THE ROYAL SOCIETY

Received 20 February 2003
Accepted 12 June 2003
Published online 13 August 2003

Genetic epidemiology, genetic maps and positional cloning

Newton E. Morton

University of Southampton, Human Genetics Division, Duthie Building, Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK (*nem@soton.ac.uk*)

Genetic epidemiology developed in the middle of the last century, focused on inherited causes of disease but with methods and results applicable to other traits and even forensics. Early success with linkage led to the localization of genes contributing to disease, and ultimately to the Human Genome Project. The discovery of millions of DNA markers has encouraged more efficient positional cloning by linkage disequilibrium (LD), using LD maps and haplotypes in ways that are rapidly evolving. This has led to large international programmes, some promising and others alarming, with laws about DNA patenting and ethical guidelines for responsible research still struggling to be born.

Keywords: linkage; linkage disequilibrium; single nucleotide polymorphism; HapMap; Biobank; ethics

1. INTRODUCTION

The first half of the twentieth century laid the foundations of population genetics under the rubric 'genetics of populations' that it retains in many languages. We owe the perception of a new science to C. C. Li (1948), whose seminal text introduced the term that was quickly accepted in the English-speaking world (Lerner 1950). His first chapter dealt with family data under incomplete ascertainment, but this was deleted in the next edition 'because the nature of the subject is not quite the same as that of the rest of this book' (Li 1955). The orphan was adopted by Neel & Schull (1954) and called 'epidemiological genetics', later replaced by 'genetic epidemiology' to include familial and group differences that, on investigation, transpire to be partly or wholly environmental (Morton *et al.* 1967). Today, population genetics has two major branches. *Evolutionary genetics* deals with processes and phylogenies for which general mathematical theories have been developed, but the events themselves were unique and took place in the past under poorly known forces of systematic pressure and chance. *Genetic epidemiology* is concerned with contemporary populations in which some replication is possible, especially with the etiology, distribution and control of disease in groups of relatives and with inherited causes of disease in populations. In recent years, the development of forensic genetics, genomics and bioinformatics has expanded both fields and may give rise to alternative disciplines at the interface between genetics and mathematics.

2. DEVELOPMENT OF GENETIC EPIDEMIOLOGY

Human genetics grew explosively in the second half of the twentieth century, freed from the incubus of eugenics and enriched by advances in cytogenetics, biochemistry and experimental genetics. These stimuli to genetic epidemiology were reinforced by opportunities for field research on mutation, modes of inheritance, linkage, inbreeding effects, population structure, and associations of blood groups and other polymorphisms with disease, conjoined with access to computers that (although primitive by today's standards) were able to support analyses more powerful than mechanical calculating machines could provide. Linkage analysis showed that a clinical entity could include multiple genetic entities (Morton 1956), and this prompted linkage maps of increasing density. The number of known polymorphisms was a limiting factor, effectively restricting linkage analysis to major alleles of high penetrancy. Blood groups and isozymes were seized avidly, but the number of polymorphic markers they provided was inadequate to investigate oligogenes of low penetrancy.

The breakthrough that with hindsight we all awaited came with DNA markers (Solomon & Bodmer 1979; Botstein *et al.* 1980). Extension from restriction fragment length polymorphisms (RFLPs) to the polymerase chain reaction (PCR) increased the number of useful markers from a few thousand to several million, most of them diallelic single nucleotide polymorphisms (SNPs) that can be exploited with minimal error in the finished genome released in 2003. The classical linkage method to localize and then clone a previously unrecognized major gene or oligogene affecting susceptibility to a particular disease (positional cloning) can now be complemented and perhaps surpassed by other methods that depend on the association of markers with phenotypes of individuals and functional assays of single loci.

As these developments are taking place, forensic genetics has been using multiallelic loci (and recently mitochondrial haplotypes) to provide evidence of guilt or innocence accordingly, as a suspect does or does not match an evidentiary sample. Although disease susceptibility is not the object of the exercise, evidence on population structure and methods for minimizing its forensic impact are shared with genetic epidemiology. Forensic genetics was introduced by Jeffreys *et al.* (1985) in a case that identified an unsuspected culprit and exonerated the

DOI 10.1098/rstb.2003.1357

© 2003 The Royal Society

disturbed but innocent man who had confessed to the crime. As technical problems in DNA typing diminished, defending barristers turned to the 'ceiling principle' that evaluated the coincidental DNA-matching probability for random individuals from a particular population by sampling many arbitrary populations, taking for each shared allele its largest frequency f in the reference samples if greater than 0.05, or 0.05 otherwise (Committee on DNA Forensic Science 1992). The match probability for genotypes at n loci would then be presented in court as $P = \Pi_i^n q_i$, where q_i is the probability of the ith genotype, of form f^2 for a homozygote and $2ff'$ for a heterozygote. The ceiling principle was not generally accepted because it has no logical basis, but it caused so much confusion that a second committee met to reject it (Committee on DNA Forensic Science 1996). Much more reliable principles were introduced that use evidence on population structure to calculate matching probabilities for random individuals and relatives (Morton 1997). However, a few problems were underestimated: admixture of DNA from two or more people, accidental or intentional failure of the chain of custody for DNA samples, and misuse of evidence from N individuals, whether taken from a DNA database or potential suspects. Suppose that all are innocent. Then the probability that at least one suspect matches the evidentiary sample by chance is $1 - (1 - P)^N$ if P is the match probability for a single individual. If P is small and N is large this is, to a close approximation, $1 - e^{-PN}$, or approximately PN if $P \ll 1/N$. If one of the individuals is selected as a defendant and found to match the evidentiary sample in an independent set of markers with random match probability P', the joint probability on the hypothesis that the match is coincidental is $[1 - (1 - P)^N]P'$, or approximately NPP' if $P \ll 1/N$. Even if the relationship between suspect and evidentiary sample is excluded, the match probability is logically different and much greater in a large database than for an innocent suspect who is subsequently typed. If the evidentiary sample is large enough, the second battery that gives P' eliminates ambiguity about N, but does not use the full evidence. Database evidence is substantially exaggerated in many trials by omitting or underestimating N. Such matters are the farthest extension of genetic epidemiology, but a logical application of its interest in sampling.

3. GENETIC MAPS

Any linear, unbranched structure in which distances between genetic entities are additive is called a *map*. Classical genetics dealt with *linkage maps* of loci with distances in recombination units called centimorgans (cM), corresponding, on average, to one crossover per 100 meioses. The distribution of crossing over differs between the sexes, and so a linkage map is sex-specific, but in some situations the male and female maps may be averaged (as in the mouse), or restricted to the homogametic sex (as in *Drosophila*). By contrast, a *cytological map* assigns loci and chromosome breakpoints (translocations, deletions, duplications and inversions) to bands made visible by staining. Localization and band length by microscopy are imprecise, and so the distance coordinate is approximated by projection on the DNA sequence in which each nucleotide in a chromosome is given an ordinal number from the end of the short arm (pter).

Linkage maps have descendants made possible by sequencing. Draft sequences are replaced by finished maps that determine gene order more reliably than linkage maps, which often assign two or more loci to the same point because there has been no certain recombinant between them in the small number of meioses on which the human linkage map is based. We are now at the point where gene order in maps should be constrained to the finished sequence as the lesser of two evils (Tapper *et al.* 2001). This is reflected in the *physical map* in which distance is measured by nucleotides. In past years this map was approximated in three ways that are now, or will soon be, of only historical interest: by *radiation-hybrid maps* in which distance was estimated from chromosome breakage, by small maps based on contigs of chromosome fragments, and by draft sequences.

Resolution of the linkage map can be enhanced by the analysis of meiotic recombination in sperm, but this has been feasible only for sequences of *ca.* 200 kb (Jeffreys *et al.* 2001). A more practical approach is through the *linkage disequilibrium (LD) map* based on the association between diallelic markers. On simple assumptions, this increases resolution by linkage in proportion to the number of generations since LD arose or was intensified by mutation, immigration or a population bottleneck. However, the LD map includes noise due to causes other than recombination (including genetic drift, gene conversion, inversion and deletion) and does not distinguish between crossovers in males and females.

These considerations lead to *map integration*, whereby for each chromosome the physical and cytogenetic maps coexist in the same database with the linkage and LD genetic maps. Each can be updated independently of the others, with interpolation in defined regions as required. The only serious obstacle is the coexistence of ephemeral databases using different symbols for loci, SNPs and other markers, represented by low-resolution colour graphics that do not convey the information required to associate specific sequences with phenotypes, bands, recombination, LD or any other property of the genome.

4. POSITIONAL CLONING

The strategy for identification of a gene of unknown function affecting risk for a genetic disease is to correlate a relevant phenotype with a sequence that is significantly different in cases and controls. This is called *positional cloning* whether by linkage or other evidence (Collins 1995), despite the fact that the finished DNA sequence provides an efficient alternative to cloning, except in short, critical regions that differ from the standard sequence. Preliminary localization may be based on cytogenetic variation, but linkage and LD are definitive. *Linkage analysis* depends on recombination in families. A few hundred highly polymorphic markers, usually microsatellites, provide a fairly efficient genome scan for detecting linkage, but the candidate regions defined in this way are generally greater than 1 cM for a major gene and much larger for oligogenes. LD is complementary. LD depends on recombination over many generations, and so resolution can be high. Families may be used, but case-control designs that do not depend on living parents and may use hypernormal controls are more efficient (Morton & Collins 1998). Diallelic markers, mostly SNPs, are approximately 100 times more frequent than multiallelic microsatellites, and more economically typed. However, the number required for an efficient genome scan is more than 100 000 and may be 10–100 times greater (Botstein & Risch 2003), reflecting both the large number of recombinants over many generations and the evolutionary variance in their distribution.

Developments in data and analysis will provide more cost-effective SNP typing, and haplotypes may provide more efficient use of LD. Alternatively, phenotypes defined on many individuals may be replaced by expression of candidate loci measured precisely in a small number of somatic cell cultures from appropriate tissues, which may not include the leucocytes and buccal cells commonly used for DNA samples. Whatever form these assays take, they will encounter the universal problem of distinguishing between a causal marker and a highly associated predictive marker, presumably by the same statistical tests currently used for more conventional analysis of LD.

Linkage analysis of major genes is based on good estimates of genetic parameters, allowing for ascertainment through one or more probands (Morton 1959). These parameters include gene frequency, dominance, penetrance and the variance due to other familial factors. At appropriate significance levels (corresponding to $p < 0.001$) many hundreds of loci have been mapped by linkage, with very few false claims (type I errors). The extension of analysis to oligogenes encounters low power, due, in part, to uncertainty about genetic parameters. This led to trial of a (usually unstated) number of alternative models, choosing the one with the highest nominal likelihood and, by contrast, to 'non-parametric' (more precisely 'weakly parametric') methods that reduce genetic parameters to variance components. These work well with quantitative phenotypes and random families, less well in affected relative pairs, and poorly under more complicated ascertainment (Zhang *et al.* 2002*b*).

Even when candidate regions determined from linkage are valid, they are usually large unless reduced by analysis of LD with SNPs. This has led to ambitious proposals for whole genome scans (Risch & Merikangas 1996). Although generally accepted in principle, scanning the genome encounters several problems. If association is tested separately for each SNP, the large number of tests requires a correction as severe as for a very large forensic database. If SNPs are tested jointly, their autocorrelations lead either to unrealistically stringent assumptions about population history or to composite likelihood, and therefore support intervals are imprecise (Devlin *et al.* 1996). If an evolutionary model with Bayesian assumptions is introduced to predict the autocorrelations, ignorance of the population parameters (time since the last bottleneck, effective population sizes, mutation, recombination and admixture) invalidates their estimates unless the location of a causal SNP is known and the posterior probability is mistaken as a prior for the same data (Morris *et al.* 2000). The kilobase scale is only roughly proportional to LD, and therefore has less power for positional cloning. Finally, SNPs fall into blocks with high LD and low haplotype diversity, punctuated by steps with low LD and high haplotype diversity (Daly *et al.* 2001). Blocks are associated with low recombination and sometimes with selective sweeps that increase the frequency of an advantageous allele and the haplotype with which it is associated. Steps are associated with high recombination in the one study that examined crossing over efficiently (Jeffreys *et al.* 2001), but simulation has shown that steps may be generated even in regions of lower recombination. It is a plausible but unproven hypothesis that skilfully analysed haplotypes, defined by an unspecified number of SNPs, may be most efficient for positional cloning by LD.

Experience with positional cloning of rare major genes has been encouraging. Pedigrees with two or more affected relatives allow the inference of susceptible haplotypes if the disease within families is *monophyletic* (inherited from the same founder haplotype). In practice, the number of case haplotypes is enriched, making conventional association metrics inappropriate (Devlin & Risch 1995). This led to an association probability ρ that allows for case enrichment (Collins & Morton 1998) and has an evolutionary interpretation (Morton *et al.* 2001). In random samples with no enrichment ρ equals one of several definitions of D', a two-valued metric originally proposed by Lewontin (1964) as a pair of maxima (one not a probability and the other not perceived as a probability) with no evolutionary interpretation (Weiss & Clark 2002). Success in localizing major genes with ρ encouraged its application to oligogenes where a causal SNP cannot be assigned with confidence to a haplotype but may, nevertheless, be significantly associated in pedigrees, random samples or case control studies. During this transition, pairs of SNPs at distance d kilobases were used to fit the Malecot equation $\rho = (1 - L)Me^{-\varepsilon d} + L$ (Collins *et al.* 1999), where the asymptote may be predicted from allele frequencies and sample size (Maniatis *et al.* 2002). This prediction amalgamates all pairs at the same distance, providing a model and a graph but not a map. It fits estimates from SNPs at low resolution adequately, but does not mirror blocks and steps for dense SNPs and is therefore obsolete. However, a high-resolution model is obtained by replacing εd with $\Sigma\varepsilon_i d_i$, where ε_i is the local value for adjacent SNPs at distance d_i in the unique interval i (Zhang *et al.* 2002*c*). This transformation creates a map in *LD units* (LDU), where 1 LDU corresponds to the number of kilobases in which substantial LD is conserved (the swept radius). At high density a block is defined by $\varepsilon_i = 0$, whereas a step may have $\varepsilon_i d_i > 1$. An LD map measured in LDU fits estimates of ρ much better than other association metrics, and much better than early applications of the Malecot model that measured distance in kilobases or occasionally on the recombination scale in centimorgans.

It would be costly if positional cloning required a population-specific LD map instead of a cosmopolitan map of a major ethnic group or even of our species. Fortunately, a single cosmopolitan map can be scaled to represent local populations efficiently, using only the three Malecot parameters required to fit local pairwise LD to a standard map in LDU (Lonjou *et al.* 2003). This has two significant advantages: first, it accommodates the hope that the number of SNPs selected for positional cloning may be much smaller than the number used to create a standard map; second, it makes the creation of a high-density local LD map extravagant for genome scanning, although useful in candidate regions. These inferences are based on small draft sequences at low resolution and must be confirmed in other data, if only to identify regions refractory to scaling because of stochastic events or selective sweeps. An even better fit to a local map may be obtained by using the cosmopolitan estimates ε_i as trial values, making a single LD map for our species a practical and satisfying goal.

5. HAPLOTYPES

Whereas human linkage analysis began with markers from blood groups and isozymes, LD became useful 30 years later when DNA markers began to provide the high density required for positional cloning. High efficiency could not be reached before a finished DNA sequence was released in 2003. Inevitably, the technical problems are only beginning to be addressed. The first is how to use SNPs most effectively. A genome scan can detect disease genes whose function was not anticipated, but an efficient scan may require more than one million SNPs (Botstein & Risch 2003). A candidate region suggested by linkage or function requires a much smaller number of SNPs, but its result may be negative or at best reveal a locus of minor effect. Currently, a popular approach to this problem is to select a relatively small number of 'haplotype-tagging SNPs (htSNPs)' that parsimoniously identify common haplotypes in a specified haploset but perform less well with diplotypes (Johnson *et al.* 2001; Zhang *et al.* 2002*a*; Stram *et al.* 2003). Some relation to power for positional cloning is assumed but not proven, and unpublished observations do not confirm any relation to power. The pursuit of htSNPs raises four subsidiary problems: (i) what definition of a block is optimal; (ii) what selection algorithm best balances cost and power; (iii) what allowance (if any) should be made for block length, which may vary from 1 to 200 kb and between populations; and (iv) what provision should be made for steps between blocks where LD decreases abruptly, primarily reflecting recombinational hot spots? This complex dilemma suggests that a useful algorithm would accommodate haplotype blocks and steps without requiring their definition, which could be imposed if there were evidence that it improves positional cloning.

Whether or not htSNPs and precise, but arbitrary, block definition prove useful, haplotypes will undoubtedly play a major role in positional cloning. Under certain conditions, haplotypes may be determined with virtually no error (Y chromosomes, mitochondria, the X chromosome in males, deletion heterozygotes, other monosomics spontaneous or induced, and diplotypes with, at most, one heterozygous marker). Even diplotypes with two or more heterozygous markers can be haplotyped in a proportion of children when genotypes of relatives are known. However, haplotyping of individual diplotypes with, at most, family information, as required for positional cloning, is much more demanding than merely estimating haplotype frequencies in a panmictic population. A vector of haplotype frequencies is required for each diplotype, weighting each SNP by its LD information on the assumption of a medial causal SNP. Although simple in principle, there are practical problems. Some computer programs restrict haplotype inference to unambiguous diplotypes. Other programs allow maximal uncertainty by imputing untyped SNPs. At present there is no way to allow for this imputation, which must be balanced against decreased sample size with increasing number of SNPs if imputing is not used. It is conventional to maximize likelihood by the E–M algorithm. This is not parsimonious, and often sample gene frequencies less than the reciprocal of the haplotype count are retained, either because the E–M algorithm (which converges slowly near zero) has not yet converged, or because Hardy–Weinberg or other assumptions are violated. The probability that the E–M algorithm gives a non-parsimonious solution increases with the number of SNPs. One solution is to favour parsimony by the Akaike criterion, minimizing $-2 \ln l + 2k$, where $\ln l$ is the logarithm of likelihood and k is the number of haplotypes with non-zero frequencies (Akaike 1974). These problems must be solved before haplotype mapping enters a tournament with LD mapping, where a composite likelihood is based on multiple single markers. In the absence of a definitive program for haplotype estimation, the serious investigator will borrow solutions from public programmes where available and otherwise modify them or introduce novel solutions using classical statistical theory, with tests of the critical assumptions. The number of SNPs in the most efficient haploset is small (Zaykin *et al.* 2002), and therefore much less than the number of SNPs in a large block. In ways that cannot yet be foreseen despite frenetic research, the analysis must exploit both the efficiency of LD maps and the power of haplotypes covering blocks and steps.

6. HapMap

Annotation for haplotypes of variable size has not been defined or shown to be efficient for positional cloning, but is the currently popular goal. Inspired by the success of the Human Genome Project, and by hope that haplotype blocks and steps will revolutionize positional cloning, an international HapMap Project has undertaken to type 400 000–600 000 SNPs (including no more than 5% of the SNPs contributing to disease) in three ethnic groups. These comprise approximately 30 Yoruban trios, 30 CEPH trios from northwestern Europe, and 90 unrelated Asians, equally divided between Japanese and Chinese (Adam 2001). A satellite study of approximately 50 genomic regions in eight additional populations is planned after mid-2004. The emphasis will be on inferring haplotype frequencies with unspecified numbers of SNPs. Inevitably, the summed haplotype frequency vectors required for positional cloning in any sample will give estimates different from any HapMap standard, which is therefore redundant. The choice of samples is arbitrary (e.g. of one Nigerian tribe instead of a large expatriate population with more diversity). Plans for analysis and representation of LD and haplotypes, including definition of maps, blocks and haplotypes are unsettled. There is a disturbing tendency to make arbitrary deviations from random selection of SNPs in each sample, for example by truncation of the minor allele frequency, even to the extent of choosing only SNPs common in all populations and thereby excluding those causal for regionally adaptive traits, such as skin colour and disease resistance. The initial cost estimate is US$100 million contributed largely by Japan (US$24 million), China (US$10 million), Canada (US$10 million), Britain (US$16 million) and the US (US$37 million). This does not include the eight satellite populations, and the total cost will undoubtedly be greater. The project is administered by a steering committee and advised by working groups representing populations, molecular and statistical analysis, and ethical, legal and social issues (ELSI).

The most intriguing legal issue is the perceived threat that some pharmaceutical companies in some countries will apply for, and secure, patents on haplotypes for particular loci, including haplotypes identified by the HapMap Project, which intends its results to be public. Already a patent application for nine haplotypes defined by seven widely spaced SNPs in the angiotensin receptor 1 gene (AGTR1), and for a database that contains haplotype data on that gene, has been filed in the US Patent Office. Can a haplotype that existed thousands of years before its discovery by familiar methods (made largely by other researchers) be considered a patentable invention? Why should human haplotypes be at risk for a hazard that sequences made public by the Human Genome Project escaped? In any event, legal advice (understandably timid) considers it possible that patent offices may approve such applications unless the HapMap Project defensively patents the haplotypes it will make public, delaying release of these data for 18 months. In the face of opposition by some companies, it is apparently not feasible to take the direct course of laws or high court rulings that bits of the human genome are not a human invention, and therefore not patentable. Because of uncertainties in the HapMap Project, many geneticists were sceptical of its promise before the patenting issue arose (Couzin 2002; Lai *et al.* 2002). A range of opinions about the utility of haplotype annotation will persist even after ownership of our genome is settled by dismissal of the patent application or restriction of its scope. Meanwhile, research to use efficiently the available evidence on LD mapping for positional cloning will continue within and outside the HapMap Project, whatever its product may be. Presentation only of raw data presupposes other groups with the interest, expertise and funding to create a product more immediately useful for positional cloning. The choice to be made of different representations of LD maps or haplotype annotation will make or break HapMap.

7. BIOBANK

Whereas the case for HapMap is not yet proven, it is better justified than the Biobank Project launched in Britain. The research design for a 15 year longitudinal study will use medical records, lifestyle questionnaires, routine biochemistry and a 50 ml blood sample from at least 500 000 British people in an effort to detect the interaction between genes and the environment for common diseases yet to be specified. Biobank is especially embarrassing to genetic epidemiologists because it violates some of our best-tested principles. The management structure of a central 'hub' with regional 'spokes' for recruitment and data collection was tried in the NIH Perinatal Study a generation ago, when US$100 million was a large expenditure indeed. Because supervision was ponderous and the research net was cast so widely, the results were nil. By contrast, under earlier and more stimulating conditions two unlinked major loci were shown to interact, the Le^b antigen in saliva requiring both Le+ and Se+ (Ceppellini 1955). This was the first and clearest example of epistasis in our species. Interaction at the glycoprotein level is more complicated, requiring H specificity for expression of Le^b (Marr *et al.* 1967). The moral is that interaction has a fair chance of being detected in humans only after the main effects of both loci are established, and the same principle applies to gene–environment interaction. There is another weakness of interaction studies in complex inheritance, where penetrancy is low. Mather & Jinks (1982) argued that the scale of gene action is unknown, and therefore the investigator should seek a transformation to make effects additive. This is especially to be recommended when, as in the data to be collected in Biobank, the genes are unspecified and the environment poorly measured.

These and other problems are manifest in a report on a workshop held after Biobank was recommended for funding, but before specific plans were made (MRC 2001). Frances Rawle pointed out that each additional 4 min of interview per participant would cost £1 million. Initial cost estimates were based on 1.5 h of nurse time per volunteer and at least 500 000 volunteers, requiring 119 full-time equivalent research nurses over 5 years, preceding at least a 5 year follow-up on the diminished number of participants. David Porteous estimated £100 million for biochemical studies, not allowing for development of new tests for which the conditions for storage might be inadequate. He entertained the possibility that 'technology development may allow for cell line generation from frozen whole blood', a speculation that cannot at present be rejected or accepted. If expression studies continue to progress, demanding cells other than lymphocytes, the experience with population studies before 1980 will be repeated. It was then good practice to preserve serum and red cells but to discard the buffy coat, leaving inadequate amounts of DNA when that became the object of study. A faint hope of viable lymphocytes from frozen blood is poor preparation for DNA typing, and no preparation at all for expression studies. Paul Burton discussed proposals for a larger cohort, older age at recruitment, and inclusion of spouse and family members, concluding that 'further discussion is clearly required'. Other speakers emphasized the need for additional data on cardiovascular, metabolic, mental health, neurology, cancer, respiratory, infective and musculoskeletal diseases. The general discussion did not address the value of nutritional and other studies confounded with social class; they have been inconsistent in identifying the main effects and so far worthless for interaction studies. The consensus was that the 'costs of meeting the academic objectives of the study might be in excess of £1000 per participant' or more than £500 million if the study is limited as then proposed.

The third report of the Parliamentary Select Committee on Science of Technology expressed concern that Biobank is 'politically driven', without the confidence of the scientific community, and funded at the expense of better research. The report concludes that the Medical Research Council 'has mismanaged its funds', reflecting a wider loss of confidence in the ability of the MRC to keep Britain internationally competitive in biomedical research

(McDowell 2002). Two successive MRC reviews have failed to fund a substantial proportion of highest-graded applications but have approved a huge rise in less productive institutional overheads. The pressure on British science is bound to increase, because the current budget estimate of £58 million for Biobank barely covers lifestyle questionnaires and routine biochemistry, without any clinical or genetic information or lavish creation and staffing of dedicated research space.

Although the Biobank proposal is more vulnerable to criticism than the HapMap project and is expected to cost more than 50 times the British contribution to HapMap, it has been treated more gently by scientists, perhaps because its development is less public and the impact on creative biomedical research in Britain is incalculable. I do not know anyone who is enthusiastic about any aspect of Biobank, although it may have few detractors in countries that hope to recruit British scientists. Two statisticians with experience in genetic epidemiology limited their comments to the advantages of a case-control design over the proposed cohort study (Clayton & McKeigue 2001). Their carefully reasoned conclusion is all the more damaging: 'The requirement that modest risk ratios should be detected, and stringent criteria for statistical significance should be adopted when large numbers of loci are tested, necessitates studies more powerful than any hitherto considered. Since cohort studies sufficiently large for this purpose are unlikely to be practicable, except for a few common diseases, proposals for very large cohort studies of genetic associations should be critically examined against alternatives. The prospects for epidemiology in the post-genome era depend on understanding how to use genetic associations to test hypotheses about causal pathways, rather than on modelling the joint effects of genotype and environment', the number of possible combinations of which require a heavy Bonferroni correction (Agresti 1990).

The same concern is elaborated by three authors who span genetics and epidemiology (Wright *et al.* 2002). They begin with a quotation: 'The most effective disease-related . . . (genetic) association studies are carried out in selective samples of individuals or families at high risk relative to the average risk in the population'. Then they note that 'The goal [of Biobank] is not gene discovery but the identification of interactions between identified genes and environmental factors of public health significance.... The study raises interesting questions as to the specific or general utility of such a large resource'. They summarize the small number of proposed gene–environment interactions in human disease, using a broad definition that includes beta haemoglobin (HBB) with malarial infection, phenylketonuria (PAH) with dietary phenylalanine, and lactose intolerance (LTC) with dietary milk. All the examples involve major genes with narrowly defined phenotypes and environment. They caution that 'only a small minority of environmental influences are measurable' and 'established risk factors such as inappropriate diet, physical inactivity, and tobacco explain 50–75% of the population incidence' in cardiovascular disease, 'but it is not clear that genetic variance in the population has any part at all in such temporal trends'. They argue that a serious attempt to implicate genetic effects would require family-based case-control studies and much better definition of the environment, with little hope that interactions with specific genes could be distinguished.

All these concerns have been ignored in the too rapid, too thoughtless plans for Biobank. Huge studies such as the Human Genome Project, while glamorously successful, are not a good model for genetic epidemiology in areas where the goals are murky, the methods quickly outmoded and the judgement of experts unfavourable. The future of genetic epidemiology lies, like other sciences, with hypothesis-driven research. It is too late to revise Biobank to conform cost-effectively to this principle. A mistake like Biobank does not happen more than once in a generation.

8. ETHICAL ISSUES

On a higher plane, genetic epidemiology poses ethical and social problems. The possibility that patenting of haplotypes may lead to privatization of the human genome has been discussed. That would cripple biomedical research and its applications to disease prevention and treatment. Other issues spring, not from industry, but from the 'uniquely human endeavour of studying some individuals for the possible benefit of others' (Federman *et al.* 2003). Increasingly this is recognized to span three groups of individuals: *subjects* in clinical trials (CIOMS 1993), *patients* receiving accepted treatment or counselling (WHO 1998; Marks 2003), and *participants* in low-risk research not covered by the other two categories. Unfortunately *participant* has also been used in a much more general sense that includes subjects. Most of the literature attempts this generality, with an emphasis on clinical experience in the United States. *Participants* in the strict sense used here have not been considered internationally or separately from high-risk research for 35 years (WHO 1968). A report on the special case of research in the social, behavioural and economic sciences (SBES) is promised this year (Marrett 2003), but low-risk research in human biology remains unrepresented. It includes disease-orientated studies of potential benefit to the sample, their relatives and other populations, as well as studies of human diversity relevant to evolutionary genetics and of interest to the population under study and a wider audience, but not immediately health-related. The maximum hazard is venepuncture, at the same low-level risk as the interviews and questionnaires of the social sciences.

There have been profound changes in the definitions of informed consent and other procedures for responsible research since the earlier WHO report. Following the SBES recommendation, participants in low-risk research should not be regarded as human subjects, but as a special category. Anonymized data collected for other purposes under appropriate safeguards should be exempt from further review, as recognized by the US National Institutes of Health. The special problems associated with DNA samples must be addressed. SBES recommendation 5 is that the review boards 'should consider a variety of procedures for obtaining informed consent and grant waivers of written consent when to do otherwise would inhibit useful SBES research with no appreciable added protection for the participants'. This should be extended to low-risk biomedical research. 'A related concern about written consent procedures (which may also apply to other forms of consent) is that they may not convey what research participants need to know to make an informed decision to participate in a research study and to understand that their participation is voluntary. Research has documented the difficulties of understanding the benefits, harms, and risks of harm of biomedical research as described in consent forms, which are often highly technical in nature'. There is controversy about whether informed consent to a DNA sample should provide the participant with a choice about destroying the sample after a fixed time. Destruction benefits no one and, as yet, the issue has been raised by ethicists but not by participants. Case law has established no such right, which clearly does not apply to DNA from deceased 'participants'. Does the community to which a participant belonged have a right to order the sample destroyed? Who speaks for the community and establishes that the participant belonged to it? Is the community a village, tribe or region? Does a tribe that never surrendered to an invading state retain sovereign powers that allow it to determine the action of its members? These are not hypothetical questions but eminently practical and excitedly discussed in certain situations, where a researcher who violates the changing norm in ceremony but not in principle may be harshly criticized, dead or alive (Morton 2001; ASHG Commentary 2002). There should be a serious attempt to resolve such issues as far as possible in an international context by a respected organization not associated with a particular government. Left to fester as they are now, these problems will cripple population studies that do some good and no harm. A nation that discourages beneficent research cannot solve problems raised by an ageing and increasing population, diminishing resources, evolving pathogens, and competition from more enterprising societies. To let research on human populations languish because the formalism of guidelines is in flux is not merely wrong, it is also a costly mistake.

REFERENCES

Adam, D. 2001 Genetics group targets disease markers in the human sequence. *Nature* **412**, 105.

Agresti, A. 1990 *Categorical data analysis*. New York: Wiley.

Akaike, H. 1974 A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19**, 716–723.

ASHG Commentary 2002 Response to allegations against James V. Neel in *Darkness in El Dorado*, by Patrick Tierney. *Am. J. Hum. Genet.* **70**, 1–70.

Botstein, D. & Risch, N. 2003 Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat. Genet.* **33**, 228–237.

Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. 1980 Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331.

Ceppellini, R. 1955 On the genetics of secretor and Lewis characters: a family study. In *Proc. 5th Congr. Int. Soc. Blood Transf. Paris*, pp. 207–211.

CIOMS 1993 *International ethical guidelines for biomedical research involving human subjects*. Geneva: Council for International Organizations of Medical Sciences.

Clayton, D. & McKeigue, P. M. 2001 Epidemiological methods for studying genes and environmental factors in common diseases. *The Lancet* **358**, 1356–1360.

Collins, A. & Morton, N. E. 1998 Mapping a disease locus by allele association. *Proc. Natl Acad. Sci. USA* **95**, 1741–1745.

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

Collins, F. S. 1995 Positional cloning moves from perditional to traditional. *Nat. Genet.* **9**, 347–350.

Committee on DNA Forensic Science 1992 *DNA technology in forensic science*. Washington, DC: National Research Council, National Academy Press.

Committee on DNA Forensic Science 1996 *The evaluation of forensic DNA evidence*. Washington, DC: National Research Council, National Academy Press.

Couzin, J. 2002 New mapping project splits the community. *Science* **296**, 1391–1393.

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

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

Devlin, B., Risch, N. & Roeder, K. 1996 Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics* **15**, 1–16.

Federman, D. D., Hanna, K. E. & Rodriguez, L. L. (eds) 2003 *Responsible research. A systems approach to protecting research participants*. Washington, DC: Institute of Medicine, National Academy Press.

Jeffreys, A. J., Wilson, V. & Their, S. L. 1985 Individual-specific 'fingerprints' of human DNA. *Nature* **316**, 75–79.

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

Johnson, G. C. (and 20 others) 2001 Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**, 233–237.

Lai, E., Bowman, C., Bonsal, A., Hughes, A., Mosteller, M. & Roses, A. D. 2002 Medical applications of haplotype-based SNP maps: learning to walk before we run. *Nat. Genet.* **32**, 353.

Lerner, I. M. 1950 *Population genetics and animal improvement*. Cambridge University Press.

Lewontin, R. C. 1964 The interaction of selection and linkage. I. General considerations. *Genetics* **49**, 49–67.

Li, C. C. 1948 *An introduction to population genetics*. Peking: National Peking University Press.

Li, C. C. 1955 *Population genetics*. University of Chicago Press.

Lonjou, C., Zhang, W., Collins, A., Tapper, W. J., Elahi, E., Maniatis, N. & Morton, N. E. 2003 Linkage disequilibrium in human populations. *Proc. Natl Acad. Sci. USA* **100**, 6069–6074.

McDowell, N. 2002 Top projects suffer as medical funding falters. *Nature* **418**, 714.

Maniatis, N., Collins, A., Xu, C.-F., McCarthy, L. C., Hewett, D. R., Tapper, W., Ennis, S., Ke, X. & Morton, N. E. 2002 The first linkage disequilibrium (LD) maps: delineation of hot and cold spots by diplotype analysis. *Proc. Natl Acad. Sci. USA* **99**, 2228–2233.

Marks, P. 2003 The evolution of the doctrine of consent. *Clin. Med.* **3**, 45–47.

Marr, A. M. S., Donald, A. S. R., Watkins, W. M. & Morgan, W. T. J. 1967 Molecular and genetic aspects of human blood-group Le^b specificity. *Nature* **215**, 1345–1349.

Marrett, C. 2003 Appendix B. Protecting participants in social, behavioural, and economic science research: issues, current problems, and potential solutions, report from the panel on IRBs, surveys and social science research. In *Responsible research. A systems approach to protecting research participants* (ed. D. D. Federman, K. E. Hanna & L. L. Rodriguez), pp. 236–248. Washington, DC: Institute of Medicine, National Academy Press.

Mather, K. & Jinks, J. L. 1982 *Biometrical genetics*. London: Chapman & Hall.

Morris, A. P., Whittaker, J. C. & Balding, D. J. 2000 Bayesian fine-scale mapping of disease loci by hidden Markov models. *Am. J. Hum. Genet.* **67**, 155–169.

Morton, N. E. 1956 The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. *Am. J. Hum. Genet.* **8**, 80–96.

Morton, N. E. 1959 Genetic tests under incomplete ascertainment. *Am. J. Hum. Genet.* **11**, 1–16.

Morton, N. E. 1997 The forensic DNA endgame. *Jurimetrics* **37**, 477–494.

Morton, N. E. 2001 Darkness in El Dorado: human genetics on trial. *J. Genet.* **80**, 45–52.

Morton, N. E., Chung, C. S. & Mi, M.-P. 1967 *Genetics of interracial crosses in Hawaii S*. Basel: Karger.

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

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

MRC 2001 Report of the UK population biomedical collection protocol development workshop. London: Medical Research Council.

Neel, J. V. & Schull, W. J. 1954 *Human heredity*. University of Chicago Press.

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

Solomon, E. & Bodmer, W. F. 1979 Evolution of sickle variant gene. *The Lancet* **1**, 923.

Stram, D. O., Haiman, C. A., Hirschhorn, J. N., Altshuler, D., Kolonel, L. N., Henderson, B. E. & Pike, M. C. 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. Heredity* **55**, 27–36.

Tapper, W., Morton, N. E., Dunham, I., Ke, X. & Collins, A. 2001 A sequence-based map of chromosome 22. *Genome Res.* **11**, 1290–1295.

Weiss, K. M. & Clark, A. G. 2002 Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**, 19–24.

WHO 1968 Research on human population genetics. World Health Organization technical report. Series no. 387, Geneva.

WHO 1998 Proposed international guidelines on ethical issues in medical genetics and genetic services. World Health Organization Human Genetics Programme, Geneva.

Wright, A. F., Carothers, A. D. & Campbell, H. 2002 Gene–environment interactions: the Biobank UK study. *Pharmacogenomics J.* **2**, 75–82.

Zaykin, D. V., Westfall, P. H., Young, S. S., Karnoub, M. A., Wagner, M. J. & Ehm, M. G. 2002 Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum. Heredity* **53**, 79–81.

Zhang, K., Calabrese, P., Nordborg, M. & Sun, F. 2002*a* Haplotype block structure and its applications to association studies. *Am. J. Hum. Genet.* **71**, 1386–1394.

Zhang, W., Collins, A., Lonjou, C. & Morton, N. E. 2002*b* A linkage tournament: affection status, parametric analysis, multivariate traits, and enhancements to variance components and relative pairs. *A. Hum. Genet.* **66**, 87–98.

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