

A peer-reviewed version of this preprint was published in PeerJ on 4 November 2014.

[View the peer-reviewed version](https://peerj.com/articles/660) (peerj.com/articles/660), which is the preferred citable publication unless you specifically need to cite this preprint.

Zhang X, Mu W, Liu C, Zhang W. 2014. Ancestry-informative markers for African Americans based on the Affymetrix Pan-African genotyping array. PeerJ 2:e660 <https://doi.org/10.7717/peerj.660>

Ancestry-informative markers for African Americans based on the Affymetrix Pan-African genotyping array

Genetic admixture has been utilized as a tool for identifying loci associated with complex traits and diseases in recently admixed populations such as African Americans. In particular, admixture mapping is an efficient approach to identifying genetic basis for those complex diseases with substantial racial or ethnic disparities. Though current advances in admixture mapping algorithms may utilize the entire panel of SNPs, providing ancestry-informative markers (AIMs) that can differentiate parental populations and estimate ancestry proportions in an admixed population may particularly benefit admixture mapping in studies of limited samples, help identify unsuitable individuals (e.g., through genotyping the most informative ancestry markers) before starting large genome-wide association studies (GWAS), or guide larger scale targeted deep re-sequencing for determining specific disease-causing variants. Defining panels of AIMs based on commercial, high-throughput genotyping platforms will facilitate the utilization of these platforms for simultaneous admixture mapping of complex traits and diseases, in addition to conventional GWAS. Here, we describe AIMs detected based on the Shannon Information Content (SIC) or F_{st} for African Americans with genome-wide coverage that were selected from ~2.3 million single nucleotide polymorphisms (SNPs) covered by the Affymetrix Axiom Pan-African array, a newly developed genotyping platform optimized for individuals of African ancestry.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44

**Ancestry-Informative Markers for African Americans Based on the
Affymetrix Pan-African Genotyping Array**

Xu Zhang^{1,2}, Wenbo Mu³, Cong Liu³, and Wei Zhang^{1,3}

¹The Affiliated Hospital of Medical School, Ningbo University, Ningbo, Zhejiang Province, China
²Section of Hematology/Oncology, Department of Medicine; ³Department of Pediatrics, University of Illinois, Chicago, Illinois, USA

***Correspondence to:**
Wei Zhang, Ph.D., 840 S. Wood St., 1200 CSB (M/C 856), Chicago, IL 60612, USA; Tel: +1 (312) 413-2024; E-mail: weizhang.chicago@gmail.com

Running Title: AIMs for African Americans

Abstract

45 Genetic admixture has been utilized as a tool for identifying loci associated with complex traits
46 and diseases in recently admixed populations such as African Americans. In particular, admixture
47 mapping is an efficient approach to identifying genetic basis for those complex diseases with
48 substantial racial or ethnic disparities. Though current advances in admixture mapping algorithms
49 may utilize the entire panel of SNPs, providing ancestry-informative markers (AIMs) that can
50 differentiate parental populations and estimate ancestry proportions in an admixed population
51 may particularly benefit admixture mapping in studies of limited samples, help identify
52 unsuitable individuals (e.g., through genotyping the most informative ancestry markers) before
53 starting large genome-wide association studies (GWAS), or guide larger scale targeted deep re-
54 sequencing for determining specific disease-causing variants. Defining panels of AIMs based on
55 commercial, high-throughput genotyping platforms will facilitate the utilization of these
56 platforms for simultaneous admixture mapping of complex traits and diseases, in addition to
57 conventional GWAS. Here, we describe AIMs detected based on the Shannon Information
58 Content (SIC) or F_{st} for African Americans with genome-wide coverage that were selected from
59 ~2.3 million single nucleotide polymorphisms (SNPs) covered by the Affymetrix Axiom Pan-
60 African array, a newly developed genotyping platform optimized for individuals of African
61 ancestry.

62

63

64

65

66

67 **Introduction**

68

69

70 High throughput genotyping arrays have facilitated genome-wide association studies
(GWAS) on complex traits (Hindorff et al. 2009) including risks for common, complex diseases

71 and drug response. In contrast to a conventional GWAS in a homogeneous parental populations
72 (e.g., Caucasians), admixture mapping or mapping by admixture linkage disequilibrium (MALD)
73 has begun to be demonstrated as a powerful tool for identifying disease-causing genetic variants
74 in recently admixed populations, such as African Americans that have both West African and
75 European American ancestry (McKeigue 2005). For example, recent admixture mapping studies
76 have identified loci associated with disease risks such as prostate cancer (Ricks-Santi et al. 2012),
77 lung cancer (Schwartz et al. 2011), and traits like blood pressure/obesity (Shetty et al. 2012) in
78 African Americans. Admixture mapping assumes that near a disease causing genetic variant there
79 will be enhanced ancestry from the population that has greater risk of getting the disease
80 (Patterson et al. 2004). Therefore, by calculating the proportion of ancestry along the genome,
81 one could use that information to identify disease causing loci in an admixed population with low
82 resolution. Subsequent fine mapping restricted to the identified genomic regions may greatly
83 increase the power of the study.

84
85 It has been demonstrated that 1,500–2,500 ancestry-informative markers (AIMs) with
86 genome-wide coverage would be sufficient (Winkler et al. 2010) to identify the ancestral
87 chromosome segments for recently admixed populations. To leverage on the power of admixture
88 mapping in African American for identifying disease causing genetic variants that may explain
89 health disparities between populations, panels of AIMs have been proposed for commercially-
90 available high throughput genotyping arrays including the Affymetrix SNP 6.0 and Illumina 1M
91 (Chen et al. 2010; Tandon et al. 2011). These genotyping arrays however are likely biased to
92 genetic variations detected from Caucasian samples. The Affymetrix Pan-African array, which
93 interrogates approximately 2.3 million SNPs, was designed for a much greater coverage of
94 genetic variations in African individuals. A panel of AIMs based on the Pan-African array may
95 enhance the distinguishing of parental populations as well as improve genome coverage. Recent

96 advances in statistical genetics have begun to allow admixture mapping utilizing the entire panel
97 of genotyped SNPs (Baran et al. 2012; Churchhouse & Marchini 2013; Maples et al. 2013),
98 however, we reasoned that providing a panel of AIMs may particularly benefit studies of a
99 limited sample size, help identify unsuitable individuals by genotyping the most informative
100 markers before starting a large GWAS, or guide larger scale targeted re-sequencing projects to
101 pinpoint causal variants. We describe here AIMs identified for the Affymetrix Pan-African array
102 based on Shannon Information Content (SIC) or F_{st} for African Americans using the 1000
103 Genomes Project (Abecasis et al. 2010) data as references for parental populations.

104

105 **Materials and Methods**

106 *SNPs covered on the Pan-African array*

107 The Affymetrix Axiom Genome-Wide Pan AFR Genotyping platform (Pan-African array)
108 (Affymetrix, Inc., Santa Clara, California) covers ~2.3 million SNPs optimized for individuals of
109 African ancestry. The Pan-African array was designed to offer $\geq 90\%$ coverage of SNPs on the
110 Yoruba genome with minor allele frequency (MAF) greater than 2%. Annotations for the Pan-
111 African array can be accessed at the Affymetrix website (<http://www.affymetrix.com/>). As a
112 platform optimized for individuals of African individuals, the Pan-African array has been
113 extensively validated in African populations from the HapMap Project (Altshuler et al. 2010),
114 including the Luhya from western Kenya (LWK), Maasai from eastern Kenya (MWK), Yoruba
115 from Ibadan, Nigeria (YRI), and the African Ancestry in the Southwest USA (ASW) (Lu et al.
116 2011). This platform offers high genomic coverage ($>85\%$) in admixed populations with West
117 African ancestry, thus particularly suitable for genome-wide scans in African American
118 individuals (admixture of African and European populations).

119 *Obtaining allele frequency and genetic map distances on parental populations*

120 Genotypes for 2176716 SNPs covered by the Pan-African array were extracted from the 1000
 121 Genomes Project (Abecasis et al. 2010) Phase I data for the 85 CEU (Caucasian residents from
 122 Utah, USA) and 88 YRI unrelated samples, representing the two major parental populations for
 123 African Americans (Western Africans and Europeans). Genome-wide genetic map distances of
 124 SNPs for genome assembly GRCh37 (Frazer et al. 2007) were downloaded from the website
 125 (http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps).

126 *Selection of ancestry-informative markers*

127 We aimed to pick the SNPs that were expected to provide the highest mutual information
 128 content to ancestry or fixation index (i.e., F_{st} , a measure of population differentiation due to
 129 genetic structure) in the genome using an iterative procedure, conditional on the observed allele
 130 frequencies in the 1000 Genomes Project CEU and YRI samples.

131 (a) *Calculation of mutual information content:* Allele frequencies for the CEU and YRI samples
 132 were used to calculate the Shannon Information Content (SIC) for each SNP using a formula
 133 from previous studies (Smith et al. 2004; Tandon et al. 2011),

$$134 \quad \text{SIC} = - \sum_{i=0}^1 (a_{i0} + a_{i1}) \ln(a_{i0} + a_{i1}) - \sum_{j=0}^1 (a_{0j} + a_{1j}) \ln(a_{0j} + a_{1j}) + \sum_{i=0}^1 \sum_{j=0}^1 a_{ij} \ln(a_{ij})$$

135 , where $a_{00} = (1 - m) \times p^{\text{YRI}}$, $a_{01} = (m \times p^{\text{CEU}})$, $a_{10} = (1 - m) \times (1 - p^{\text{YRI}})$, and

136 $a_{11} = m \times (1 - p^{\text{CEU}})$. Here, p^{CEU} and p^{YRI} are the allele frequencies in the CEU (European) and

137 YRI (African) samples, and m is the proportion of European ancestry in African Americans,
 138 which was set to 0.20 following the same assumption of 20% European ancestry (Tandon et al.
 139 2011). Notably, SNP selection was found not very sensitive to the choice of m (Smith et al.
 140 2004). In addition, the F_{st} was also computed for each of the 2176716 SNPs between the two
 141 parental populations based on Wright's approximate formula (Wright 1950),

$$142 \quad F_{ST} = (H_T - H_S) / H_T$$

143 , where H_T represents expected heterozygosity per locus of the total population and H_S represents
144 expected heterozygosity of a subpopulation.

145 (b) *Selection of AIMs*: We aimed to detect AIMs that are not packed around certain genomic
146 regions due to linkage disequilibrium (LD), thus being more representative of the genome. Since
147 LD declines gradually with increased genetic distance (Shifman et al. 2003), we assume each
148 SNP is not in LD with distant SNPs more than 0.25 cM (~250 kb) away, similar to what was used
149 in previous publications (Tandon et al. 2011). We selected AIMs using an iterative procedure for
150 each chromosome: 1) SNPs were ranked based on SIC; 2) SNP with the highest SIC was selected
151 as a candidate AIM; 3) Any SNPs within 0.25 cM or within 250 kb of the selected SNP were
152 excluded; 4) Steps 2 and 3 were repeated until no more SNPs left. To avoid densely packed
153 markers, no more than 8 candidate AIMs were selected within any 4 cM region. This procedure
154 ensured a good coverage of AIMs across the entire genome. The quality of the detected candidate
155 AIMs was examined using the build-in data quality checking procedure of ANCESTRYMAP 2.0
156 (Patterson et al. 2004) for extracting top “bad” markers, for which allele counts for the ancestral
157 (African and European) genotypes appeared to be grossly inconsistent with counts on the 56
158 unrelated ASW samples from 1000 Genome Project (Abecasis et al. 2010). After applying the
159 ANCESTRYMAP quality checks, we obtained the final panel of AIMs. We also repeated the
160 same selection procedure using F_{st} to identify a companion panel of AIMs. Supplemental Tables 1
161 and 2 contain detailed information on the final AIMs.

162 *Evaluation of the detected AIMs for the Pan-African array*

163 The informativeness of the AIMs was evaluated at each SNP using the ANCESTRYMAP-
164 generated *rpower* value, which is a measure of uncertainty in ancestry inference at a given locus.
165 Specifically, *rpower* is the expected value of the squared correlation between inferred and true
166 ancestry (Patterson et al. 2004). In addition, proportion of variance explained (PVE) by the first
167 principal component (PC) using the detected AIMs on the CEU, YRI, and ASW samples was

168 compared with PVE's from previously published AIMs (based on Affymetrix SNP 6.0 and
169 Illumina 1M arrays) (Tandon et al. 2011) as well as 1000 random sets of SNPs.

170

171

172

173 **Results and Discussion**

174 Given that the Pan-African array was population-optimized, this platform is expected to
175 offer higher coverage of genetic variation for individuals of African ancestry than previous
176 platforms mostly designed based on Caucasians. Genotyping using the Affymetrix Pan-African
177 array will provide opportunities for performing admixture mapping in African Americans to
178 detect genetic variants associated with those traits that exhibit disparities between parental
179 populations, for instance certain cancers (Schwartz et al. 2011). The primary result from this
180 study was a panel of SNPs based on the Pan-African array. We acknowledge that with recent
181 advances in statistical genetics, admixture mapping in African Americans may not rely on a
182 limited number of AIMs any more (Baran et al. 2012; Churchhouse & Marchini 2013; Maples et
183 al. 2013). We propose that some applications for our detected AIMs could include: 1) to facilitate
184 admixture mapping in limited samples; 2) to help identify problematic individuals through
185 genotyping some top-ranking AIMs before starting a large GWAS; 3) to guide targeted re-
186 sequencing projects that may not have genome-wide genotypic data.

187 Using an iterative selection algorithm, a total of 6011 candidate AIMs were detected
188 based on SIC, which can measure the uncertainty in genome-wide ancestry or ancestry at a given
189 locus (Tandon et al. 2011). We further examined the quality of these candidates using the build-in
190 checking procedure of ANCESTRYMAP (Patterson et al. 2004) and identified a final set of AIMs
191 with 5995 SNPs based on SIC. We also repeated the same analysis using F_{st} to identify a
192 companion panel of 6012 after ANCESTRYMAP checking from 6034 detected candidate SNPs.

193 The selected AIMs with rs numbers, genomic positions, reference alleles, alternative alleles, and
194 allele counts in the CEU or YRI samples are shown in supplemental materials. Overall, AIMs
195 based on SIC and F_{st} performed consistently with each other. The average *rpower* (i.e., average
196 ancestry information) of the AIMs based on SIC or F_{st} was 0.85 (**Figure 1A**), compared to ~0.81
197 for previous AIMs detected for Affymetrix SNP 6.0 and Illumina 1M arrays (Tandon et al. 2011).
198 The average proportion of European ancestry in ASW was estimated to be 0.25 and 0.24 and the
199 average generations of admixture was estimated to be 5.4 and 5.5 using the AIMs based on SIC
200 and F_{st} , respectively, consistent with previous estimation (Tandon et al. 2011).

201 The availability of dense genetic variation data from the HapMap Project (HapMap 2003;
202 HapMap 2005) allows a genome-wide analysis of population differentiation. In particular, the
203 CEU (European) and YRI (African) samples represented the two major parental populations of
204 African Americans. Our major criteria of identifying AIMs were designed 1) to enrich SNPs with
205 higher information content (or F_{st}) between the CEU and YRI samples; and 2) to have a
206 comprehensive genomic coverage. The genome-wide iterative scan for AIMs based on a genetic
207 distance bin in a size of 0.25 cM, guaranteed a comprehensive coverage of the entire human
208 genome, as well as limit the possibility that the identified AIMs are in strong LD in a particular
209 genomic region, as described in previous publications (Chen et al. 2010; Tandon et al. 2011). The
210 final AIMs are those SNPs with the highest SIC (or F_{st}) separated by at least the distance of 0.25
211 cM (~250 kb) between the two parental populations. the detected AIMs were able to recapture the
212 most prominent population structures by being tested on the combined HapMap CEU, YRI, and
213 ASW samples (**Figure 1B**). A simulation analysis demonstrated that the detected AIMs based on
214 the Pan-African array explained substantially higher proportion of variance by the first PCs in the
215 same population than random sets of SNPs in the human genome (**Figure 1C**). Though our
216 analysis showed that the AIMs detected based on SIC and F_{st} performed consistently, given some

217 potential problems of F_{st} , in particular its dependency on within-population diversity (Sherwin
218 2010), we generally recommend the use of the final panel of AIMs detected based on SIC.

219 The assumption of no LD based on 0.25 cM (~250 kb) could be stringent and cause loss
220 of some informative SNPs, given that the average distance of LD decay between SNP pairs is
221 around 20-30 kb across diverse populations, with generally shorter distance in African Americans
222 (Shifman et al. 2003). Nevertheless, this cutoff was chosen to balance between minimizing the
223 possibility of LD and the comprehensive genomic coverage of AIMs (Tandon et al. 2011).

224 In summary, the Affymetrix Pan-African array provides a population-optimized
225 genotyping platform for GWAS in individuals of African ancestry. The genotypic data profiled by
226 this platform also offers opportunities for admixture mapping in African Americans, a recently
227 admixed population, for certain complex traits and disease susceptibilities with disparities
228 between parental populations. The AIMs we described in this study represent the most
229 informative sets of unlinked markers that can be an important resource to facilitate such
230 applications based on this new tool.

231

232 **References**

- 233 Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, and
234 McVean GA. 2010. A map of human genome variation from population-scale sequencing.
235 *Nature* 467:1061-1073.
- 236 Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E,
237 Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de
238 Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A,
239 Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs
240 RA, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C,
241 McCarroll SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price
242 AL, Soranzo N, Bonnen PE, Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F,
243 Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner
244 SF, Zhang Q, Ghorri MJ, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF,
245 Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster
246 MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO,
247 Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, and McEwen JE. 2010.
248 Integrating common and rare genetic variation in diverse human populations. *Nature*
249 467:52-58.

- 250 Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, Rodriguez-Cintron W,
251 Chapela R, Ford JG, Avila PC, Rodriguez-Santana J, Burchard EG, and Halperin E. 2012.
252 Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*
253 28:1359-1367.
- 254 Chen G, Shriner D, Zhou J, Doumatey A, Huang H, Gerry NP, Herbert A, Christman MF, Chen Y,
255 Dunston GM, Faruque MU, Rotimi CN, and Adeyemo A. 2010. Development of
256 admixture mapping panels for African Americans from commercial high-density SNP
257 arrays. *BMC Genomics* 11:417.
- 258 Churchhouse C, and Marchini J. 2013. Multiway admixture deconvolution using phased or
259 unphased ancestral panels. *Genet Epidemiol* 37:1-12.
- 260 Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A,
261 Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao
262 Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B,
263 Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A,
264 Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M,
265 Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X,
266 He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM,
267 Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR,
268 Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS,
269 Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD,
270 Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK,
271 Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A,
272 Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R,
273 Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de
274 Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A,
275 Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D,
276 Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE,
277 Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas
278 DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J,
279 Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris
280 AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C,
281 Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi
282 I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A,
283 Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW,
284 Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM,
285 Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly
286 MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM,
287 Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z,
288 Han H, Kang L, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K,
289 Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE,
290 Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS,
291 Kennedy K, Jamieson R, and Stewart J. 2007. A second generation human haplotype map
292 of over 3.1 million SNPs. *Nature* 449:851-861.
- 293 HapMap. 2003. The International HapMap Project. *Nature* 426:789-796.
- 294 HapMap. 2005. A haplotype map of the human genome. *Nature* 437:1299-1320.
- 295 Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA.
296 2009. Potential etiologic and functional implications of genome-wide association loci for
297 human diseases and traits. *Proc Natl Acad Sci U S A* 106:9362-9367.

298 Lu Y, Patterson N, Zhan Y, Mallick S, and Reich D. 2011. Technical design document for a SNP
299 array that is optimized for population genetics. ftp://ftp.cephb.fr/hgdp_supp10/.
300 Maples BK, Gravel S, Kenny EE, and Bustamante CD. 2013. RFMix: a discriminative modeling
301 approach for rapid and robust local-ancestry inference. *Am J Hum Genet* 93:278-288.
302 McKeigue PM. 2005. Prospects for admixture mapping of complex traits. *Am J Hum Genet* 76:1-
303 7.
304 Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL,
305 Smith MW, O'Brien SJ, Altshuler D, Daly MJ, and Reich D. 2004. Methods for high-
306 density admixture mapping of disease genes. *Am J Hum Genet* 74:979-1000.
307 Ricks-Santi LJ, Apprey V, Mason T, Wilson B, Abbas M, Hernandez W, Hooker S, Doura M,
308 Bonney G, Dunston G, Kittles R, and Ahaghotu C. 2012. Identification of genetic risk
309 associated with prostate cancer using ancestry informative markers. *Prostate Cancer*
310 *Prostatic Dis* 15:359-364.
311 Schwartz AG, Wenzlaff AS, Bock CH, Ruterbusch JJ, Chen W, Cote ML, Artis AS, Van Dyke
312 AL, Land SJ, Harris CC, Pine SR, Spitz MR, Amos CI, Levin AM, and McKeigue PM.
313 2011. Admixture mapping of lung cancer in 1812 African-Americans. *Carcinogenesis*
314 32:312-317.
315 Sherwin W. 2010. Entropy and information approaches to genetic diversity and its expression:
316 genomic geography. *Entropy* 12:1765-1798.
317 Shetty PB, Tang H, Tayo BO, Morrison AC, Hanis CL, Rao DC, Young JH, Fox ER, Boerwinkle
318 E, Cooper RS, Risch NJ, and Zhu X. 2012. Variants in CXADR and F2RL1 are associated
319 with blood pressure and obesity in African-Americans in regions identified through
320 admixture mapping. *J Hypertens* 30:1970-1976.
321 Shifman S, Kuypers J, Kokoris M, Yakir B, and Darvasi A. 2003. Linkage disequilibrium patterns
322 of the human genome across populations. *Hum Mol Genet* 12:771-776.
323 Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing
324 BD, Malasky MJ, Scafe C, Le E, De Jager PL, Mignault AA, Yi Z, De The G, Essex M,
325 Sankale JL, Moore JH, Poku K, Phair JP, Goedert JJ, Vlahov D, Williams SM, Tishkoff
326 SA, Winkler CA, De La Vega FM, Woodage T, Sninsky JJ, Hafler DA, Altshuler D,
327 Gilbert DA, O'Brien SJ, and Reich D. 2004. A high-density admixture map for disease
328 gene discovery in african americans. *Am J Hum Genet* 74:1001-1013.
329 Tandon A, Patterson N, and Reich D. 2011. Ancestry informative marker panels for African
330 Americans based on subsets of commercially available SNP arrays. *Genet Epidemiol*
331 35:80-83.
332 Winkler CA, Nelson GW, and Smith MW. 2010. Admixture mapping comes of age. *Annu Rev*
333 *Genomics Hum Genet* 11:65-89.
334 Wright S. 1950. Genetical structure of populations. *Nature* 166:247-249.
335
336

Figure 1

Evaluation analysis of ancestry-informative markers.

(A) The *rpower* distributions for AIMs selected based on SIC and F_{st} . The average *rpower* is 0.85 (sd= 0.06) for both lists. (B) Principal components analysis on the 1000 Genomes Project CEU, YRI and ASW panels (n=85 88, 56 unrelated samples, respectively) using the AIMs detected based on SIC. (C) Comparison of the proportion of variance explained (PVE) by the first PCs derived from the CEU, YRI, and ASW samples. The histogram shows the distribution from 1000 randomly-sampled sets of SNPs according to the number of AIMs (based on SIC) on each chromosome. Circles denote real PVE observations for each panel of AIMs: AIMs selected by SIC (5885 SNPs) and F_{st} (6012 SNPs) from Pan-African array, AIMs selected from Affymetrix SNP 6.0 (4290 SNPs), and Illumina 1M (4285 SNPs), respectively.

