Genetic Epidemiology 35:80–83 (2011)

Ancestry Informative Marker Panels for African Americans Based on Subsets of Commercially Available SNP Arrays

Arti Tandon,[1,2]* Nick Patterson,[2] and David Reich[1,2]

[1]*Department of Genetics, Harvard Medical School, Boston, Massachusetts*
[2]*Broad Institute of Harvard and MIT, Cambridge, Massachusetts*

Admixture mapping is a widely used method for localizing disease genes in African Americans. Most current methods for inferring ancestry at each locus in the genome use a few thousand single nucleotide polymorphisms (SNPs) that are very different in frequency between West Africans and European Americans, and that are required to not be in linkage disequilibrium in the ancestral populations. Modern SNP arrays provide data on hundreds of thousands of SNPs per sample, and to use these to infer ancestry, using many of the standard methods, it is necessary to choose subsets of the SNPs for analysis. Here we present panels of about 4,300 ancestry informative markers (AIMs) that are subsets respectively of SNPs on the Illumina 1 M, Illumina 650, Illumina 610, Affymetrix 6.0 and Affymetrix 5.0 arrays. To validate the usefulness of these panels, we applied them to samples that are different from the ones used to select the SNPs. The panels provide about 80% of the maximum information about African or European ancestry, even with up to 10% missing data. *Genet. Epidemiol.* 35:80–83, 2011.

Key words: ancestry informative markers; admixture mapping; African American disease studies

Contract grant sponsor: NIH; Contract grant number: U01-HG004168; Contract grant sponsor: Burroughs Wellcome Career Development Award.
*Correspondence to: Arti Tandon, Department of Genetics, Harvard Medical School, Boston, MA 02115.
E-mail: atandon@broadinstitute.org
Received 1 April 2010; Revised 20 August 2010
Published online 10 December 2010 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/gepi.20550

INTRODUCTION

The recent reduction in cost for genome wide association studies has led to the collection of genotypes in hundreds of thousands of samples, including tens of thousands of African Americans. To maximize statistical power to detect disease risk factors in African Americans, it is important to compare the frequencies of alleles in cases to those in controls, controlling for whether an individual has inherited 0, 1 or 2 alleles of European ancestry at each locus. Searching for regions of the genome where cases have an unusual amount of one ancestry compared with the genomic average—the ''admixture mapping'' signal—can contribute information about disease gene localization above and beyond the case-control information, and has been used to identify susceptibility factors for multiple sclerosis [Reich et al., 2005], prostate cancer [Freedman et al., 2006] and end stage renal disease [Kao et al., 2008; Kopp et al., 2008].

The best established methods for inferring ancestry in African Americans use data from panels of ancestry informative markers (AIMs) whose alleles are highly differentiated in the ancestral African and European populations and are not in linkage disequilibrium (LD) with each other in the ancestral populations [Parra et al., 1998]. We and others have shown that by genotyping as few as a thousand such markers spaced across the genome, and combining the data across loci using a Hidden Markov Model, one can obtain useful inferences of ancestry at each locus genome-wide [McKeigue, 1998; Patterson et al., 2004; Hoggart et al., 2004; Montana and Pritchard, 2004; Zhu et al., 2004]. This idea has been implemented in the ANCESTRYMAP [Patterson et al., 2004], ADMIXMAP [Hoggart et al., 2004] and MALDSOFT [Montana and Pritchard, 2004] software packages. Our group has developed three iteratively improved panels of markers for admixture mapping in African Americans [Freedman et al., 2006; Smith et al., 2004; Reich et al., 2007], the last of which is a 1,509 single nucleotide polymorphism (SNP) panel (Phase 3 panel) that is available from Illumina for the GoldenGate platform (http://www.illumina.com/products/african_american_admixture_panel.ilmn♯). Another excellent map was generated by Tian et al. [2006].

In an era when commercial SNP genotyping arrays can produce data for hundreds of thousands of genetic markers at a cost that is not much greater than that of genotyping a couple of thousand SNPs, most data on African Americans is collected on hundreds of thousands of markers, instead of a panel of thousands of AIMs. Recent methods such as SABER [Tang et al., 2006], LAMP [Sankararaman et al., 2008b], SWITCH [Sankararaman et al., 2008a], HAPAA [Sundquist et al., 2008] and HAPMIX [Price et al., 2009] use the dense SNP data to make higher resolution inferences of locus-specific ancestry than is possible with the panels presented in this report. However, these methods have not been exhaustively tested in a disease gene mapping context, and in fact, inaccurate modeling of LD may produce systematically biased estimates of ancestry possibly leading to false-positive associations in some of these methods [Price et al., 2008]. An alternative approach is to simply select a

subset of markers from the array that are highly differentiated, and that are separated enough that they are not in LD with each other in the ancestral populations, making it possible to use well-established methods like ANCESTRYMAP, ADMIXMAP and MALDSOFT. Here we present appropriate subsets of five commercial SNP arrays: Illumina 650Y, Illumina 610-Quad, Illumina1M duo, Affymetrix 5.0 and Illumina 6.0.

METHODS

To choose SNPs for each of the commercial panels, we used the data from the International Haplotype Map (http://hapmap.ncbi.nlm.nih.gov/cgi-perl/gbrowse/hapmap3r2_B36/): 55 unrelated European Americans from Utah (CEU) and 54 unrelated Yoruba from Ibadan, Nigeria (YRI) that had been genotyped in all three rounds of the HapMap project [The International HapMap Consortium, 2003]. We restricted analysis to 1,440,616 SNPs that had been genotyped on the Illumina 1M duo and the Affymetrix 6.0 arrays as part of HapMap Phase 3. For each SNP on our array, we calculated the mutual information between the SNP and ancestry, similar to Smith et al. [2004] and Bercovici et al. [2008]. We iteratively selected the next most informative SNP in the genome, conditional on the previously chosen set of SNPs, and continued until no more SNPs were available that satisfied the criteria. The detailed steps were as follows:

SNP PICKING ALGORITHM

(a) *Picking the SNP with the highest expected mutual information content.* We used the Shannon Information Content (SIC) formula [Smith et al., 2004] to pick the SNP that was expected to provide the highest mutual information content to ancestry in the genome, conditional on the observed allele frequencies in the HapMap CEU and YRI samples, (p^{EA} and p^{WA} respectively) and the SNPs that had previously been chosen (Equation 1):

$$\mathrm{SIC} = -\sum_{i=0}^{1}(a_{i0}+a_{i1})\ln(a_{i0}+a_{i1}) - \sum_{j=0}^{1}(a_{0j}+a_{1j})\ln(a_{0j}+a_{1j}) + \sum_{i=0}^{1}\sum_{j=0}^{1} a_{ij}\ln(a_{ij}) \quad (1)$$

Here, $a_{00} = (1-m)p^{WA}$, $a_{01} = mp^{EA}$, $a_{10} = (1-m)(1-p^{WA})$ and $a_{11} = m(1-p^{EA})$, where m is the European mixture proportion (assumed to be 20%). For simplicity in ranking SNPs, we assumed that each candidate SNP was in complete admixture LD with all previously chosen SNPs for 4 centimorgans (cM) in either direction, and was not in admixture LD with more distant SNPs. In practice, admixture LD declines gradually with genetic distance, and a somewhat better ranking of markers could be obtained if we fully modeled this. We note that our empirical evaluation of map informativeness (see Results) properly accounts for the decline of admixture LD with distance, and so the only effect of this simplification is on the prioritization of SNPs for the map.

(b) *Encouraging redundancy in the map*: To deal with the possibility of missing data, we calculated SIC for each SNP after "flattening" the allele frequency estimates in the CEU and YRI by treating them as 20% less divergent than they actually were.

(c) *Excluding SNPs*: We did not pick SNPs that satisfied any of the following criteria: (i) They were within 0.25 Mb or 0.25 centimorgans (cM) of any previously selected SNP. (ii) More than eight other SNPs had previously been chosen within a 2 cM window of the candidate SNP. These exclusion criteria were very aggressive, and resulted in the elimination of many SNPs that in fact could have added marginally to the information content of our map. Our philosophy in building the map was to minimize the possibility that there was no LD among SNPs in the ancestral populations (as this is known to contribute to false-positives), even at the expense of a slight diminishment in map quality. As described below, we also implemented a further step that decreased the possibility that the SNPs were in LD by directly testing for LD in the parental populations.

We applied the selection algorithm to each of the set of SNPs in the Illumina 650Y, 610 Quad, 1M duo, Affymetrix 5.0 and 6.0 arrays in turn, almost all of which were subsets of the approximately 1.4 million SNPs in HapMap3. This produced a rank-ordering of SNPs in decreasing order of expected informativeness, for each SNP array.

SNP PANEL EVALUATION

To evaluate each SNP panel's performance and further filter potentially problematic SNPs, we used the ANCESTRYMAP [Patterson et al., 2004] software, applying it to a second set of 56 European Americans (CEU) and 58 West Africans (YRI) from HapMap Phase 3, as well as to 46 unrelated African Americans from the American Southwest (ASW). These samples were genetically unrelated to those we used to choose markers for the map, ensuring the independence of samples used in marker selection and validation. ANCESTRYMAP has built in data quality checking procedures that allowed us to eliminate markers in the map for which allele counts for the ancestral (African and European) genotypes appeared to be grossly inconsistent with counts on the African-American samples, potentially reflecting genotyping problems. This involved directly measuring LD between all neighboring markers in the map, and throwing out markers that gave even weak evidence of LD based on the conservative criteria from Smith et al. [2004]. After applying the ANCESTRYMAP quality checks we obtained panels of between 4,323 and 4,345 SNPs, depending on the array we analyzed (Table I).

RESULTS

Analysis of the 46 African American (ASW) samples by ANCESTRYMAP indicates that the average European ancestry proportion in the African American samples is $21.1 \pm 8.3\%$, and that the estimated number of generations since admixture averaging across lineages is 5.9 ± 1.3, in the range of what has been estimated for previous studies in African Americans. ANCESTRYMAP also estimates parameters corresponding to how genetically close the modern parental populations are to the true ancestors of the admixed samples. Our estimated F_{ST} between the true African American ancestral population and the Yoruba is <0.0001, and our estimated F_{ST} between the true European American ancestral population and the CEU is <0.0002, highlighting the effectiveness of using the YRI and CEU samples as surrogates for the ancestral populations of

TABLE I. Subsets of different SNP arrays that are useful for local ancestry inference in African Americans

SNP panel	No. of SNPs	Average African-European frequency difference (%)	Average ancestry information (rpower)[a]	Average ancestry information (rpower) with 10% fewer SNPs[a]	Average ancestry information (rpower) with 20% fewer SNPs[a]	(Minimum–maximum) rpower genome-wide	Percent of genome with rpower >60%
HAPADMIX	4,345	60	0.822±0.07	0.788±0.08	0.763±0.08	(0.18–0.99)	98.8%
Illumina 650	4,327	57	0.803±0.07	0.760±0.08	0.738±0.08	(0.17–0.98)	98.4%
Illumina 610	4,331	56	0.800±0.08	0.760±0.08	0.734±0.08	(0.16–0.99)	98%
Affymetrix 5.0	4,325	53	0.772±0.08	0.729±0.08	0.702±0.09	(0.08–0.97)	96.7%
Affymetrix 6.0	4,323	57	0.806±0.07	0.768±0.08	0.74±0.08	(0.14–0.99)	98.6%
Illumina 1M duo	4,332	59	0.816±0.07	0.78±0.08	0.754±0.08	(0.20–0.99)	98.7%
Tian et al. panel	3,234	58	0.768±0.08	0.761±0.09	0.753±0.10	(0.36–0.94)	95.6%
Phase 3 panel	1,396	71	0.689±0.08	0.685±0.09	0.681±0.1	(0.32–0.97)	87.3%

[a]We report ±1 standard deviation.

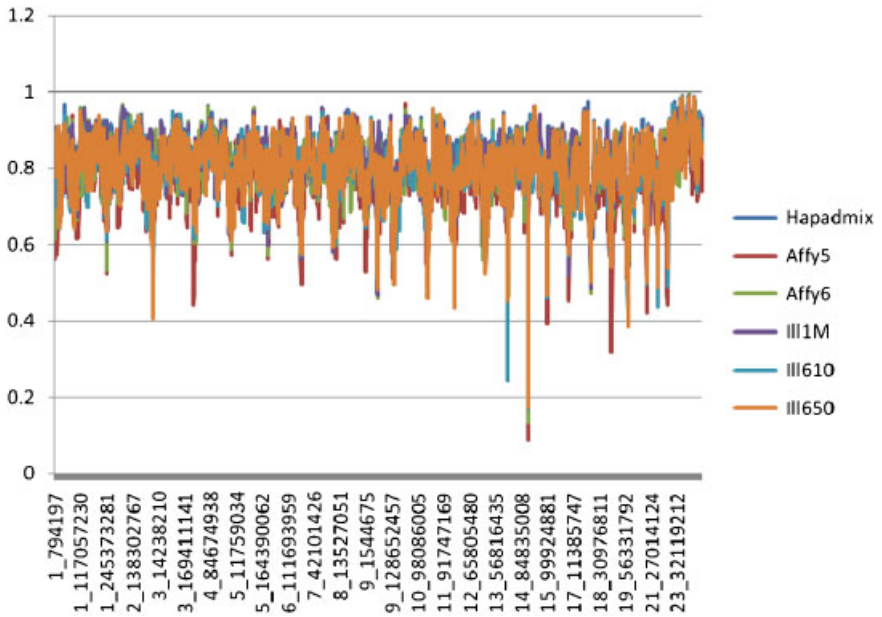

Fig. 1. Estimated information about ancestry extracted at each locus in the genome, based on using each of the panels discussed in this paper and validating its performance on 46 African Americans from the Southwest United States. The average information for the HAPADMIX map of 4,345 markers is 82%, which is much higher than previously reported maps. The only regions of substantially reduced information content occur at the telomeres. The local information content for each of the panels is highly correlated.

African Americans. The fact that the Yoruba provide such a good proxy for the ancestral African population of African Americans may at first seem surprising given that it is known that the African ancestors of African Americans came from many parts of Africa. However, this finding is in fact a straightforward consequence of the fact that the allele frequencies of many populations in West Africa are very similar, a legacy of the large effective population sizes that appear to have persisted for many tens of thousands of years in West Africa [Patterson et al., 2004; Smith et al., 2004].

To evaluate the informativeness of the map at each locus in the genome, we took advantage of the fact that the ANCESTRYMAP software provides a calculation of the "rpower", which is a measure of uncertainty in ancestry inference at a given locus, and specifically, is the expected value of the squared correlation between inferred and true ancestry. For the panel based on all >1.4 million SNPs, which we refer to as the HAPADMIX panel, the average rpower is about 82%. For the least informative map (the subset of the Affy 5.0 array), we infer it to have a high average rpower at 77% (Table I). The high level of redundancy that is built into these maps is demonstrated by the fact that when we re-ran our analysis with 10 and 20% of the markers randomly dropped (to simulate the effect of a high levels of missing data and markers that fail genotyping quality control or design criteria), our rpower was still excellent, ranging from 73 to 79% for 10% missing data and 70 to 76% for 20% missing data depending on the array (Table I). These results should be contrasted with the average rpower (68% in samples that were genotyped for another study) of the Phase 3 map (http://www.illumina.com/products/african_american_admixture_panel.ilmn♯), the average rpower of 76% of the Tian et al. map [Tian

et al., 2006], and the 98% rpower that comes from running HAPMIX [Price et al., 2008] on data from hundreds of thousands of SNPs, taking advantage of the information that is available from markers in LD. A distribution of rpower across the genome for our HAPADMIX and other panels is shown in Figure 1, and indicates that rpower is fairly uniform across the genome. The plots show that almost the only regions where the information content are low are at the ends of the chromosomes, where ancestry information can only be determined based on data from one side of each SNP. Another interesting feature of these plots is that information content rarely exceeds 90%.

DISCUSSION

We believe that the 90% maximum of the rpower reflects the fact that we did not allow markers to be packed at a density of more than one per 0.25 cM or one per 0.25 Mb. It will be difficult to improve rpower further by harvesting additional markers from the 1000 Genomes Project (http://www.1000genomes.org), as most AIMs are likely to have been discovered by earlier SNP discovery projects, since they are markers that are highly different in frequency between West Africans and Europeans and hence are likely to have already been discovered in past SNP discovery efforts that have included samples of both African and European ancestry.

The SNP panels that we have generated are publicly available at our website (http://genepath.med.harvard.edu/~reich/AACommlPanels.html). These panels provide the most informative sets of unlinked AIMs of which we are aware, and are powerful resources for scientists wishing to estimate genome-wide ancestry in African Americans, as well as for using local ancestry as a covariate in disease gene mapping studies.

ACKNOWLEDGMENTS

This work was supported by NIH grants U01-HG004168, and by a Burroughs Wellcome Career Development Award in the Biomedical Sciences to D.R. We thank Alkes Price for critical comments.

REFERENCES

Bercovici S, Geiger D, Shlush L, Skorecki K, Templeton A. 2008. Panel construction for mapping in admixed populations via expected mutual information. Genome Res 18:661–667.

Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, Penney K, Steen RG, Ardlie K, John EM, Oakley-Girvan I, Whittemore AS, Cooney KA, Ingles SA, Altshuler D, Henderson BE, Reich D. 2006. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Natl Acad Sci USA 103:14068–14073.

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.

Kao WH, Klag MJ, Meoni LA, Reich D, Berthier-Schaad Y, Li M, Coresh J, Patterson N, Tandon A, Powe NR, Fink NE, Sadler JH, Weir MR, Abboud HE, Adler SG, Divers J, Iyengar SK, Freedman BI, Kimmel PL, Knowler WC, Kohn OF, Kramp K, Leehey DJ, Nicholas SB, Pahl MV, Schelling JR, Sedor JR, Thornley-Brown D, Winkler CA, Smith MW, Parekh RS, Family Investigation of Nephropathy and Diabetes Research Group. 2008. MYH9 is associated with nondiabetic end-stage renal disease in African Americans. Nat Genet 40:1185–1192.

Kopp JB, Smith MW, Nelson GW, Johnson RC, Freedman BI, Bowden DW, Oleksyk T, McKenzie LM, Kajiyama H, Ahuja TS, Berns JS, Briggs W, Cho ME, Dart RA, Kimmel PL, Korbet SM, Michel DM, Mokrzycki MH, Schelling JR, Simon E, Trachtman H, Vlahov D, Winkler CA. 2008. MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. Nat Genet 40:1175–1184.

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.

Montana G, Pritchard JK. 2004. Statistical tests for admixture mapping with case-control and cases-only data. Am J Hum Genet 75:771–789.

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.

Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly M, Reich D. 2004. Methods for high-density admixture mapping of disease genes. Am J Hum Genet 74:1001–1013.

Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, Ge D, Rotter JI, Torres E, Taylor KD, Goldstein DB, Reich D. 2008. Long-Range LD Can Confound Genome Scans in Admixed Populations. Am J Hum Genet 83:132–135.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet 5:e1000519.

Reich D, Patterson N, 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 gene for multiple sclerosis susceptibility. Nat Genet 37:1113–1118.

Reich D, Patterson N, Ramesh V, De Jager PL, McDonald GJ, Tandon A, Choy E, Hu D, Tamraz B, Pawlikowska L, Wassel-Fyr C, Huntsman S, Waliszewska A, Rossin E, Li R, Garcia M, Reiner A, Ferrell R, Cummings S, Kwok PY, Harris T, Zmuda JM, Ziv E; Health, Aging and Body Composition (Health ABC) Study. 2007. Admixture mapping of an allele affecting interleukin 6 soluble receptor and interleukin 6 levels. Am J Hum Genet 80:716–726.

Sankararaman S, Kimmel G, Halperin E, Jordan MI. 2008a. On the inference of ancestries in admixed populations. Genome Res 18:668–675.

Sankararaman S, Sridhar S, Kimmel G, Halperin E. 2008b. Estimating local ancestry in admixed populations. Am J Hum Genet 82:290–303.

Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, Malasky MJ, Scafe C, Le E, De Jager P, Yi Z, De Thé G, Essex M, Kanki PJ, Moore JH, Poku K, Phair JP, Goedert JJ, Vlahov D, Williams SM, Tishkoff SA, Winkler CA, De La Vega FM, Woodage T, Sninsky JJ, Hafler DA, Altshuler D, Gilbert DA, O'Brien SJ, Reich D. 2004. A high density admixture map for disease gene discovery in African Americans. Am J Hum Genet 74:979–1000.

Sundquist A, Fratkin E, Do CB, Batzoglou S. 2008. Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Res 18:767–782.

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.

The International HapMap Consortium. 2003. The International HapMap Project. Nature 426:789–796.

Tian C, Hinds DA, Shigeta R, Kittles R, Ballinger DG, Seldin MF. 2006. A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. Am J Hum Genet 79:640–649.

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.