Genome-wide Scan of 29,141 African Americans Finds No Evidence of Directional Selection since Admixture

Gaurav Bhatia,[1,2,*] Arti Tandon,[2,3] Nick Patterson,[2] Melinda C. Aldrich,[4,5,6] Christine B. Ambrosone,[7] Christopher Amos,[8] Elisa V. Bandera,[9] Sonja I. Berndt,[10] Leslie Bernstein,[11] William J. Blot,[4,5,12] Cathryn H. Bock,[13] Neil Caporaso,[10] Graham Casey,[14] Sandra L. Deming,[4,5] W. Ryan Diver,[15] Susan M. Gapstur,[15] Elizabeth M. Gillanders,[16] Curtis C. Harris,[17] Brian E. Henderson,[14] Sue A. Ingles,[14] William Isaacs,[18] Phillip L. De Jager,[2,3,19] Esther M. John,[20,21] Rick A. Kittles,[22] Emma Larkin,[23] Lorna H. McNeill,[24,25] Robert C. Millikan,[26,27,36] Adam Murphy,[28] Christine Neslund-Dudas,[29] Sarah Nyante,[26,27] Michael F. Press,[14] Jorge L. Rodriguez-Gil,[30] Benjamin A. Rybicki,[29] Ann G. Schwartz,[13] Lisa B. Signorello,[4,5,12] Margaret Spitz,[8] Sara S. Strom,[31] Margaret A. Tucker,[10] John K. Wiencke,[32] John S. Witte,[33] Xifeng Wu,[8] Yuko Yamamura,[31] Krista A. Zanetti,[16,17] Wei Zheng,[4,5] Regina G. Ziegler,[10] Stephen J. Chanock,[10] Christopher A. Haiman,[14] David Reich,[2,3,35] and Alkes L. Price[2,34,35]

The extent of recent selection in admixed populations is currently an unresolved question. We scanned the genomes of 29,141 African Americans and failed to find any genome-wide-significant deviations in local ancestry, indicating no evidence of selection influencing ancestry after admixture. A recent analysis of data from 1,890 African Americans reported that there was evidence of selection in African Americans after their ancestors left Africa, both before and after admixture. Selection after admixture was reported on the basis of deviations in local ancestry, and selection before admixture was reported on the basis of allele-frequency differences between African Americans and African populations. The local-ancestry deviations reported by the previous study did not replicate in our very large sample, and we show that such deviations were expected purely by chance, given the number of hypotheses tested. We further show that the previous study's conclusion of selection in African Americans before admixture is also subject to doubt. This is because the F_{ST} statistics they used were inflated and because true signals of unusual allele-frequency differences between African Americans and African populations would be best explained by selection that occurred in Africa prior to migration to the Americas.

Admixed populations, such as African Americans and Latinos, are formed by the mixing of genetically differentiated ancestral populations. Alleles that are highly differentiated between the ancestral populations and advantageous in the admixed population are expected to rise in frequency after admixture, causing local ancestry to deviate from the genome-wide average.[1] These deviations have been interpreted as a signal of the action of natural selection since admixture.[2–4] We note that sampling noise, genetic drift after admixture, and small systematic biases in local-ancestry

[1]Division of Health, Science, and Technology, the Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA 02139, USA; [2]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA; [3]Harvard Medical School, New Research Building, 77 Avenue Louis Pasteur, Boston, MA 02115, USA; [4]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Nashville, TN 37203, USA; [5]Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN 37203, USA; [6]Department of Thoracic Surgery, Vanderbilt University School of Medicine, Nashville, TN 37203, USA; [7]Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY 14263, USA; [8]Section of Biostatistics and Epidemiology, Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Hanover, NH 03766, USA; [9]Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08903, USA; [10]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA; [11]Division of Cancer Etiology, Department of Population Sciences, Beckman Research Institute, City of Hope, CA 91010, USA; [12]International Epidemiology Institute, Rockville, MD 20850, USA; [13]Karmanos Cancer Institute and Department of Oncology, Wayne State University of Medicine, Detroit, MI 48201, USA; [14]Departments of Preventive Medicine and Pathology, Keck School of Medicine, University of Southern California Norris Comprehensive Cancer Center, Los Angeles, CA 90033, USA; [15]Epidemiology Research Program, American Cancer Society, Atlanta, GA 30303, USA; [16]Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD 20892, USA; [17]Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, Bethesda, MD 20892, USA; [18]James Buchanan Brady Urological Institute, Johns Hopkins Hospital and Medical Institutions, Baltimore, MD 21287, USA; [19]Program in Translational NeuroPsychiatric Genomics, Institute for the Neurosciences, Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA; [20]Cancer Prevention Institute of California, Fremont, CA 94538, USA; [21]Stanford Cancer Center, Stanford Medicine, Stanford, CA 94305, USA; [22]Department of Medicine, University of Illinois at Chicago, Chicago, IL 60607, USA; [23]Division of Allergy, Pulmonary, and Critical Care, Department of Medicine, Vanderbilt University Medical Center, 6100 Medical Center East, Nashville, TN 37232-8300, USA; [24]Department of Health Disparities Research, Cancer Prevention and Population Sciences, the University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA; [25]Center for Community Implementation and Dissemination Research, Duncan Family Institute, the University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA; [26]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, USA; [27]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, USA; [28]Department of Urology, Northwestern University, Chicago, IL 60611, USA; [29]Department of Public Health Sciences, Henry Ford Hospital, Detroit, MI 48202, USA; [30]Sylvester Comprehensive Cancer Center and Department of Epidemiology and Public Health, University of Miami Miller School of Medicine, Miami, FL 33136, USA; [31]Department of Epidemiology, the University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA; [32]University of California, San Francisco, San Francisco, CA 94158, USA; [33]Departments of Epidemiology and Biostatistics and Urology, Institute for Human Genetics, University of California, San Francisco, San Francisco, CA 94158, USA; [34]Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA
[35]These authors contributed equally to this work
[36]In memoriam
*Correspondence: gbhatia@mit.edu

http://dx.doi.org/10.1016/j.ajhg.2014.08.011. ©2014 by The American Society of Human Genetics. All rights reserved.

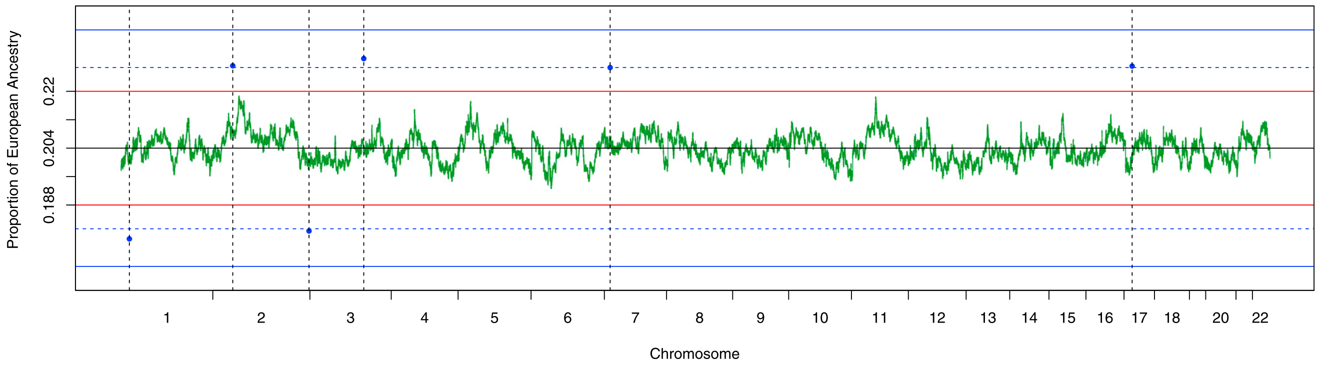

Figure 1. Ancestry at Each Location in the Genome in 29,141 African Americans
This figure gives the proportion of European ancestry at each of the 118,006 SNPs common to all cohorts. The black line indicates the genome-wide average proportion of European ancestry. The red and blue lines indicate the threshold for genome-wide significance ($p < 10^{-5}$) in our study and in the Jin et al. study,[9] respectively. The dashed blue line indicates the significance threshold ($p < 2.7 \times 10^{-3}$) that was actually used in the Jin et al. study.[9] The SD was computed empirically over all SNPs. It is clear that no region attained genome-wide significance in our scan. For the six loci reported under selection in Jin et al.,[9] dashed vertical lines indicate their location, and blue points indicate their deviation in local ancestry. These deviations are reported in relation to the genome-wide average ancestry proportion in our study. None of the six reported loci exceeded the threshold for genome-wide significance ($p < 10^{-5}$) for the Jin et al.[9] study (blue lines).

inference[5,6] will also produce deviations in local ancestry, making it important to account for these factors before concluding that natural selection has occurred.

To better understand deviations in local ancestry as a signal of selection, we simulated the evolution of local ancestry in an admixed population. The population was created seven generations ago with ancestral proportions of 80% and 20% from two ancestral populations to mimic an idealized demographic history of African Americans. We simulated neutral evolution with a variety of effective population sizes (N_e) over seven generations by using a recombination map built from African American data.[7] For each value of N_e, we assessed the variance in local ancestry, minimum detectable selection coefficient, and effective number of statistical tests (see Table S1, available online). Our results suggest that genetic drift can contribute significantly to the variance in average local ancestry (as a function of N_e) and thus reduce power to detect selection. We note that small systematic biases in local-ancestry inference will also contribute to this variance and have a similar effect on power.

In light of these simulated results, we sought to investigate possible recent selection in African Americans. We performed an admixture scan for unusual deviations in local ancestry in 29,141 African Americans from five cohorts from the African American Lung Cancer Consortium (AALCC), African American Breast Cancer Consortium (AABCC), African American Prostate Cancer Consortium (AAPCC), Children's Hospital of Philadelphia, and Candidate Gene Association Resource (CARe). Sample sizes and genotyping arrays are listed in Table S2. We note that the AALCC, AABCC, and AAPCC cohorts consist of disease-affected individuals and control subjects, but phenotype information was not available in the current study. The inclusion of affected individuals could produce false-positive signals of selection as a result of admixture associations with disease but is unlikely to produce false-negative selection signals, which would only occur if admixture association and selection each caused local-ancestry deviations that perfectly negated each other at the same locus.

We filtered the data to remove genotyping artifacts, related individuals, and individuals with very little European or African ancestry (see Table S2). To estimate local ancestry, we used HapMap3 CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) and YRI (Yoruba in Ibadan, Nigeria) haplotypes as ancestral populations in HAPMIX.[8] The average local ancestry at each locus was calculated as an average of the local-ancestry estimates across all samples. Because of issues with ancestry inference at the ends of chromosomes, we removed the first and last 2 Mb of each chromosome from analysis. We note that in these regions, three loci (which do not overlap any previously published loci[9]) did show significant deviations in local ancestry, but these are very likely to be artifacts (see Appendix A). We subsequently focused on local-ancestry estimates for 118,007 SNPs in the intersection of all cohorts. Because of the extent of admixture linkage disequilibrium (LD) in African Americans, the number of markers required to tag the entire genome is approximately 2,000–3,000,[1,10,11] making our use of 118,000 markers sufficient to tag local ancestry genome-wide.

The average proportion of European ancestry over all samples and all SNPs was 0.204 (SD = 0.0036 across SNPs). On the basis of the extent of admixture LD in African Americans,[1,10,12] we defined a genome-wide-significant signal of selection as a local-ancestry deviation greater than 4.42 SDs ($p < 10^{-5}$), corresponding to 5,000 hypotheses tested.[1] We used the empirical SD because the theoretical SD can be affected by genetic drift after admixture, cryptic relatedness, and other factors that are difficult to quantify (see Tables S1 and S3). Figure 1 displays the

Table 1. Comparison of Deviations in Average Local Ancestry

	Jin et al.[9]		Current Study	
Region	**Deviation**	**Nominal p Value**	**Deviation**	**Nominal p Value**
Chr1: 17,409,539–21,604,321	−0.025	7.43×10^{-4}	−0.004	0.55
Chr2: 241,750,403–242,568,618	−0.023	2.07×10^{-3}	−0.006	0.44
Chr2: 37,451,925–37,508,581	0.023	2.16×10^{-3}	0.005	0.51
Chr3: 116,930,811–118,313,302	0.025	8.58×10^{-4}	−0.002	0.83
Chr6: 163,653,158–163,653,428	0.023	2.70×10^{-3}	0.004	0.60
Chr16: 61,214,438–61,242,497	0.023	2.26×10^{-3}	0.006	0.41

We list the six regions reported by Jin et al.[9] to have unusual deviations in local ancestry and compare these to our scan. The deviation is in the proportion of European local ancestry. None of the six regions replicated at nominal significance ($p < 0.05$) in our analysis. All positions are from UCSC Genome Browser build hg18.

average local ancestry at each SNP and indicates no genome-wide-significant deviation in local ancestry. We note that our simulations suggest that the actual effective number of independent hypotheses might be closer to 1,000–1,500 (see Table S1). However, our results remain null even if we correct for only 1,000 hypotheses tested (4.06 SDs; $p < 5 \times 10^{-5}$). Additionally, our results remain null in a smaller sample of 23,000 individuals with more extensive genomic coverage of 461,000 SNPs (see Appendix A).

To better understand the implications of these results, we evaluated the range of selection coefficients that we would have high power to detect. Assuming a normal sampling distribution of observed average local ancestry, we can solve for the true average local ancestry (γ_L) that would constitute a genome-wide-significant signal of selection. In our case, we had 95% power to detect selection at loci where $\gamma_L < 0.183$ or $\gamma_L > 0.225$. Assuming seven generations since admixture,[8] we performed a grid search of possible values of the selection coefficient for local ancestry (s_{anc}) to find those that would produce these values of γ_L and obtained an estimate of 0.019, providing an upper bound on the strength of selection since admixture. Similarly, our simulations with $N_e = 50{,}000$ (whose variance in average local ancestry was similar to the variance observed in real data; see Table S1) also indicated a minimum detectable selection coefficient of approximately 0.019 (see Figure S1). We note that, in general, the selection coefficient per local-ancestry block (s_{anc}) will be lower than the selection coefficient per allele (s), and an s_{anc} of 0.019 could correspond to a large value of s, representing strong selection. The conversion between s and s_{anc} will depend on allele frequencies in European and African populations (see Table S4).

Our results suggest that selection stronger than $s_{anc} > 0.019$ since admixture can be ruled out, and they contrast with a report of six loci as targets of selection after admixture in a recent study by Jin et al.[9] However, that study considered any deviation greater than 3 SDs ($p < 2.7 \times 10^{-3}$), corresponding to only 20 hypotheses tested, to be genome-wide significant. The six loci did not replicate at nominal significance ($p < 0.05$) in our analysis of many more samples (see Figure 1 and Table 1). When we used a threshold of 3 SDs in our data, six loci showed significant deviations. None of these overlap those reported by Jin et al. (see Table S5), suggesting that reported signals of selection after admixture are likely to be false positives because of an insufficient correction for multiple tests. For five of the six loci in Table 1, the deviation that we observed has the same sign as the previously reported deviation. This could be due to statistical chance ($p = 0.11$; one-sided Fisher's exact test), genetic drift after admixture, or small systematic biases in local-ancestry inference (see Table S6). In any case, our results show that the proportion of African ancestry at these six loci was not strongly affected by natural selection since admixture.

Allele-frequency differentiation can be a powerful test for selection.[13–18] Indeed, population differentiation between African Americans and YRI was used as the basis of 14 selection signals recently reported by Jin et al. This was described as a test for selection that occurred after the forced migration of the African ancestors of African Americans (both before and after admixture). Jin et al. ultimately concluded that selection occurred before admixture given the lack of overlap with signals of selection after admixture from deviations in local ancestry.[9] Specifically, single SNPs were ranked by an estimate of F_{ST}, and the most highly differentiated SNPs were reported as signals of selection. These single SNP estimates of F_{ST} were produced with the Weir and Cockerham[19] (WC) F_{ST} estimator. However, a concern with the use of the WC estimator for this application is that estimates can strongly depend on the ratio of sample sizes used. This can potentially result in overestimates of F_{ST} at neutral SNPs,[20] leading to false-positive signals of selection (see Table S7).

On the other hand, the Hudson estimator,[20,21] which is a simple average of the population-specific estimators of Weir and Hill,[22] does not have this bias. We assessed the magnitude of inflation of WC estimates in the loci reported by Jin et al.[9] Their analysis compared African segments of 1,890 African Americans and 113 YRI at SNPs with minor allele frequency (MAF) > 5% and reported a

Table 2. Comparison of Signals of Population Differentiation

			Jin et al.[9] data			Bhatia et al.[23] Data
SNP ID	**Region**	**Gene**	**WC F_{ST}**	**Hudson F_{ST}**	**Model-Based p Value**	**Model-Based p Value**
rs1541044	chr1: 100,125,058–100,183,875	–	0.0562	0.0439[a]	4.7×10^{-5}	0.04
rs4460629	chr1: 153,401,959–153,464,086	–	0.0692	0.0650	6.8×10^{-7}	2.1×10^{-4}
rs12094201	chr1: 236,509,336	–	0.0561	0.0489	1.7×10^{-5}	0.86
rs7642575	chr3: 31,400,165	–	0.0453	0.0393[a]	1.1×10^{-4}	0.41
rs652888	chr6: 26,554,684–33,961,049	*HLA*	0.0711	0.0627	1.1×10^{-6}	1.8×10^{-11}
rs9478984	chr6: 151,555,551–151,569,258	–	0.0545	0.0596	2.1×10^{-6}	0.02
rs10499542	chr7: 22,235,870	–	0.0461	0.0453	3.6×10^{-5}	0.35
rs304735	chr7: 79,768,487–80,482,597	*CD36*	0.0946	0.0690	3.0×10^{-7}	3.7×10^{-13}
rs2920283	chr8: 143,754,039–143,758,933	*PSCA*	0.0468	0.0532	7.6×10^{-6}	6.4×10^{-7}
rs1498487	chr11: 5,034,229–5,421,456	*HBB*	0.0617	0.0464	2.4×10^{-5}	1.7×10^{-7}
rs4883422	chr12: 7,189,594	–	0.0472	0.0461	3.0×10^{-5}	1.3×10^{-3}
rs6491096	chr13: 25,488,362	–	0.0472	0.0373[a]	1.5×10^{-4}	0.4
rs1075875	chr16: 47,595,721	–	0.0766	0.0608	1.3×10^{-6}	NA[b]
rs6015945	chr20: 59,319,574	–	0.0627	0.0550	4.3×10^{-6}	0.5

We recreated Table 2 from Jin et al.[9] by analyzing the same data with the Hudson instead of the WC estimator. We also estimated the p value at each SNP by using the reported $F_{ST} = 0.0007$ of Jin et al.[9] and a model-based approach.[24] Finally, we report the model-based p value of the most significant SNP in the region from the parallel study by Bhatia et al.[23] We note that results reported in that paper were more significant than those reported here because Bhatia et al. analyzed additional populations. All positions are from UCSC Genome Browser build hg18.

[a] These loci fell below the threshold for the 99.99th percentile (0.0452) when the Hudson estimator was used.

[b] This locus was not available (NA) because it lacked data in the Bhatia et al.[23] study.

total of 40 SNPs—the 99.99th percentile of 401,559 SNPs tested—clustered into 14 loci that had $F_{ST} > 0.0452$. Ten of these loci were previously unreported targets of natural selection, and four were reported as genome-wide significant in the parallel study of Bhatia et al.[23] (or nearly genome-wide significant in the case of *HBB*, a previously identified target of selection[24]). Of the ten novel signals, nine produced lower estimates when we used the Hudson estimator, and three fell below the Jin et al.[9] threshold ($F_{ST} > 0.0452$; see Table 2). We note that the 99.99th percentile of F_{ST} could change as a result of the switch from the WC estimator to the Hudson estimator; however, our analyses indicated that the magnitude of this change would be smaller than the decreases observed at most of the ten reported novel loci (see Appendix A), suggesting that inflated WC F_{ST} estimates might lead to false-positive signals of selection.

In addition to having issues with F_{ST} estimation, studies that simply rank the most highly differentiated SNPs between populations are unable to evaluate genome-wide significance of reported signals. On the other hand, model-based approaches[23–26] can formally assess genome-wide significance. In general, studies that use a model-based approach are well powered if sample sizes are much larger than 1 / F_{ST},[23] given that both F_{ST} and sampling noise contribute to normal variation in allele-frequency differences. In the Jin et al.[9] comparison, the sample size of YRI ($n = 113$) is much smaller than the reciprocal of F_{ST} between African Americans and YRI (1 / F_{ST} = 1,429). When re-evaluated with a model-based approach,[23,24] none of the reported SNPs achieved genome-wide significance ($p < 5 \times 10^{-8}$; see Table 2). We note that model-based approaches do require robust estimates of F_{ST}, but these are easily available from even small samples of genome-wide data. We re-examined the statistical significance of the ten novel loci reported by Jin et al.[9] in the separate data set of Bhatia et al.,[23] which included 6,209 African Americans and 756 YRI. The Bhatia et al.[23] data include nine of these ten loci, and only four of the nine loci were nominally significant ($p < 0.05$ without correction for multiple-hypothesis testing; see Table 2). Extending the analysis to all 29,141 African Americans in the current study yielded very similar results, given that the YRI sample size was the limiting factor (see Table S8). We caution that the four nominally significant loci should not be viewed as being independently replicated because genetic drift is common to both analyses such that loci in the tail of one analysis could be expected to lie in the tail of the other analysis. The lack of nominal significance at most loci in the non-independent analysis of Bhatia et al.[23] data suggests that most of the reported novel loci are false positives. We note that the results of Jin et al. and Bhatia et al. were both corrected for European admixture either locally[9] or genome-wide.[23] Our analyses (see Table S8) agree with prior results that correction for European admixture is imperative[27] and found that both corrections perform similarly in terms of power.

It is important to recognize that even robust, genome-wide-significant evidence of unusual population differentiation (e.g., at the four loci identified by both Bhatia et al.[23] and Jin et al.[9]) does not imply selection following the forced migration from Africa. The observed population differences at these loci are best explained by selection within Africa. As an example, we consider the well-studied sickle-cell variant rs334 at the *HBB* locus, where biological evidence suggests that some selection since the arrival of Africans in the Americas is likely to have occurred. Homozygotes for the recessive allele are afflicted with sickle-cell anemia, a debilitating condition that results in very low fertility. However, the minor allele at rs334 is maintained at high frequency in Africa because heterozygotes have increased malaria resistance.[28] The MAF at rs334 in African Americans is 0.050,[29] corresponding to an allele frequency of 0.063 (0.050/0.8) on African segments. Conservatively assuming the strongest possible negative selection against the minor allele, we calculate that the maximum allele-frequency difference due to selection post-Africa (after the African ancestors of African Americans migrated from Africa) would be 0.034 (see Appendix A). However, an allele-frequency difference of 0.20 at the *HBB* locus was reported between Nigerians and Gambians,[23] indicative of larger allele-frequency differences due to selection in Africa. Although these populations have a higher level of differentiation ($F_{ST} = 0.006$) than our comparison of African Americans and Nigerians ($F_{ST} = 0.001$), we note that allele-frequency differences at *HBB* are generally related to malaria endemicity and altitude as opposed to F_{ST} between the populations.[24] Thus, we believe that selection in Africa rather than post-Africa is the most likely explanation for most of the observed frequency differences between African Americans and YRI.

Overall, we conclude that there is no locus with genome-wide-significant evidence of selection influencing ancestry in African Americans after their ancestors left Africa and that genome-wide-significant evidence of population differentiation is likely to be best explained by selection in Africa. In addition, we place an upper bound on the selection that could have occurred after admixture and not be detected in our data ($s_{anc} > 0.019$). Although strong selection after admixture can be ruled out by our data, weak selection after admixture might have occurred, for example, at the *HBB* locus. Although our results contrast with previous reports[9] of selection post-Africa, this discrepancy can be explained by insufficient correction for multiple tests, usage of the WC F_{ST} estimator instead of the Hudson estimator, and the action of natural selection in Africa.

Several recent studies have investigated unusual deviations in local ancestry as a possible signal of natural selection in admixed populations. Bryc et al.[2] analyzed 365 African Americans and reported three loci with >3 SDs but correctly noted that these differences were not significant after correction for multiple tests. Jeong et al.[3] analyzed 96 Tibetan individuals (derived from admixture of Han- and Sherpa-related populations thousands of years ago) and focused on genes associated with hemoglobin levels (*EGLN1* and *EPAS1*); they found that the observed deviations (3.59 SDs and 3.74 SDs, respectively) at these candidate loci were statistically significant after correction for multiple tests. A recent study[4] used a new method of local-ancestry inference and reported three loci (including two in the *HLA* region) with very large (>20%) deviations in local ancestry in 58 Mexican (MXL) samples, but these very large deviations were not observed in consensus MXL local-ancestry calls[5,8,30,31] published by the 1000 Genomes Consortium[32] (see Table S9). Finally, recent studies[33–35] have demonstrated evidence of selection since ancient admixture with archaic human populations.

Although a number of alternate methods of detecting selection exist,[36–40] we have focused here on deviations in local ancestry and on population differentiation. We conclude with four recommendations for future studies utilizing these approaches. First, studies reporting selection since admixture on the basis of deviations in local ancestry in African Americans (or in other admixed populations with similar ages of admixture) should employ a genome-wide-significance threshold of $p < 10^{-5}$. Second, studies reporting selection on the basis of deviations in local ancestry should be cognizant of the possibility that errors in local-ancestry inference can lead to false-positive signals[1] and that reports of selection might need to be confirmed by multiple methods. Third, studies reporting selection on the basis of population differentiation and involving unequal sample sizes should not use the WC F_{ST} estimator,[19] which is susceptible to bias in this case, and instead should use the Hudson estimator.[20–22] Fourth, genome-wide significance should not be assessed on the basis of a simple ranking and instead should be assessed via robust model-based approaches.[23–26,41]

Appendix A

Systematic Deviations in Average Local Ancestry at the Ends of Chromosomes

In the analysis presented in the main text, we removed the first and last 2 Mb of each chromosome because of observed systematic deviations in these regions. When we included all available data, we did observe significant peaks in ancestry (Figure S2). These peaks resided in the first 2 Mb of chromosomes 1 and 7 and the last 2 Mb of chromosome 9. Strong evidence that these peaks were the result of inaccurate local-ancestry inference in these loci was based on (1) a high degree of heterogeneity in inferred local ancestry across cohorts (see Figure S3)—the cohorts showing significant deviations were all genotyped on the same platform (see Table S10)—and (2) unexpected reduction in the length of local-ancestry segments (measured in cM) (see Figure S4). Because of this evidence, we removed the first and last 2 Mb of each chromosome.

Impact of Number of SNPs Analyzed

To test the effect of using a relatively small set of 118,000 SNPs, we excluded the 6,000 CARe individuals who were genotyped on the Affymetrix 6.0 chip. The remaining 22,900 individuals were all genotyped on 461,000 SNPs. In this data set, which had >4-fold denser coverage, we observed no genome-wide-significant deviations in average local ancestry (maximum deviation = 3.76 SDs). This null result is consistent with our result in the full data set and with the extent of admixture LD in African Americans. Because of this admixture LD, 2,000–3,000 markers are sufficient to tag local ancestry in analyses of natural selection since admixture.[1,10,11]

Changes in Estimator Alter the 99.99th Percentile

Use of the Hudson F_{ST} estimator instead of the WC estimator results in lower estimates of F_{ST} at the loci reported by Jin et al.[9] However, it is possible that the threshold at the 99.99th percentile is also lowered by use of this estimator and that reported loci still fall at this upper tail of the distribution. To assess this effect in sample sizes similar to those of Jin et al.[9] we subsampled 2,500 African American individuals from our data, subtracted European allele frequencies from CEU,[23] and compared the result to YRI by using both the WC and Hudson F_{ST} at every SNP. According to this analysis, the 99.99th percentile of F_{ST} was 0.048 for the WC estimator and 0.046 for the Hudson estimator.

Jin et al.[9] reported a threshold of 0.0452. Even if this decreases by 0.002 as a result of using the Hudson estimator, the mean difference between the WC and Hudson F_{ST} estimates at the ten novel loci would be 0.006, and 2 of the 14 reported loci would no longer be in the 99.99th percentile (with F_{ST} estimates of 0.037 and 0.039; see Table 2).

Model of Selection at *HBB*

We assume the strongest possible negative selection against the minor allele at *HBB*, that heterozygotes have no advantage (because of much lower rates of malaria in the Americas), and that no people with sickle-cell anemia have children. From this information, we can work backward in time with the following equation:

$$p_{g+1} = \frac{p_g}{1 - p_g}, \qquad \text{Equation A1}$$

where p_g represents the sickle-cell allele g generations in the past. Assuming that $p_0 = 0.0625$[29] and that seven generations have passed since the admixture of the African and European ancestors of African Americans,[8] we have $p_1 = 0.0962$. Thus, the allele frequency in the African ancestors of African Americans seven generations ago would have been 0.096, and the maximum allele-frequency difference due to selection since the migration from Africa would have been 0.034.

Under this model, the per-allele selection coefficient is simply the allele frequency in the population—not on African segments alone—at the current generation ($s^g = \gamma p_g$, where γ is the proportion of African ancestry at the *HBB* locus during the current generation). If we assume that the proportion of local ancestry at each locus seven generations ago is equivalent to the current genome-wide average, the maximum value of this coefficient is $s = 0.796(p_7) = 0.077$. The selection coefficient per copy of African local ancestry is given by $s_{anc} = \gamma(p)^2$. That is, given that an individual carries one African chromosome at the *HBB* locus, he must also carry (1) the sickle-cell allele on this first African chromosome (with probability p), (2) a second African chromosome at this locus (with probability γ), and (3) the sickle-cell allele on that second African chromosome (with probability p). According to our model, the maximum value of this coefficient is $s_{anc} = 0.796(p_7)^2 = 0.0074$. We also explored the effect of weak negative selection against heterozygotes (h) for the sickle-cell allele on both local ancestry and allele-frequency changes following admixture. Our results suggest that only very strong negative selection against heterozygotes ($h > 0.05$) would produce a genome-wide-significant deviation in average local ancestry, whereas allele frequencies would be affected at smaller values of h (see Table S11).

Supplemental Data

Supplemental Data include 5 figures and 11 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2014.08.011.

Acknowledgments

We thank Carlos Bustamante, Bogdan Pasaniuc, and Amy L. Williams for helpful discussions and Richard Cooper for sharing data from 756 Yoruba samples. This research was funded by NIH grant R01 HG006399.

Received: May 2, 2014
Accepted: August 22, 2014
Published: September 18, 2014

References

1. Seldin, M.F., Pasaniuc, B., and Price, A.L. (2011). New approaches to disease mapping in admixed populations. Nat. Rev. Genet. *12*, 523–528.
2. Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S.A., and Bustamante, C.D. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. USA *107*, 786–791.
3. Jeong, C., Alkorta-Aranburu, G., Basnyat, B., Neupane, M., Witonsky, D.B., Pritchard, J.K., Beall, C.M., and Di Rienzo, A. (2014). Admixture facilitates genetic adaptations to high altitude in Tibet. Nat. Commun. *5*, 3281.
4. Guan, Y. (2014). Detecting structure of haplotypes and local ancestry. Genetics *196*, 625–642.
5. Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R.,

Ford, J.G., Avila, P.C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. Bioinformatics *28*, 1359–1367.

6. Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Zaitlen, N., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., et al. (2013). Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. Bioinformatics *29*, 1407–1415.
7. Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., et al. (2011). The landscape of recombination in African Americans. Nature *476*, 170–175.
8. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. *5*, e1000519.
9. Jin, W., Xu, S., Wang, H., Yu, Y., Shen, Y., Wu, B., and Jin, L. (2012). Genome-wide detection of natural selection in African Americans pre- and post-admixture. Genome Res. *22*, 519–527.
10. Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., O’Brien, S.J., Altshuler, D., et al. (2004). Methods for high-density admixture mapping of disease genes. Am. J. Hum. Genet. *74*, 979–1000.
11. Smith, M.W., and O’Brien, S.J. (2005). Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat. Rev. Genet. *6*, 623–632.
12. Smith, M.W., Patterson, N., Lautenberger, J.A., Truelove, A.L., McDonald, G.J., Waliszewska, A., Kessing, B.D., Malasky, M.J., Scafe, C., Le, E., et al. (2004). A high-density admixture map for disease gene discovery in african americans. Am. J. Hum. Genet. *74*, 1001–1013.
13. Akey, J.M., Zhang, G., Zhang, K., Jin, L., and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome Res. *12*, 1805–1814.
14. McEvoy, B.P., Montgomery, G.W., McRae, A.F., Ripatti, S., Perola, M., Spector, T.D., Cherkas, L., Ahmadi, K.R., Boomsma, D., Willemsen, G., et al. (2009). Geographical structure and differential natural selection among North European populations. Genome Res. *19*, 804–814.
15. Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., and Pritchard, J.K. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.
16. Teo, Y.-Y., Sim, X., Ong, R.T.H., Tan, A.K.S., Chen, J., Tantoso, E., Small, K.S., Ku, C.-S., Lee, E.J.D., Seielstad, M., and Chia, K.S. (2009). Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res. *19*, 2154–2162.
17. Akey, J.M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. *19*, 711–722.
18. Engelken, J., Carnero-Montoro, E., Pybus, M., Andrews, G.K., Lalueza-Fox, C., Comas, D., Sekler, I., de la Rasilla, M., Rosas, A., Stoneking, M., et al. (2014). Extreme population differences in the human zinc transporter ZIP4 (SLC39A4) are explained by positive selection in Sub-Saharan Africa. PLoS Genet. *10*, e1004128.
19. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. Evolution *38*, 1358–1370.
20. Bhatia, G., Patterson, N., Sankararaman, S., and Price, A.L. (2013). Estimating and interpreting FST: the impact of rare variants. Genome Res. *23*, 1514–1521.
21. Hudson, R.R., Slatkin, M., and Maddison, W.P. (1992). Estimation of levels of gene flow from DNA sequence data. Genetics *132*, 583–589.
22. Weir, B.S., and Hill, W.G. (2002). Estimating F-statistics. Annu. Rev. Genet. *36*, 721–750.
23. Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., et al. (2011). Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. Am. J. Hum. Genet. *89*, 368–381.
24. Ayodo, G., Price, A.L., Keinan, A., Ajwang, A., Otieno, M.F., Orago, A.S.S., Patterson, N., and Reich, D. (2007). Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. Am. J. Hum. Genet. *81*, 234–242.
25. Lewontin, R.C., and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics *74*, 175–195.
26. Price, A.L., Helgason, A., Palsson, S., Stefansson, H., St Clair, D., Andreassen, O.A., Reich, D., Kong, A., and Stefansson, K. (2009). The impact of divergence time on the nature of population structure: an example from Iceland. PLoS Genet. *5*, e1000505.
27. Huerta-Sánchez, E., Degiorgio, M., Pagani, L., Tarekegn, A., Ekong, R., Antao, T., Cardona, A., Montgomery, H.E., Cavalleri, G.L., Robbins, P.A., et al. (2013). Genetic signatures reveal high-altitude adaptation in a set of ethiopian populations. Mol. Biol. Evol. *30*, 1877–1888.
28. Aidoo, M., Terlouw, D.J., Kolczak, M.S., McElroy, P.D., ter Kuile, F.O., Kariuki, S., Nahlen, B.L., Lal, A.A., and Udhayakumar, V. (2002). Protective effects of the sickle cell gene against malaria morbidity and mortality. Lancet *359*, 1311–1312.
29. Auer, P.L., Johnsen, J.M., Johnson, A.D., Logsdon, B.A., Lange, L.A., Nalls, M.A., Zhang, G., Franceschini, N., Fox, K., Lange, E.M., et al. (2012). Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. Am. J. Hum. Genet. *91*, 794–808.
30. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
31. Churchhouse, C., and Marchini, J. (2013). Multiway admixture deconvolution using phased or unphased ancestral panels. Genet. Epidemiol. *37*, 1–12.
32. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
33. Vernot, B., and Akey, J.M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. Science *343*, 1017–1021.
34. Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The

genomic landscape of Neanderthal ancestry in present-day humans. Nature *507*, 354–357.

35. Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B.M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., et al. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature *512*, 194–197.
36. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. Science *312*, 1614–1620.
37. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol. *4*, e72.
38. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al.; International HapMap Consortium (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.
39. Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. Genome Res. *20*, 393–402.
40. Peter, B.M., Huerta-Sanchez, E., and Nielsen, R. (2012). Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genet. *8*, e1003011.
41. Grossman, S.R., Shlyakhter, I., Karlsson, E.K., Byrne, E.H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. Science *327*, 883–886.