MAPPING BY ADMIXTURE LINKAGE DISEQUILIBRIUM: ADVANCES, LIMITATIONS AND GUIDELINES

Michael W. Smith‡ and Stephen J. O'Brien**

Abstract | Mapping by admixture linkage disequilibrium (MALD) is a theoretically powerful, although unproven, approach to mapping genetic variants that are involved in human disease. MALD takes advantage of long-range haplotypes that are generated by gene flow among recently admixed ethnic groups, such as African-Americans and Latinos. Under ideal circumstances, MALD will have more power to detect some genetic variants than other types of genome-wide association study that are carried out among more ethnically homogeneous populations. It will also require 200–500 times fewer markers, providing a significant economic advantage. The MALD approach is now being applied, with results expected in the near future.

LINKAGE MAPPING
A method for localizing genes that is based on the co-inheritance of genetic markers and phenotypes in families over several generations.

ASSOCIATION STUDIES
A gene-discovery strategy that compares cases with controls to assess the contribution of genetic variants to phenotypes in specific populations.

**Laboratory of Genomic Diversity, National Cancer Institute, Frederick, Maryland 21702, USA.*
‡Basic Research Program, SAIC-Frederick, National Cancer Institute.
Correspondence to M.W.S.
e-mail: smithm@ncifcrf.gov
doi:10.1038/nrg1657
Published online 12 July 2005

Human genetics faces an important and challenging goal in identifying the genetic variants that underlie complex disease. Two main approaches are available for mapping the relevant genes and identifying the variants that cause disease in humans: LINKAGE MAPPING in families and population-based genetic ASSOCIATION STUDIES. Linkage mapping has been very successful in finding genes for rare, Mendelian, monogenic diseases for which there is a strong familial risk. However, for complex diseases that involve variants at several loci, each of which contribute small amounts to the overall genetic contribution, linkage studies mainly identify only those loci that have the strongest influence.

In theory, genetic association mapping has greater power than linkage studies to identify variants with weak effects that might contribute risk for common complex diseases[1]. Whole-genome association studies have the advantage of allowing the entire genome to be assessed for disease-associated variants, rather than analysing specific candidate genes. However, the disadvantage of such studies is that a large amount of genotyping is required. This can be reduced by using a subset of markers to report on neighbouring linked markers within the same HAPLOTYPE. The International HapMap Project has begun to define the haplotype structures of three major ethnic groups (European, African and eastern Asian populations) as a prerequisite for mounting whole-genome, haplotype-based association scans by genotyping a reduced number of tagging variants[2–4]. However, it is estimated that even after this has been completed, 300,000–1,000,000 SNPs will need to be assessed to scan the human genome using available haplotypes. Until genotyping costs decrease substantially, the HapMap approach to disease gene discovery would require several million dollars for a single association study.

Mapping by admixture linkage disequilibrium (MALD), also known as admixture mapping, is a genetic association strategy that makes use of one of the consequences of ADMIXTURE. The gene flow that takes place during admixture results in the temporary generation of long haplotype blocks, which include polymorphic variants — inducing a phenomenon known as ADMIXTURE LINKAGE DISEQUILIBRIUM (ALD). MALD has statistical power that is similar to association mapping to detect disease-associated variants that differ markedly in frequency between populations[5–8]. It uses admixture that has occurred over the past several hundred years. This places it somewhere between association mapping, in which recombination over hundreds to thousands of generations has broken apart closely linked markers, and linkage mapping, in which a few generations are studied to estimate gene locations.

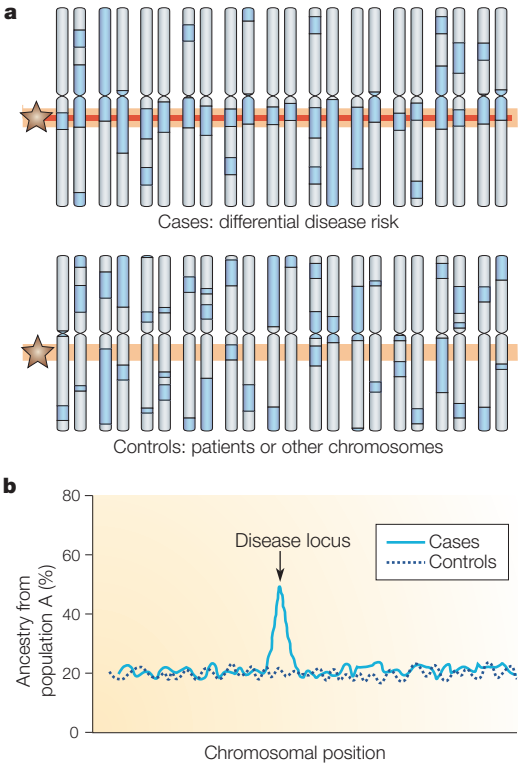

Figure 1 | **Detecting disease-associated genomic regions using mapping by admixture linkage disequilibrium. a** | The strategy that is used to assess the ancestral origin of chromosomal segments in mapping by admixture linkage disequilibrium (MALD)[7,13,15]. Genotyping MALD markers is used to assess parental ancestry across a single chromosome in multiple cases (individuals with the disease of interest) versus matched healthy controls. The region indicated by the star is derived more often from one of the parental populations only in the disease cases, indicating that this region contains a disease-susceptibility locus. In the controls, the same region has an equal probability of originating from either parental population. **b** | A theoretical example of how an admixture signal can be detected using the MALD method for a disease with a higher incidence in one parental population (population A). The proportion of ancestry from population A in multiple individuals (both with the disease (cases) and without the disease (controls)) is shown schematically for different positions on a single chromosome. An elevated ancestry proportion from population A in cases is evident at the peak (marked by an arrow), which indicates the involvement of the corresponding genomic region in the disease. The peak can be identified by the higher (or lower; not shown) level of ancestry that is seen in cases relative to the same region in controls, and/or relative to the remainder of the genome in cases (only the neighbouring chromosomal region is shown here). Part **b** is modified, with permission, from REF. 13 © (2004) The University of Chicago Press.

Unlike whole-genome association studies within genetically homogeneous populations, MALD is economically feasible today because it requires studying 2,000–3,000 markers — 200–500 times fewer than are required for whole-genome haplotype-based mapping[5–8]. In MALD, the ability to detect haplotypes that contain a disease-associated variant for a specific complex disease is maximized by analysing the markers that are most divergent between the parent populations of the admixed group (for example, African and Euopean populations, in the case of African-Americans)[5–11]. Recently, high-density short tandem repeat (STR) and SNP marker maps for MALD have been developed[10,11], along with powerful analytical tools[12–16]. It is therefore now becoming feasible to carry out MALD with sufficient statistical power and to rigorously analyse the results. Here we describe the theory behind the MALD approach, its development, and its feasibility for different diseases and populations. We also discuss statistical power, and assess the advantages and disadvantages of MALD as a tool for identifying genetic variants that are involved in complex disease.

The theory of MALD

MALD was theoretically proposed as an approach to disease-gene mapping more than a decade ago by Chakraborty and Weiss[17], who suggested that the consequences of admixture could potentially be used for gene localization. Subsequently, Briscoe *et al.*[5] and Stephens *et al.*[6] described initial parameters, algebraic simulations and guidelines for detecting linkage disequilibrium (LD) of disease genes and MALD markers with large frequency differences between the parental populations of admixed groups. A seminal advance was the concept of assessing the parental-population ancestry of individual chromosomal segments[7] across the genome of individuals in an admixed study group (see below). This stimulated the development of the robust statistical approaches discussed in this review[7,12,13,15,16,18,19].

MALD is potentially suitable for localizing disease-gene polymorphisms with substantial allele-frequency differences between ethnic groups that formed in the past few hundred years through admixture, or between the parental populations of these groups. The process of admixture can be conceptualized as generating large chromosomal segments that originate from one or the other parental population. This occurs through the process of recombination that takes place after gene flow commences between ethnic groups (see supplementary information S1 (figure)). The resulting ALD causes non-random association between loci across large distances of tens of centimorgans or more. ALD decays quickly between loci that are separated by large physical distances (those that are >30 cM apart, or are on different chromosomes), but more slowly among closely linked loci (those that are <10 cM apart)[5,6,17]. Central to the MALD approach are the resulting chromosomal blocks of discrete ancestry that can be used to track disease-gene alleles that might be included within them (see supplementary information S1 (figure)).

The idea of the MALD approach is to screen across the genome in a population of individuals of mixed ancestry, specifically in individuals who are affected by the disease of interest. The strategy looks for regions with an unusually high representation of a particular chromosomal segment from the parental population with the higher risk for the disease (FIG. 1). The reason for the high frequency of a particular chromosomal segment is due to the location of the disease-gene allele within it.

HAPLOTYPE
The sequence of a single chromosome, summarized as a unique combination of known polymorphic sites.

ADMIXTURE
The formation of a new population by interbreeding between individuals from genetically divergent parental populations, and subsequently by interbreeding between their offspring.

ADMIXTURE LINKAGE DISEQUILIBRIUM
The non-random association of genetic variants due to admixture that decays rapidly (in a few generations) between unlinked genes and more slowly between linked ones.

Table 1 | **Diseases with different risks in Africans and Europeans***

Disease or related trait	Population relative risk (African vs European)	95% Confidence interval	References
Lower relative risk in African-Americans			
Hepatitis C clearance	0.19	(0.10–0.38)	48
HIV vertical transmission	0.30	(0.10–0.90)	49
Multiple sclerosis	0.50	n.d.	50
Atrial fibrillation	0.51	(0.31–0.76)	51
Coronary artery disease	0.75	(0.60–0.95)	52
Carotid artery disease	0.62	(0.46–0.82)	52
Osteoporosis/BMD‡	Lower§	n.a.	53,54
Higher relative risk in African-Americans			
Lupus nephritis with systemic lupus erythematosus	3.13	(1.21–8.09)	55
Myeloma	3.14	(2.00–4.93)	56
Dementia	3.21	(2.18–4.73)	57
Prostate cancer	2.73	(2.13–3.52)	56
Hypertensive heart disease	2.80	(2.03–3.86)	56
Pregnancy-related death	2.65	(1.73–4.07)	58
Hypertension	2.61	(2.09–3.27)	52
Focal segmental glomerulosclerosis	2.49	(1.05–5.95)	59
Intracranial haemorrhage	2.10	(1.44–3.06)	56
Non-insulin dependent diabetes	1.99	(1.60–2.48)	52,60
End-stage renal disease	1.87	(1.47–2.39)	61
Stroke	1.57 1.30–5.00‖	(1.27–1.94) (1.00–1.61)	56 62
Hypertensive retinopathy	1.48	(1.08–2.03)	63
Lung cancer	1.48	(1.30–1.67)	56
HIV progression	1.41	(1.06–1.86)	64
Obesity/BMI	Higher§	n.a.	65
Systemic lupus erythematosus	Higher§	n.a.	66
Systemic sclerosis	Higher§	n.a.	67

*These diseases are potentially amenable to mapping by admixture linkage disequilibrium. ‡In a study of bone mineral density (BMD) in boys, race accounted for 13% of overall BMD (with a partial r^2 — the proportion of the variance explained) and as much as 37% of femoral neck BMD among the 5 sites tested[68]. §Some studies present evidence of significant differences in the direction noted without specific numerical quantification. ‖The range of increased risk of stroke analysed across a full spectrum of ages. BMI, body mass index; HIV, human immunodeficiency virus; n.a., not applicable; n.d., no data.

The MALD strategy involves 5 steps. First, a cohort for a particular disease is developed from an admixed ethnic group, ideally with large disease-incidence differences between the parental populations. Second, cases and controls are genotyped with a set of polymorphic markers that are highly informative about ancestry (for example, SNPs that differ in frequency by >60% between the two ancestral populations)[9,11]. Third, the patchwork ancestry profile across the genome of every individual is assessed[13–15] (see supplementary information S1 (figure)). Fourth, chromosomal regions that have an elevated frequency (compared with other genomic regions) of the ancestry with the higher disease incidence are identified (FIG. 1). Finally, the causative allele in the implicated chromosome segment is identified by high-density SNP-based association studies, followed by inspection of the candidate genes that lie within the region.

There are several requirements that need to be fulfilled for MALD to be feasible for a particular disease and population, which are discussed in the sections below. First, MALD-based identification of disease genes requires a measurable difference in the frequency of disease-causing alleles between the parental populations. Second, admixture ideally needs to be at least two generations old to reduce the initial disequilibrium across chromosomes and between unlinked loci, while LD within chromosomes remains strong[6]. Third, a set of markers that specifically differentiate chromosomes derived from the parental populations is needed[10,11].

Which diseases are suitable for MALD?

As indicated above, the most favourable opportunities for using MALD involve complex genetically influenced diseases for which the incidence is very different between the parental populations. Such diseases have the highest *a priori* chance of there being a difference in the frequencies of genetic risk factors between populations. However, incidence differences among ethnic groups might derive from socioeconomic or environmental factors, as well as from differing genetic risk. For MALD-based gene discovery, the likelihood of success depends on the relative influences of genetic versus environmental factors. For example, consider a multifactorial disease in which the frequencies of disease-associated alleles in the parental populations are similar, but distinct environmental or social influences result in different incidences of disease between the two. MALD would fail in this case. Conversely, the MALD approach would succeed in the case of a disease with similar incidences in the parental populations in which different disease-gene frequencies are offset by socioeconomic and/or environmental influences.

The diseases for which MALD has the highest *a priori* chance of success occur in African-Americans, a group in which there are several well-known cases of diseases that differ in incidence between the parental African and European groups. We identified 25 diseases as potential candidates for MALD-based gene discovery among African-Americans through a search of the literature (TABLE 1), including various cancers, immunological pathologies, non-insulin-dependent diabetes and multiple sclerosis. Several diseases in TABLE 1 are related in terms of their pathogenesis (for example, the different forms of heart disease and the different cancers), and might therefore share causal disease genes. The list is not exhaustive, but might perhaps stimulate the design of MALD-based gene-discovery

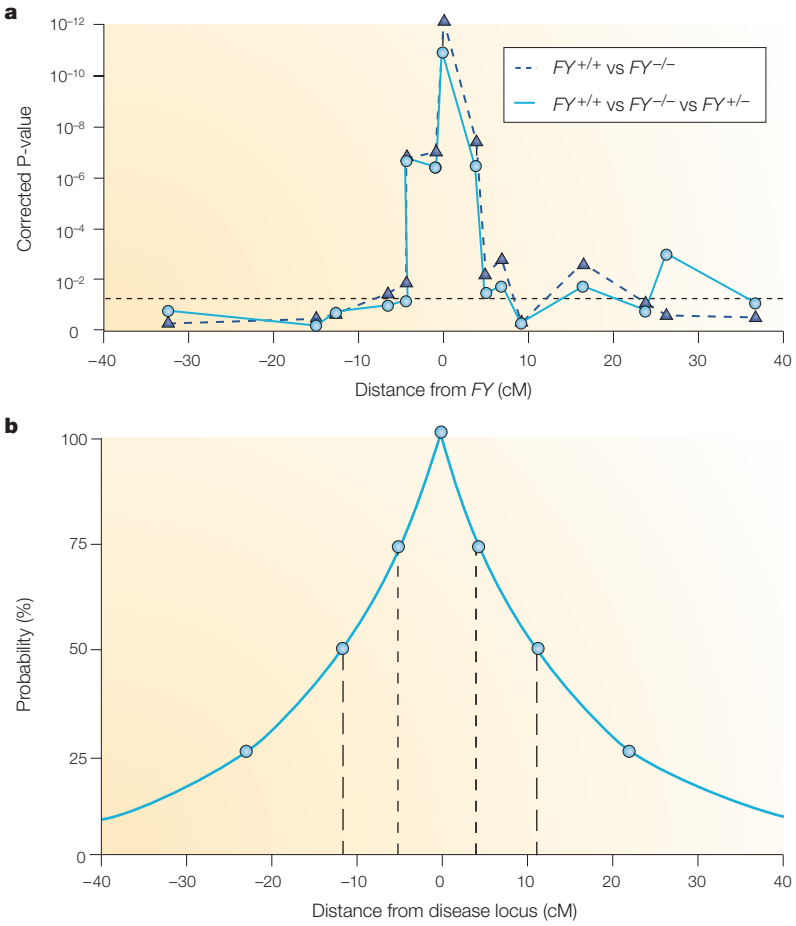

Figure 2 | **Assessment of linkage disequilibrium that is caused by admixture in African-Americans. a** | The extent of admixture linkage disequilibrium (ALD) in the Duffy blood group (*FY*) region. The alternate FIXATION of the FY^{-} allele in African populations and the FY^{+} allele in European populations — an extreme example of differentiation between the two continental groups — allows the tracking of ALD between the *FY* polymorphism and neighbouring markers. The association of flanking markers with the FY^{-} polymorphism was assessed for 15 STRs and 1 insertion–deletion (indel) polymorphism at the *AT3* (antithrombin III) gene[34]. The *x* axis shows the position of these markers relative to the *FY* locus and the *y* axis shows the strength of the association of these markers with the FY^{-} allele (expressed as a P-value corrected for the number of tests carried out). The position of each marker is indicated as a data point on the graph. The results of CONTINGENCY ANALYSIS of the two homozygous *FY* genotypes (+/+ and –/–) are shown with a dotted curve and triangles indicate the data points. An analysis that included heterozygous individuals (+/–) is also illustrated, as indicated by the solid blue curve and circles representing the data points. The dotted line indicates a corrected probability of 0.05. Modified, with permission, from REF. 34 © (2000) The University of Chicago Press. **b** | The probability of no recombination having taken place between a disease locus and a genetic marker since admixture began in African-Americans. Data are based on the genome-wide assessment of mapping by admixture linkage disequilibrium markers from Smith *et al.*[11] and the same *x*-axis scale is used as in part **a**. The probability is shown as a function of the distance of the marker from the disease locus, based on individual estimates of the number of generations since admixture[11]. Dotted lines indicate the points that correspond to 75% and 50% ancestral segment recombination from the disease locus. Modified, with permission, from REF. 11 © (2004) The University of Chicago Press.

projects in African-American groups for these and other complex diseases.

The extent of admixture and MALD feasibility

A level of admixture that is greater than 10% is needed for MALD to be feasible, and it is therefore important to assess the levels of admixture in study populations. Recent efforts have estimated the extent of admixture in various ethnic groups using LEAST-SQUARED, MAXIMUM LIKELIHOOD, and BAYESIAN analytical approaches[20–25]. Although there is genetic heterogeneity within admixed groups, and any generalized description of a population is necessarily simplified, MALD-based studies can still provide valuable information. African-American populations are on average approximately 80% African and 20% European in genetic origin (ranging from 1–90% European in different individuals). Latino populations show a rough average of 50% European and 50% Native American origin, although some Latino populations have large African contributions, especially in the Caribbean basin and parts of South America. The European contribution to Latino ancestry can range from 33–95%, Native American from 0–58% and western African from 0–29%, with proportions that vary greatly from population to population[26–30]. There is also evidence of admixture in Australian Aborigines[31] and among populations of the Pacific islands, such as Hawaii[32].

The extent of ALD, which is related to the degree and dynamics of admixture, is also important for determining the feasibility of MALD, as it determines the size of ALD blocks on chromosomes. The extent of ALD within admixed populations is dependent on the number of generations of gene flow, the variation in contribution from the two populations over time and, especially, the degree of recent mixing with either parental population[33]. The simplest view of ALD dynamics involves a single generation of admixture followed by PANMIXIA (see supplementary information S1 (figure)). However, a more realistic situation is one of continuous admixture[6,33]. Continuous admixture results in a wider range of LD segment-sizes derived from the ancestral populations, and ensures that chromosomal block sizes of at least a few centimorgans, which are necessary for the use of MALD for gene discovery, are maintained.

An empirical test of the extent of ALD in the African-American population came from analysis of the DUFFY BLOOD GROUP gene (*FY*), which shows large allele-frequency differences between ethnic groups[20,34]. Analysis of STRs in the *FY* region in a group of African-Americans showed that LD extends across a 30-cM region, but is strongest for a flanking interval of 5–10 cM, which is centred on the *FY* gene itself[34] (FIG. 2a). Other studies have demonstrated ALD of a similar extent when examining large blocks of other genomic regions[13,14]. Analysis of a group of 3,011 MALD markers across the genome revealed an average LD block-size of 12 cM in an African-American population sample[11,13] (FIG. 2b), indicating an average of 6–7 generations of admixture in this ethnic group[11,13]. These findings of a moderate number of generations since admixture began, and the extended levels of ALD that are detectable across tens of centimorgans, indicate that MALD is feasible in African-Americans and, by direct extension and supposition, in many less-well-characterized admixed populations.

FIXATION
Fixation occurs when a specific allele at a locus is found exclusively in one population but in another, an alternative allele is exclusively present.

Marker maps for MALD

Maps for African-American populations. A key requirement for MALD in African-Americans is a set of genetic markers that provide information about whether the regions of the genome under study have an African or European origin. In 2001, Smith *et al.*[10] put forward a MALD marker map of 744 STRs for the identification of regions with large frequency differences between the two parental populations. More recently, we have developed a set of 3,011 SNP markers that can be used for MALD-based disease-gene discovery in African-Americans[11]. This MALD SNP map was generated after searching approximately 450,000 SNP allele-frequency estimates to find markers with the greatest frequency differences between African-Americans and European-Americans. Evenly spaced SNP markers (with a mean interval of 1 cM) with the largest SHANNON INFORMATION CONTENT (SIC; also known as mutual information; see supplementary information S2 (box)) — which have values of at least 0.035 — were selected. These markers were genotyped in an independent population of individuals from western African, European-American and African-American descent, along with a limited sample of groups with Asian and Native American origins[11].

The SNP marker set for MALD can be reduced to an analysis set of 2,154 SNPs (with an average density of 1 SNP per 1.5 cM) by eliminating the markers that show LD with other SNPs in the parental European or western African populations[11]. Currently, genotyping platforms produced by Illumina and Parallele allow 1,500 to 20,000 genotype determinations with a single assay, and the array-based Affymetrix technology can genotype 500,000 random SNPs. Our first-generation African-American MALD marker set[11] conveniently fits on the Parallele or Illumina platforms.

We estimate that the current MALD SNP map would provide approximately 50–70% of the maximum information about ancestry. In other words, the informativeness of population-discriminating MALD markers would need to be roughly doubled to carry out a fully informative genome scan[11]. The information provided by the current MALD map is comparable to that achieved with the standard Marshfield map that is used for family-based linkage studies, which contains ~300 microsatellite markers. Whole-genome MALD mapping in African-Americans is therefore just as practical and powerful today as linkage mapping has been for the past decade[8,35].

The resources for developing MALD maps for African-American populations will increase significantly in the near future. As described above, the first set of SNPs for MALD was derived from 450,000 allele-frequency estimates[11]. The HapMap project[2] will have genotyped nearly five-million markers by December 2005 and Perlegen has recently released information on 1,586,383 SNPs, with frequencies measured both in European-Americans and African-Americans[36]. Together, these millions of extra SNP frequencies will be mined rapidly to build the next-generation MALD map for African-American populations and other admixed groups. This should make it possible to obtain 85–90% informativeness about whether an individual has African or European ancestry at a specific locus, but will still require the genotyping of just two to three thousand MALD markers.

Maps for other populations. In terms of developing MALD markers, the challenges are greater for other admixed populations compared with African-Americans. In these other groups, the extent of admixture is not as well understood and dense haplotype maps are only just becoming available or being developed. One approach to building maps for these populations is to select informative SNPs with a high SIC from 'parental' populations for Latinos, Mexican-Americans, Hawaiians, Aborigines and other admixed groups.

Alternatively, Pritchard and colleagues[12,35] recently suggested that developing MALD markers with large differences in frequency between parental populations will become less important as dense genotyping becomes more economical. For example, consider a genome scan of an admixed study group using 45,000 SNP markers, which have not been selected for population differences, at a density of at least 1 SNP per 0.1 cM (now feasible with 100,000 SNP Affymetrix array-genotyping platforms). Here, ancestry assignments for chromosomal segments could be achieved by using the incremental increase in information obtained by examining large numbers of random markers. Genotyping using these dense SNP sets, with parental population samples as a reference, would allow the determination of chromosomal segmental ancestry, albeit at ten times the cost of using 3,000 MALD markers at a 1 cM density. Nonetheless, this approach would have the efficacy and power of MALD, and would be feasible for populations that are less well defined than African-Americans, and for which no MALD marker set is available.

Apart from the cost of the high-density random SNP approach to admixture mapping, another concern is the use of numerous linked markers that are in LD in the parental populations. Although this potential problem was not evident in the simulated data presented by Montana and Pritchard[12], the inclusion in genome scans of markers that show even weak LD in ancestral populations could contribute to false-positive associations. This is because they violate the assumption of independence between adjacent markers (D. Reich and N. Patterson, personal communication). Continuing admixture can exacerbate this effect. Because of this concern, we excluded SNP markers that are in LD in the parental populations from MALD disease-gene ascertainment studies carried out using our low-density SNP map[11]. Assessment of markers that are in LD in parental populations, and their exclusion, would be much more difficult using randomly selected, dense SNPs. Nevertheless, as dense HapMap data become available from representatives of the parental populations that formed many admixed groups, the prospects for developing efficacious MALD marker maps are excellent.

CONTINGENCY ANALYSIS
A chi-squared analysis of the numbers of observations to test for differences between categories in a data table.

LEAST-SQUARED APPROACH
A statistical estimation technique that estimates parameters on the basis of minimizing the square of the differences between a model and the observations.

MAXIMUM LIKELIHOOD APPROACH
A method for estimating parameter values in a model that have the highest probability of explaining the data observed.

BAYESIAN APPROACH
A statistical methodology that takes prior knowledge into account.

PANMIXIA
The process by which individuals in a population choose each other as mates with equal likelihood.

DUFFY BLOOD GROUP
Encoded by the *FY* gene, this is an antigen expressed on red blood cells that is a scavenger receptor for chemokines and also serves as a receptor for the malarial parasite, *Plasmodium vivax*.

SHANNON INFORMATION CONTENT
A measure that is used to quantify the informativeness of a marker or set of markers for determining the ancestral state of a chromosomal segment or locus.

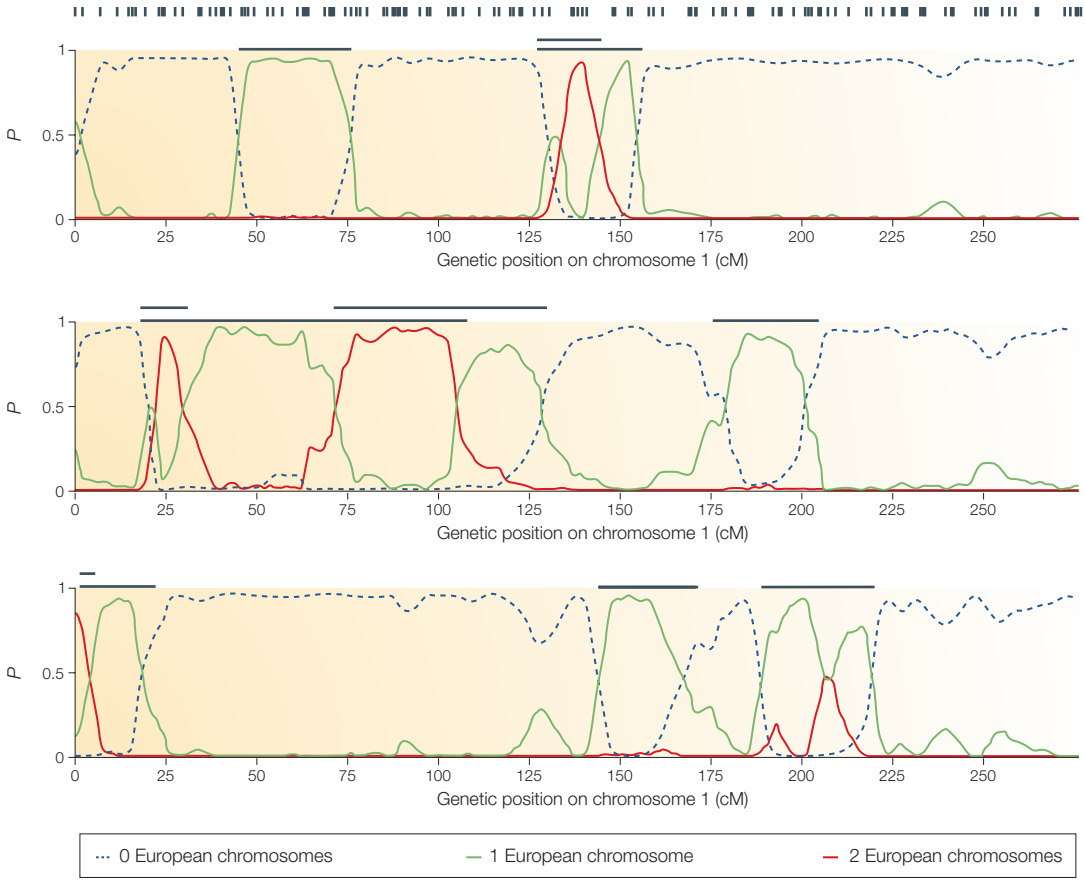

Figure 3 | **Ancestry-assessment estimates using the ANCESTRYMAP algorithm.** The figure illustrates ancestry-assessment estimates of the number of European chromosome segments in African-American individuals using ANCESTRYMAP. Assessments were carried out for representative African-Americans across chromosome 1 (REF. 11); results from three individuals are shown here. Mapping by admixture linkage disequilibrium mapping markers were genotyped, and the position of each marker on the map is indicated by the hatch marks across the top of the figure. Each individual was assessed for the probability (*P*) of having 0, 1 or 2 European chromosomes at positions along both copies of chromosome 1. The bars across the top of each plot indicate segments of European ancestry. Modified, with permission, from REF. 11 © (2004) The University of Chicago Press.

HIDDEN MARKOV MODEL (HMM). A statistical model of a sequence of events for which the probability of an event occurring depends on previous and subsequent events occurring. It is useful in admixture mapping as a complex and interdependent model can be calculated to fit the segmental nature of admixed chromosomes.

MARKOV CHAIN MONTE CARLO (MCMC). The distributions underlying the hidden Markov model are extremely complex, making their direct estimation a huge task. This is simplified in MCMC analysis by generating averages of the expectations from the underlying distributions to model and analyse the results of admixture mapping.

Analytical approaches for MALD

Gene discovery by MALD depends on the ability to detect the departure of a causative disease allele from genetic equilibrium in an admixed study population. Early proposals suggested that disease genes could be identified on the basis of their being in LD with MALD markers that show a departure from population-equilibrium frequencies among groups of people that are affected by the disease[5,6,17]. However, more recently, innovative detection strategies have used chromosomal ancestry-assignment algorithms that allow the quantitative assessment of the ancestry of individual loci along each chromosome (FIG. 3). Five research groups have developed HIDDEN MARKOV MODEL (HMM) methodologies to make robust estimates of segmental ancestry using parental-population allele-frequency estimates for each locus and its adjacent markers, although the individual loci might only be moderately informative about ancestry[7,12,13,15,16,19]. In theory, an HMM approach extracts the maximum information about ancestry that is available in a data set.

A simple HMM analysis requires that the precise frequencies of the alleles that correspond to the markers examined are known in the parental populations. Errors in frequency estimates for ancestral populations can lead to false-positive inferences of association with the disease[13]. For example, if the variant that is more common in European-Americans is estimated to have an erroneously high frequency in Europeans, African-Americans will be inferred to have an unusually low amount of European ancestry in this section of the genome, which is exactly what would be expected for a disease gene — a false-positive association. To address the uncertainty in allele-frequency estimates in parental populations explicitly, three of the new HMM approaches (see supplementary information S3 (table)) have nested the HMM inside a MARKOV CHAIN MONTE CARLO (MCMC) analysis[12,13,19]. This nesting corrects for the uncertainty about parental-population allele frequencies by taking into account those frequencies found in the admixed population.

Supplementary information S3 (table) compares the three combined MCMC/HMM methods. The methods

Table 2 | **Sample sizes needed for MALD in African-Americans***

Allele frequency difference	Allele frequency ratio (African: European)	ARR = 0.5	ARR = 1.5	ARR = 2.5
60%				
	70%:10%	457	1,418	320
	10%:70%	666	1,104	181
	80%:20%	396	1,532	375
	20%:80%	592	1,205	223
	90%:30%	339	1,650	435
	30%:90%	522	1,310	270
50%				
	60%:10%	729	1,925	407
	10%:60%	988	1,556	248
	70%:20%	636	2,084	482
	20%:70%	880	1,699	307
	80%:30%	550	2,249	563
	30%:80%	777	1,849	372
40%				
	50%:10%	1,257	2,830	556
	10%:50%	1,590	2,378	367
	60%:20%	1,104	3,071	666
	20%:60%	1,418	2,600	457
	70%:30%	961	3,322	786
	30%:70%	1,256	2,831	557

*Sample sizes are shown for 80% statistical power. Power calculations assume that the mapping by admixture linkage disequilibrium (MALD) map used is 100% informative in determining ancestry at each locus. The available MALD map[11] is estimated as being 50–80% informative; so sample-size estimates would need to be increased by 25–100%. See REFS 12,15,16 for more details and power calculations. ARR, allelic relative risk.

differ in several aspects: the approaches taken to test for association, their ability to examine the X chromosome, whether they allow for quantitative trait analysis, their ability to test for association in ancestral populations and the speed of the software. They also vary in whether PARAMETRIC or NON-PARAMETRIC test statistics are calculated. Extensive testing by simulation has been carried out for the ANCESTRYMAP software, and to a lesser extent for the other two methods. Overall, ANCESTRYMAP provides a promising method of analysis with a favourable software speed, but it is likely that all MALD software will improve markedly as the method is applied to real samples and diseases.

The extensive simulation tests of ANCESTRYMAP indicate that the MCMC-based admixture mapping is not prone to false positives[13] (see also the series of simulations by Montana and Pritchard[12]). Because of the complexity of the MCMC approach and the lack of internal checks in the MCMC software, Patterson *et al.*[13] carried out exhaustive tests of the software, executing more than 2,000 simulations[13]. This was made possible by the high speed of the analysis (a scan of thousands of markers and thousands of individuals can be carried out in less than an hour using a Pentium 4 PC running a LINUX operating system). The results of these simulations indicate that the ANCESTRYMAP software provides analytical options that are sufficiently robust for use in disease-association scans that minimize false-positive association signals.

PARAMETRIC TESTS
Statistical tests which use models that make assumptions about the distributions of sample values and parameters.

NON-PARAMETRIC TESTS
Statistical procedures that are not based on models or assumptions pertaining to the distribution of the variable.

Efficiency of cases only and shared controls

A potential advantage of the MALD study design is the prospect of screening affected individuals (cases) without the need to also screen unaffected individuals (matched controls)[12,13,19]. The case-only approach is feasible because a MALD association signal derives from an increased gene-segment frequency compared with the average frequency of genomic segments elsewhere in the genome of the cases themselves — an internal genomic 'control' (FIG. 1b). However, it is still useful to genotype control individuals for the same region to ensure that a deviation from the average level of European ancestry is seen in cases only, and not in controls. A similar deviation from the rest of the genomic segments in controls would indicate a problem with the methodology, or with the genomic region being studied[12,13,19].

Although there is an extra cost associated with genotyping controls, it is possible to use the same MALD control samples for different studies, reducing the overall cost when parallel admixture studies are carried out using the same marker set. Although epidemiologists warn that comparing cases with a generic set of controls (that is, one that is not matched to the cases under study) can generate false-positive associations, MALD analyses are specifically designed to adjust for differences in ancestry, so mismatched controls should not be of great concern[13,19,25].

Sample sizes required for MALD

Because suitable SNP markers and methods are now available for MALD in African-Americans[11], the question naturally arises as to the number of cases and controls that would be needed to identify a disease gene. Several authors have addressed this issue in detail[13,16,19,37] with answers that seem to be contradictory, but actually focus on different components of statistical-power estimates. Statistical power depends on the quantitative ability of MALD markers to differentiate between chromosomal segments from the parental populations, the number of disease genes that contribute to the disease, and the difference in the frequency of the causal allele between Africans and Europeans (the population relative-risk; TABLE 1). It also depends on the allelic relative risk (ARR) for disease-gene alleles — that is, the strength of the influence of each allele or genotype on the disease phenotype.

In TABLE 2, we estimate the sample sizes that are required for MALD genome scans of African-American individuals. Estimates are shown as a function of disease-allele frequency for different ARRs. These values show the need for larger sample sizes when the disease-associated allele is more common in

Box 1 | **Criteria for declaring significance in a MALD study**

- **The Bayesian statistic for detecting genome-wide significant association that was suggested by Patterson *et al.*[13] should be >2 (a similar criterion can be used for the methods of Hoggart *et al.*[19], and Montana and Pritchard[12]).**
- **The deviation of European ancestry compared with the genome average should be seen in cases only, and not in controls** (FIGS 1a,b)**.**
- **The signal should remain when the marker that contributes most strongly to disease is removed.**
- **Markers that are in linkage disequilibrium with each other in ancestral European and western African populations should be excluded from the mapping by admixture linkage disequilibrium (MALD) marker set.**
- **The region of association should be statistically significant based on two different Markov chain Monte Carlo analysis-software packages.**
- **The P-values for case–control association studies should be obtained by carrying out** PERMUTATION TESTING**. The statistic at the disease locus must be more extreme, and therefore more significant, than for any other locus throughout the genome in 100 random permutations of the case and control labels.**
- **The statistic for association should increase in significance when marker density at the locus is increased, or when more affected samples are added to the study.**

These criteria are adapted from those presented by Reich and Patterson[47].

Africans than Europeans, given the higher frequency of African versus European ancestry among African-Americans. The power estimates are necessarily based on simplifying assumptions, such as assumptions of the actual disease-gene allele frequencies in both parental populations, the extent of admixture, ARR and the informative value of the MALD markers used in determining the ancestry of chromosome segments.

In general, it is feasible to carry out MALD with 80% statistical power using a sample size of 440 or fewer cases when there is a strong ARR of 2.5 and underlying allelic differences of 0.6. This increases to 340–670 cases for an ARR of 0.5, and to 1,100–1,650 cases for an ARR of 1.5 (TABLE 2). We note that in practice the number of cases examined will have to be increased to achieve the desired power to account for the less-than-full informativeness of the markers examined. However, the several hundred times fewer markers necessary for a MALD analysis compares favourably with the absolute number of markers examined in a whole-genome scan in terms of requiring less statistical testing, thereby increasing the power of a MALD scan.

The question of which parental population has the higher frequency of the disease allele is not a great concern for MALD studies in African-Americans. Usually, only a few hundred more cases are required for mapping a variant that has a higher frequency in Africans than in Europeans, both under the range of conditions examined here when ARR >1 and those that have been presented previously[11]. Considering these power calculations and the ease of high-throughput genotyping, the extra genotyping and effort necessary to organize patient collections for diseases that are more prevalent in Africans is just as feasible an approach for MALD analysis as the converse approach, making the mapping strategy feasible for diseases with higher frequencies in either ethnic group.

PERMUTATION TESTING
An approach in which the actual data are randomized many times to generate a distribution of outcomes, so that the fraction of observations with values that are more extreme than the outcome that is observed with the real data reflects the statistical significance.

Identifying the disease-causing variants

Identifying a region that contains a disease gene is only the first step in the MALD gene-discovery process[35,38]. The average extent of ALD in African-Americans (FIG. 2a) indicates that a typical peak of significant admixture association will be about two thirds of the width of a significant peak in a family-based linkage study — that is, it is likely to be about 10 cM and will typically encompass ~100 genes. Closing in on the causal genetic locus in a 10-cM region would require extensive SNP haplotyping, which would include candidate-gene assessment. This would then be followed by a targeted case–control association study for the disease phenotype using extra markers from the implicated chromosomal segment.

Limitations and guidelines

There are some limitations and caveats of the MALD approach that need to be considered. First, MALD will not be able to localize all risk variants for a disease. In particular, it will miss disease-gene alleles that have a similar frequency in the two parental populations. In addition, situations in which false-positive associations occur can result from using this method with some programs and approaches when parental-population allele frequencies are not accurately estimated, when LD between nearby markers in the parental populations is not recognized or when other unrecognized complexities occur. The nature of the MCMC and HMM methods requires that a putative association must satisfy several stringent criteria (summarized in BOX 1) before it can be confirmed as being implicated in the disease that is being studied.

Following these guidelines, using MCMC/HMM methods and the accumulated SNP markers from projects such as the HapMap, MALD is now feasible for populations of African-Americans, and potentially for Latino populations. Application to Pacific Islanders, Australian Aborigines and other admixed groups will depend on further work to quantify the

degree of admixture in these populations and the confounding effects of bottlenecks and other population processes. With powerful SNP marker sets that are able to differentiate between chromosomal regions of different continental origins, MALD should be able to detect small differences in allele frequency (10–20%) to discover the genetic variants that underlie disease.

Ethical considerations

Great care should be taken in applying the MALD approach in the context of racial sensitivities and the potential misuse of genetic information in racial and social characterizations[39–41]. In our opinion, understanding ethnic differences in disease incidence is important in medical studies, and this has been discussed extensively in the literature [42–44]. However, studies of diseases that differ in their prevalence between ethnic groups must be carried out using the highest standards of personal privacy, with attention to the troubled history of studying disease in a racial context. All proposed studies should also be designed carefully to avoid misinterpretation, over-interpretation, misapplication of conclusions, discrimination and prejudicial inferences.

Our view is that MALD-based association studies should focus on those that are likely to benefit traditionally understudied non-European ethnic groups. For example, the high power of MALD to identify alleles that are associated with diseases that differ in prevalence between parental populations offers the promise to move some minority groups from the margins of complex disease genetics to the focus of these studies. It also has the potential to increase the chances of success in finding genes that are important for disease risk in these populations. With this in mind, we hope that the advent of MALD as a powerful method for finding disease genes will be seen as a positive development for many minority groups, including African-Americans and Latinos, towards improving community health.

Although MALD is sometimes seen as a potentially ethically problematic approach, there are actually important ways in which it can break down concepts of race. For example, the drug BiDil, which is used to prevent heart failure, was recently designated by the US Food and Drug Administration (FDA) as the first drug approved only for use in a specific racial group — African-Americans[45]. This increased the official scope for using racial categories to guide the prescription of drugs. MALD, however, could potentially pinpoint the genetic factor (if one exists) that makes BiDil effective for some African-American individuals but not for others. If such a variant were identified, it would then be possible to prescribe the drug not on the basis of racial categories, but instead according to individual genotype in all races. Finally, we emphasize that although the potential benefits of MALD are considerable, great care must be taken in using this approach in communities that historically distrust scientific research involving participants from minority groups.

Conclusions

Genome-wide association scans using the MALD approach and involving large numbers of samples are now being carried out. Such studies have become feasible because of their lower costs in comparison with other types of association study and because they can be carried out using current genotyping technologies. At the current cost of a few hundred US dollars per individual, MALD mapping is practical and is being pursued in several laboratories, including our own. The MALD approach is a useful complement to linkage studies and HapMap-based association scans. Significantly, this approach addresses health issues in minority communities that are historically underserved.

Over the next few years, the main focus of MALD studies will be to explore empirically whether this is a practical approach to discovering disease genes. The potential of MALD for genetic analysis awaits demonstration through proof-of-principle studies that successfully discover disease genes using this method and translate this promising approach into medical benefits.

1. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
2. Gibbs, R. A. *et al.* The International HapMap Project. *Nature* **426**, 789–796 (2003).
3. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108 (2005).
4. Wang, W. Y., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nature Rev. Genet.* **6**, 109–118 (2005).
5. Briscoe, D., Stephens, J. C. & O'Brien, S. J. Linkage disequilibrium in admixed populations: applications in gene mapping. *J. Hered.* **85**, 59–63 (1994).
6. Stephens, J. C., Briscoe, D. & O'Brien, S. J. Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *Am. J. Hum. Genet.* **55**, 809–824 (1994).
This is an early description of the power of MALD for identifying disease genes by exploring models of gene discovery. This paper describes the characteristics of suitable populations and the consequences of admixture for linkage disequilibrium.
7. McKeigue, P. M. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *Am. J. Hum. Genet.* **63**, 241–251 (1998).
This paper conceptualizes and describes the idea of gene mapping using admixture analysis based on chromosome segments that are derived from ancestral populations.
8. Darvasi, A. & Shifman, S. The beauty of admixture. *Nature Genet.* **37**, 118–119 (2005).
9. Dean, M. *et al.* Polymorphic admixture typing in human ethnic populations. *Am. J. Hum. Genet.* **55**, 788–808 (1994).
10. Smith, M. W. *et al.* Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. *Am. J. Hum. Genet.* **69**, 1080–1094 (2001).
11. Smith, M. W. *et al.* A high-density admixture map for disease gene discovery in African Americans. *Am. J. Hum. Genet.* **74**, 1001–1013 (2004).
A genome-wide set of markers for MALD-based gene discovery in African-Americans is described in this paper. An average of 6 generations since admixture and average European chromosomal segments with block sizes of 11 cM are estimated.
12. Montana, G. & Pritchard, J. K. Statistical tests for admixture mapping with case–control and cases-only data. *Am. J. Hum. Genet.* **75**, 771–789 (2004).
The authors describe the use of the program STRUCTURE and MALDsoft for admixture mapping with simulated data. They suggest that a large set of random SNPs can be used to discover disease genes nearly as well as a much smaller set of markers that are enriched for MALD-based gene discovery.
13. Patterson, N. *et al.* Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74**, 979–1000 (2004).
This paper describes a rapid gene-mapping algorithm (ANCESTRYMAP) for MALD that uses MCMC and hidden Markov chain methodologies that are capable of whole-genome admixture scans. It extensively models the sample sizes that are necessary for gene discovery using MALD.
14. Seldin, M. F. *et al.* Putative ancestral origins of chromosomal segments in individual African Americans: implications for admixture mapping. *Genome Res.* **14**, 1076–1084 (2004).
15. Zhang, C., Chen, K., Seldin, M. F. & Li, H. A hidden Markov modeling approach for admixture mapping based on case–control data. *Genet. Epidemiol.* **27**, 225–239 (2004).

16. Zhu, X., Cooper, R. S. & Elston, R. C. Linkage analysis of a complex disease through use of admixed populations. *Am. J. Hum. Genet.* **74**, 1136–1153 (2004).
17. Chakraborty, R. & Weiss, K. M. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl Acad. Sci. USA* **85**, 9119–9123 (1988).
A classic reference that describes admixture and its potential use in finding traits of interest.
18. Hoggart, C. J. *et al.* Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet.* **72**, 1492–1504 (2003).
19. Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G. & McKeigue, P. M. Design and analysis of admixture mapping studies. *Am. J. Hum. Genet.* **74**, 965–978 (2004).
This article describes a methodology for admixture analysis and makes sample-size estimates for MALD-based gene discovery.
20. Parra, E. J. *et al.* Estimating African American admixture proportions by use of population-specific alleles. *Am. J. Hum. Genet.* **63**, 1839–1851 (1998).
21. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
22. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data. Linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
23. Halder, I. & Shriver, M. D. Measuring and using admixture to study the genetics of complex disease. *Hum. Genomics* **1**, 52–62 (2003).
24. Wang, J. Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* **164**, 747–765 (2003).
25. Reiner, A. P. *et al.* Population structure, admixture, and aging-related phenotypes in African American adults: the cardiovascular health study. *Am. J. Hum. Genet.* **76**, 463–477 (2005).
26. Madrigal, L. *et al.* Ethnicity, gene flow, and population subdivision in Limon, Costa Rica. *Am. J. Phys. Anthropol.* **114**, 99–108 (2001).
27. Bertoni, B., Budowle, B., Sans, M., Barton, S. A. & Chakraborty, R. Admixture in Hispanics: distribution of ancestral population contributions in the continental United States. *Hum. Biol.* **75**, 1–11 (2003).
28. Bonilla, C., Shriver, M. D., Parra, E. J., Jones, A. & Fernandez, J. R. Ancestral proportions and their association with skin pigmentation and bone mineral density in Puerto Rican women from New York city. *Hum. Genet.* (2004).
29. Collins-Schramm, H. E. *et al.* Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. *Hum. Genet.* **114**, 263–271 (2004).
30. Parra, E. J. *et al.* Relation of type 2 diabetes to individual admixture and candidate gene polymorphisms in the Hispanic American population of San Luis Valley, Colorado. *J. Med. Genet.* **41**, e116 (2004).
31. Rousham, E. K. & Gracey, M. Factors affecting birthweight of rural Australian Aborigines. *Ann. Hum. Biol.* **29**, 363–372 (2002).
32. Grandinetti, A. *et al.* Relationship between plasma glucose concentrations and Native Hawaiian ancestry: the Native Hawaiian Health Research Project. *Int. J. Obes. Relat. Metab. Disord.* **26**, 778–782 (2002).
33. Pfaff, C. L. *et al.* Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am. J. Hum. Genet.* **68**, 198–207 (2001).
34. Lautenberger, J. A., Stephens, J. C., O'Brien, S. J. & Smith, M. W. Significant admixture linkage disequilibrium across 30 cM around the *FY* locus in African Americans. *Am. J. Hum. Genet.* **66**, 969–978 (2000).
35. Zhu, X. *et al.* Admixture mapping for hypertension loci with genome-scan markers. *Nature Genet.* **37**, 177–181 (2005).
36. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
37. McKeigue, P. M. Prospects for admixture mapping of complex traits. *Am. J. Hum. Genet.* **76**, 1–7 (2004).
38. Mountain, J. L. & Risch, N. Assessing genetic contributions to phenotypic differences among 'racial' and 'ethnic' groups. *Nature Genet.* **36**, S48–S53 (2004).
39. Schwartz, R. S. Racial profiling in medical research. *N. Engl. J. Med.* **344**, 1392–1393 (2001).
40. Braun, L. Race, ethnicity, and health: can genetics explain disparities? *Perspect. Biol. Med.* **45**, 159–174 (2002).
41. Pearce, N., Foliaki, S., Sporle, A. & Cunningham, C. Genetics, race, ethnicity, and health. *BMJ* **328**, 1070–1072 (2004).
42. Risch, N., Burchard, E., Ziv, E. & Tang, H. Categorization of humans in biomedical research: genes, race and disease. *Genome Biol.* **3**, 1–12 (2002).
43. Burchard, E. G. *et al.* The importance of race and ethnic background in biomedical research and clinical practice. *N. Engl. J. Med.* **348**, 1170–1175 (2003).
44. Bamshad, M., Wooding, S., Salisbury, B. A. & Stephens, J. C. Deconstructing the relationship between genetics and race. *Nature Rev. Genet.* **5**, 598–609 (2004).
45. Taylor, A. L. *et al.* Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *N. Engl. J. Med.* **351**, 2049–2057 (2004).
46. McEliece, R. *The Theory of Information and Coding. Encyclopaedia of Math and its Applications* Vol. 3 (Addison Wesley, 1977).
47. Reich, D. & Patterson, N. Pitfalls and prospects for admixture mapping. *Philos. Trans. R. Soc. B* (in the press).
48. Thomas, D. L. *et al.* The natural history of hepatitis C virus infection: host, viral, and environmental factors. *JAMA* **284**, 450–456 (2000).
49. Tess, B. H., Rodrigues, L. C., Newell, M. L., Dunn, D. T. & Lago, T. D. Breastfeeding, genetic, obstetric and other risk factors associated with mother-to-child transmission of HIV-1 in Sao Paulo State, Brazil. Sao Paulo collaborative study for vertical transmission of HIV-1. *Aids* **12**, 513–520 (1998).
50. Hogancamp, W. E., Rodriguez, M. & Weinshenker, B. G. The epidemiology of multiple sclerosis. *Mayo Clin. Proc.* **72**, 871–878 (1997).
51. Ruo, B., Capra, A. M., Jensvold, N. G. & Go, A. S. Racial variation in the prevalence of atrial fibrillation among patients with heart failure: the Epidemiology, Practice, Outcomes, and Costs of Heart Failure (EPOCH) study. *J. Am. Coll. Cardiol.* **43**, 429–435 (2004).
52. Gupta, V. *et al.* Racial differences in thoracic aorta atherosclerosis among ischemic stroke patients. *Stroke* **34**, 408–412 (2003).
53. Bohannon, A. D. Osteoporosis and African American women. *J. Womens Health Gend. Based Med.* **8**, 609–615 (1999).
54. Finkelstein, J. S. *et al.* Ethnic variation in bone density in premenopausal and early perimenopausal women: effects of anthropometric and lifestyle factors. *J. Clin. Endocrinol. Metab.* **87**, 3057–3067 (2002).
55. Bastian, H. M. *et al.* Systemic lupus erythematosus in three ethnic groups. XII. Risk factors for lupus nephritis after diagnosis. *Lupus* **11**, 152–160 (2002).
56. Davey Smith, G., Neaton, J. D., Wentworth, D., Stamler, R. & Stamler, J. Mortality differences between black and white men in the USA: contribution of income and other risk factors among men screened for the MRFIT. *Lancet* **351**, 934–939 (1998).
57. Demirovic, J. *et al.* Prevalence of dementia in three ethnic groups: the South Florida program on aging and health. *Ann. Epidemiol.* **13**, 472–478 (2003).
58. Harper, M. A. *et al.* Racial disparity in pregnancy-related mortality following a live birth outcome. *Ann. Epidemiol.* **14**, 274–279 (2004).
59. Kopp, J. B. & Winkler, C. HIV-associated nephropathy in African Americans. *Kidney Int.* S43–S49 (2003).
60. Songer, T. J. & Zimmet, P. Z. Epidemiology of type II diabetes: an international perspective. *Pharmacoeconomics* **8** (Suppl. 1), 1–11 (1995).
61. Klag, M. J. *et al.* End-stage renal disease in African-American and white men. 16-year MRFIT findings. *JAMA* **277**, 1293–1298 (1997).
62. Kissela, B. *et al.* Stroke in a biracial population: the excess burden of stroke among blacks. *Stroke* **35**, 426–431 (2004).
63. Wong, T. Y. *et al.* Racial differences in the prevalence of hypertensive retinopathy. *Hypertension* **41**, 1086–1091 (2003).
64. McGinnis, K. A. *et al.* Understanding racial disparities in HIV using data from the veterans aging cohort 3-site study and VA administrative data. *Am. J. Public Health* **93**, 1728–1733 (2003).
65. Hodge, A. M. & Zimmet, P. Z. The epidemiology of obesity. *Baillieres Clin. Endocrinol. Metab.* **8**, 577–599 (1994).
66. Molokhia, M. & McKeigue, P. Risk for rheumatic disease in relation to ethnicity and admixture. *Arthritis Res.* **2**, 115–125 (2000).
67. Reveille, J. D. Ethnicity and race and systemic sclerosis: how it affects susceptibility, severity, antibody genetics, and clinical manifestations. *Curr. Rheumatol. Rep.* **5**, 160–167 (2003).
68. Wright, N. M., Papadea, N., Veldhuis, J. D. & Bell, N. H. Growth hormone secretion and bone mineral density in prepubertal black and white boys. *Calcif. Tissue Int.* **70**, 146–152 (2002).

Acknowledgements

We thank J. Coresh, M. Dean, L. Kao, M. Klag, J. Lind, G. Nelson, T. Oleksyk, S. Shrestha and C. Winkler for many discussions regarding MALD. D. Reich and N. Patterson were helpful in discussing the merits of the available software, making the power calculations presented and discussing the future of applying MALD to gene discovery. The content of this publication does not necessarily reflect the views or policies of the US Department of Health and Human Services, nor does the mention of trade names, commercial products or organizations imply endorsement by the US government. The project included in this manuscript has been funded in whole or in part with federal funds from the National Cancer Institute and US National Institutes of Health.

Competing interests statement

The authors declare no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
AT3 | *FY*

FURTHER INFORMATION

ADMIXMAP software: http://www.ucd.ie/genepi/software.html
ANCESTRYMAP software: http://genepath.med.harvard.edu/~reich
Family linkage mapping resources at the Marshfield Center for Medical Genetics: http://research.marshfieldclinic.org/genetics
International HapMap Project:: http://www.HapMap.org
MALDsoft software: http://pritch.bsd.uchicago.edu/software.html

SUPPLEMENTARY INFORMATION

See online article: S1 (figure) | S2 (box) | S3 (table)
Access to this links box is available online.

Author biographies

Michael W. Smith is a Principal Investigator in the Laboratory of Genomic Diversity at the National Cancer Institute in Frederick, Maryland, USA. His research focus is on disease-gene discovery, using both association approaches and mapping by admixture linkage disequilibrium. He is interested in understanding natural variation in humans in relation to individual phenotypes and the history of the human genome.

Stephen J. O'Brien is the Chief of the Laboratory of Genomic Diversity at the National Cancer Institute in Frederick, Maryland, USA. His principal research focus involves the discovery and characterization of human genes that regulate response to infectious diseases, particularly AIDS, hepatitis and cancer. He also studies the evolution of mammalian genome organization through studies of comparative genomics of the domestic cat, their wild felid relatives, and their infectious disease agents.

Online summary

- Mapping by admixture linkage disequilibrium (MALD) is a genetic strategy for discovering genes that underlie complex diseases. The method is based on differences in disease-gene frequency between the parental racial groups of admixed populations.
- A MALD-based full-genome scan can be carried out using a few thousand markers that are able to differentiate, to a high degree, between chromosomal ancestries in relation to the parental populations. This enables the discovery of regions that harbour genes associated with complex diseases.
- Differences in the proportion of admixture for a particular chromosomal segment between cases and controls can implicate a region that is several centimorgans in size as being involved in a disease. This can also be done using cases only, by looking for differences in admixture proportions between specific locations and the rest of the genome in the same individual.
- MALD-based genome scans are already possible in African-Americans, and are now underway. These studies are using a published set of MALD markers that are highly enriched for the ability to differentiate between chromosomal segments derived from African and European ancestors. The marker set will improve as frequency data accumulate from the HapMap project.
- MALD scans in other groups (Latinos, Pacific Islanders and other admixed populations) will become possible in the near future as appropriate markers are mined from HapMap allele frequencies.
- Proof of the efficacy of MALD awaits its successful application among African-Americans for potentially amenable diseases, such as prostate cancer, multiple sclerosis and end-stage renal disease. If these studies are successful, MALD could then be applied to other groups over the next few years.

Online links

Entrez:
AT3
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=Graphics&list_uids=462

FY
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=Graphics&list_uids=2532

ADMIXMAP software:
http://www.ucd.ie/genepi/software.html

ANCESTRYMAP software:
http://genepath.med.harvard.edu/~reich

Family Linkage Mapping resources at the Marshfield Center for Medical Genetics:
http://research.marshfieldclinic.org/genetics

International HapMap Project:
http://www.HapMap.org

MALDsoft software:
http://pritch.bsd.uchicago.edu/software.html