Fouchet R, Genetet B (1976) Idiopathic hemochromatosis and HLA-antigens. Diabete Metab 2:113–118

Smith CAB (1953) The detection of linkage in human genetics. J R Stat Soc B 15:153–192

Smithies O (1955) Zone electrophoresis in starch gels: group variations in the serum proteins of normal human adults. Biochem J 61:629–641

Solomon E, Bodmer WF (1979) Evolution of sickle variant gene. Lancet 1(8122):923

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100:9440–9445

Suarez BK, Rice J, Reich T (1978) The generalized sib pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 42:87–94

Sved JA (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor Popul Biol 2:125–141

Tapper W, Collins A, Gibson J, Maniatis N, Ennis S, Morton NE (2005) A map of the human genome in linkage disequilibrium units. Proc Natl Acad Sci 102:11835–11839

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

Thomas DC, Haile RW, Duggan D (2005) Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet 77:337–345

Tjio JH, Levan A (1956) The chromosome number of man. Hereditas 42:1–6

Watson JD, Crick FHC (1953) Molecular structure of nucleic acids. Nature 171:737–738

Watson J, Cyranoski D (2005) Beset by practical hurdles, U.K. BioBank moves at sluggish pace. Nat Med 11:696

Yanase T (1997) Human genetics: past, present, and future, with special reference to major trends in Japan. Jpn J Hum Genet 42:265–316

Zhang W, Collins A, Maniatis N, Tapper W, Morton NE (2002) Properties of linkage disequilibrium maps. Proc Natl Acad Sci USA 99:17004–17007

Zhang W, Collins A, Morton NE (2004) Does haplotype diversity predict power for association mapping of disease susceptibility? Hum Genet 115:157–164