ARTICLE

A Genomewide Single-Nucleotide–Polymorphism Panel with High Ancestry Information for African American Admixture Mapping

Chao Tian, David A. Hinds, Russell Shigeta, Rick Kittles, Dennis G. Ballinger, and Michael F. Seldin

Admixture mapping requires a genomewide panel of relatively evenly spaced markers that can distinguish the ancestral origins of chromosomal segments in admixed individuals. Through use of the results of the International HapMap Project and specific selection criteria, the current study has examined the ability of selected single-nucleotide polymorphisms (SNPs) to extract continental ancestry information in African American subjects and to explore parameters for admixture mapping. Genotyping of two linguistically diverse West African populations (Bini and Kanuri Nigerians, who are Niger-Congo [Bantu] and Nilo-Saharan speakers, respectively), European Americans, and African Americans validated a genomewide set of >4,000 SNP ancestry-informative markers with mean and median F_{ST} values >0.59 and mean and median Fisher's information content >2.5. This set of SNPs extracted a larger amount of ancestry information in African Americans than previously reported SNP panels and provides nearly uniform coverage of the genome. Moreover, in the current study, simulations show that this more informative panel improves power for admixture mapping in African Americans when ethnicity risk ratios are modest. This is particularly important in the application of admixture mapping in complex genetic diseases for which only modest ethnicity risk ratios of relevant susceptibility genes are expected.

Admixture mapping is a potentially powerful approach for mapping disease-susceptibility loci in human complex diseases in admixed populations of parental populations of different continental origin.[1–6] The approach is based on the assumption that some susceptibility variants will be associated with continental ancestry and that this association can be discerned, in admixed populations, by examining linkage to ancestry. Theoretically, when admixture between continental populations has occurred relatively recently, the chromosomal segments derived from the parental populations can be deduced from the different gametic allele frequencies in the admixed population. This gene flow between genetically distinct populations also results in admixture linkage disequilibrium (LD) among loci that have different allele frequencies in the founding populations. In African Americans (AFA) with ~20% European ancestry and ~80% West African (WAFR) ancestry, significant LD can be detectable for as much as 30 cM.[7–10] In addition, studies have shown analytic evidence of ancestry-definable chromosomal segments in AFA[4,6]

The attractions of admixture mapping are that it has higher statistical power than family linkage studies,[11] it requires 200–500-fold fewer markers than do association studies for a whole-genome scan, and it is less susceptible to allelic heterogeneity, which can confound genetic studies.[12] Admixture mapping has the important advantage over general association studies of not being deterred by multiple independent mutational events that may have occurred—provided the events have accumulated differentially between the continents—since only an allele's ancestral identity is used. Recently, two studies have identified putative chromosomal regions linked to disease traits through use of admixture-mapping approaches, further highlighting interest in extending these methods.[13,14] The key requirements for admixture mapping are appropriate analytic tools and a set of ancestry-informative markers (AIMs) that can distinguish between the continental populations. It is important to note that, for AIMs, the allele-frequency differences (δs) between continental populations are an order of magnitude greater than those among continental subpopulations.[15,16]

Both the development of algorithms for admixture mapping and the identification of markers for defining continental ancestry have progressed rapidly.[1–5] Currently, the most useful admixture-mapping algorithms rely on hidden Markov models (HMMs) for determining ancestry linkage. The HMM approach is designed to infer the unobserved local ancestries for each individual and is specifically designed to take advantage of the multipoint information from linked markers. The transition probabilities in HMM, simulated by Poisson arrivals, provide an approximation of the correlations in ancestry between linked markers. Although the actual underlying model is never known, simulation studies have shown the ability of these methods to discern ancestry linkage in a variety of admixture models and conditions. The current study further examines the power and limitations of this approach in the context of a new, more extensive panel of validated AIMs for use in studies of ancestry linkage in AFA.

Identification of markers for admixture mapping has accelerated as a by-product of large-scale genomic projects and genotyping platforms.[17–19] For admixture mapping in

From the Rowe Program in Human Genetics, Departments of Biochemistry and Medicine, University of California–Davis (UC-Davis), Davis (C.T.; R.S.; M.F.S.); Perlegen Sciences, Mountain View, CA (D.A.H.; D.G.B.); and Comprehensive Cancer Center, Ohio State University, Columbus (R.K.)

Received June 14, 2006; accepted for publication July 25, 2006; electronically published August 15, 2006.

Address for correspondence and reprints: Dr. Michael F. Seldin, Room 4453, Tupper Hall, Department of Biological Chemistry: Medicine, One Shields Avenue, University of California, Davis, Davis, CA 95616. E-mail: mfseldin@ucdavis.edu

Am. J. Hum. Genet. 2006;79:640–649. 0002-9297/2006/7904-0006$15.00

AFA, a set of ~1,500 AIMs has been validated and used for admixture mapping, with a provocative result suggesting a new important susceptibility region for multiple sclerosis.[4,13] With the recent completion of HapMap phase 1[20,21] and phase 2, the opportunity to develop a more comprehensive and informative panel of AIMs is evident. With use of the HapMap SNP frequencies as a primary screen, the current study validates a more comprehensive SNP AIM panel, with increased power for admixture mapping.

Methods

Populations Samples

DNA samples or genotyping results for 24 European Americans (EURA), 60 CEPH Europeans (CEU), 72 WAFR, 60 Yoruban West Africans (YRI), and 96 AFA were included in this study. These populations were based on self-identified ethnic affiliation. The EURA were from New York City; the WAFR were collected in Nigeria and were either (1) Bini, a Niger-Congo group of Bantu speakers from Edo State (24 subjects), or (2) Kanuri, a group of Nilo-Saharan speakers from the Lake Chad region of northern Nigeria (48 subjects). The CEU and YRI were the HapMap panel genotypes,[20] and the AFA DNA samples were obtained from Coriell Institute for Medical Research. The other DNA samples and blood samples were obtained from all individuals, according to protocols and informed-consent procedures approved by institutional review boards, and were labeled with an anonymous code number or, in the case of the AFA, under approved procedures. The subjects studied were all healthy and unrelated.

Statistical Methods

Population admixture proportions were determined using (1) the weighted least-square methods[22] applied in the program ADMIX.PAS, (2) the Bayesian clustering algorithms developed by Pritchard and applied in the program STRUCTURE version 2.1,[23,24] and (3) another Bayesian clustering algorithm, ADMIXMAP.[25] Individual admixture proportions were determined using STRUCTURE 2.1 and ADMIXMAP. For STRUCTURE, each analysis was performed without any prior population assignment and was performed at least three times, with similar results, with >5,000 replicates and 2,000 burn-in cycles under the admixture model. We used the "infer α" option, with a separate α estimated for each population (where α is the Dirichlet parameter for degree of admixture). Runs were performed under the $\lambda = 1$ option, where λ parametizes the allele-frequency prior and is based on the Dirichlet distribution of allele frequencies. The log likelihood of each analysis at varying number of population groups (k) is also estimated in the STRUCTURE analysis and, as expected, favored two population groups in the AFA. For analyses using different values of k ($k = 2, k = 3, \ldots k = 6$), at least 95% of the ancestry in the AFA population was derived from two clusters that corresponded to the WAFR and EURA clusters. For ADMIXMAP, 23,000 iterations and 2,000 burn-in cycles were used under the random mating model. The runs were performed under prior allele-frequency estimation, with use of the results of the parental allele–frequency determinations. The number of generations was allowed to vary and thus was determined for each gamete by the Markov chain–Monte Carlo (MCMC) algorithm.

Admixture mapping on simulated data sets was performed using several computational algorithms: the MALDSOFT algorithm[5] applied to STRUCTURE[24] results, ADMIXMAP,[25] and ANCESTRYMAP.[4] For MALDSOFT/STRUCTURE, we used the same parameters described above, using the linkage option. For ADMIXMAP, most runs were performed using 2,000 iterations and 400 burn-in cycles. Similar results were obtained using longer runs (23,000 iterations and 2,000 burn-in cycles), and monitoring of ergodic averages showed that the sampler had run long enough for the posterior means to have been estimated accurately. For ANCESTRYMAP, 400 iterations and 200 burn-in cycles were used; longer runs produced similar results. A normalized score of 4.0 for ADMIXMAP and STRUCTURE/MALDSOFT was found to approximate a conservative α level that was based on large numbers of simulations. For ANCESTRYMAP in the case-only algorithm, a LOD score of 4.0 was similarly used as an appropriate α level, and, for case-control, the level corresponded to a Z score of 4.0. For ANCESTRYMAP in the case-only algorithm, a LOD score of 4.0 was similarly used as an appropriate α level, and, for case-control, the level corresponded to a Z score of 4.0.

F_{ST} was determined using Genetix software, which applies the Weir and Cockerham algorithm.[26] Hardy-Weinberg equilibrium was examined using an exact test implemented in the FINETTI software, which can be accessed interactively at the Institut für Humangenetik Web site. The measures of informativeness of each SNP—I_n, I_a, ORCA, and Fisher's information content (FIC)—were determined using the algorithms described elsewhere.[27,28] The Perl script used for I_n, I_a, and ORCA was kindly provided by Dr. Noah Rosenberg. For estimating FIC in the HapMap data, the AFA allele frequency was based on 0.8 African and 0.2 European contributions. For the final FIC determinations, the actual allele frequencies in the AFA sample set were used. The FIC scale varies from 0 (no information) to a maximum of 6.5 (when alleles are fixed in opposite directions in the parental population and the allele frequencies in the admixed populations are at a 0.8:0.2 proportion). Importantly, the FIC uses the admixture proportion in determination of the relative information of each SNP.

Selection of SNPs from HapMap Data

SNPs were chosen from the genotyping results of the HapMap Project, to obtain a genomewide set of AIMs for distinguishing between WAFR and European origin. Genotypes from 60 unrelated subjects (parents) from YRI and 60 unrelated CEU were available for ~4 million SNPs in the combined phase 1 and phase 2 HapMap results. Initial examination of these sets identified >300,000 SNPs with an F_{ST} >0.25 and an FIC >1.0. The FIC allows selection of markers that are particularly informative in an admixed population in which the contribution of one parental population is substantially greater than that of the other parental population. It favors selection of markers that are closer to fixation in the parental population with the greater contribution (West African in the current study). With use of the FIC measurement and a computational algorithm that we developed, ~7,000 SNPs were selected from this set of 300,000 SNPs by choosing a maximum of 4 SNPs in 2-Mb windows, with a minimum distance of 100 kb between SNPs. Additional SNPs were added in regions with lower informativeness, SNPs were thinned in regions of high informativeness, and SNPs that failed assay design algorithms for the Perlegen Sciences lithographic array platform were replaced with other informative SNPs, to complete the set of 5,362 SNPs.

Genotyping

The genotyping platform employed high-density oligonucleotide, photolithographic microarrays (DNA chips) as described elsewhere.[29] In brief, 25-bp oligonucleotides were designed to include 24 features per SNP, corresponding to forward- and reverse-strand tilings for sequences complementary to each of two SNP alleles. Genomic DNA was PCR amplified, labeled, and then hybridized to a microarray containing these oligonucleotides. After a series of reactions to develop the signal, the hybridization of the labeled sample to the chip was detected using a confocal laser scanner, and each SNP genotype was determined from signal intensities by use of algorithms described elsewhere.[29]

Validation and Exclusion Methods

To validate the genomewide panel of SNP AIMs, secondary SNP screening was performed using the 5,362 autosome and X-chromosome SNPs selected from the HapMap results, as described above. Typing was performed using genomic DNA from 24 EURA, 24 Bini WAFR, and 48 Kanuri WAFR subjects. Of the initial SNPs, 416 SNPs were excluded from further analysis because of absent or incomplete typing (<90% complete) on this genotyping platform. Another 22 SNPs were found to be monomorphic with use of this platform. Nine SNPs were nearly fixed in the same direction in EURA and WAFR populations and may represent errors in the initial HapMap compilation, with respect to the allele frequency in the African or European populations. For establishing a final genomewide AIM set, additional SNPs were then excluded on the basis of several criteria: (1) >20% difference in the allele frequency either in comparison of the HapMap CEUR and the new EURA sample (139 SNPs excluded) or in comparison of the HapMap YRI sample and the new Bantu-speaking Bini WAFR sample (14 SNPs excluded); (2) δ >20% between Bini and Kanuri WAFR subjects (11 SNPs excluded); (3) $F_{ST} > 0.1$ and $\delta > 0.15$ between HapMap CEUR and new EURA (3 SNPs excluded); (4) $F_{ST} > 0.1$ and allele frequency >0.15 between HapMap YRI and Bini-speaking WAFR subjects (14 SNPs excluded); (5) FIC > 0.1 and allele frequency >0.15 between Bini and Kanuri WAFR subjects (18 SNPs excluded), (6) an FIC value of <1.0, calculated using the new EURA, WAFR, and AFA genotypes (340 SNPs excluded); (7) an FIC <1.5 with use of the new genotyping results if the new FIC-typing results showed an absolute difference of >1.0 from the HapMap results (40 SNPs excluded); (8) a combined WAFR/EURA $F_{ST} < 0.35$ (44 SNPs excluded); and (9) genotyping results in any of the parental groups that are not in Hardy-Weinberg equilibrium (with use of criteria of exact *P* value <.005) (70 SNPs excluded). The purpose of these filters was to eliminate SNPs with substantially heterogeneous allele frequencies within populations of the same continental origin. These exclusions resulted in a final genomewide set of 4,222 SNP AIMs (see our Rich Text Format [RTF] file [online only]).

Genetic Maps

For the current studies, the analyses are, in part, dependent on the position of SNPs on genetic maps. Four genetic maps were used: deCODE,[30] Marshfield,[31] Rutgers,[32] and a modified physical map. The Marshfield map uses genotyping data from selected families from CEPH. The Rutgers genetic map combines genotyping data from both the CEPH and deCODE families. For the deCODE, Marshfield, and Rutgers maps, the position of each SNP was determined by interpolation with use of markers that were both on the genetic map and for which an unambiguous physical map position was available in National Center for Biotechnology Information Build 35. Any markers that were not in the same relative order in both the genetic and physical maps were omitted as anchors for the interpolation of the genetic positions of the SNPs. The modified physical map was considered the physical map that directly reflects the genetic map (1×10^6 bp = 1 cM), except for deletion of the pericentromeric regions, in which no recombination is known to occur.

SNP AIM Subsets

Smaller subsets of SNP AIMs examined in this study were derived from the 4,222 SNP AIM set. Sets of 2,221 SNPs and 1,110 SNPs were chosen simply by selecting every other or every fourth SNP AIM in order of chromosomal position. A set of 2,000 SNPs enriched for information content was obtained by choosing the three most informative SNPs in each 5-cM segment of the deCODE map and then by empirical testing for informativeness in admixture mapping. SNPs were then added in regions where the informativeness was low and were removed from regions with high informativeness.

Simulations and Power Estimations

Simulations were performed by modification of a program developed elsewhere.[2] Chromosomes were simulated using a continuous-gene-flow model based on 80:20 European:African admixture for 6 generations (conditions similar to that estimated for the AFA population).[4,6,8] The WAFR and EURA allele frequencies for each marker were as we determined (see our RTF file [online only]) or as reported elsewhere[33] for the comparison of AIM sets. The simulations were performed using different ethnicity risk ratios (ERRs), defined as "the risk in the admixed population conferred by one parental group compared with the other parental group." The results of simulations were analyzed using three different computational algorithms.[4,5,25] For assessment of admixture information, genomewide analyses were performed in each analysis. For power estimation, simulations of three chromosomes were performed to decrease the computational time required for these analyses. Testing of each program showed that the power estimation was similar when three chromosomes, rather than the entire genome, were used. To examine the effects of LD in parental populations, we added markers to both the simulated parental and admixed chromosomes with identical genotypes (complete LD) to an adjacent marker. We tested 16 different positions on chromosomes 6 and 8 for this aspect of the study.

Estimation of Ancestry across Each Chromosomal Region

For the 96 AFA with real genotyping data and for simulated sets of 96 AFA or 700 AFA, we used STRUCTURE output to estimate the contribution of WAFR and EURA parental populations at each marker. The log probability ratio (WAFR vs. EURA) for each marker was estimated as described elsewhere.[6] Only markers with log likelihood probability ratio >2 WAFR or EURA were scored.

Results

Screening and Validation of European/West African SNP AIMs

To develop a more comprehensive genomewide SNP AIM panel for admixture mapping, 5,400 SNPs were selected on the basis of analysis of HapMap results (see the "Methods" section). These putative AIM SNPs were genotyped on additional samples from EURA (24 subjects), two WAFR (24 Niger-Congo speakers; 48 Nilo-Saharan speakers), and AFA (96 subjects) populations. Exclusion criteria (described in the "Methods" section) resulted in a genomewide panel of 4,222 SNPs. The new typing results were then combined with the HapMap results that provided 60 additional unrelated European subjects and 60 additional WAFR subjects. The AIMs distinguishing WAFR and EURA had the following mean (median) values: $\delta = 0.593$ (0.592), F_{ST} 0.601 (0.596), and FIC 2.695 (2.52). For this set, the mean (median) values of intrapopulation differences were ~2 orders of magnitude smaller: CEU versus new EURA, mean (median) $\delta = 0.059$ (0.051), mean (median) $F_{ST} = 0.00$ (0.00); YRI versus new Bantu speakers, mean (median) $\delta = 0.026$ (0.017), $F_{ST} = 0.003$ (0.003); all Bantu speaking versus Nilo-Saharan speakers, mean (median) $\delta = 0.035$ (0.025), $F_{ST} = 0.013$ (0.004). As expected, for each SNP AIM in this set, the allele frequency in the Coriell set of AFA subjects was between the WAFR and EURA allele frequencies (fig. 1). The average number of generations since admixture was also estimated in the AFA individuals. With use of the STRUCTURE algorithm, the mean number of generations was 6.66. For ADMIXMAP, which examines each gamete separately, the mean (median) number of generations was 7.30 (6.7). These estimates were in rough agreement with previous estimates in other AFA subjects.[4,6]

Assessment of Genomewide Information Content

In addition to examining the individual information content of each SNP, we assessed the ability to extract admixture-mapping information using ADMIXMAP. This algorithm determines the ability to assign ancestry along the chromosome in the admixed population as a function of the amount of observed variance compared with total variance. The extracted information in the AFA subjects is determined on the basis of the ancestry information content of the markers, the genetic map, and the empirical assessment of the admixture model (number of generations since admixture for each gamete). For estimation of the admixture-mapping information, we used the mean extracted information between adjacent SNPs as the information content of each interval. This estimated admixture-mapping information was calculated for the entire set of 4,222 SNP AIMs and for smaller subsets of AIMs, with use of four different genetic maps (fig. 2 and table 1). When the interpolated deCODE map was considered (see the "Methods" section), the 4,222 SNP AIMs extract >60% of the admixture information for >98% of the genome and >70% for >90% of the genome. With use of 2,000 SNPs selected for informativeness, the coverage was only marginally decreased for these levels of extracted admixture information. However, when genomic regions with more-complete admixture information was considered, there was a much larger difference in the genomic coverage (e.g., at the level of a minimum of 80% extracted information, the 4,222 SNPs provided coverage for >66% of the genome, whereas the 2,000-marker set provided this level of extracted information for 35% of the genome). With use of the Marshfield genetic map (which is based on selected CEPH pedigrees) or the Rutgers genetic map (which is a combination of deCODE and CEPH data), the coverage was similar to that with use of the deCODE map (table 1). Coverage for the map based directly on physical mapping distances was greater for each SNP AIM set (table 1). This is probably due to our initial SNP selection methods, which excluded SNPs within 100-kb intervals, and the relatively smaller number of SNPs in physical regions with very high meiotic recombination.

A comparison was also performed between a SNP AIM set published elsewhere[19] and the current SNP set (table

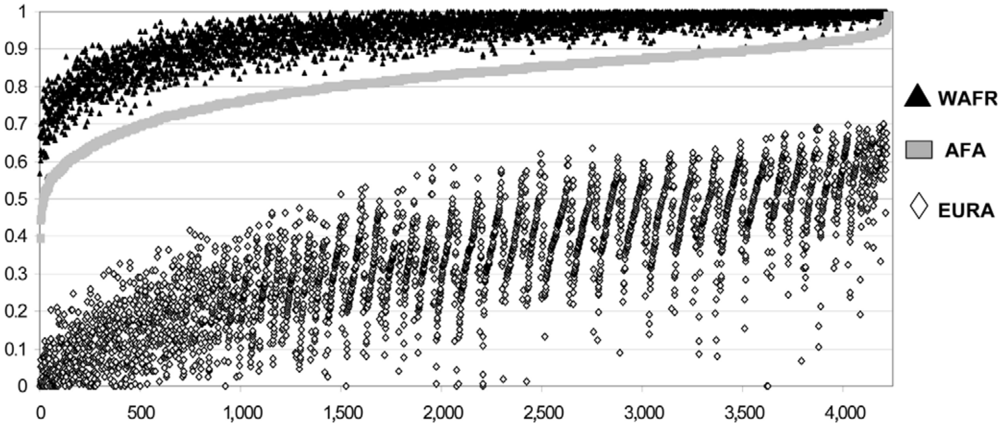

Figure 1. δ of EURA/WAFR SNP AIMs. The SNP AIMs were arranged in ascending allele frequency in the AFA population, and, for each SNP, the WAFR SNP allele was chosen as the higher-frequency allele. Since AIM SNPs were chosen to maximize FIC, the majority of SNPs are close to fixation in the WAFR subjects. The data are based on nearly complete genotyping results in 84 EURA, 142 WAFR, and 96 AFA subjects.

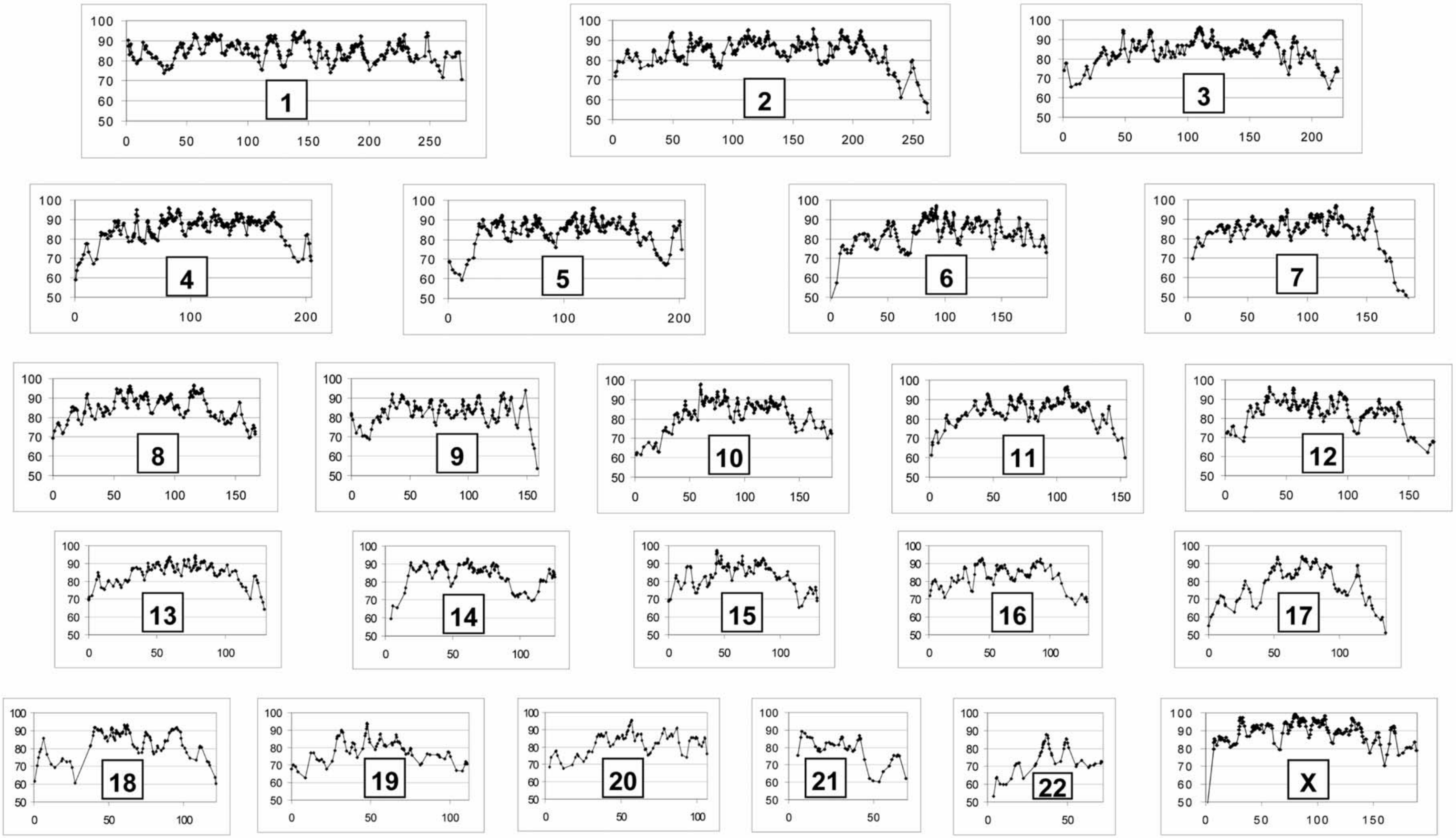

Figure 2. Admixture-mapping distribution for each chromosome. The admixture-mapping information (ordinate) is shown for each position on the deCODE sex-averaged map. The information was determined using ADMIXMAP analysis of genotyping results of 4,222 SNPs typed in the AFA samples (96 subjects).

Table 1. Assessment of Genomic Coverage of SNP AIMs

Sample Set,[a] SNP Set,[b] and Genetic Map	Admixture Mapping–Information Extraction[c] (%)					
	>.50	>.60	>.70	>.75	>.80	>.85
96 AFA:						
4,222 CS:						
deCODE	99.8	98.9	90.7	81.5	66.6	38.8
Marshfield	99.8	98.4	91.0	81.3	64.5	37.5
Rutgers	100.0	98.8	90.5	81.3	67.4	38.4
Modified physical	99.9	99.5	95.9	90.9	80.9	57.1
2,111 CS:						
deCODE	97.5	89.6	70.4	52.3	26.3	11.6
Marshfield	97.9	90.1	68.5	51.0	28.4	11.1
Rutgers	96.8	88.8	69.9	51.2	29.7	10.6
Modified physical	99.9	98.2	87.5	67.7	33.6	10.0
2,000 CS:						
deCODE	99.9	98.1	86.3	66.6	34.9	10.2
Marshfield	99.7	97.5	84.8	61.5	33.2	9.4
Rutgers	99.4	97.8	84.4	64.4	34.1	9.4
Modified physical	100.0	99.7	90.6	65.0	30.7	9.6
1,056 CS:						
deCODE	86.6	63.0	28.6	14.1	6.8	3.8
Marshfield	86.3	62.9	28.4	15.7	6.9	4.5
Rutgers	83.4	60.7	25.9	12.0	3.9	1.4
Modified physical	97.1	82.0	30.8	12.4	5.8	3.0
700 Simulation:						
4,222 CS:						
deCODE	99.7	98.9	90.2	81.1	65.9	43.3
2,154 PR						
deCODE	89.6	68.3	29.1	12.8	4.4	1.5

[a] The results were determined either from the actual genotypes of 96 AFA or on the basis of simulation of the admixed AFA population.

[b] The different SNP sets were derived either from the current study (CS) or from previous results (PR).[18] The 2,000-CS SNP set was selected for informativeness for the deCODE genetic map (see text for details).

[c] The percentage of the genome covered at different levels of admixture-mapping information for different SNP sets based on different genetic maps. The level of admixture mapping was determined using ADMIXMAP (see text).

1). For this comparison, we used the published allele frequencies for the AIMs and our allele frequencies in identical simulations with realistic modeling conditions (see the "Methods" section). The current 4,222-SNP set showed similar results regardless of whether the actual AFA genotyping data were examined or an AFA data set was generated by our admixture model and simulations (see the "Methods" section) and showed a substantial increase in admixture mapping–information compared with the previous SNP set (table 1).

Analysis of Power Using Simulated Data Sets

To examine the relative efficacy of admixture mapping using different densities of AIMs, we examined simulation models using different ERRs. For each ERR, the power was examined for different marker sets with different abilities to extract information. The disease allele was placed in the middle of a 20-cM interval on chromosome 7, where the complete SNP AIM panel provided nearly complete (~0.9) admixture-mapping information. The marker-allele frequencies and map positions (deCODE) were those in our current SNP data set for this initial level of admixture information. For examination of the power with lower admixture-mapping information, the number of SNPs was decreased throughout the genome and was empirically examined, to provide a plateau of admixture information at the desired levels (0.8, 0.6, and 0.5) for this region containing the modeled disease allele. For the lowest level of admixture information (0.5), the FIC of SNPs in the region of the modeled region was also decreased, to achieve the lower admixture information. For each model, both case-only and case-control analyses were performed (fig. 3). Similar to previous reports for admixture mapping in AFA, the power is substantially higher in case-only compared with case-control analyses for each model. For the case-only analysis, power is substantially increased for the SNP sets with higher admixture information. The falloff in power was most pronounced when the admixture information was 0.5, where ~50% more power is observed for the higher information sets (0.8 and 0.9) at an ERR 1.5. For case-control studies, a modest difference (20%) was observed at ERR = 1.75.

Examination of the Effects of LD in the Parental Population

Previous studies[33,34] have raised an issue concerning possible false-positive results in admixture mapping when there is LD present in the parental population. Although the SNP panel in the current study was chosen such that all

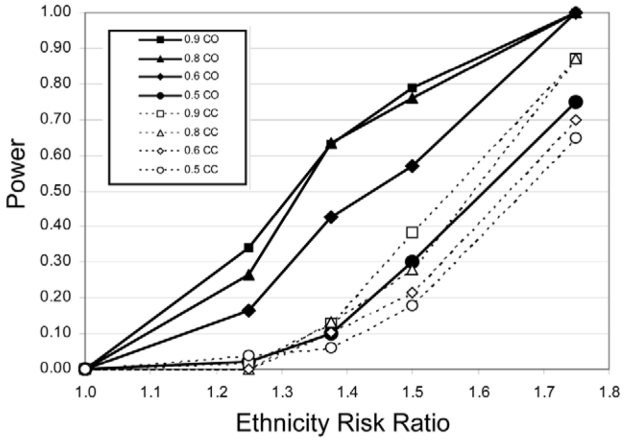

Figure 3. Power for admixture mapping as a function of admixture-mapping information. The power was determined from simulations with 700 cases, 700 controls, and SNP sets with admixture information corresponding to the legend for the SNP set used (see the "Analysis of Power Using Simulated Data Sets" and "Methods" sections). The power curves were determined using the ADMIXMAP program and deCODE genetic map for either case-only (CO) or case-control (CC) analyses. The appropriate α level for these analyses was a normalized score of 4.0, which was based on extensive simulations. The results are based on a minimum of 50 separate simulations and analysis for each measurement. Similar results were obtained using ANCESTRYMAP and MALDSOFT algorithms in more-limited analyses (data not shown).

SNPs were separated by a minimum of 100 kb, residual LD is still present in both parental group representatives. In the 4,222-SNP panel, there are 396 adjacent SNPs with r^2 values >0.2, 189 SNPs with r^2 values >0.5, and 85 SNPs with r^2 values >0.8 in either or both parental population groups. To ascertain whether LD causes false-positive results, we added SNPs in LD to regions of chromosomes 6 and 8 in our simulations, for a modeled locus (ERR = 1.75) on chromosome 7. Admixture mapping was then performed using both case-only and case-control algorithms with use of ADMIXMAP, STRUCTURE, and MALDSOFT. For ANCESTRYMAP case-only analyses, false-positive peaks at least as great as the true-positive peaks were observed at the chromosomal position with complete LD in 4 of 16 simulations. In contrast, no false-positive peaks (normalized score >4.0) were observed in the same simulations with the ADMIXMAP or MALDSOFT algorithms for case-only analyses. For the case-control algorithms, no false-positive peaks were observed with use of any of these three algorithms.

Population and Individual Admixture Assessment

The current data provide a rich source of information for examining admixture in AFA subjects. STRUCTURE analysis strongly favored two populations (see the "Methods" section), and all analyses were performed using this assumption. For population assessment, each chromosome was examined using both weighted least-square testing and the Bayesian cluster methods used in STRUCTURE and ADMIXMAP. For the 96 AFA individuals, the mean EURA contribution to the autosomal chromosomes was similar with use of least-square and MCMC algorithms: least square (SE) 0.211 (0.003), STRUCTURE 0.216 (0.015), and ADMIXMAP 0.216 (0.016). The X chromosome showed a smaller European contribution (least-means square 0.136).

In addition, genomewide assessment of admixture in individuals was examined using STRUCTURE. To provide a measure of the accuracy of these estimations and the number of SNP AIMs required to determine individual accuracy, we examined nonoverlapping SNP sets containing different numbers of SNPs. Each set was chosen to contain evenly distributed SNP AIMs with FIC values >2.5. The correlation of individual admixture was determined for four independent sets of SNP AIMs examined in each of five SNP sets that contained either 20, 40, 80, 160, or 999 autosomal SNP AIMs (table 2). These analyses showed strong correlations between results for individual sets of SNPs. When 160 markers were used, each set had an average 90% Bayesian CI of <10% (ancestry uncertain), and all comparison of the four independent sets had r^2 values >0.9.

Finally, we examined the ancestry assignment across each autosomal chromosome in the 96 AFA individuals, with use of results obtained from STRUCTURE analyses (see the "Methods" section). For this analysis, we scored each marker position along each chromosome as "WAFR origin" if the ln ratio for the probability of WAFR:EURA origin was >2.0 and as "EURA origin" if the ratio EURA: WAFR origin was >2.0. Examination of the genome showed, as expected, a much greater fraction of assigned segments from WAFR than from EURA (mean ratio of assigned chromosomal segments was 4.45:1 WAFR:EURA). When each marker position along each chromosome was examined, there was substantial variation in the ancestry assignment. This estimate of chromosomal ancestry is illustrated for chromosome 1 (fig. 4) and was determined for each chromosome (fig. 5). However, this variation (mean SD = 1.0) was similar to those observed when simulations were performed using the same sample size of 96 individuals (mean SD = 1.4). Thus, on the basis of the admixture model (continuous gene flow and number of generations), there is no evidence of a skewed distribution of ancestry along the chromosome that is independent of stochastic distribution.

Table 2. Individual Admixture Assessment

				r^2	
No. of Markers[a]	FIC	F_{ST}	90% CL[b]	Four Sets[c]	All SNPs[d]
20	3.26–3.35	.66–.73	.232	.56–.72	.75–.84
40	3.12–3.62	.65–.70	.178	.71–.81	.85–.87
80	3.23–3.44	.65–.68	.132	.85–.90	.90–.94
160	3.23–3.46	.66–.68	.096	.93–.96	.96–.96
999	2.56–2.62	.59–.60	.045	.99–.99	.98–.98

[a] Four sets of nonoverlapping markers were chosen for each set, with mean FIC >2.5. The mean FIC range and mean F_{ST} range for the four sets are provided in the adjacent columns.

[b] The mean 90% Bayesian confidence limits (CL) for individual admixture proportion (0–1 scale) is shown.

[c] The range of the correlation coefficients (r^2) for comparisons among four different sets of markers.

[d] The range of the correlation coefficients (r^2) for each of the four sets, with the entire set of 3,997 autosomal SNP AIMs.

Discussion

The current study was undertaken as part of a project to apply admixture mapping in the AFA population. Although previous studies have defined SNP AIMs, it was clear that the results of the HapMap studies could be used to develop a more comprehensive marker set for admixture mapping. Not surprisingly, using SNPs selected from this much larger initial screen, we were able to develop a set of SNPs with greater ability to extract ancestry information and to enhance admixture-mapping power. A novel feature of this panel is the high proportion of markers that are close to fixation in the WAFR parental population (fig. 1). Such markers (1) provide more information in the AFA subjects in whom the individual admixture proportions are close to 20:80 EURA:WAFR and (2) reflect our selection strategy favoring high FIC values rather than high F_{ST} values as the initial selection criterion (see the "Methods" section). Most of this increased power could be realized using an optimal subset of 2,000 AIM SNPs that captures a majority

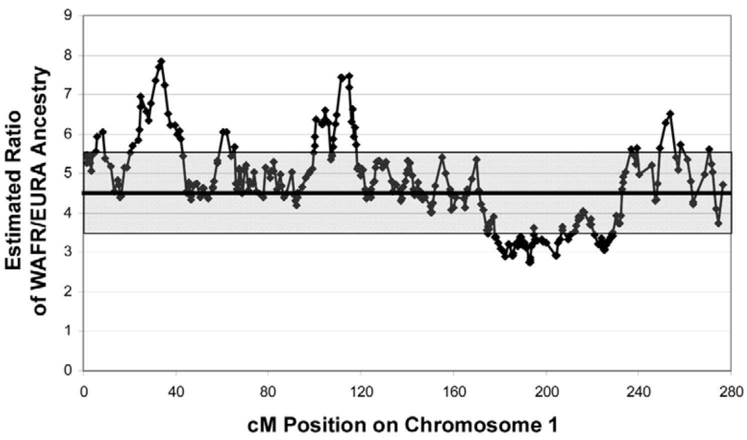

Figure 4. Estimated ratio of WAFR/EURA ancestry across chromosome 1 in 96 AFA individuals. The ancestry ratio was determined by scoring each marker on the basis of the assignment of ancestry (WAFR vs. EURA) in each predicted gamete (see the "Methods" section). For chromosome 1, a mean of 179 of a possible 192 chromosomes for each marker were scored as "WAFR" or "EURA" ancestry, with use of a log likelihood probability ratio >2. The figure shows the autosomal mean ± 1 SD (*gray rectangle*).

of the admixture-mapping information (table 1). The SNPs used for these studies are provided (see our RTF file [online only]).

To assess the information for admixture mapping, as well the power of this approach, a genetic map is required. Since available analyses suggest that there are differences in meiotic recombination frequency across different genomic intervals in different human populations,[35] there is obviously some uncertainty in the choice of an appropriate genetic map. In the present study, we compared four maps: the deCODE, Marshfield, and Rutgers sex-averaged maps that are based on microsatellite data in specific European-derived populations[30–32] and a map in which the genetic distances directly corresponded to physical distances (with correction for the absence of recombination in the pericentromeric region of each chromosome). Since similar power was achieved for four different genetic maps, our results suggest that this issue will probably not be a major deterrent in the application of admixture-mapping methods.

In this study, we examined the ability of SNP sets to define the admixture proportion in both AFA populations and in each individual studied. The results showed that the determination of individual admixture proportions is robust when 160 EURA/WAFR SNP AIMs are used. When individual chromosomes were examined, the EURA contribution to the X chromosome was significantly less than that for each of the autosomal chromosomes, which is consistent with previous observations of decreased EURA contribution to mitochondrial DNA and increased contribution to the Y chromosome,[36] as well as with historical data suggesting sex-biased gene flow.

For the 96 AFA individuals genotyped in this study, chromosomal regions showed substantial fluctuation in the estimated average WAFR/EURA parental ancestry. However, simulated AFA chromosomes showed similar variations, which suggests that most of these deviations are due to small sample size and the admixture model (continuous gene flow and small number of generations). These fluctuations are minimized when sample sizes are substantially larger (in simulations, the SD decreases from 1.4 to 0.66 when 700 subjects rather than 96 subjects are examined). However, these results emphasize the importance of sample size in genetic studies in AFA populations, in that the type II errors will be prevalent in such studies without large sample sizes. Although, sample size is important in all candidate-gene studies, the problem is more severe in admixed populations (especially under continuous-gene-flow models) and is not controlled by structured association methods. Studies with larger numbers of AFA controls will be needed to determine whether there are any genomic regions subject to other factors, such as segregation distortion that may require additional consideration in evaluating case-only admixture-mapping studies. The current study is encouraging, in that these results do not violate the key assumption of case-only algorithms, which presumes homogeneity of ancestry frequencies across the genome in the admixed population.

The complete 4,222-SNP AIM set or the 2,000-SNP AIM set provided in this study will enable genomewide studies

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics.*

Figure 5. Estimated ratio of WAFR/EURA ancestry across each chromosome in 96 AFA individuals. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics.*

with greater power and more-uniform genomic coverage than previous AIM panels. Admixture mapping in the context of a genomewide association test with ~500,000 SNPs has been suggested as an alternative approach. The advantage of performing a dense-genome SNP screen is that additional association results not due to ancestry could be examined in the same study, and loci can be examined for disease LD, controlling for ancestry linkage even within the ancestry-linkage intervals. However, it should be noted that the cost of this genomewide association test is currently ~10-fold more expensive than a genomewide SNP AIM initial study with the use of 2,000 SNP AIMs.

Elsewhere, Reich and Patterson[34] and Smith and O'Brien[33] have suggested that LD in the parental populations might limit the application of admixture mapping with use of a genomewide SNP-association panel. However, as shown in our simulations, with complete LD in the parental populations, this appears to be a substantial problem only in the case-only algorithm in ANCESTRYMAP and does not appear to result in false-positive signals in the case-only algorithms applied in both ADMIXMAP and MALDSOFT. When false-positive signals due to LD were observed in ANCESTRYMAP, we also observed aberrant ancestry assignment in the corresponding chromosomal region. This indicated that the false-positive results were not due to differences in the application of statistical tests (likelihood ratio compared with score test). ANCESTRYMAP fits the admixture generation of each individual rather than each gamete, as performed by ADMIXMAP. However, when this latter option was tested in ADMIXMAP (fixing each individual admixture generation), the false-positive results due to LD were still not observed. We could not discern any difference in the HMM algorithms, which suggests that the explanation is probably due to differences in implementation. Regardless of the explanation, the absence of false-positive tests in ADMIXMAP and STRUCTURE in the presence of LD in parental populations has practical implications, in that it should enable more-robust admixture mapping by not requiring exclusion of markers in partial LD with each other in the parental populations. Additional testing of the sensitivity and specificity of these findings in a variety of admixture and marker conditions is warranted.

A new Markov-HMM (MHMM) algorithm has recently been developed that accounts for LD in parental populations.[37] However, the power and hence efficacy of using whole-genome SNP-association test panels in admixture mapping is not yet clear. The available 500,000-SNP sets will have much smaller numbers of SNPs with AIM characteristics compared with panels chosen from ~4.0 million SNPs. In particular, SNPs that are not close to fixation in the WAFR population may result in false estimations of ancestry if the allele frequencies are under- or overestimated in the putative parental populations tested. We have shown, for particular AIMs, that the allelic variation is relatively small within WAFR populations[15] and have extended these observations in the current study, in which the δs between Niger-Congo (Bantu)–speaking and Nilo-Saharan–speaking WAFRs was small. This is not surprising, since most of the chosen AIMs are close to being fixed in the WAFR populations (fig. 1). However, this limited variation in WAFR populations may not be true for SNPs not close to fixation and may result in more ambiguity in parental-ancestry assignment despite the large numbers of SNPs used in these genomewide association panels. The possible future addition of AIM SNPs to these genomewide association panels would, of course, allow for the best of both scenarios, and the choice of study design may then be driven entirely by cost.

In conclusion, admixture mapping has advanced rapidly over the past several years and should, together with other approaches, help solve the difficult problem of unraveling the genetics of many complex diseases. The current study provides an improved set of AFA SNP AIMs that should be useful in a variety of studies of diseases with higher prevalence in African populations (e.g., lupus [MIM #152700], prostate cancer [MIM #176807], and diabetic nephropathy [MIM #603933]) or European populations (e.g., multiple sclerosis [MIM #126200] and osteoporosis [MIM #166710]). Similarly, panels of SNP AIMs for European versus American Indian ancestry will be available shortly and should provide similar power for admixture mapping in Mexican American populations.

Acknowledgments

This work was supported by National Institutes of Health grant R01 AR050267. We also acknowledge the excellent technical and infrastructural support of our Perlegen and UC-Davis colleagues.

Web Resources

The URLs for data presented herein are as follows:

Institut für Humangenetik, http://ihg.gsf.de/cgi-bin/hw/hwa1.pl (for the FINETTI software)

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for lupus, prostate cancer, diabetic nephropathy, multiple sclerosis, and osteoporosis)

References

1. McKeigue PM (2005) Prospects for admixture mapping of complex traits. Am J Hum Genet 76:1–7
2. Zhang C, Chen K, Seldin MF, Li H (2004) A hidden Markov modeling approach for admixture mapping based on case-control data. Genet Epidemiol 27:225–239
3. Zhu X, Cooper RS, Elston RC (2004) Linkage analysis of a complex disease through use of admixed populations. Am J Hum Genet 74:1136–1153
4. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly MJ, Reich D (2004) Methods for high-density admixture mapping of disease genes. Am J Hum Genet 74:979–1000
5. Montana G, Pritchard JK (2004) Statistical tests for admixture mapping with case-control and cases-only data. Am J Hum Genet 75:771–789
6. Seldin MF, Morii T, Collins-Schramm HE, Chima B, Kittles R,

Criswell LA, Li H (2004) Putative ancestral origins of chromosomal segments in individual African Americans: implications for admixture mapping. Genome Res 14:1076–1084
7. Stephens JC, Briscoe D, O'Brien SJ (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. Am J Hum Genet 55:809–824
8. Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, Hutchinson RG, Ferrell RE, Boerwinkle E, Shriver MD (2001) Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet 68:198–207
9. Collins-Schramm HE, Chima B, Operario DJ, Criswell LA, Seldin MF (2003) Markers informative for ancestry demonstrate consistent megabase-length linkage disequilibrium in the African American population. Hum Genet 113:211–219
10. Lautenberger JA, Stephens JC, O'Brien SJ, Smith MW (2000) Significant admixture linkage disequilibrium across 30 cM around the FY locus in African Americans. Am J Hum Genet 66:969–978
11. McKeigue PM (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 63:241–251
12. Terwilliger JD, Weiss KM (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr Opin Biotechnol 9:578–594
13. Reich D, Patterson N, De Jager PL, McDonald GJ, Waliszewska A, Tandon A, Lincoln RR, DeLoa C, Fruhan SA, Cabre P, Bera O, Semana G, Kelly MA, Francis DA, Ardlie K, Khan O, Cree BA, Hauser SL, Oksenberg JR, Hafler DA (2005) A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. Nat Genet 37:1113–1118
14. Zhu X, Luke A, Cooper RS, Quertermous T, Hanis C, Mosley T, Gu CC, Tang H, Rao DC, Risch N, Weder A (2005) Admixture mapping for hypertension loci with genome-scan markers. Nat Genet 37:177–181
15. Collins-Schramm HE, Kittles RA, Operario DJ, Weber JL, Criswell LA, Cooper RS, Seldin MF (2002) Markers that discriminate between European and African ancestry show limited variation within Africa. Hum Genet 111:566–569
16. Collins-Schramm HE, Chima B, Morii T, Wah K, Figueroa Y, Criswell LA Hanson RL Knowler WC, Silva G, Belmont JW, Seldin MF (2004) 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
17. Collins-Schramm HE, Phillips CM, Operario DJ, Lee JS, Weber JL, Hanson RL, Knowler WC, Cooper R, Li H, Seldin MF (2002) Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. Am J Hum Genet 70:737–750
18. Smith MW, Lautenberger JA, Shin HD, Chretien J-P, Shrestha S, Gilbert DA, O'Brien SJ (2001) Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. Am J Hum Genet 69:1080–1094
19. Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, et al (2004) A high-density admixture map for disease gene discovery in African Americans. Am J Hum Genet 74:1001–1013
20. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. Nature 437:1299–1320
21. Thorisson GA, Smith AV, Krishnan L, Stein LD (2005) The International HapMap Project Web site. Genome Res 15:1592–1593
22. Long JC (1991) The genetic structure of admixed populations. Genetics 127:417–428
23. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
24. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587
25. Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM (2004) Design and analysis of admixture mapping studies. Am J Hum Genet 74:965–978
26. Weir B Cockerham C (1984) Estimating *F*-statistics for the analysis of population structure. Evolution 38:1358–1370
27. Pfaff CL, Barnholtz-Sloan J, Wagner JK, Long JC (2004) Information on ancestry from genetic markers. Genet Epidemiol 26:305–315
28. Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402–1422
29. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079
30. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247
31. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet 63: 861–869
32. Kong X, Murphy K, Raj T, He C, White PS, Matise TC (2004) A combined linkage-physical map of the human genome. Am J Hum Genet 75:1143–1148 (erratum 76:373)
33. Smith MW, O'Brien SJ (2005) Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat Rev Genet 6:623–632
34. Reich D, Patterson N (2005) Will admixture mapping work to find disease genes? Philos Trans R Soc Lond B Biol Sci 360: 1605–1607
35. Evans DM, Cardon LR (2005) A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. Am J Hum Genet 76: 681–687
36. Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63:1839–1851
37. Tang H, Coram M, Wang P, Zhu X, Risch N (2006) Reconstructing genetic ancestry blocks in admixed individuals. Am J Hum Genet 79:1–12