Measuring and using admixture to study the genetics of complex diseases

*Indrani Halder and Mark D. Shriver**

Department of Anthropology, Pennsylvania State University, University Park, PA 16801, USA
**Correspondence to*: Tel: +1 814 863 1078; Fax: +1 814 863 1474; E-mail: mds17@psu.edu

Date received (in revised form): 25th August 2003

Abstract

Admixture is an important evolutionary force that can and should be used in efforts to apply genomic data and technology to the study of complex disease genetics. Admixture linkage disequilibrium (ALD) is created by the process of admixture and, in recently admixed populations, extends for substantial distances (of the order of 10 to 20 cM). The amount of ALD generated depends on the level of admixture, ancestry information content of markers and the admixture dynamics of the population, and thus influences admixture mapping (AM). The authors discuss different models of admixture and how these can have an impact on the success of AM studies. Selection of markers is important, since markers informative for parental population ancestry are required and these are uncommon. Rarely does the process of admixture result in a population that is uniform for individual admixture levels, but instead there is substantial population stratification. This stratification can be understood as variation in individual admixtures and can be both a source of statistical power for ancestry–phenotype correlation studies as well as a confounder in causing false-positives in gene association studies. Methods to detect and control for stratification in case/control and AM studies are reviewed, along with recent studies showing individual ancestry–phenotype correlations. Using skin pigmentation as a model phenotype, implications of AM in complex disease gene mapping studies are discussed. Finally, the article discusses some limitations of this approach that should be considered when designing an effective AM study.

Keywords: *complex diseases, admixture linkage disequilibrium (ALD), admixture mapping (AM), biogeographical ancestry (BGA), structure, phenotype–ancestry correlation*

Introduction

Genetic analysis of phenotypes and diseases has traditionally followed two approaches: family-based linkage analysis and population-based association studies. While in linkage analysis it is the co-segregation of alleles in families that is measured, population-based studies use non-random associations between phenotypes and alleles in populations to identify causative genes. Linkage analysis has proven to be immensely successful as a means of identifying genes for a number of single gene diseases with simple Mendelian inheritance (eg see OMIM database). Complex diseases are multifactorial, polygenic and often characterised by late age of onset, incomplete penetrance, locus heterogeneity and environmental exposures and, despite significant efforts, have not been amenable to family-based mapping.

Linkage disequilibrium (LD) is an important aspect of genetic association studies and is generated in a population through mutation, selection, drift, non-random mating and admixture.[1] Allelic associations due to LD are significant and are correlated with physical distance within small genomic regions but decay over time due to recombination.[2–4] LD-based association studies have been successful in both fine scale mapping[5,6] and initial disease gene mapping in homogeneous populations that have undergone recent bottlenecks (eg Hirschsprung disease in Mennonites,[7] Bardet–Beidle syndrome in Bedouins[8]). Allelic associations can result either from direct functional effects of the alleles tested or indirectly through non-random associations between the allele measured and nearby functional alleles. Since functional alleles in most genes are still unknown and are indeed an object of the research, LD is an important feature of how genes can be screened for alleles that alter disease risk. Thus, there has been substantial focus on the extent of LD across the genome and the definition of statistical methods for disease gene mapping using LD.[9–11] In large cosmopolitan populations, however, LD may be difficult to detect when the mutation is old, since the amount of remaining LD may be small. Additionally, false-positive associations due to population stratification are important confounders in LD-based association studies.

Admixture studies and their use in disease gene mapping

Intermixture between previously isolated populations leads to the creation of admixed populations. The process of admixture

itself creates LD between all loci, linked and unlinked, that have different allele frequencies in the parental populations. The magnitude of admixture linkage disequilibrium (ALD) in an admixed population depends on the allele frequency differential between the parental populations, the level of admixture, the admixture dynamics, the time since admixture and the recombination rate between the loci.[12] While ALD between unlinked markers decays rapidly (within two to four generations), ALD between linked markers decays more slowly. The exponential decrease in ALD with genetic distance facilitates the differentiation of ALD that is high between markers that are close together and genetically linked, from ALD generated at unlinked loci. Thus, if the parental populations differ in a trait or disease due to different frequencies of risk alleles, it should be possible to identify the loci containing these alleles using admixture mapping (AM).[12–14]

Many US residents can trace their genetic ancestry to more than one continent. The European colonial period that started in the late 1400s brought together in the New World populations that had been geographically isolated, namely, Europeans, West Africans and Native Americans. Given the recent and common origin of all human populations, this admixture had only a small average effect on the gene pools of these new populations. In other words, for most genomic regions, the pre-colonial (or parental) populations had similar allele frequencies and, at these, admixture was of little consequence. At some other loci, however, there had been some change in allele frequency in the time since the separation of parental populations and it is at these loci where admixture has had an important effect. Since populations like African Americans, African Caribbeans and Mexican Americans were formed in the recent past, allelic associations in these populations that were created by admixture extend over large distances. Admixed populations represent a useful resource for mapping complex-disease genes by using this long-range ALD,[12] which requires fewer markers to screen the genome than other populations or approaches. Understanding the genetic consequences of admixture is important because it can be both a confounding factor and a source of statistical power in gene identification studies.

Two models of admixture dynamics have been described to represent the extremes of the process by which an admixed population is formed: the continuous gene flow (CGF) model and the hybrid isolation (HI) model.[15,16] In the HI model, admixture occurs immediately in a single generation without further contribution from either parental population, hence, ALD is generated in a single generation and gradually decays in successive generations through independent assortment and recombination between loci. Few false-positive results are thus expected in an association study under the HI model. Alternatively, the CGF model represents a situation where admixture occurs at a steady rate in each generation, with contributions from one (or all) of the parental populations into the admixed population. ALD under the CGF model increases in each generation, since new admixture is constantly occurring. A point will be reached, however (when the admixture proportion = 0.5), where continued admixture will actually decrease the ALD, since added gene flow will result in the conversion of the admixed population into the introgressing parental population. Figure 1 shows the amount of ALD expected under these two models for linked and unlinked loci. For both models, association between markers is inversely correlated with the genetic distance between them. Simulation studies have shown that populations that have a demographic history more consistent with the CGF model of admixture retain ALD over larger chromosomal regions and show significant associations between unlinked marker loci.[15] While associations between unlinked markers could potentially lead to false-positives, conditioning upon parental admixture allows the distinction between associations arising due to true linkage and those due to CGF stratification to be made, thereby providing greater power for detecting ALD over larger chromosomal distances.[15]

There are several ways in which admixture can be an important resource in the elucidation of genetic factors that contribute to the risk of common disease. Common diseases often have environmental components to their risk, and the clinical phenotype results from currently unknown interactions between environmental factors and underlying genotypes. Decomposing the sources of variation is thus important in order accurately to understand the aetiology of the trait. It is possible to distinguish between the genetic and environmental explanations for ethnic differences in disease risk (and investigating the mode of inheritance), by studying the relationship of disease risk to individual admixture.[14,17–19] For example, recent studies have demonstrated a strong relationship between proportional West African ancestry and the risk of systemic lupus erythematosus in admixed populations in Trinidad.[18] Several common diseases (eg hypertension, diabetes, obesity, prostate cancer and osteoporosis) have differences in risk among population groups (see Table 1). In situations where these differences have a genetic basis, genes underlying these differences can be identified by testing for locus ancestry by conditioning on parental admixture. As detailed by Shriver *et al.*, this approach has a greater statistical power than family linkage studies for mapping polygenic traits.[14] Estimates of biogeographical ancestry (BGA), the proportional ancestry levels of an individual, can be used in conjunction with measured environmental effects for investigating the roles of environmental and inherited risks underlying complex traits.[18–20] It is important to recognise that associations between individual admixture and disease risk might reflect correlations between BGA and socio-cultural variables and exposures. For example, hypothetically, if BGA and years of education were to be correlated, hypertension might be correlated with BGA, even though the causal risk factor was years of education or vice versa.

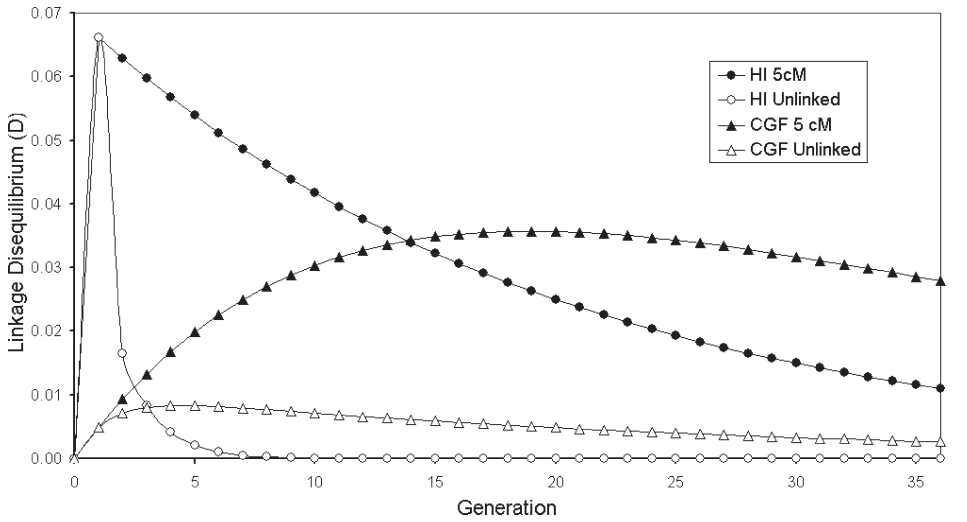

Figure 1. The amount of admixture linkage disequilibrium (ALD) expected under the continuous gene flow (CGF) and hybrid isolation (HI) models of admixture for unlinked loci and loci linked at 5 cM. The results shown are for two loci with $\delta = 0.54$ and 0.49, and with 50 per cent admixture in the first generation for the HI model and 1.9 per cent admixture for 36 generations under the CGF model (equivalent to 50 per cent total). ALD under the HI model decreases for both linked and unlinked loci, whereas ALD under the CGF model for both linked and unlinked loci increases initially and then decreases (adapted from Pfaff *et al.*, 2001[15])

Marker choice for admixture mapping

Admixture-based methods rely on using suitable markers and estimates of allele frequencies from appropriately identified parental populations. Since ALD is fairly new and extends over larger distances, fewer markers are required for AM studies. Markers informative for ancestry have been used in several contexts and have been referred to as 'ideal,'[21] 'private'[22] and 'unique'.[23] Informativeness of such markers can be measured as the allele frequency differential (δ), which is the absolute value of the difference of a particular allele between populations.[12,24] Microsatellites and insertion/deletion polymorphisms with $\delta > 0.3$ were recently called 'ethnic-difference markers' (EDMs)[25] suitable for mapping by admixture linkage disequilibrium (MALD). Additionally, markers with high δ and very high log likelihood allelic ratio (LLAR) between populations have been designated 'population specific alleles' (PSAs).[26] This report followed from earlier work where markers with large allele frequency difference were identified to be appropriate for admixture studies,[27,28] and most (>95 per cent) of the arbitrarily identified biallelic markers had $\delta < 50$ per cent.[24] Thus, the authors proposed that ideal PSAs should have $\delta > 50$ per cent and also indicated that for multiallelic loci, a composite δ could be estimated as one half the summation of the absolute value of allelic frequency differences for all alleles at that locus.[26] It has also been shown that markers with lower δ values, of approximately 30 per cent, can provide up to 80 per cent power for detecting associations at distances of 5 cM with a large enough sample size ($N = 1{,}000$).[15]

Pfaff *et al.*,[15] suggested referring to markers suitable for admixture studies as 'ancestry informative markers' (AIMs), given that the central feature of these markers is the ancestry information content (f).[29] The present authors agree that the term AIM more accurately describes these markers and does so using language that is less likely to be misunderstood and misinterpreted.[14,17,28] Marker information content 'f' denotes the locus-specific F_{st} and is a value representative of the differentiation between two populations at a single locus. This is equivalent to Wahlund's standardised variance for allele frequency. Simulation studies for estimating the information content of markers with varying levels of f have shown that for 1,000 markers with average information content for ancestry at 40 per cent between two ancestral subpopulations, approximately 80 per cent of the information about ancestry can be extracted from an initial genome screen.[13,29] After initial identification of regions showing admixture, more markers can be typed in these regions to increase extraction of information to nearly 100 per cent.

It is well established, however, that only 5–15 per cent of the total genetic variation results from differences among human populations.[30–32] Moreover, most alleles are shared between populations, and alleles common in one population are also common in other populations. Thus, most genetic markers are unaffected by admixture and it is imperative to choose markers that show high levels of δ (and f) between the parental populations. Recent studies by several groups have focused on identifying panels of markers suitable for admixture studies. One notable study screened 744 microsatellite markers

Table 1. Diseases with possible genetic components based on ethnic differences in disease rates and hence amenable to admixture mapping

Disease	High-risk groups	Low-risk groups	Relative risk ratio	Reference(s)
Obesity	African women Native Americans South Asians (central adiposity,) Pacific Islanders, Aboriginal Australians	Europeans	2:4	[64,65]
Non-insulin dependent diabetes (NIDDM)	South Asians, West Africans, Peninsular Arabs, Pacific Islanders and Native Americans	Europeans	4:7	[66,67]
Hypertension	African Americans, West Africans	Europeans	2:3	[68,69]
Coronary heart disease	South Asians	West African men	2:4	[70,71]
End-stage renal disease	Native Americans and African populations	Europeans	N/A	[72]
Dementia	Europeans	African Americans, Hispanic Americans	N/A	[73]
Autoimmune diseases: Systemic lupus erythematosus	 West Africans Native Americans	 Europeans Europeans	N/A	[55]
Skin cancer	Europeans		N/A	[74]
Lung cancer	Africans	European Americans, Chinese, Japanese		[75] [76]
Prostate cancer	Africans and African Americans			[77]
Multiple sclerosis	Europeans	Chinese, Japanese, African Americans, Turkmens, Uzbeks, Native Siberians, New Zealand Maoris	N/A	[78]
Osteoporosis	European Americans	African Americans	N/A	[79]

N/A = not available.

for composite δ values and LLAR in four different populations and identified a genome spanning set of 315 markers (average spacing 10 cM, $\delta \geq 0.3$) for mapping in African Americans and 214 markers (average spacing of 16 cM, $\delta \geq 0.25$) for mapping in Hispanics.[33] A DNA pooling method was used to identify 151 AIMs (microsatellites and short insertion/deletion polymorphisms), with $\delta > 0.3$ for mapping in Mexican American populations to distinguish between European-American and Native-American contributions.[25] Ninety-seven AIMs were identified for mapping in African-American populations[25] that show limited variation within Africa.[34] The authors' group has reported AIMs over the past few years.[14,17,26,35,36] Additional resources are available for obtaining marker frequency, and genotype and haplotype information, from The SNP Consortium (TSC; http://snp.cshl.org), the National Center for Biotechnology Information's 'dbSNP' website (http://www.ncbi.nlm.nih.gov/SNP), the Marshfield Database (http://research.marshfieldclinic.org/genetics/Default.htm) and the ongoing HapMap project.

Admixed populations and admixture proportions

Since the amount of ALD created is proportional to the level of admixture in a population, it is important briefly to review studies on admixture levels across populations. Those populations that are likely to be useful for admixture studies include African Americans, Mexican Americans, Cubans and Puerto Ricans in the USA, African Caribbeans, various Latin American populations, various groups in Central and South America and the Caribbean islands, Anglo Indians in India and 'coloured' populations of South Africa. Various statistical approaches have been used to estimate admixture proportions in these

populations and have been reviewed in detail elsewhere.[37] These include a least squares method, a weighted least squares method[16,38,39] and likelihood methods.[38,40] A recent review of admixture studies and admixture proportions of various Latin American populations is provided by Sans.[41]. African Americans are a well-studied group with substantial European and West African contributions and a smaller Native American contribution.[27,35,42,43] A survey of current literature indicates that European admixture ranges from 3.5 per cent in the Gullah Sea Islanders of South Carolina,[35] to 28 per cent in New Orleans.[35] Admixture estimates in African-American populations can be highly variable across the USA, which is likely to reflect local variation in the demographic histories and social norms.

US Hispanics form a complex socio-political conglomerate including Puerto Ricans, Cubans, Spanish Americans, Mexican Americans. Various groups from Central and South America can also be studied using ancestry AM. The proportional contributions from parental Europeans are estimated to be the largest, followed by a substantial Native American ancestry and varying amounts of West African ancestry.[16,17,44] In a sample of Mexican Americans from Arizona, the admixture estimates obtained using a weighted least squares method showed 29 ± 4 per cent Native American, 68 ± 5 per cent European and 3 ± 2 per cent West African contribution.[16] A recent study reports the following estimates for a Hispanic population from the San Luis Valley, Colorado: 62.7 ± 2.1 per cent European, 34.1 ± 1.9 per cent Native American, 3.2 ± 1.5 per cent West African.[17] In Puerto Ricans from New York City, the estimates obtained were 53.3 ± 2.8 per cent European, 29.1 ± 2.3 per cent West African, 17.6 ± 2.4 per cent Native American.[17] In a separate Mexican-American population sample from California, European ancestry was estimated to be 60 per cent and Native American contribution was estimated at 40 per cent.[25] As with African-American populations, there is substantial variation across populations. From these results, it is evident that, when studying any new admixed population sample, it is important to accurately determine the proportional contributions and not to rely on previously obtained estimates from a similar population. Additionally, it is instructive to have information on the levels of stratification related to admixture that are present in the population under consideration.[15]

Ancestry–phenotype correlations; phenotype and complex disease gene mapping

Traits and diseases more prevalent in one population than in others are amenable to admixture analysis and some examples are listed in Table 1. Most of the diseases shown in this Table have a complex aetiology affected by multiple genes and environmental factors. Earlier studies[45,46] focused on admixed populations as units of analysis in exploring relationships between ancestry and phenotypes.[12] These authors showed that non-insulin-dependent (Type 2) diabetes mellitus prevalence is correlated with admixture proportions among a selection of populations with varying levels of Native American ancestry. Data like these provide compelling evidence for frequency differences in risk modifying alleles, but such data have not been collected for many diseases. Another related approach is to test for individual admixture–phenotype correlations within an admixed population. Correlations between ancestry and phenotypes have been detected and reported by various authors.[14,17–19,44,45,47]

A prerequisite for testing ancestry/phenotype correlations is the presence of stratification related to admixture, which will be evident in variation in individual ancestry levels. Figure 2 shows the distribution of BGA estimates from three examples of Hispanic population samples, Puerto Ricans from New York, Mexicans from Tlapa, Mexico and Hispanics from the San Luis Valley, Colorado.[17] Substantial variation is observed in all three samples. With the San Luis Valley group, more variability is observed on the European–Native American axis, while the New York group is more variable on the European–West African axis. Following the argument of Chakraborty and Weiss,[48] admixture proportions should be correlated with diseases/traits that differ in populations due to underlying genetic differences. In each of these population samples, strong positive correlation was observed between individual ancestry and skin pigmentation measured as melanin index 'M' or lightness index 'L' (Figures 3A, 3B and 3C). A significant negative correlation was also observed between the proportion of West African ancestry and bone mineral density (BMD) in the Puerto Rican sample.[17] Proportion West African ancestry and skin pigmentation (measured as melanin index) in individuals is also correlated in African Americans from Washington DC and African Caribbeans from the UK, but not in European Americans from State College, Pennsylvania (Figure 4).[14] Recently, correlations have been observed between proportion West African ancestry and lower insulin sensitivity, higher fasting insulin and acute insulin response to glucose in a combined sample of African-American and European-American children.[20] In a separate sample of African-American females, West African admixture is associated with body mass index, fat mass, fat-free mass and BMD.[19] It is important to keep in mind that ancestry–phenotype correlations are dependent on both the existence of functional alleles at different frequencies in parental populations, and significant stratification related to admixture. Although most admixed populations tested to date are structured, there is variation in the amount of stratification present, and this structure should be tested for explicitly when investigating a new population.[15,42,49]

Methods developed for admixture analyses/study design

Theoretical and experimental studies have explored the parameters that characterise and affect admixture

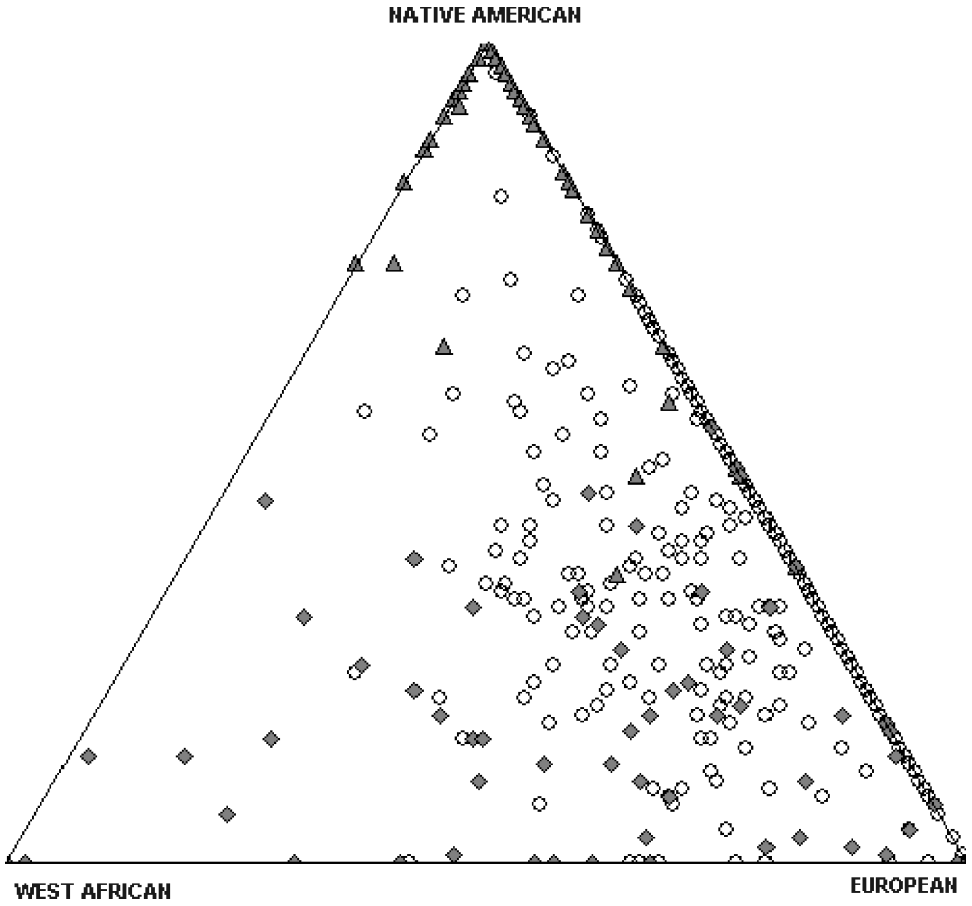

Figure 2. Triangle plot showing biogeographical ancestry of three Hispanic populations. Each vertex represents a parental population, which for this plot are Europeans, West Africans and Native Americans. The three populations shown are Hispanics from the San Luis Valley (blank circles), Puerto Ricans from New York City (grey diamonds) and Mexicans from Tlapa, Mexico (grey triangles) (adapted from Bonilla, 2003[17])

studies.[15,24,28,35,42,50,51] The acronym MALD was proposed[28,50] to designate the mapping method proposed originally by Chakraborty and Weiss, which exploited the long range allelic associations created through ALD.[12] Parameters critical for MALD include the genetic distance between markers and disease locus (θ); number of generations since admixture (t); proportion of admixture (m) from one parental population; the allele frequency differential (δ) between parental populations; and sample size (N).[12,28,52] Simulation studies suggest that sample sizes of 200–300 patients, typed for 200–300 evenly spaced markers, each having allele frequency differentials >0.3, have a >95 per cent chance of locating the causative gene, when there has been no new admixture from the parental population in the last four generations and no other sources of population structure or sample heterogeneity.[28,50]

Other approaches proposed for using admixture include a method based on the transmission disequilibrium test (TDT)[53] that assesses excess transmission of alleles derived from high-risk ancestors to affected offspring of parents who are heterozygous at the marker locus, containing one allele from each of two ancestral populations.[52] A second TDT-based likelihood approach was developed that compared the transmission of haplotypes with non-transmission in affected offspring in an admixed population following a multipoint method. It obtained a likelihood statistic to determine the significance of various models under different scenarios.[54]

One fundamental limitation of MALD as initially described and in its early extensions, is the effects of stratification on causing false-positive association.[12,24,28] The TDT is one means of correcting for this stratification. Another is by conditioning on parental admixture.[29] Marker data at all loci are combined to estimate ancestry of alleles at each locus. When allelic ancestry at marker loci is known, this approach is analogous to a linkage analysis, hence the term AM is more appropriate than MALD for describing this method and to

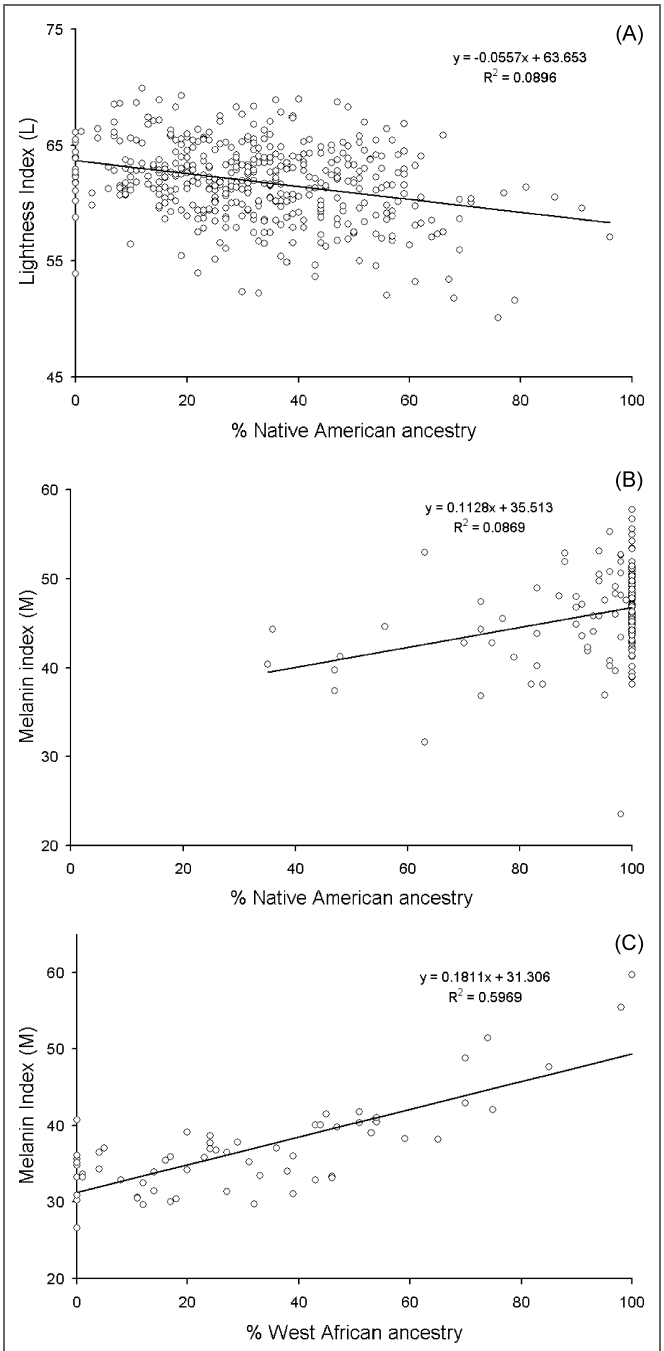

Figure 3. The relationship between proportional ancestry and skin pigmentation in three Hispanic populations. For all populations, proportional ancestry was estimated using the maximum likelihood (ML) method (adapted from Bonilla, 2003).[17] (A) Percent Native American ancestry versus lightness index (L) in Hispanics from the San Luis Valley, Colorado (ancestry estimated using 22 AIMs). (B) Percent Native American ancestry versus melanin index in Mexicans from Tlapa, Mexico (ancestry estimates using 29 AIMs). (C) Percent African ancestry versus melanin index (M) in Puerto Ricans from New York City (ancestry estimated using 35 AIMs)

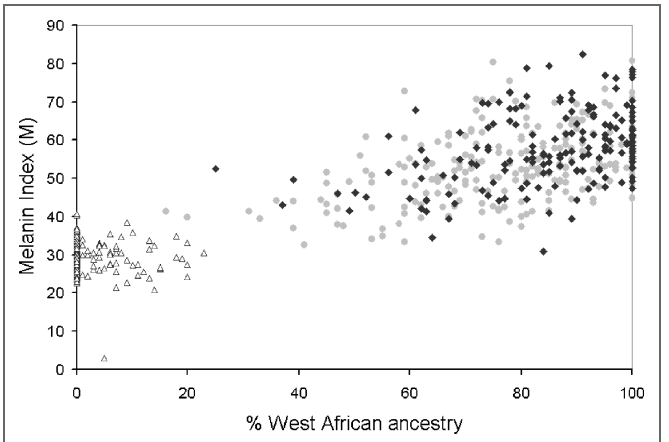

Figure 4. The relationship between percent African ancestry and skin pigmentation in three populations. Percent African Ancestry (obtained using 34 AIMs and calculated by the maximum likelihood (ML) method) and the melanin index (M) are shown for three populations, European Americans from State College, Pennsylvania (diamonds), African Americans from Washington, DC and State College, Pennsylvania (squares) and African Caribbeans from Britain (triangles). (With permission from Shriver *et al.*, 2003[14])

distinguish it from LD approaches.[13,14,29] The underlying variation in ancestry of chromosomes of mixed descent is modelled to extract all of the information about linkage that is generated by admixture. For example, where a locus is assumed to account for variation in skin pigmentation between two parental groups, eg West Africans and Europeans, individuals can be classified according to whether they have 0, 1 or 2 alleles of West African descent at this locus. By comparing these three groups for mean pigmentation level, holding all other factors constant, variation in pigmentation can be observed depending upon the number of alleles of West African ancestry in an individual. Controlling for parental admixture eliminates association of the trait with ancestry at unlinked loci. By removing the background effects of ancestry, it is possible to observe the locus-specific effects on a trait/disease.[14,17] Allelic ancestry at a locus is inferred from the marker by using the conditional probability of each allelic state given the ancestry-specific allele frequencies. A complex hierarchical model with many nuisance parameters is used to model the distribution of admixture in the population. This is implemented using the ADMIXMAP program (at http://www.lshtm.ac.uk/eph.eu/GeneticEpidemiologyGroup/htm), which follows a Bayesian approach with Markov chain simulation, and incorporates the admixture of each individual's parents and the random variation of ancestry on chromosomes inherited from each of the parents in the model.[13,14,29]

Variation in individual admixture introduces population stratification, which in turn can inflate the number of

Table 2. Diseases showing ancestry–phenotype correlation

Phenotype	Population studied	Association observed	Test statistic reported	Reference
Non-insulin-dependent diabetes mellitus (NIDDM)	Mexican Americans and Pima Indians	Amerindian ancestry with NIDDM	Kendall's $\tau = 0.848 \pm 0.221$, [$p \approx 8.1 \times 10^5$]	[48]
NIDDM	Mexican Americans	Amerindian ancestry with NIDDM	0.943[c] ($p < 0.001$)	[63]
1) Body mass index (BMI) 2) Plasma glucose 3) NIDDM	Pima Indians	European admixture with BMI, plasma glucose, 2-hour glucose	0.455[b] (95% CI: 0.301–0.688)	[47]
NIDDM	Mexican Americans	Native American ancestry with NIDDM prevalence	N/A	[45]
Skin pigmentation (reflectrometry)	1) African Americans 2) Afro-Caribbeans 3) European Americans	Melanin index versus % African ancestry	1) 0.21[a], ($p < 0.0001$) 2) 0.16[a] ($p < 0.0001$) 3) 0.001[a] ($p =$ NS)	[14] * Mapped phenotype to two loci: TYR and OCA as candidates which influence normal pigmentation variation
Systemic lupus erythematosus (SLE)	Caribbeans (without Indian or Chinese ancestry)	SLE and African Ancestry	28.4 (95% CI: 1.7–485 after SES adjustment[b])	[18]
Insulin-related phenotypes 1) Insulin sensitivity (S_I), 2) Fasting insulin (FA), 3) Acute insulin response (AIR)	African American Europeans Americans	African admixture (ADM): 1) with S_I 2) with FA 3) with AIR	 1) ($p < 0.001$)[a] 2) ($p < 0.01$)[a] 3) ($p < 0.001$)[a]	[20]
Oxygen capacity	Quechua natives	Positive: Spanish admixture with large VO_2 at high altitudes	0.8[a]	[80]
Bone mineral density (BMD)	Puerto Ricans from New York	Positive: European admixture with lower BMD	0.065[a] ($p = 0.042$)	[17]
Skin pigmentation (Lightness index)	Hispanics from the San Luis Valley, Colorado	Positive: Proportional European ancestry with increased Lightness	0.0821[a] ($p < 0.001$)	[17]

a = R^2; b = risk ratio; c = rank-order correlation.

significant associations that are observed[53,55,56] and is a potential confounder in association studies.[29,57–59] Various statistical approaches have been developed to detect and control for stratification within a population sample.[14,15,17,42,60–62] For example, the D_t/D_0 test examines the relationship between the observed LD and the predicted ALD between unlinked marker pairs for detecting structure within the sample. Using individual ancestry as a conditioning variable in analysis of variance tests, it is possible to eliminate association of the trait with unlinked alleles.[14,17] The Bayesian approaches implemented by McKeigue *et al.* and Pritchard *et al.*[13,61] offer an advantage over classical maximum likelihood based methods[44,63] by allowing for missing genotype and ancestry data and modelling admixture hierarchically. Methods have been developed to control for parental admixture[29] and to account for uncertain BGA estimation.[59]

Recent studies and future directions

Several theoretical and practical studies indicate that AM approaches promise to be suitable for identifying genes causing complex diseases. Methodological advancements have been made to offset the potential problems arising from association between unlinked loci by conditioning on parental admixture,[13,29] and to detect and correct for population stratification.[59,60] Use of Bayesian AM[13,29,59] can take into consideration various uncertainties, including missing data values for estimating admixture proportions, and can overcome problems arising out of mis-specification of parental allele frequencies and promises to be an effective tool for admixture studies. This method, which is different from the classical disequilibrium-based approach that is more commonly used, is perhaps more suitable for disease gene mapping in admixed populations and has already been successfully used for mapping.[14] Table 2 summarises recent studies showing associations between ancestry and phenotypes/diseases and instances where AM was used to identify genes. Currently, the primary impediment to exhaustive AM genome scans is the lack of verified AIM panels. Sufficient numbers of markers are available as candidate AIMs, but effort and resources are required to confirm these markers and to generate accurate parental allele frequencies. Efforts are currently underway in several laboratories to identify more AIMs for this purpose. It seems inevitable that more such studies will be carried out in the near future to utilise the immense potential of this approach.

Acknowledgments

We thank Dr Paul McKeigue and Dr Esteban Parra for helpful discussions on the subject. We also acknowledge helpful comments from an unknown reviewer. This work was supported in part by grants from NIH/NIDDK (DK53958) and NIH/NHGRI (HG02154) to M.D.S.

References

1. Hartl, D.L. and Clark, A.G. (1997), *Principles of Population Genetics*, Sinauer Associates, Sunderland, MA.
2. Jorde, L.B. (1995), 'Linkage disequilibrium as a gene-mapping tool', *Am. J. Hum. Genet.* Vol. 56, pp. 11–14.
3. Huttley, G.A., Smith, M.W., Carrington, M. and O'Brien, S.J. (1999), 'A scan for linkage disequilibrium across the human genome', *Genetics* Vol. 152, pp. 1711–1722.
4. Ardlie, K.G., Kruglyak, L. and Seielstad, M. (2002), 'Patterns of linkage disequilibrium in the human genome', *Nat. Rev. Genet.*, Vol. 3, pp. 299–309; Erratum in: *Nat. Rev. Genet.*, Vol. 3, p. 566.
5. Kerem, E., Reisman, J., Corey, M. *et al.* (1989), 'Prediction of mortality in patients with cystic fibrosis', *N. Engl. J. Med.*, Vol. 326, pp. 1187–1191.
6. MacDonald, M.E., Vonsattel, J.P., Shrinidhi, J. *et al.* (1992), 'Evidence for the GluR6 gene associated with younger onset age of Huntington's disease', *Neurology*, Vol. 53, pp. 1330–1332.
7. Puffenberger, E.G., Kaufmann, E.R., Bolk, S. *et al.* (1994), 'Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22', *Hum. Mol. Genet.*, Vol. 3, pp. 1217–1225.
8. Sheffield, V.C., Carmi, R., Kwitek-Black, A. *et al.* (1994), 'Identification of a Bardet–Beidle syndrome locus on chromosome 3 and evaluation of an efficient approach to homozygosity mapping', *Hum. Mol. Genet.*, Vol. 3, pp. 1331–1335.
9. Risch, N. and Merikangas, K. (1996), 'The future of genetic studies of complex human diseases', *Science*, Vol. 273, pp. 1516–1517.
10. Pritchard, J.K. and Przeworski, M. (2001), 'Linkage disequilibrium in humans: Models and data', *Am. J. Hum. Genet.* Vol. 69, pp. 1–14.
11. Weiss, K.M. and Clark, A.G. (2002), 'Linkage disequilibrium and the mapping of complex human traits', *Trends Genet.* Vol. 18, pp. 19–24.
12. Chakraborty, R. and Weiss, K.M. (1988), 'Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci', *Genetics* Vol. 85, pp. 9119–9123.
13. McKeigue, P.M., Carpenter, J., Parra, E.J. and Shriver, M.D. (2000), 'Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: Application to African-American populations', *Ann. Hum. Genet.* Vol. 64, pp. 171–186.
14. Shriver, M.D., Parra, E.J., Dios, S. *et al.* (2003), 'Skin pigmentation, biogeographical ancestry and admixture mapping', *Hum. Genet.* Vol. 112, pp. 387–399.
15. Pfaff, C.L., Parra, E.J., Bonilla, C. *et al.* (2001), 'Population structure in admixed populations: Effects of admixture dynamics on the pattern of linkage disequilibrium', *Am. J. Hum. Genet.* Vol. 68, pp. 198–207.
16. Long, J.C. (1991), 'The genetic structure of admixed populations', *Genetics* Vol. 127, pp. 417–428.
17. Bonilla, C., 'Admixture in three Hispanic populations: Ancestry proportions, population structure, and gene mapping', PhD Thesis, Department of Anthropology, The Pennsylvania State University, University Park, PA, USA.
18. Molokhia, M., Hoggart, C.J., Patrick, A.L. *et al.* (2003), 'Relation of risk of systemic lupus erythematosus to West African admixture in a Caribbean population', *Hum. Genet.* Vol. 112, pp. 310–318.
19. Fernández, J.R., Shriver, M.D., Beasley, T.M. *et al.* (2003), 'Association of African genetic admixture with resting metabolic rate and obesity among African American women', *Obesity Res.*, Vol. 11, No. 7, pp. 904–911.
20. Gower, B.A., Fernandez, J.R., Beasley, T.M. *et al.* (2003), 'Using genetic admixture to explain racial differences in insulin-related phenotypes', *Diabetes* Vol. 52, pp. 1047–1051.
21. Reed, T.E. (1973), 'Number of gene loci required for accurate estimation of ancestral population proportions in individual human hybrids', *Science* Vol. 244, pp. 575–576.
22. Neel, J.V. (1974), 'Developments in monitoring human populations for mutation rates', *Mutat. Res.* Vol. 26, pp. 319–328.

23. Chakraborty, R., Kamboh, M.I. and Ferrell, R.E. (1991), 'Unique alleles in admixed populations: A strategy for determining hereditary population differences of disease frequencies', *Ethn. Dis.* Vol. 1, pp. 245–256.
24. Dean, M., Stephens, J.C., Winkler, C. *et al.* (1994), 'Polymorphic admixture typing in human ethnic populations', *Am. J. Hum. Genet.* Vol. 55, pp. 788–808.
25. Collins-Schramm, H.E., Phillips, C.M., Operario, D.J. *et al.* (2002), 'Ethnic-difference markers for use in mapping by admixture linage disequilibrium', *Am. J. Hum. Genet.* Vol. 70, pp. 737–750.
26. Shriver, M.D., Smith, M.W., Jin, L. *et al.* (1997), 'Ethnic-affiliation estimation by use of population-specific DNA markers', *Am. J. Hum. Genet.* Vol. 60, pp. 957–964.
27. Chakraborty, R., Kamboh, M.I., Nwankwo, M. and Ferrell, R.E. (1992), 'Caucasian genes in American blacks: New data', *Am. J. Hum. Genet.* Vol. 50, pp. 145–155.
28. Stephens, J.C., Briscoe, D. and O'Brien, S.J. (1994), 'Mapping by admixture linkage disequilibrium in human populations: Limits and guidelines', *Am. J. Hum. Genet.* Vol. 55, pp. 809–824.
29. McKeigue, P.M. (1998), '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.* Vol. 63, pp. 241–251.
30. Nei, M. (1987), *Molecular Population Genetics*, Columbia University Press, New York, NY.
31. Cavalli-Sforza, L., Menozzi, P. and Piazza, A. (1994), *The History and Geography of Human Genes*, Princeton University Press, Princeton, NJ.
32. Deka, R., Shriver, M.D., Yu, L.M. *et al.* (1995), 'Intra- and inter-population diversity at short tandem repeat loci in diverse populations of the world', *Electrophoresis* Vol. 16, pp. 1659–1664.
33. Smith, M.W., Lautenberger, J.A., Shin, H.D. *et al.* (2001), 'Markers for mapping by admixture linkage disequilibrium in African-American and Hispanic Populations', *Am. J. Hum. Genet.* Vol. 69, pp. 1080–1094.
34. Collins-Schramm, H.E., Kittles, R.A., Operario, D.J. *et al.* (2002), 'Markers that discriminate between European and African ancestry show limited variation within Africa', *Hum. Genet.* Vol. 111, pp. 566–569.
35. Parra, E.J., Marcini, A., Akey, J. *et al.* (1998), 'Estimating African American admixture proportions by use of population specific alleles', *Am. J. Hum. Genet.* Vol. 63, pp. 1839–1851.
36. Akey, J., Zhang, G., Jin, L. and Shriver, M.D. (2002), 'Interrogating a high-density SNP map for signatures of natural selection', *Genome Res.* Vol. 12, pp. 1805–1814.
37. Chakraborty, R. (1986), 'Gene admixture in human populations: Models and predictions', *Yearb. Phys. Anthropol.* Vol. 29, pp. 1–43.
38. Elston, R.C. (1971), 'The estimation of admixture in racial hybrids', *Ann. Hum. Genet.* Vol. 35, pp. 9–17.
39. Long, J.C. and Smouse, P.E. (1983), 'Intertribal gene flow between the Ye'cuana and Yanomama: Genetic analysis of an admixed village', *Am. J. Phys. Anthropol.* Vol. 61, pp. 411–422.
40. Chikhi, L., Bruford, M.W. and Beaumont, M.A. (2001), 'Estimation of admixture proportions: A likelihood-based approach using Markov chain Monte Carlo', *Genetics* Vol. 158, pp. 1347–1362.
41. Sans, M. (2000), 'Admixture studies in Latin America: From the 20th to the 21st century', *Hum. Biol.* Vol. 72, pp. 155–177.
42. Parra, E.J., Kittles, R.A., Argyropoulos, G. *et al.* (2001), 'Ancestral proportions and admixture dynamics in geographically defined African-Americans living in South Carolina', *Am. J. Phys. Anthropol.* Vol. 114, pp. 18–29.
43. Destro-Bisol, G., Maviglia, R., Caglia, A. *et al.* (1999), 'Estimating European admixture in African Americans by using microsatellites and a microsatellite haplotype (CD4/Alu)', *Hum. Genet.* Vol. 104, pp. 149–157.
44. Hanis, C.L., Chakraborty, R., Ferrell, R.E. and Schull, W.J. (1986), 'Individual admixture estimates: Disease associations and individual risk of diabetes and gallbladder disease among Mexican-Americans in Starr County, Texas', *Am. J. Phys. Anthropol.* Vol. 70, pp. 433–441.
45. Gardner, Jr., L.I., Stern, M.P., Haffner, S.M. *et al.* (1984), 'Prevalence of diabetes in Mexican Americans. Relationship to percent of gene pool derived from Native American sources', *Diabetes* Vol. 33, pp. 86–92.
46. Long, J.C., Williams, R.C., McAuley, J.E. *et al.* (1991), 'Genetic variation in Arizona Mexican Americans: Estimation and interpretation of admixture proportions', *Am. J. Phys. Anthropol.* Vol. 84, pp. 141–157.
47. Williams, R.C., Long, J.C., Hanson, R.L. *et al.* (2000), 'Individual estimates of European genetic admixture associated with lower body-mass index, plasma glucose, and prevalence of type 2 diabetes in Pima Indians', *Am. J. Hum. Genet.* Vol. 66, pp. 527–538.
48. Chakraborty, R. and Weiss, K.M. (1986), 'Frequencies of complex diseases in hybrid populations', *Am. J. Phys. Anthropol.* Vol. 70, pp. 489–503.
49. Pfaff, C.L. (2001), 'Estimating admixture dynamics: Implications for mapping genes', PhD Thesis, Department of Anthropology, The Pennsylvania State University, University Park, PA, USA.
50. Briscoe, D., Stephens, J.C. and O'Brien, S.J. (1994), 'Linkage disequilibrium in admixed populations: Applications in gene mapping', *J. Hered.* Vol. 85, pp. 59–63.
51. Lautenberger, J.A., Stephens, J.C., O'Brien, S.J. and Smith, M.W. (2000), 'Significant admixture linkage disequilibrium across 30 cM around the FY locus in African Americans', *Am. J. Hum. Genet.* Vol. 66, pp. 969–978.
52. McKeigue, P.M. (1997), 'Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations', *Am J. Hum. Genet.* Vol. 60, pp. 188–196.
53. Ewens, W.J. and Spielman, R.S. (1995), 'The transmission/disequilibrium test: History, subdivision, and admixture', *Am. J. Hum. Genet.* Vol. 57, pp. 455–464.
54. Zheng, C. and Elston, R.C. (1999), 'Multipoint linkage disequilibrium mapping with particular reference to the African-American population', *Genet. Epidemiol.* Vol. 17, pp. 79–101.
55. Molokhia, M. and McKeigue, P.M. (2000), 'Risk for rheumatic disease in relation to ethnicity and admixture', *Arthritis Res.* Vol. 2, pp. 115–125.
56. Rybicki, B.A., Iyengar, S.K., Harris, T. *et al.* (2002), 'The distribution of long range admixture linkage disequilibrium in an African-American population', *Hum. Hered.* Vol. 53, pp. 187–196.
57. Lander, E.S. and Schork, N.J. (1994), 'Genetic dissection of complex traits', *Science* Vol. 265, pp. 2037–2048.
58. Thomas, D.C. and Witte, J.S. (2002), 'Point: population stratification: A problem for case-control studies of candidate-gene associations?', *Cancer Epidemiol. Biomarkers Prev.* Vol. 11, pp. 505–512.
59. Hoggart, C.J., Parra, E.J., Shriver, M.D. *et al.* (2003), 'Control of confounding of genetic associations in stratified populations', *Am. J. Hum. Genet.* Vol. 72, pp. 1492–1504.
60. Devlin, B. and Roeder, K. (1999), 'Genomic control for association studies', *Biometrics* Vol. 55, pp. 997–1004.
61. Pritchard, J.K., Stephens, M., Rosenberg, N.A. and Donnelly, P. (2000), 'Association mapping in structured populations', *Am. J. Hum. Genet.* Vol. 67, pp. 170–181.
62. Reich, R.E. and Goldstein, D.B. (2001), 'Detecting association in a case-control study while correcting for population stratification', *Genet. Epidemiol.* Vol. 20, pp. 4–16.
63. Chakraborty, R., Ferrell, R.E., Stern, M.P. *et al.* (1986), 'Relationship of prevalence of non-insulin-dependent diabetes mellitus to Amerindian admixture in the Mexican Americans of San Antonio, Texas', *Genet. Epidemiol.* Vol. 3, pp. 435–454.
64. McKeigue, P.M., Shah, B. and Marmot, M.G. (1991), 'Relation of central obesity and insulin resistance with high diabetes prevalence and cardiovascular risk in South Asians', *Lancet* Vol. 337, pp. 382–386.
65. Hodge, A.M. and Zimmet, P.Z. (1994), 'The epidemiology of obesity', *Baillieres Clin. Endocrinol. Metab.* Vol. 8, pp. 577–599.
66. Songer, T.J. and Zimmet, P.Z. (1995), 'Epidemiology of type II diabetes: An international perspective', *Pharmacoeconomics* Vol. 8 (Suppl. 1), pp. 1–11.
67. Martinez, N.C. (1993), 'Diabetes and minority populations. Focus on Mexican Americans', *Nurs. Clin. North Am.* Vol. 28, pp. 87–95.

68. Douglas, J.G., Thibonnier, M. and Wright, Jr., J.T. (1996), 'Essential hypertension: Racial/ethnic differences in pathophysiology', *J. Assoc. Acad. Minor. Phys.* Vol. 7, pp. 16–21.
69. Gaines, K. and Burke, G. (1995), 'Ethnic differences in stroke: Black–white differences in the United States population. SECORDS Investigators. Southeastern Consortium on Racial Differences in Stroke', *Neuroepidemiology* Vol. 14, pp. 209–239.
70. Zoratti, R. (1998), 'A review on ethnic differences in plasma triglycerides and high-density-lipoprotein cholesterol: Is the lipid pattern the key factor for the low coronary heart disease rate in people of African origin?', *Eur. J. Epidemiol.* Vol. 14, pp. 9–21.
71. McKeigue, P.M., Miller, G.J. and Marmot, M.G. (1989), 'Coronary heart disease in south Asians overseas: A review', *J. Clin. Epidemiol.* Vol. 42, pp. 597–609.
72. Ferguson, R. and Morrissey, E. (1993), 'Risk factors for end-stage renal disease among minorities', *Transplant. Proc.* Vol. 25, pp. 2415–2420.
73. Hargrave, R., Stoeklin, M., Haan, M. and Reed, B. (2000), 'Clinical aspects of dementia in African-American, Hispanic, and white patients', *J. Nat. Med. Assoc.* Vol. 92, pp. 15–21.
74. Boni, R., Schuster, C., Nehrhoff, B. and Burg, G. (2002), 'Epidemiology of skin cancer', *Neuroendocrinol. Lett.* Vol. 23(Suppl. 2), pp. 48–51.
75. Schwartz, A.G. and Swanson, G.M. (1997), 'Lung carcinoma in African Americans and whites. A population-based study in metropolitan Detroit, Michigan', *Cancer* Vol. 79, pp. 45–52.
76. Shimizu, H., Wu, A.H., Koo, L.C. *et al.* (1985), 'Lung cancer in women living in the Pacific Basin area', *Nat. Cancer Inst. Monogr.* Vol. 69, pp. 197–201.
77. Hoffman, R.M., Gilliland, F.D., Eley, J.W. *et al.* (2001), 'Racial and ethnic differences in advanced-stage prostate cancer: The Prostate Cancer Outcomes Study', *J. Nat. Cancer Inst.* Vol. 93, pp. 388–395.
78. Rosati, G. (2001), 'The prevalence of multiple sclerosis in the world: An update', *Neurol. Sci.* Vol. 22, pp. 117–139.
79. Bohannon, A.D. (1999), 'Osteoporosis and African American women', *J. Womens Health Gend. Based Med.* Vol. 8, pp. 609–615.
80. Brutsaert, T.D., Parra, E.J., Shriver, M.D. *et al.* (2003), 'Spanish genetic admixture is associated with larger VO_{2max} decrement from sea level to 4,338 meters in Peruvian Quechua', *J. Appl. Physiol.* Vol. 95, No. 2, pp. 519–528.