Estimating Evolutionary Parameters and Detecting
Signals of Natural Selection from Genetic Data

by

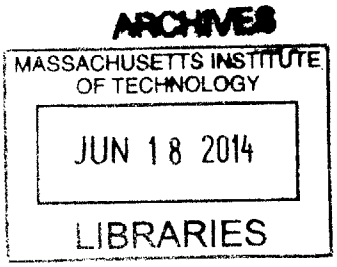
Gaurav Bhatia

M.S. Computer Science
University of California, San Diego (2009)

SUBMITTED TO THE HARVARD-MIT DIVISION OF HEALTH SCIENCE AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014



Signature redacted

Signature of Author.....
Harvard-MIT Division of Health Sciences and Technology
May 15, 2014

Signature redacted

Certified by.....
Alkes L. Price
Assistant Professor of Biostatistics and Epidemiology
Thesis Supervisor

Signature redacted

Accepted by.....
Emery N. Brown, MD, PhD
Director, Harvard-MIT Program in Health Sciences and Technology
Professor of Computational Neuroscience and Health Sciences and Technology

Estimating Evolutionary Parameters and Detecting
Signals of Natural Selection from Genetic Data

by

Gaurav Bhatia

Submitted to the Harvard-MIT Division of Health Science and Technology ON May 8, 2014 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy.

Abstract

Even prior to the elucidation of the structure of DNA, the theoretical foundations of population genetics had been well developed. Advances made by Sewall Wright, John B.S. Haldane, and Ronald A. Fisher form the basis with which we understand the statistical dynamics of evolution and inheritance. Using this foundation, recent advances in DNA profiling technologies have enabled genome-wide analysis of thousands of individuals from a diverse array of human populations. These new analyses can answer fundamental questions about human population differences, natural selection, and admixture. However, with this deluge of newly available data, confusion about statistical methods may lead to misleading conclusions about human population history and natural selection. We view it as imperative to put analyses of population differences on sound statistical footing.

In the course of this thesis, we have developed methods and reanalyzed existing results in two related areas: the detection of natural selection and estimation of genetic distance. Throughout our work, we have strived for statistical rigor, attempting to understand variation in previously reported results and provide a resource for other researchers in our field. Where necessary, we have made simplifying assumptions about evolutionary processes but have attempted to state these clearly and validate their reasonableness using simulations. Our efforts have culminated in three projects that will be described in the subsequent chapters: (1) A model based approach to detect natural selection in 3 populations (2) A protocol to generate consistent estimates of F_{ST} and, (3) Reanalysis of previously reports of selection in African Americans since the arrival of their ancestors in the Americas.

We note that our work is just part of a rich literature on population and evolutionary genetics. We have attempted to cite this literature in detail and have published our own methods to enable others to utilize and improve upon them.

Acknowledgements

I am very grateful to my advisor, Alkes Price, for his thoughtful guidance and support throughout this process. I feel incredibly lucky to have interacted with Alkes. The clarity of his scientific insights is now my standard for thinking about difficult scientific questions. His energetic approach to science has provided an example that I will attempt to emulate throughout my career. In addition to this individual awesomeness, Alkes has given me the opportunity to interact with truly brilliant thinkers.

Nick Patterson, who co-led our work on F_{ST} (Chapter 2), and David Reich, who is the co-senior author on our work reanalyzing previously reported signals of selection (Chapter 4), were instrumental in the completion of this work. Together with Shamil Sunyaev, these four individuals have provided me with invaluable scientific guidance as my thesis committee.

In addition, former and current lab members Bogdan, Noah, Sasha, Bjarni, Po-Ru, Pier, Chia-Yen, Tristan, and Hilary have been great collaborators. Countless interesting discussions have occurred because of their intellect and curiosity.

To my family, this is one of many experiences that I would never have had if not for your sacrifice, love and support. I know that I owe you too much to ever repay.

Table of Contents

Chapter 1: Introduction

From the 2001 cost of ~$3 billion, the cost of generating a single human genome sequence has decreased by nearly a factor of 1 million [1]. This decrease has been mirrored to a lesser extent in genotyping technologies [2], enabling private companies [3,4] to offer genome-wide genotyping direct-to-consumer at increasingly reduced cost. At the point of this writing, 23andme and AncestryDNA were offering a genome-wide profile for $99. As a point of a comparison, genetic testing of a single locus or a small number of loci in the medical context can cost "more than $2000" [5]. Though this difference is partially explained by different technologies and tolerance for errors, the revolutionary potential of ultra low-cost DNA profiling is clear.

Scientifically, the revolutionary impact of these technologies has already been felt. Since the release of a human genome reference [6], analyses of large data-sets of human genetic variation have enabled numerous large scale genome-wide association studies [7-9] that have identified thousands of genetic associations with common phenotypes [10]. These associations have provided some insight into the genetic architecture of human disease [11,12], while raising new questions [13].

Beyond medical genetics, the availability of whole genome data from large numbers of individuals has enabled testing of human population genetic hypotheses as never before. Large consortia [14-19] have made data from a diverse array of populations available as a resource for population geneticists. These data and others have been used to detect natural selection [20-25], and answer questions about human demographic history [26-30]. Indeed, these data present an orthogonal line of evidence for comparison to the historical, archaeological and linguistic record [28,31].

In this thesis, we focus on analysis of genome-wide data for the purpose of assessing genetic distance between populations and detecting natural selection. Our goal is to explain discrepancies in previous reports and contribute to a methodological basis for future studies. A particular focus of this work is the analysis of allele frequency differences between different populations. These differences, quantified by F_{ST}, are the result of neutral genetic drift as well as natural selection and can give insight into both. F_{ST} is a quantity originally described by Sewall Wright (1949) and Gustave Malécot (1948), which is used to quantify the genetic distance between pairs of populations. Specifically, F_{ST} is the "correlation between random gametes, drawn from the same subpopulation, relative to the total" [32].

Unfortunately, in this original description, Wright did not clearly define the "total" population, or give a means of estimating F_{ST} from observed data. Subsequent authors have attempted to clarify this[33-37], though there has been significant disagreement about both topics. Addition to the confusion, F_{ST} has been related to a large number of additional quantities, such as divergence time [38], coalescent time [39], migration rates [32], and heterozygosity [33,40]. Beyond clarifying F_{ST}'s definition and estimation, its value as a metric of differentiation has also been the subject of recent debate [41,42]. Considering all of this, the fact that published estimates of F_{ST} from genotype data [16] were nearly double those from sequence data [17] was difficult to evaluate. In Chapter 2 of this thesis and our related publication [43], we show that the difference between these estimates is largely artifactual and provide a protocol that can be used to produce consistent estimates of F_{ST} across studies.

In addition to its use in quantifying genetic distances, F_{ST} can be used to calibrate selection statistics. Specifically, under certain assumptions, F_{ST} functions as a parameter in a statistical model of genetic drift. To detect selection, this as used as a null model of allele frequency differences between populations[44-47] and violations of this model represent likely targets of selection. We explored this approach in the context three

populations with majority West African ancestry: African Americans, Gambians, and Nigerians. In addition to testing all pairs of populations for evidence of selection, we developed an approach to reconstruct a tree of the three populations and test for selection using the reconstructed tree as a null model. This approach may increase power to detect selection and resolution of the specific population that is subject to selection. We used these approaches to corroborate previously published targets of selection [46,48-50] and provide evidence for a novel target of selection. This analysis is described in Chapter 3.

In Chapter 4, we reanalyze a previous report of natural selection in African Americans since the arrival of their ancestors in the Americas [51]. Given the small number of generations since admixture between the African ancestors of African Americans and Europeans [52], any detected selection would have to be extremely strong. This suggestion garnered much attention, including a *New York Times* article suggesting, "the harsh new world... apparently brought genetic change". Considering this high profile, and that previous reports of such recent natural selection[53] may have been false positives [54], we sought to examine the claims of Jin and colleagues [51] in detail. Building on our work in Chapter 2, we show that estimates of allele frequency differences (i.e. F_{ST}) are inflated due to artifacts of the estimation method. In addition, any selection that did occur is more parsimoniously explained by selection in Africa, as opposed to selection in the Americas. The second line of evidence used by Jin and colleagues is an excess of African or European ancestry at loci under selection. However, we show that reported loci do not achieve an appropriate genome-wide significance threshold [55] and, thus, are likely to be false positives. Overall, we conclude that no study has provided evidence of selection since admixture in African Americans.

Finally, in Chapter 5, we summarize these findings and discuss future directions.

References

1. National Human Genome Research Institute DNA Sequencing Costs. Genome.Gov.

2. Manolio, T.A. Genome-Wide Association (GWA) Studies. Google.com.

3. 23andMe - Genetic Testing for Ancestry; DNA Test. 23andme.com.

4. DNA Tests for Ethnicity & Genealogical DNA testing at AncestryDNA. Ancestry.com.

5. Genetics Home Reference, N.L.O.M. What is the cost of genetic testing, and how long does it take to get the results? Ghr.Nlm.Nih.Gov.

6. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., and FitzHugh, W. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

7. Lango-Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.

8. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.

9. International Schizophrenia Consortium, Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.

10. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc.Natl.Acad.Sci.U.S.a. *106*, 9362–9367.

11. Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. Trends Genet. *17*, 502–510.

12. Wang, W.Y.S., Barratt, B.J., Clayton, D.G., and Todd, J.A. (2005). Genome-wide association studies: theoretical and practical concerns. Nature Reviews. Genetics *6*, 109–118.

13. Maher, B. (2008). Personal genomes: The case of the missing heritability. Nature *456*, 18–21.

14. International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.

15. International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

16. International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

17. 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

18. 1000 Genomes Project Consortium, 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. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

19. Cavalli-Sforza, L.L. (2005). The Human Genome Diversity Project: past, present and future. Nat.Rev.Genet. *6*, 333–340.

20. 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.

21. Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.

22. Akey, J.M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. *19*, 711–722.

23. 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., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.

24. 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.

25. 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. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.

26. Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. (2009). Reconstructing Indian population history. Nature *461*, 489–494.

27. Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science *336*, 740–743.

28. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., Mesa, N., et al. (2012). Reconstructing Native American population history. Nature *488*, 370–374.

29. Pickrell, J.K., and Pritchard, J.K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet *8*, e1002967.

30. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S.,

Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.

31. Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1994). The History and Geography of Human Genes.

32. Wright, S. (1949). THE GENETICAL STRUCTURE OF POPULATIONS. Ann. Hum. Genet. *15*, 323–354.

33. Nei, M. (1973). Analysis of Gene Diversity in Subdivided Populations. Proceedings of the National Academy of Sciences *70*, 3321–3323.

34. Nei, M. (1986). Definition and Estimation of Fixation Indices. Evolution *40*, 643–645.

35. Cockerham, C.C. (1969). Variance of Gene Frequencies. Evolution *23*, 72–84.

36. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. Evolution *38*, 1358–1370.

37. Weir, B.S., and Hill, W.G. (2002). Estimating F-statistics. Annu. Rev. Genet. *36*, 721–750.

38. Cavalli-Sforza, L.L., and Bodmer, W.F. (1971). The genetics of human populations (San Francisco: W.H. Freeman).

39. Slatkin, M. (1991). Inbreeding coefficients and coalescence times. Genetics Research *58*, 167.

40. Hudson, R.R., Slatkin, M., and Maddison, W.P. (1992). Estimation of levels of gene flow from DNA sequence data. Genetics *132*, 583–589.

41. Jost, L. (2008). GST and its relatives do not measure differentiation. Mol. Ecol. *17*, 4015–4026.

42. Ryman, N., and Leimar, O. (2009). G(ST) is still a useful measure of genetic differentiation - a comment on Jost's D. Mol. Ecol. *18*, 2084–7–discussion2088–91.

43. Bhatia, G., Patterson, N., Sankararaman, S., and Price, A.L. (2013). Estimating and interpreting FST: The impact of rare variants. Genome Res.

44. 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.

45. Nicholson, G., Smith, A.V., Jonsson, F., Gústafsson, Ó., Stefansson, K., and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *64*, 695–715.

46. 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.

47. 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.

48. Hedrick, P.W., and Thomson, G. (1983). Evidence for balancing selection at HLA. Genetics *104*, 449–456.

49. Fry, A.E., Ghansa, A., Small, K.S., Palma, A., Auburn, S., Diakite, M., Green, A., Campino, S., Teo, Y.Y., Clark, T.G., et al. (2009). Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. Human Molecular Genetics *18*, 2683–2692.

50. Kwiatkowski, D.P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. Am. J. Hum. Genet. *77*, 171–192.

51. 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.

52. 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.

53. Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E.G., and Risch, N.J. (2007). Recent genetic selection in the ancestral admixture of Puerto Ricans. Am. J. Hum. Genet. *81*, 626–633.

54. Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., Taylor, K.D., et al. (2008). Long-Range LD Can Confound Genome Scans in Admixed Populations. The American Journal of Human Genetics *83*, 132–135.

55. Seldin, M.F., Pasaniuc, B., and Price, A.L. (2011). New approaches to disease mapping in admixed populations. Nature Reviews. Genetics *12*, 523–528.

Chapter 2: Estimating and Interpreting F_{ST}: The Impact of Rare Variants

In a pair of seminal papers, Sewall Wright and Gustave Malécot introduced F_{ST} as a measure of structure in natural populations. In the decades that followed, a number of papers provided differing definitions, estimation methods, and interpretations beyond Wright's. While this diversity in methods has enabled many studies in genetics, it has also introduced confusion about how to estimate F_{ST} from available data.

Considering this confusion, wide variation in published estimates of F_{ST} for pairs of HapMap populations is a cause for concern. These estimates changed—in some cases more than two-fold—when comparing estimates from genotyping arrays to those from sequence data [1,2]. Indeed, changes in F_{ST} from sequencing data might be expected data due to population genetic factors affecting rare variants. While rare variants do influence the result, we show that this is largely through differences in estimation methods. Correcting for this yields estimates of F_{ST} that are much more concordant between sequence and genotype data.

These differences relate to three specific issues: (1) estimating F_{ST} for a single SNP, (2) combining estimates of F_{ST} across multiple SNPs, and (3) selecting the set of SNPs used in the computation. Changes in each of these aspects of estimation may result in F_{ST} estimates that are highly divergent from one another. Here, we clarify these issues and propose solutions.

Introduction

Since its introduction by Sewall Wright and Gustave Malécot [3,4], F_{ST} estimation [5,6] has become a key component of studies of population structure in humans [1,2,7,8] and other species [3,4,9-12]. Additionally, F_{ST} is a mainstay of much of the literature on detecting directional natural selection [13-22]. Despite a recent debate about the utility of F_{ST} and related measures [23,24], F_{ST} continues to be widely used by population geneticists [25-27].

Despite its widespread use in genetic studies, confusion remains about what F_{ST} is and how to estimate it. Beyond Wright's original description of F_{ST} as a ratio of variances, F_{ST} has been conceptually defined in many ways [4,28-32]. Additionally, multiple estimators for F_{ST} have been described in the literature [5,29,30,33-35], often making the correct choice of estimator unclear.

With this diversity of definition and estimation in mind, we consider estimates of F_{ST} published by the 1000 Genomes Consortium [1] of 0.052 for European and East Asian populations and 0.071 for European and West African populations. These are less than half of the published estimates, 0.111 and 0.156, from HapMap3 data [2] and may be the result of demography that differentially impacts F_{ST} at rare variants. These estimates have been subsequently used to simulate properties of recent rare variants [36] making it imperative to know if this reduction in F_{ST} is a meaningful result of the inclusion of rare variants or merely an artifact of estimation.

To answer these questions, we examine the issues surrounding F_{ST} estimated on data containing rare variants. We focus our attention on F_{ST} estimation in the context of comparing two populations—potentially with differing amounts of drift since the populations split—using a series of bi-allelic SNPs. Considering this general scenario, we employ the definition of [35] which allows for population-specific F_{ST}. Using this definition, we divide the issues surrounding estimation into three categories and examine them using both simulated and 1000 Genomes data:

(1) Estimating F_{ST} for a single SNP. For two populations with different population-specific F_{ST}, we are interested in estimating the average of these population-specific F_{ST}'s. We examine several estimators in the limit of very large sample sizes. For different estimators this limit can vary with changes in sample sizes, potentially leading to inconsistency across studies, or consistently overestimate F_{ST}. We suggest an estimator that is shown via simulation and application to empirical data to avoid these issues.

(2) Combining estimates of F_{ST} across multiple SNPs. We show that an average of single-SNP estimates of F_{ST} can result in a large reduction in the genome-wide estimate of F_{ST}, particularly in data sets that include rare variants. We show how to compute an average that avoids this problem.

(3) Dependence of F_{ST} on the set of SNPs analyzed. We explore the effects of SNP ascertainment schemes [33] and demonstrate the effects of demographic events on rare and common variants. We show that the underlying F_{ST} parameter *is a function of the set of SNPs analyzed*, in addition to the populations studied.

We conclude that F_{ST} estimates reported by the 1000 Genomes Consortium [1] are a consequence of the estimation method that was applied and are not informative for human demographic history. Correcting for differences in estimation method yields F_{ST} estimates of 0.106 for Europeans and East Asians and 0.139 for Europeans and West Africans. Despite a large increase relative to the results reported by the 1000 Genomes Consortium, these remain slightly reduced relative to the corresponding estimates from HapMap3. The small reduction in F_{ST} relative to HapMap3 is due not to the inclusion of rare variants, but to the ascertainment of SNPs in HapMap3 that excluded less differentiated common variants. In fact, when ascertaining within one of the two populations studied, rare variants actually have *higher* F_{ST} estimates than common

variants. This is consistent with bottlenecks having a more significant effect on F_{ST} estimates than recent expansion. Overall, our results contradict a recent statement "among human populations, F_{ST} is typically estimated to be <0.1" by [36] that was based on results from [1].

All together, our results suggest that, in the setting of rare variants, a careful protocol for producing F_{ST} estimates is warranted. We provide such a protocol.

Results

Theory

Defining F_{ST}

Wright [4,37] defined F_{ST} as the correlation of randomly drawn gametes from the same population, relative to the total population. However, he did not clearly specify the "total population", leaving subsequent authors to interpret it's meaning. For Nei[29], the "total" population is the combination of the two population samples. This means, that F_{ST} quantifies drift relative to an average of the two population samples. For Cockerham, Weir, and Hill[32,35], the "total" population is the most recent common ancestral population to the two populations being considered. We agree with these authors that F_{ST} is a parameter of the evolutionary process, and not a statistic from observed samples as Nei described.

Confusion surrounding F_{ST} also comes from the differing set of assumptions made by each definition. For example, the definitions of Cockerham, and Weir and Hill all assume that studied SNPs were polymorphic in the ancestral population. Additionally, in a two-population comparison, Weir and Cockerham [5,32] defined a single F_{ST} for both populations studied. This assumes that the two populations have experienced identical amounts of drift since splitting, which may be unrealistic in many real data sets. Weir and Hill [35] generalized this, allowing for each population to have it's own F_{ST}. Because we believe that the definition of [35] (WH) is sufficiently general to apply to real world scenarios, we use this definition throughout our manuscript. WH define F_{ST} for a single population as:

$$\begin{aligned} E[p_i^s \mid p_{anc}^s] &= p_{anc}^s \\ Var(p_i^s \mid p_{anc}^s) &= F_{ST}^i p_{anc}^s (1 - p_{anc}^s) \end{aligned} \qquad (1)$$

Where p_i^s is the allele frequency of the derived allele in population *i*, at SNP *s*, and p_{anc}^s is the allele frequency of the derived allele in the ancestral population at SNP *s*.
We analyze F_{ST} estimates, using the WH definition, in the context of comparing two populations each with it's own population-specific F_{ST}. Below, we focus on how these estimates will vary with changes in estimation method.

In addition to the definitions described above, F_{ST} has been related to divergence time, coalescent times, and migration rates. Additionally, likelihood based definitions view F_{ST} as a parameter in the distribution (e.g. normal, beta) of allele frequencies in current populations. We give a short overview of these definitions, and the assumptions that they depend on in the Supplementary Material of Bhatia et al.[50].

Choice of F_{ST} Estimator

While estimators of F_{ST} handle issues related to finite sample size, we are interested in their behavior in the limit of large sample sizes, or the "quantity being estimated". Most published estimates of F_{ST} are produced using the Weir and

Cockerham (WC) [5] (>8,000 citations), or Nei [29] (>5,500 citations) estimators. However, we do not recommend these estimators.

The WC estimator was for the case of populations with identical F_{ST}, and if it is used when F_{ST} is not identical for both populations, we demonstrate that the WC quantity being estimated becomes dependent on the ratio of sample sizes (see Methods), M according to:

$$\hat{F}_{ST}^{WC} \rightarrow \frac{(F_{ST}^1 + F_{ST}^2)}{(F_{ST}^1 + F_{ST}^2 + 2\frac{1}{(M+1)}[M(1-F_{ST}^1)+(1-F_{ST}^2)])} \quad (2)$$

We note that this variation with sample size is not due to any flaw in WC estimator, but rather due to the use of the WC estimator for a purpose different than what was intended.

In the setting described by the Weir and Hill definition, the Nei estimator will consistently overestimate F_{ST} and the degree of overestimation will depend upon the magnitude of F_{ST} itself (see Methods):

$$\hat{F}_{ST}^{Nei} \rightarrow \frac{(F_{ST}^1 + F_{ST}^2)}{2 - \frac{(F_{ST}^1 + F_{ST}^2)}{2}} \quad (3)$$

We note that this result, with a maximum value of 2, makes it impossible to view F_{ST} as a correlation.

We also analyze a different estimator of F_{ST} motivated by Hudson [30,38], which produces estimates that are independent of sample sizes even when F_{ST} is not identical across populations. We provide comparisons of this estimator to the WC when applied to simulated (See Supplementary Material of Bhatia et al.[50]) and empirical data (see below). We note that while Hudson did not explicitly provide an estimator of F_{ST}, he did describe a method of estimation that corresponds to the estimator that we explicitly provide here (see Supplementary Material of Bhatia et al.[50]). Thus, we refer to this estimator as the Hudson estimator. Hudson estimates correspond to a simple average of the population specific F_{ST} estimates as given by:

$$\hat{F}_{ST}^{Hudson} \rightarrow \frac{(F_{ST}^1 + F_{ST}^2)}{2} \quad (4)$$

We note that the Hudson estimator is a simple average of the population-specific estimators proposed by Weir and Hill 2002.

Combining estimates of F_{ST} across multiple SNPs

We investigate two approaches for combining estimates of F_{ST} across multiple SNPs. In the first approach, variance components—the numerator and denominator—are averaged separately and the genome-wide estimate of F_{ST} is a "ratio of averages" [2,5]. In the second approach, single SNP estimates of F_{ST} are averaged across SNPs. The resulting "average of ratios" is reported as the genome-wide estimate [1] (see Methods).

In the context of the WH definition, the numerator of the Hudson F_{ST} estimator (See Methods) is an unbiased estimator of the variance between populations. The denominator is an unbiased estimator of the total variance in the ancestral population. However, this does not mean that the ratio of the estimators is itself an unbiased estimator of F_{ST}. We are not aware of any unbiased estimator.

While, an unbiased estimator is not available, F_{ST} estimates produced using a ratio of these two unbiased estimates will be asymptotically consistent in the sense that

they will converge to the correct underlying value as the number of independent SNPs increases. This is the basis of our recommendation F_{ST} be estimated as a ratio of averages.

We analyze the effects of choosing an average of ratios in coalescent simulations detailed in the Supplementary Material of Bhatia et al.[50].

Dependence of F_{ST} on the set of SNPs analyzed

It is well known that population genetic factors can cause variation in F_{ST} estimates, and that ascertainment scheme can alter the properties of studied SNPs [39,40]. For example, selection can result in differences between F_{ST} estimated on genic and nongenic SNPs [41-43]; complex demography can cause F_{ST} to vary with SNP allele frequency [44] (see below). Indeed, variation in F_{ST} estimates between ascertained classes of SNPs can be used to test a variety of hypotheses about population history [45,46]. This usage of F_{ST} demonstrates that there is no single correct ascertainment scheme, as F_{ST} is a parameter of both the populations *and* the set of SNPs that are used in the computation.

Though there is no single correct ascertainment scheme, ascertainment in an out-group may have desirable properties. Outgroup ascertainment guarantees that studied SNPs were polymorphic in the most recent common ancestral population (ignoring recurrent mutation), satisfying an assumption made in the Weir and Hill definition. This leads estimates of F_{ST} to be independent of allele-frequency and depend upon time since divergence according to a simple equation (see Supplementary Equation s1).

While we view these as desirable properties, if no reasonable out-group sample is available, it may become necessary to choose SNPs that are polymorphic in one, both, or either of the populations studied. These choices will affect the estimate of F_{ST} produced and may explain discrepancies in F_{ST} estimates across studies of the same populations.

We explore the effects of various ascertainment schemes on F_{ST} estimates across the allele frequency spectrum in a variety of simulated demographic scenarios (See Supplementary Material of Bhatia et al.[50]).

Other Estimation Methods

In addition to the methods that we consider above, we have also analyzed several additional methods. The moment-based estimator of Weir and Hill (2002) (WH) introduced population-specific estimates of F_{ST}. Weir and Hill recommend a sample size weighted average of these estimates, which may result in wide variation with sample size. However, one could report these estimates independently or perform a simple average of these estimates.

A separate maximum-likelihood estimator of Weir and Hill 2002 (WH-ML) is based upon a normal approximation to genetic drift. However, the equations given in [35] are not applicable to the general case of unequal sample size, and the authors recommend that estimates be "simply averaged across loci" causing WH-ML estimates to vary widely with the inclusion of rare variants.

The Bayesian method of [34] (Holsinger) approximates the distribution of allele frequencies as a beta distribution. Our simulations suggest that Holsinger estimates increase dramatically if rare SNPs are analyzed.

We also evaluated two estimators based on the beta-binomial likelihood using point estimates for the allele frequency in the ancestral population (DJB, personal communication). These estimates perform well for small values of F_{ST} but do poorly as

F_{ST} increases. It may be possible to improve on these methods by integrating over the distribution of ancestral allele frequencies, and this is an area of active research.

We describe results from these methods in greater detail in the Supplementary Material of Bhatia et al.[50].

Analysis of 1000 Genomes Data

We analyzed data from 1000 Genomes populations [1] to illustrate the effects of changes in each of the aspects of estimation described above. We focus largely on the comparison of Utah residents of European ancestry (CEU) and Chinese individuals from Beijing (CHB), as the Yoruba in Ibadan, Nigeria (YRI) sample functions as a natural outgroup for ascertainment of SNPs. This ascertainment has desirable properties (see above).

Choice of F_{ST} Estimator

Estimates of F_{ST} for CEU and CHB are 0.106 (s.e. 0.0006), 0.112 (s.e. 0.0006), and 0.107 (s.e. 0.0006) for the WC, Nei, and Hudson estimators, respectively. These estimates were produced over SNPs ascertained as polymorphic in YRI. The higher Nei estimate is expected. In addition, sample sizes for CEU (85 individuals) and CHB (97 individuals) are similar so we do not expect WC and Hudson estimates to differ.

In order to investigate the effects of sample size variation we selected 14 individuals—the size of the smallest sample (Iberian populations in Spain; IBS) in the 1000 Genomes Consortium data—from both CEU and CHB to produce populations CEU14 and CHB14. Hudson F_{ST} estimates for CEU14 and CHB are similar to those for CHB14 and CEU (see Table 1). However, WC estimates are 0.114 (s.e. 0.0006) and 0.107 (s.e. 0.0006) for CEU14 vs. CHB and CHB14 vs. CEU, respectively. The difference between these estimates is statistically significant (>8 standard errors). To verify that this difference is not due to different sets of polymorphic SNPs, we re-estimated F_{ST} restricting to SNPs that were polymorphic in YRI and at least one of CEU14 or CHB14. Re-estimated values of F_{ST} were similar to those above and WC estimates remained discordant (data not shown).

The effect of sample size variation is further exacerbated when ascertainment is performed within the populations studied. For example, in comparing IBS—with a sample size of only 14 individuals—to YRI, no reasonable out-group population exists in the 1000 Genomes data. If we ascertain within one of these populations, WC estimates are 0.121 and 0.144 for ascertainment in YRI and IBS, respectively. These estimates—computed using identical populations and *even identical individuals*—are highly divergent at >25 standard errors apart, whereas Hudson estimates are much more stable (see Table 1). This underscores that F_{ST} estimates can vary substantially based on the choice of estimator

Regardless of choice of estimator, our estimates of F_{ST} from 1000 Genomes data are relatively close to previously reported values of F_{ST} (see Table S1 of Bhatia et al.[50] for all populations). This suggests that while the choice of estimator can impact the resulting value of F_{ST}, it does not explain the disparate results reported by the 1000 Genomes consortium, and other aspects of estimation may be involved. We consider these in the sections below.

Combining Estimates of F_{ST} across Multiple SNPs

From 1000 Genomes data, we estimated F_{ST} for CEU and CHB as 0.106 (s.e. 0.0006) and 0.072 (s.e. 0.0003) for the ratio of averages and average of ratios, respectively. These estimates were produced over SNPs ascertained as polymorphic in YRI. This suggests that the result reported by the 1000 Genomes consortium (0.052)

may be partially explained by the large reduction in F_{ST} obtained by use of an average of ratios. These results are replicated for several comparisons of populations included in the 1000 Genomes data (see Table 2).

To explore the effect of the rare variants included in sequence data we compared our results to those obtained using HapMap3 genotypes. We obtain F_{ST} estimates for CEU and CHB of 0.110 (s.e. 0.0010) and 0.089 (s.e. 0.0006) using the ratio of averages and average of ratios, respectively. This suggests that the inclusion of rare variants with low single-SNP F_{ST} estimates in the 1000 Genomes data tends to exacerbate the discrepancy produced by the average of ratios. We expect that this discrepancy will grow with sample sizes and sequencing depth (see Figure S2 of Bhatia et al.[50]). Ultimately, using the average of ratios may make estimates incomparable across studies and unrelated to population demographic history.

While the use of the average of ratios clearly results in lower estimates of F_{ST}, these estimates are not as low as those published by the 1000 Genomes Consortium. Below, we explore the possibility that the remaining discrepancy can be accounted for by differences in the set of SNPs analyzed.

Dependence of F_{ST} on the Set of SNPs Analyzed

When estimating F_{ST} for CEU and CHB, we compared the effects of ascertaining in YRI (YRI-ascertainment) versus ascertaining SNPs that were polymorphic in CEU, CHB, both populations, or either population (see Table 3). When using an average of ratios, our estimates of F_{ST} were approximately 0.103 for all of these modified ascertainment schemes. These can be compared to an F_{ST} of 0.106 produced from YRI-ascertainment in 1000 Genomes Data or 0.110 in HapMap3 data. Though statistically significant, these results suggest that the effects of modified ascertainment are not very large when analyzing human populations using a ratio of averages. This indicates that reasonable estimates of F_{ST} may be produced when comparing populations without access to an out-group.

However, when using an average of ratios and including all SNPs polymorphic in either CEU or CHB, our estimate changed from 0.072 to 0.047 (s.e. 0.0002), which is similar to the result reported by the 1000 Genomes Consortium. This suggests that much of the discrepancy between previously published estimates of F_{ST} for CEU and CHB and the published 1000 Genomes estimate is explained by using the average of ratios and an ascertainment scheme that includes all SNPs that are polymorphic in either of the two populations. These results are replicated for comparisons of continental populations included in the 1000 Genomes data as we obtained values of 0.056, and 0.063 for comparisons of CEU-YRI and CHB-YRI, respectively.

Separately, we note that when comparing CEU to CHB on the 1000 Genomes data we observed *larger* F_{ST} estimates of 0.108 for the lowest frequency SNPs ($0.0 < MAF <= 0.05$) versus estimates of 0.103 for the most common SNPs ($0.45 < MAF < 0.5$), when ascertaining in CEU. These estimates were 0.131 and 0.097 when ascertaining in CHB (See Figure 1). Increased F_{ST} for rare variants suggests that bottlenecks are likely to be a stronger influence on F_{ST} estimates for CEU and CHB than recent expansions. Our results also indicate that bottlenecks in the population history of CHB are likely to be stronger than those in the population history of CEU, consistent with the findings of [38]. This is in contrast to the much lower F_{ST} estimates reported on sequence data by the 1000 Genomes consortium, which might suggest that expansions are a stronger influence on F_{ST} at rare SNPs.

Under a simple demographic history (i.e. without migration or admixture), this dependence on minor allele frequency is expected to disappear when ascertaining SNPs in an out-group. When ascertaining in YRI we do not observe any significant

dependence on frequency, which suggests that YRI is a reasonable out-group for the comparison for CEU and CHB.

We note that when ascertaining in YRI, our genome-wide estimate of F_{ST} (0.106) is lower than estimated from HapMap3 (0.110). To investigate whether this difference is due to non-random ascertainment of HapMap3 SNPs, we sampled 10 subsets of SNPs from the 1000 Genomes data that matched the allele frequency spectrum of HapMap3 SNPs (see Supplementary Material of Bhatia et al.[50]). We estimated F_{ST} for CEU and CHB in each of these subsets ranging from 0.106-0.107 (s.e. 0.0010). This suggests that HapMap3 SNPs are more highly differentiated than random SNPs, consistent with previous findings on the effects of ascertainment on genotyping arrays [40,41].

Recommendations

Choice of F_{ST} Estimator

Because the Hudson estimator is not sensitive to the ratio of sample sizes and does not systematically overestimate F_{ST}, we recommend it be used to estimate F_{ST} for pairs of populations. The proposed estimator for F_{ST} and a corresponding block-jackknife estimator for standard error of F_{ST} are implemented in the EIGENSOFT software package (see Web Resources).

Combining Estimates of F_{ST} across Multiple SNPs

Using an average of ratios will result in large reductions in F_{ST} estimates. This effect will be exacerbated when estimating F_{ST} from sequence data. Therefore, we recommend using a ratio of averages.

Dependence of F_{ST} on the Set of SNPs Analyzed

Estimating F_{ST} from SNPs ascertained in an out-group has the valuable properties that (1) F_{ST} estimates are expected to be independent of allele frequency in the out-group and (2) F_{ST} estimates will relate to divergence time according to Supplementary Equation s1 if there has been no migration or admixture. However, data from a reasonable out-group is not always available. Additionally, comparison of F_{ST} between ascertained classes of SNPs (e.g. genic vs. nongenic) can be used to test a variety of hypotheses about population history. Thus, we recommend that future publications of F_{ST} estimates include details of the ascertainment scheme used, including the proportion of SNPs that are polymorphic in each sample.

Discussion

The use of F_{ST} to quantify the genetic distance between populations and to assess differentiation at individual SNPs is widespread. Here, we point out several challenges surrounding F_{ST}, and provide a protocol for its robust estimation in the case of two populations and bi-allelic SNPs. We show that the estimator of F_{ST}, the method of combining estimates across SNPs and the scheme for SNP ascertainment can impact the resulting estimate of F_{ST}. An inappropriate choice for any of these aspects of estimation can lead to widely disparate estimates of F_{ST}, especially in a setting of large numbers of rare variants.

Indeed, the F_{ST} estimate 0.052 for CEU and CHB reported by the 1000 Genomes Consortium [1] underscores the need for a careful analysis. Utilizing the careful protocol set out here, we provide an estimate of 0.106 for CEU and CHB on 1000 Genomes data, which is close to our estimate of 0.110 on HapMap3 [2] data. Additionally, we show that when ascertaining for SNPs in one of the two populations studied, rare variants have higher F_{ST} estimates than common variants. This is the exact opposite of the results suggested by the 1000 Genomes data. The difference between these two results

changes the conclusions that are drawn about the role of demography in shaping the patterns of differentiation between human populations.

In addition to altering genome-wide estimates of F_{ST}, the choice of estimator can introduce inflation at the level of single SNP estimates. This inflation can be arbitrarily large and may call into question the results of any study of selection that uses the WC estimator [5] to produce single-SNP F_{ST} estimates for two populations with large differences in the sample sizes (See Supplementary Material of Bhatia et al.[50]).

Another concern about F_{ST} was considered by [23], who showed that as heterozygosity becomes large F_{ST} will naturally approach 0—indicating low differentiation—even if all alleles at a locus are population private. In an effort to avoid this problem Jost introduced D as an alternate measure of differentiation. However, it has been suggested that Jost's D shares the same problems as F_{ST}, and that these problems are sometimes even more pronounced for Jost's D [24]. In any case, F_{ST} and related measures "unquestionably provide important insights into population structure" [23], particularly for species such as humans in which heterozygosity is relatively low.

In conclusion, we recommend the use of the Hudson estimator [30,38] of F_{ST} that is independent of sample size. We demonstrate that a ratio of averages is an appropriate method for combining these estimates across multiple SNPs. We also show the value of estimating F_{ST} from SNPs ascertained in an out-group, though we do not view this as a necessity. We do recommend, however, that future publications of F_{ST} estimates include details of the ascertainment of SNPs.

Methods

Weir and Cockerham's F_{ST} (WC)

Definition

Weir and Cockerham [5] used the definition provided by Cockerham [32] of F_{ST} as a ratio of the variance between populations to the total variance in the ancestral population. We analyze this definition in the Supplementary Material of Bhatia et al.[50].

Estimator

In the setting of population specific F_{ST}, described by the WH definition, the WC estimator will result in estimates that vary with the ratio of sample sizes (see Supplementary Material of Bhatia et al.[50] for details). For the case of 2 populations and biallelic SNPs, the WC estimator is:

$$\hat{F}_{ST}^{WC} = 1 - \frac{2\frac{n_1 n_2}{n_1+n_2}\frac{1}{n_1+n_2-2}[n_1\tilde{p}_1(1-\tilde{p}_1)+n_2\tilde{p}_2(1-\tilde{p}_2)]}{\frac{n_1 n_2}{n_1+n_2}(\tilde{p}_1-\tilde{p}_2)^2+(2\frac{n_1 n_2}{n_1+n_2}-1)\frac{1}{n_1+n_2-2}[n_1\tilde{p}_1(1-\tilde{p}_1)+n_2\tilde{p}_2(1-\tilde{p}_2)]} \quad (5)$$

where n_i is the sample size and $\tilde{p}_i$ is the sample allele frequency in population i for $i \in \{1,2\}$. Then, in the limit of large sample sizes ($n_i - 1 \approx n_i$), we can assume that sample allele frequencies become close to population allele frequencies ($\tilde{p}_i \to p_i$). We analyze the estimator as the sample sizes increase, but their ratio goes to a constant M (see Supplementary Material of Bhatia et al.[50] for a derivation). In this case, we show (see Supplementary Material of Bhatia et al.[50]) that the estimate tends toward equation 1 (see Results).

If the sample sizes are equal, $M = 1$, then the estimate becomes

$$\hat{F}_{ST}^{WC} \rightarrow \frac{(F_{ST}^1 + F_{ST}^2)}{2}$$

Also, when F_{ST} is identical for both populations, i.e. $F_{ST}^1 = F_{ST}^2 = F_{ST}$, it is straightforward to see that $\hat{F}_{ST} \rightarrow F_{ST}$, i.e. the estimate will not depend upon the ratio of sample sizes (M). We note that if F_{ST} is identical across populations, weighting by sample sizes will reduce the variance of the estimator. This was the intent of Weir and Cockerham. If the sample sizes are unequal or this assumption does not hold, however, the estimate will depend upon the ratio of sample sizes underlying the limit. Given the complexity of human population history, it is unlikely that this assumption will hold in general. This means that even if large numbers of samples and SNPs are used to estimate F_{ST} for a pair of populations, this estimate may not be comparable across studies with different sample sizes.

We note that when F_{ST} is not identical for both populations, it is possible to estimate F_{ST} separately for each population (i.e. $\hat{F}_{ST}^1, \hat{F}_{ST}^2$) (Weir and Hill 2002). Estimates for these produced according to the method given in (Weir and Hill 2002) will not depend on sample size. We focus here on estimating F_{ST} for a pair of populations, as this is a very common use when analyzing human genetic data.

Nei's F_{ST}

Definition

Nei [33] defined F_{ST} (he used the term G_{ST}) based upon the sample gene diversity between and within populations as

$$F_{ST} = \frac{D'_{ST}}{H_T} \tag{6}$$

where D'_{ST} is the average gene diversity between populations and H_T is the diversity in the average of the two population samples. We consider this definition in detail in the Supplementary Material of Bhatia et al.[50].

Estimator

In the case of two populations and biallelic SNPs, Nei's estimator is

$$\hat{F}_{ST}^{Nei} = \frac{(\tilde{p}_1 - \tilde{p}_2)^2}{2\tilde{p}_{avg}(1 - \tilde{p}_{avg})} \tag{7}$$

where

$$\tilde{p}_{avg} = \frac{\tilde{p}_1 + \tilde{p}_2}{2}$$

and $\tilde{p}_i$ is the sample allele frequency in population i for $i \in \{1,2\}$. We note that this is Nei's updated estimator and, for the case of two populations, differs from the estimator given in [29,47] by a factor of 2. We use the estimator given in [33] as it is most closely related to the other estimators considered. This is identical to the estimator used in several recent papers [21,48].

Using the definition of (Weir and Hill 2002) we show (see Supplementary Material of Bhatia et al.[50]) that estimates made using Nei's estimator will tend toward equation 2 (see Results), with a maximum value of 2 as $F_{ST}^1 \rightarrow 1, F_{ST}^1 \rightarrow 1$. This overestimates the average of population-specific F_{ST} values and alters the relation from this average of F_{ST} values to divergence time (see Supplementary Material of Bhatia et al.[50]). Estimates of

F_{ST} given for the Nei estimator were generated using the proposed estimator for the numerator (see Supplementary Material of Bhatia et al.[50]) and a simple estimator for the denominator.

Hudson's F_{ST}

Definition

Hudson [30] defined F_{ST} in terms of heterozygosity. The fundamental difference between these estimators is that for Hudson, the total variance is based upon the ancestral population and not the current sample.

Estimator

Hudson's estimator for F_{ST} is given by

$$\hat{F}_{ST}^{Hudson} = 1 - \frac{H_w}{H_b} \tag{8}$$

where H_w is the mean number of differences within populations, and H_b is the mean number of differences between populations. While Hudson did not give explicit equations for H_w and H_b, we cast his description into an explicit estimator (see Supplementary Material of Bhatia et al.[50] for a derivation). The estimator that we analyze is:

$$\hat{F}_{ST}^{Hudson} = \frac{(\tilde{p}_1 - \tilde{p}_2)^2 - \frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1 - 1} - \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2 - 1}}{\tilde{p}_1(1-\tilde{p}_2) + \tilde{p}_2(1-\tilde{p}_1)} \tag{9}$$

where n_i is the sample size and $\tilde{p}_i$ is the sample allele frequency in population *i* for $i \in \{1,2\}$. Analyzing this estimator using the definition of (Weir and Hill 2002) we show (see Supplementary Material of Bhatia et al.[50]) that F_{ST} estimated using Hudson's estimator will tend toward equation 3 (see Results) which is exactly the average of population-specific F_{ST} values that we seek to estimate. This emerges naturally, as the proposed estimator is the simple average of the population specific estimators given in (Weir and Hill 2002). This estimator has the desirable properties that it is (1) independent of sample composition and (2) does not overestimate F_{ST} (it has a maximum value of 1). We recommend its use to produce estimates of F_{ST} for two populations.

Estimating F_{ST} From Multiple SNPs

The Hudson estimator is asymptotically consistent as the estimators of the variance components involved in the computation of F_{ST} are unbiased in the context of the WH definition. However, as their quotient is not an unbiased estimator of F_{ST}, use of an average of ratios will, in general, result in a biased estimate.

As many rare variants discovered by deep sequencing are population specific, we analyze the effect of this approach in the presence of many such variants. Consider a rare SNP with $p_1 = \varepsilon, p_2 = 0$. This yields a single SNP $F_{ST} = \varepsilon$. An estimate produced using an average of ratios will be highly sensitive to rare SNPs of this type and is likely to exhibit dependence on both the sequencing depth and sample size used in the analysis (see Figure S2 of Bhatia et al.[50]).

Previous works have examined this choice and advocated for the use of a ratio of averages [5,49]. However, in describing the WH-ML method, Weir and Hill recommend that

estimates be "simply averaged over loci." We believe that use of an average of ratios can account for the bulk of the discrepancy between the estimates of F_{ST} from [1] and previously published estimates [2] (see Results).

Effects Of SNP Ascertainment

In relating quantities being estimated from current populations to parameters of the evolutionary model we have calculated expected values given the allele frequency in the ancestral population. This implicitly performs an ascertainment of SNPs that are polymorphic in the ancestral population or, equivalently, in an out-group population. Provided there is no migration or admixture between populations, the relationship between F_{ST} and divergence time is given by equation s12.

This relationship is accounts for changes in effective population size (i.e. bottlenecks or expansions) in the demographic history of the populations being compared. Additionally, ascertainment in an out-group renders the estimate independent of the allele frequency spectrum in the out-group. Therefore, with this type of ascertainment scheme, estimates should be concordant regardless of whether they are produced from rare or common SNPs.

While ascertainment in an out-group has several helpful properties, in many practical circumstances no data from a reasonable out-group is available. In these instances, F_{ST} can be estimated using SNPs ascertained in either one of the populations under study. However, in these instances estimates are *not* expected to be independent of allele frequency spectrum or complex demographic scenarios.

Acknowledgements

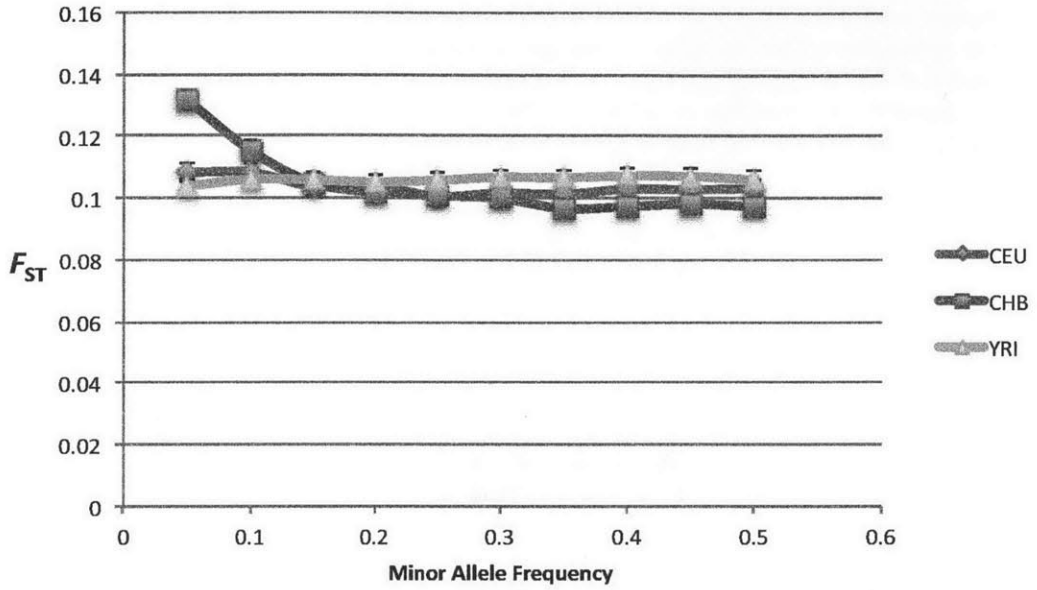
We are grateful to D. Reich, S. Sunyaev, S. Myers, N. Zaitlen, J. Wilson, A. Keinan, and W.G. Hill for helpful discussions. We thank W. Jin for helpful discussions and providing data. This research was funded by NIH grants T32 HG002295 (G.B.), R01 HG006399 (N.P. and A.L.P.), R03 HG006170 (G.B. and A.L.P.) and R01 MH084676 (A.L.P.).

Web Resources

EIGENSOFT 4.2 http://www.hsph.harvard.edu/faculty/alkes-price/software/
TreeSelect 1.1 http://www.hsph.harvard.edu/faculty/alkes-price/software/
Hickory 1.1 http://darwin.eeb.uconn.edu/hickory/hickory.html

Figures

Figure 1



Allele frequency dependence of F_{ST} under different ascertainment schemes. This shows F_{ST} for CEU and CHB as a function of allele frequency when ascertaining in either CEU, CHB, and YRI. The increased F_{ST} for rare variants is consistent with bottlenecks being a stronger force on F_{ST} for CEU and CHB than recent expansion. In fact, this is consistent with a stronger bottleneck in the population history of CHB. We note that this frequency dependence disappears when ascertaining in YRI suggesting that YRI is a reasonable out-group for the comparison of CEU and CHB

Tables

Table 1.

		F_{ST} Estimator					
		WC		Nei		Hudson	
Comparison	# Of SNPs	Est.	Std. Error	Est.	Std. Error	Est.	Std. Error
CEUvCHB	7799780	0.107	5.70E-04	0.112	6.36E-04	0.106	5.69E-04
CEUvYRI	17814120	0.139	4.97E-04	0.149	5.79E-04	0.139	5.00E-04
CHBvYRI	17814120	0.163	5.85E-04	0.175	6.84E-04	0.161	5.78E-04
CEUvCHB14	7215431	0.107	6.10E-04	0.113	7.16E-04	0.106	6.36E-04
CHBvCEU14	7465953	0.114	6.49E-04	0.114	7.12E-04	0.107	6.32E-04
IBSvYRI	17814120	0.121	4.37E-04	0.145	6.02E-04	0.131	6.73E-04
YRIvIBS*	7709984	0.144	8.06E-04	0.141	7.77E-04	0.134	8.43E-04

*In this case ascertainment was performed in the IBS sample. In all other cases, ascertainment was performed in YRI.

F_{ST} estimates for pairs of populations in 1000 Genomes. Unless otherwise specified SNPs were ascertained as polymorphic in YRI. These estimates are concordant with results reported on common SNPs [2] than the results reported by the 1000 Genomes consortium [1]. Even so, we note that the choice of F_{ST} estimator impacts the resulting estimate. This is evident when comparing CEU14—14 individuals sampled from the CEU population—to CHB, and CHB to CEU14. Though these estimates are produced using overlapping sets of SNPs and individuals, the estimates are statistically significantly different when produced using the WC estimator. This difference is underscored when comparing the YRI and IBS populations. The small sample from the IBS population causes WC estimates to change significantly depending on ascertainment in IBS (line 4) or YRI (line 5). The number of SNPs listed indicates the number of SNPs that were polymorphic in the ascertained population (usually YRI) and at least one of the populations studied.

Table 2

Comparison	Ratio of Averages			
	1000 Genomes		HapMap 3	
	Est.	Std. Error	Est.	Std. Error
CEU-YRI	0.139	5.00E-04	0.156	9.73E-04
CEU-CHB	0.106	5.69E-04	0.110	9.61E-04
CHB-YRI	0.161	5.78E-04	0.183	1.13E-03
Comparison	Average of Ratios			
	1000 Genomes		HapMap 3	
	Est.	Std. Error	Est.	Std. Error
CEU-YRI	0.063	1.53E-04	0.124	6.23E-04
CEU-CHB	0.072	3.04E-04	0.089	6.35E-04
CHB-YRI	0.070	1.70E-04	0.141	6.93E-04

A comparison of the F_{ST} estimated using 1000 Genomes and HapMap data by either using a ratio of averages or an average of ratios. It is clear that the average of ratios of F_{ST} results in a significant underestimate of F_{ST} and use of an average of ratios approach can explain the bulk of the discrepancy between the F_{ST} reported by the 1000 Genomes Consortium and previously reported estimates. The ratio of averages estimates are much more concordant with estimates on HapMap data. We believe that discrepancies between these different data sets are due to the different set of SNPs used in the computation. Finally, use of the average of ratios results in a smaller reduction when applied to HapMap3 data. This is consistent with an average of ratios being sensitive to rare variants that are, in general, excluded from the HapMap set of SNPs.

Table 3

Polymorphic In	Ratio of Averages		Average of Ratios	
CEU	0.104	6.19E-04	0.056	2.55E-04
CHB	0.104	6.40E-04	0.057	2.74E-04
CEU AND CHB	0.104	7.25E-04	0.078	4.49E-04
CEU OR CHB	0.103	5.64E-04	**0.047**	1.87E-04

Assessing the effect of ascertainment schemes and combination methods on the resulting F_{ST} estimate for CEU and CHB. When using a ratio of averages, modified ascertainment results in a small, though statistically significant, difference from a value 0.106 obtained using YRI-ascertainment. The effect is much larger when employing an average of ratios, and the bolded cell indicates that a permissive ascertainment scheme coupled with an average of ratios can produce a value similar to the estimate of F_{ST} for CEU and CHB published by the 1000 Genomes Consortium.

References

1. 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

2. International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

3. Malécot, G. (1948). Les mathématiques de l'hérédié. Masson.

4. Wright, S. (1949). THE GENETICAL STRUCTURE OF POPULATIONS. Ann. Hum. Genet. *15*, 323–354.

5. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. Evolution *38*, 1358–1370.

6. Holsinger, K.E., and Weir, B.S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nature Reviews. Genetics *10*, 639–650.

7. International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

8. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.

9. Selander, R.K., and Hudson, R.O. (1976). Animal Population Structure Under Close Inbreeding: The Land Snail Rumina in Southern France. The American Naturalist *110*, 695–718.

10. Guries, R.P., and Ledig, F.T. (1982). Genetic Diversity and Population Structure in Pitch Pine (Pinus rigida Mill.). Evolution *36*, 387–402.

11. Ellstrand, N.C., and Elam, D.R. (1993). Population Genetic Consequences of Small Population Size: Implications for Plant Conservation. Annual Review of Ecology and Systematics *24 IS* , 217–242.

12. Palumbi, S.R., and Baker, C.S. (1994). Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. Molecular Biology and Evolution *11*, 426–435.

13. 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.

14. Nicholson, G., Smith, A.V., Jonsson, F., Gústafsson, Ó., Stefansson, K., and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *64*, 695–715.

15. Beaumont, M.A., and Balding, D.J. (2004). Identifying adaptive genetic divergence

among populations from genome scans. Mol. Ecol. *13*, 969–980.

16. Foll, M., and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics *180*, 977–993.

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. 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.

19. 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., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.

20. 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., et al. (2009). Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res. *19*, 2154–2162.

21. 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.

22. 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.

23. Jost, L. (2008). GST and its relatives do not measure differentiation. Mol. Ecol. *17*, 4015–4026.

24. Ryman, N., and Leimar, O. (2009). G(ST) is still a useful measure of genetic differentiation - a comment on Jost's D. Mol. Ecol. *18*, 2084–7–discussion2088–91.

25. Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., Gong, X., Wang, H., Shen, Y., Pan, X., et al. (2009). Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am. J. Hum. Genet. *85*, 762–774.

26. Edelaar, P., Alonso, D., Lagerveld, S., Senar, J.C., and Björklund, M. (2012). Population differentiation and restricted gene flow in Spanish crossbills: not isolation-by-distance but isolation-by-ecology. Journal of Evolutionary Biology *25*, 417–430.

27. Hangartner, S., Laurila, A., and Räsänen, K. (2012). Adaptive divergence in moor frog (Rana arvalis) populations along an acidification gradient: inferences from Q(st) - F(st) correlations. Evolution *66*, 867–881.

28. Slatkin, M. (1991). Inbreeding coefficients and coalescence times. Genetics

Research *58*, 167.

29. Nei, M. (1973). Analysis of Gene Diversity in Subdivided Populations. Proceedings of the National Academy of Sciences *70*, 3321–3323.

30. Hudson, R.R., Slatkin, M., and Maddison, W.P. (1992). Estimation of levels of gene flow from DNA sequence data. Genetics *132*, 583–589.

31. Cavalli-Sforza, L.L., and Bodmer, W.F. (1971). The genetics of human populations (San Francisco: W.H. Freeman).

32. Cockerham, C.C. (1969). Variance of Gene Frequencies. Evolution *23*, 72–84.

33. Nei, M. (1986). Definition and Estimation of Fixation Indices. Evolution *40*, 643–645.

34. Holsinger, K.E. (1999). Analysis of Genetic Diversity in Geographically Structured Populations: A Bayesian Perspective. Hereditas *130*, 245–255.

35. Weir, B.S., and Hill, W.G. (2002). Estimating F-statistics. Annu. Rev. Genet. *36*, 721–750.

36. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nature Genetics *44*, 243–246.

37. Wright, S. (1950). Genetical structure of populations. Nature *166*, 247–249.

38. Keinan, A., Mullikin, J.C., Patterson, N., and Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nature Genetics *39*, 1251–1255.

39. Ramírez-Soriano, A., and Calafell, F. (2008). FABSIM: a software for generating FST distributions with various ascertainment biases. Bioinformatics *24*, 2790–2791.

40. Albrechtsen, A., Nielsen, F.C., and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. Molecular Biology and Evolution *27*, 2534–2547.

41. Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H., and Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. Genome Research *15*, 1496–1502.

42. Barreiro, L.B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. Nature Genetics *40*, 340–345.

43. Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., 1000 Genomes Project, Sella, G., and Przeworski, M. (2011). Classic selective sweeps were rare in recent human evolution. Science *331*, 920–924.

44. Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome Res. *15*, 1576–1583.

45. Weir, B.S., Cardon, L.R., Anderson, A.D., Nielsen, D.M., and Hill, W.G. (2005). Measures of human population structure show heterogeneity among genomic regions. Genome Res. *15*, 1468–1476.

46. McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet *5*, e1000471.

47. Nei, M., and Chesser, R.K. (1983). Estimation of fixation indices and gene diversities. Ann. Hum. Genet. *47*, 253–259.

48. 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.

49. Reynolds, J., Weir, B.S., and Cockerham, C.C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics *105*, 767–779.

50. Bhatia, G., Patterson, N., Sankararaman, S., and Price, A.L. (2013). Estimating and interpreting FST: The impact of rare variants. Genome Res.

Chapter 3: Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection

The study of recent natural selection in human populations has important applications to human population history and medicine. Positive natural selection drives the increase in beneficial alleles and plays a role in explaining diversity across human populations. By discovering traits subject to positive selection we can better understand the population level response to environmental pressures including infectious disease.

Our study examines unusual population differentiation between three large datasets to detect natural selection. The populations examined: African Americans, Nigerians, and Gambians, are genetically close to one another ($F_{ST} < 0.01$ for all pairs), allowing us to detect selection even with moderate changes in allele frequency. We also develop a tree-based method to pinpoint the population in which selection occurred, incorporating information across populations.

Our genome-wide significant results corroborate loci previously reported to be under selection in Africans including *HBB* and *CD36*. At the *HLA* locus on chromosome 6, results suggest the existence of multiple, independent targets of population-specific selective pressure. In addition, we report a genome-wide significant ($P=1.36 \times 10^{-11}$) signal of selection in the Prostate Stem Cell Antigen (*PSCA*) gene. The most significantly differentiated marker in our analysis, rs2920283, is highly differentiated in both Africa and East Asia and has prior genome-wide significant associations to bladder and gastric cancers.

Introduction

The study of recent natural selection in humans has important applications to human population history and medicine. Previous studies have reported selection at loci associated with susceptibility to falciparum malaria[1-3], vivax malaria[4], Lassa virus [5], end-stage kidney disease[6], tuberculosis and HIV/AIDS[7-9]. Indeed, it has been suggested that signals of selection at malaria loci are "only the tip of the iceberg"[10]. Signals of selection fit into three main categories: unusually long, recent haplotypes; deviations from the expected allele frequency spectrum; and unusual population differentiation[11]. Signals of the first two types are only expected under the "selective sweep" model of selection[1, 12]. This model assumes that a novel or very rare variant is subject to selection and then "sweeps" to high frequency, carrying hitchhiking variants and long haplotypes with it. If, however, selection acts on a common or "standing" variant, as has been suggested in recent studies[13-15], these tests would be unlikely to uncover a signal. Therefore, a key advantage of our approach, based on unusual population differentiation, is the ability to detect selection on standing variation[16]. Additionally, while other approaches based on population differentiation simply report top-ranked loci, our study of selection allows for the assessment of genome-wide significance.

Many prior studies of unusual population differentiation have focused on comparing continental populations[2, 17-19]. Because of large genetic distances (F_{ST})[20] these studies may be better suited to understanding population history rather than detecting selection[21]. Studies of population differentiation to detect selection are maximally powered when comparing closely related populations that have large effective population size, with data from a large number of individuals (> $1/F_{ST}$). This approach has been applied genome-wide to comparisons of closely related populations within Europe and within East Asia[22, 23], and to candidate loci of closely related populations within Africa[24]. Now, the availability of genome-wide data from > 12,000 individuals of African-American, Nigerian and Gambian ancestry makes it possible to proceed with genome-wide application of this approach in Africa.

To accomplish this analysis, we have developed a tree-based method, which incorporates information from all 3 populations in order to increase power to detect selection and enable resolution of the population subject to selection. However, both African-American[25] and Gambian[26, 27] populations have significant European-related admixture. While it is possible to perform a study of population differentiation between admixed populations, our method minimizes F_{ST} and maximizes power by accounting for this admixture. Additionally, we sought to increase coverage of selected loci y performing imputation using a combined reference panel of Europeans (CEU) and Yoruba (YRI) from the HapMap 3 Project[28]. We note that our method bears similarity to the Locus Specific Branch Length (LSBL)[29] method, though our statistic follows a well-defined distribution under the null model of no selection. This allows for the evaluation of genome-wide significance, as opposed to the ranking of loci produced by most genome-wide scans for selection[30].

We applied this approach and detected genome-wide significant signals at previously established targets of selection in *CD36*[31] [MIM 173510], *HBB*[24, 32, 33] [MIM 141900]—both reported targets of selection due to malaria—and *HLA*[7, 34, 35] [MIM 142800], which has a major role in immunity, including in malaria resistance[10, 33]. In addition, by combining evidence of extreme population differentiation within Africa and within East Asia, we have identified a genome-wide significant locus under selection in the Prostate Stem Cell Antigen (*PSCA*) [MIM 602470] gene ($P=1.36\times10^{-11}$). The most significantly differentiated marker at this locus, rs2920283, is also highly differentiated in our analysis of East Asian populations. This SNP is tightly linked to a nonsynonymous coding variant that has previous genome-wide significant associations to bladder[36] and

gastric cancers[37]. The *PSCA* markers are common in all continental populations, indicating a likely instance of selection on standing variation.

In addition, at the *HLA* locus, we observe multiple signals of differentiation. While selection at *HLA* is unsurprising given its role in immunity and many disease associations[7], we note that several markers that are highly differentiated on one branch of the tree do not show significant differentiation on other branches. This evidence is consistent with multiple, population-specific selective pressures at the *HLA* locus.

Methods

CARe Data set

Our main African-American (AA) dataset consists of 6209 unrelated individuals genotyped on the Affymetrix 6.0 array as previously described[38]. These individuals were genotyped as part of one of the ARIC, CARDIA, CFS, JHS or MESA cohorts in the CARe consortium[39]. The ARIC study is a prospective population-based study of atherosclerosis and cardiovascular diseases in 15,792 men and women, including 11,478 non-Hispanic whites and 4,314 African Americans, drawn from 4 U.S. communities (suburban Minneapolis, Minnesota; Washington County, Maryland; Forsyth County, North Carolina, and Jackson, Mississippi). The CARDIA study is a prospective, multi-center investigation of the natural history and etiology of cardiovascular disease in African Americans and whites 18-30 years of age at the time of initial examination. The initial examination included 5,115 participants selectively recruited to represent proportionate racial, gender, age, and education groups from four communities: Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA. The Cleveland Family Study (CFS) is a family-based, longitudinal study designed to characterize the genetic and non-genetic risk factors for sleep apnea. In total, 2534 individuals (46% African American) from 352 families were studied on up to 4 occasions over a period of 16 years (1990-2006). The Jackson Heart Study (JHS) is a prospective population-based study to seek the causes of the high prevalence of common complex diseases among African Americans in the Jackson, Mississippi metropolitan area, including cardiovascular disease, type-2 diabetes, obesity, chronic kidney disease, and stroke. The Multi-Ethnic Study of Atherosclerosis (MESA) is a study of the characteristics of subclinical cardiovascular disease (disease detected non-invasively before it has produced clinical signs and symptoms) and the risk factors that predict progression to clinically overt cardiovascular disease or progression of the subclinical disease.

Other Data sets

Additionally, we analyze 756 Nigerian (NIG) individuals genotyped on the Affymetrix 6.0 array as well as a Gambian dataset of 2946 individuals from the WTCCC-TB study[27] genotyped on the Affymetrix 500k array. For quality control we have utilized separate datasets of 757 African Americans genotyped on the Affymetrix 6.0 array and 2556 Gambians genotyped on the Affymetrix 500k array as part of the MalariaGen study[26]. Finally, to account for European-related admixture we used a dataset of 1178 European (EA) individuals genotyped on the Affymetrix 6.0 array. We also analyzed genome-wide data from the International HapMap 3 Project[28]. For our analysis, we considered only unrelated individuals. The population panel consisted of 113 Yoruba from Ibadan, Nigeria (YRI), 112 individuals of northwestern European ancestry (CEU), 84 Han Chinese from Beijing (CHB), 85 Chinese in Metropolitan Denver, Colorado (CHD), 86 Japanese from Tokyo (JPT), 88 Toscans from Italy (TSI), and 90 Luhya from Webuye, Kenya (LWK). We analyzed all autosomal SNPs in our pairwise comparisons. Finally, for the purpose of illustration of population differentiation on a global scale at loci of interest, we examined allele frequencies in the 52 distinct ethnic groups genotyped as part of the Human Genome Diversity Project[40]. Appropriate sample consent and IRB approval was obtained in all cases.

Quality Control

In order to limit the possibility that assay artifacts in our data cause spurious signals of selection[41], we compared each of our large data sets with an independent dataset, genotyping individuals drawn from the same population. That is, we compared our

primary African-American dataset with an African-American dataset from a separate study, and similarly for our Nigerian and Gambian datasets. We then excluded all markers that showed significant ($P < 10^{-6}$) population differentiation between the two datasets. This is a conservative approach because it excludes markers with assay artifacts in either of the two datasets. In order to further eliminate assay artifacts, we only reported loci that contained at least 2 SNPs with $P < 10^{-6}$ within 1 Mb of each other. Both our pairwise and tree based methods depend upon an assumption of a normal distribution of allele frequency differences[42]. This assumption is not likely to hold when alleles are very rare (see Figure S1 of Bhatia et al.[62]). Therefore, SNPs with an average MAF < 5% were excluded from reported results.

Two Populations

Our approach to detecting unusual population differentiation over a set of SNPs genotyped in a pair of populations proceeds in two steps. The first step is to estimate the degree of differentiation between the two populations. Wright's F_{ST} is a measure of genetic drift and can be used for this purpose. Let $D_s = p_1^s - p_2^s$ represent the allele frequency difference at SNP s between population 1 and population 2.

If population 1 and 2 are identical then D_s is approximately normally distributed with mean 0 and variance

$$\sigma_{D_s}^2 = p_{avg}^s(1-p_{avg}^s)(1/N_1 + 1/N_2).$$

We note that this variance is only due to using finite size samples. Here, N_i is the sample size for population i and p_{avg}^s may be a simple or sample size weighted average of p_1^s and p_2^s. If population 1 and 2 are genetically differentiated then D_s is again approximately normally distributed with mean 0 and variance

$$\sigma_{D_s}^2 = p_{avg}^s(1-p_{avg}^s)(2F_{ST} + 1/N_1 + 1/N_2)$$

In this formulation both genetic drift and sampling error provide components of the variance. From this, we can estimate F_{ST} using a method of moments. We note that this is not the standard estimator of F_{ST} and was chosen because it guarantees a correct statistic *in expectation* ($\lambda_{GC} = 1$) when evaluated as below.

While it is possible to test for significant allele frequency differences without accounting for F_{ST} by using a χ^2 test based on a 2x2 contingency table[23, 43], this is not a test for selection[22]. In particular, this approach tests a null hypothesis that the allele frequency is *identical* in the two populations. This implies that neither drift nor selection has taken place. However, when comparing genetically differentiated populations, we expect differences to accrue due to genetic drift and should not be surprised to see this hypothesis rejected. On the other hand, a test for population differences such as ours can test the null hypothesis that the observed allele frequency difference can be accounted for by drift alone. This removes drift as an alternate explanation and gives stronger evidence of a selective event. To illustrate this, we reexamined the most highly differentiated marker in a recent study of East Asians[23]. This marker has a P-value of 2.4×10^{-13} for the null hypothesis of "neither drift nor selection", and a P-value of 1.32×10^{-8} under our test for "no selection." While this remains genome-wide significant, less differentiated markers reported in this type of analysis may not be convincing cases of selection.

The second step in detecting selection using two populations is to evaluate a statistic for population differentiation at every available marker. Our statistic is based upon a likelihood ratio test. The null model assumes that the observed allele frequency

differences are solely due to genetic drift and sampling error.

$$L_{NULL} = \frac{1}{\sqrt{2\pi\sigma_{D_s}^2}} \cdot e^{\frac{-D_s^2}{2\sigma_{D_s}^2}}$$

The causal model allows an arbitrary amount of differentiation between the populations to be attributed to selection.

$$L_{CAUSAL} = \max_{SEL} \frac{1}{\sqrt{2\pi\sigma_{D_s}^2}} \cdot e^{\frac{-(D_s+SEL)^2}{2\sigma_{D_s}^2}}$$

where *SEL* denotes the allele frequency difference attributed to selection. This gives a likelihood ratio test as below.

$$LRT = \frac{\max_{SEL} \frac{1}{\sqrt{2\pi\sigma_{D_s}^2}} \cdot e^{\frac{-(D_s+SEL)^2}{2\sigma_{D_s}^2}}}{\frac{1}{\sqrt{2\pi\sigma_{D_s}^2}} \cdot e^{\frac{-D_s^2}{2\sigma_{D_s}^2}}}$$

Maximizing over SEL gives $SEL = -D_s$, then we have

$$2\ln(LRT) = \frac{D_s^2}{\sigma_{D_s}^2},$$

which is a χ^2 1 d.f. statistic. In order to verify that this statistic gave the correct null distribution we performed neutral simulations (see Table S1 of Bhatia et al.[62]). Additionally, we sought to investigate the power of such a test to detect selection when one of the two populations was under selection, or when both populations were under selection with differing selection coefficients. Our simulations (see Table S2 of Bhatia et al.[62]) show that, as expected, this test is highly sensitive to the difference between the selection coefficients in the two populations. This indicates that maximal power is obtained when comparing closely related populations subject to differing environmental pressures.

Finally, we note that a normal distribution is an approximation of the true distribution of allele frequency differences under neutral drift. We evaluated the validity of this approximation by comparing the cumulative distribution function under the normal approximation to the distribution obtained using Kimura theory (see Figure S1 of Bhatia et al.[62]). While the normal approximation breaks down for rare variation (MAF < 0.05) and high genetic drift (F_{ST} > 0.01), it appears reasonable for the range of allele frequencies and genetic drift that are under consideration here.

Multiple Populations

We can generalize the analysis of unusual allele frequency differentiation between a pair of populations to multiple populations in an unrooted tree. That is, we can consider each population to be a leaf-node in an unrooted tree that describes the patterns of population divergence without knowing the order of divergence events in time. Then if we can reconstruct the tree from the observed populations we can begin the work of detecting selection in the tree. This approach presents a variety of challenges relative to the pairwise test.

We must select an unrooted tree topology, estimate the branch lengths, and develop a statistic to use on the resulting tree. As the number of populations increases, each of these steps becomes increasingly difficult. Indeed, the number of possible unrooted tree topologies given n populations is $(2n-5)!/[2^{(n-3)}(n-3)!]$ and this does not begin to consider the possible branch length assignments to each of these topologies. While the literature on tree estimation in the context of multiple populations is relatively well developed[44], we consider the simpler case of $n=3$ for this study. This allows us to analyze each of these problems in discrete steps. For larger n, the analysis may have to be combined.

Given $n=3$ there is a single, star-shaped topology for an unrooted tree. In order to estimate the branch lengths, we utilized a pseudo-likelihood model considering all pairs of populations involved.

In our approach we consider pairwise differences between each pair of populations. We can define a pairwise variance, $\sigma^2_{i,j}=\hat{p}^s_c(1-\hat{p}^s_c)(2F^i_{st}+2F^j_{st}+1/N_i+1/N_j)$ where the F_{ST} between the populations is represented by a sum of the branch lengths F^i_{ST} and F^j_{ST}. If we assume independence of the pairwise differences, this gives a pseudo-likelihood

$$l^s(p^s_1,p^s_2,p^s_3,\vec{F}_{ST})=\prod_{i=1}^{3}\prod_{j=i+1}^{3}\frac{1}{\sqrt{2\pi\sigma^2_{i,j}}}\cdot e^{\frac{-(p^s_i-p^s_j)^2}{2\sigma^2_{i,j}}}$$

We used gradient ascent to find a local maximum likelihood estimate for $\vec{F}_{ST}$ over all SNPs *s*. This gives results that closely recapitulate previous estimates of F_{ST}. Once the branch lengths are estimated, we can estimate the allele frequency at our central node using a branch length weighted average.

$$\hat{p}^s_c=\frac{\sum_i \frac{p^s_i}{(2F^i_{ST}+1/N_i)}}{\sum_i \frac{1}{(2F^i_{ST}+1/N_i)}}$$

Given an estimate of the allele frequencies at the central node, we devise a test for selection akin to our pairwise test for population differentiation. In particular, we first re-estimate F_{ST} between each population and the allele frequencies at the central node. Once this is done, we formulate our statistic based upon the likelihood ratio test. We note that this test focuses on selection at any *single* branch in the tree, and each branch can be tested in turn, provided that the appropriate multiple testing correction is paid as a penalty.

In our null model we assume that all population differentiation is the result of genetic drift.

$$L_{NULL}=\prod_{i=1}^{3}\frac{1}{\sqrt{2\pi\sigma^2_{D_i}}}\cdot e^{\frac{-D_i^2}{2\sigma^2_{D_i}}}$$

where $D_i = p_i - \hat{p}_c$ and $\sigma^2_{D_i} = \hat{p}^s_c(1-\hat{p}^s_c)(2F^i_{ST} + 1/N_i)$

In our causal model, we allow an arbitrary amount of differentiation on one branch to be attributed to selection. Therefore, we have

$$L_{CAUSAL} = \max_{SEL} \frac{1}{\sqrt{2\pi\sigma^2_{D_{iSEL}}}} \cdot e^{\frac{-(D_{iSEL}+SEL)^2}{2\sigma^2_{D_{iSEL}}}} \prod_{i \neq i_{SEL}} \frac{1}{\sqrt{2\pi\sigma^2_{D_i}}} \cdot e^{\frac{-D_i^2}{2\sigma^2_{D_i}}},$$

where i_{SEL} represents the branch on which we are allowing an arbitrary amount of differentiation due to selection. F^i_{ST} in both of these equations is re-estimated using the central allele frequencies estimated in the prior step. This re-estimation guarantees that we have a correct statistic *in expectation* ($\lambda_{GC} = 1$). The test becomes akin to our pairwise test for population differentiation and we have

$$LRT = \frac{\max_{SEL} \frac{1}{\sqrt{2\pi\sigma^2_{D_{iSEL}}}} \cdot e^{\frac{-(D_{iSEL}+SEL)^2}{2\sigma^2_{D_{iSEL}}}} \prod_{i \neq i_{SEL}} \frac{1}{\sqrt{2\pi\sigma^2_{D_i}}} \cdot e^{\frac{-D_i^2}{2\sigma^2_{D_i}}}}{\prod_{i=1}^{3} \frac{1}{\sqrt{2\pi\sigma^2_{D_i}}} \cdot e^{\frac{-D_i^2}{2\sigma^2_{D_i}}}}$$

and

$$2\ln(LRT) = \frac{D^2_{i_{SEL}}}{\sigma^2_{D_{iSEL}}}$$

This is a χ^2 1 degree of freedom statistic. At a first glance, this approach passes the sanity checks of giving no additional power when one of the branch lengths is very large relative to the others, and of giving additional power when a large differentiation is replicated over multiple branches. Software implementing our methods is publicly available (TreeSelect software; see Web Resources).

We note that pairwise comparisons between our main datasets was performed but yielded nothing that was fundamentally different from our tree-based results. As such only the tree-based results are reported in the main text.

Our genome-wide significance threshold for this analysis is based on 10^6 markers tested for 3 branches of the tree, with a corrected significance level of $\alpha < 0.05$. Using a standard Bonferroni correction this gives a nominal significance level $P < 1.67 \times 10^{-8}$. In our analysis of additional populations we only included the comparison between East Asian populations because allele frequency differences in East Asia are independent of allele frequency differences in our tree-based analysis. This is not the case for differentiation between African populations (LWK vs. YRI)—as Nigerians are represented in the tree—nor European populations (CEU vs. TSI)—as Europeans are used to correct for European-like admixture. However, we conservatively correct for 3 additional tests, as though all comparisons were performed. This gives a nominal significance level of $P < 5.56 \times 10^{-9}$.

Controlling for Admixture

In order to maximize power, we sought to minimize genetic distance between our

populations by accounting for European-related admixture in our African-American and Gambian datasets. A simple example of comparing an admixed population to an unadmixed population is the comparison of African Americans (AA) to Nigerians (YRI). The African admixed component of AA individuals has been shown to have $F_{ST} < 10^{-3}$ with respect to YRI[28, 45]. However, European admixture in AA individuals increases the observed value of F_{ST} to 0.0075 and results in a less powerful test. We address this by producing estimates of the "pseudo-unadmixed" allele frequencies, where $p^s_{AA'} = \frac{p^s_{AA} - \alpha_{AA} p^s_{EUR}}{(1-\alpha_{AA})}$. The parameter α_{AA} can then be estimated to minimize F_{ST} with Nigerians. This process was performed separately for African-American and Gambian datasets.

The allele frequencies in European-related admixture were estimated from our 1178 European individuals. These individuals were split into two equally sized datasets used to produce estimates of the "pseudo-unadmixed" allele frequencies for African Americans and Gambians, respectively.

Population Differences By SNP Class

In order to test for enrichment of highly differentiated SNPs based upon annotated functional class, we partitioned the SNPs according to predicted functional impact[46]. We assigned SNPs to be either genic or nongenic and further subdivided genic SNPs into either synonymous or nonsynonymous categories (all nongenic SNPs were categorized as synonymous). We tested for an excess of highly differentiated markers (P< 0.0001) in genic vs. nongenic SNPs and in nonsynonymous vs. synonymous by using a χ^2 test on a 2x2 contingency table. We used the dbSNP classification for function-class annotations and assigned intronic, 5' UTR, 3' UTR, synonymous, nonsynonymous and splice site mutations as genic.

We also sought to evaluate variation in F_{ST} across the genome by comparing estimates of F_{ST} between genic and nongenic SNPs. To explore this further, we partitioned the SNPs according to evidence for background selection as estimated by the previously described *B* parameter[47]. We binned SNPs according to the estimate of *B* (0 <= *B* <= 1) at the SNP, using 10 equally sized bins for *B*. Because of the change in F_{ST} according to bin reported statistics for differentiation were calculated separately for each bin. However, reported values for F_{ST} are genome-wide averages.

Imputation

We used the MaCH[48] software package to perform imputation of the HapMap3 SNP set in each of our datasets. Our European dataset was used to create the pseudo-unadmixed datasets of African Americans and Gambians. The imputation process proceeded in three steps. First, the model parameters were estimated using a subset of 300 individuals from each dataset. The input files for the reference CEU and YRI panels were downloaded from the MaCH website. Next, the imputation was performed 300 individuals at a time and parallelized on a large computing cluster. Finally, once the imputation was complete, we performed quality control on the results using $\hat{r}^2$ as our quality metric[48]. Only SNPs that had $\hat{r}^2 > 0.6$ in the combined set of individuals were retained.

Results

Population structure in African-American, Nigerian and Gambian populations
500 individuals from each of our African-American, Nigerian, and Gambian datasets were studied together with 500 European individuals via PCA with EIGENSOFT[49] (see Figure 1). The PCA was performed on the basis of 309,373 autosomal SNPs shared by all individuals. As expected, European and Nigerian individuals form tight clusters that are separated by PC1. The African-American individuals form a cline between these two clusters indicating varying degrees of European admixture in African-American individuals. We note that while several African-American individuals come very close to the Nigerian cluster, there remains a non-zero distance between all African-American individuals and the Nigerian cluster. This is consistent with a small, but measurable, F_{ST} between the African ancestors of African Americans and Nigerians. The Gambian individuals are separated from Europeans on PC1 and from the Nigerians on PC2. We label each of the Gambian individuals with their subpopulation label (Mandinka, Jola, Fula, Wolloff) and note the existence of cryptic population structure within the Gambia. Several Fula individuals show significant evidence of European-related admixture by their position on PC1. Additionally, the four subpopulations form overlapping but distinguishable clusters along PC2.

We further investigated population structure by estimating the pairwise F_{ST} between each pair of populations (see Table 1a). However, we sought to increase power and decrease genetic distance between our populations by accounting for the significant European-related admixture. We produced new 'pseudo-unadmixed' populations by subtracting European allele frequencies, weighted by admixture proportion, from both the African-American and Gambian datasets (see Methods). We computed admixture proportions α_{AA} and α_{GAM} to minimize the pairwise F_{ST} estimates between each 'pseudo-unadmixed' population and the Nigerians (see Table 1b). This reduced F_{ST} between African Americans and Nigerians from an estimate of 0.0075 to an estimate of 0.0011. We calculated an α_{AA} of 0.20 and an α_{GAM} of 0.02 consistent with prior estimates. We also examined pairwise genetic distances in the Gambia (see Table 2). The lowest F_{ST} was estimated between Mandinka and Wolloff subpopulations (F_{ST} = 0.0005) and the highest between the Fula and Jola subpopulations (F_{ST} =0.005). These values are consistent with prior estimates[26, 27] and indicate that studies of selection using population differentiation within the Gambia may be a fruitful endeavor. However, given current sample sizes such a study is unlikely to be well-powered.

In order to validate our use of imputed data we compared F_{ST} estimates between pairs of imputed datasets to those observed between genotyped datasets. Pairwise F_{ST} estimates were 0.0048, 0.0012, and 0.0066 for genotyped SNPs in African Americans vs. Gambians, African Americans vs. Nigerians and Nigerians vs. Gambians, respectively. The corresponding estimates for all SNPs (genotyped + imputed) were 0.0044, 0.0011, and 0.0058. This close concordance, and the absence of peaks of population differentiation containing only imputed SNPs, suggests that our reported results do not contain spurious signals due to imputation. All reported results are on data imputed with a combined HapMap 3[28] reference panel of CEU and YRI.

Signals of selection in African-ancestry populations
Our tree-based method evaluates selection on a set of markers from multiple populations in two steps (see Methods). In the first, an unrooted tree of populations is estimated. This tree is intended to explain the observed amount of divergence between

each pair of populations. With three populations, this is a "star" shaped topology where each population is a leaf node connected to a single internal population by a branch. The length of this branch operates similar to Wright's F_{ST} and represents the genetic distance between the leaf population and the internal population (see Figure 2). Following our subtraction of European-related admixture, we estimated the tree for our three data sets in each of 10 bins based on the strength of background selection. For the tree connecting African Americans, Nigerians, and Gambians (see Figure 2b) we estimate branch lengths of 0.0005, 0.0006, and 0.0046. These are closely concordant to the pairwise results for F_{ST}.

Once the tree is estimated, we can evaluate a statistic for selection at every marker common to all datasets. This statistic enables resolution of the population subject to the selective pressure and can give additional power to detect loci under selection relative to pairwise comparisons.

Q-Q plots comparing observed and expected p-values indicate an excess of highly differentiated markers (Figure 3). The proportion of markers with $P < 0.0001$ is 0.0005. After excluding loci with genome-wide significant evidence of selection the proportion of markers with $P < 0.0001$ is 0.0002. This excess is suggestive of additional selected loci beyond the genome-wide significant signals we describe here. We note that genetic drift at rare and low frequency SNPs (MAF < 5%) is unlikely to be well described in our model and these SNPs are not included in the analysis. Our threshold for genome-wide significance in this analysis was $P < 1.67 \times 10^{-8}$ (see Methods).

A genome-wide significant signal (see Figure 4) at *CD36*[2, 24, 31] is present on both the Nigerian ($P=2.32 \times 10^{-09}$) and African-American ($P=7.05 \times 10^{-09}$) branches of the tree. Additionally, we note a highly suggestive signal for selection at the *HBB*[32, 50] locus ($P=6.15 \times 10^{-08}$) on chromosome 11. Selection at both of these loci has been previously detected using population differentiation between African populations ascertained based on malaria exposure [24]. The finding of selection at these loci in a genome-wide scan without ascertainment of populations further corroborates the power of our approach (see Table 3 for all signals).

Natural selection at *HBB* is likely due to the well-known association in which heterozygotes for the sickle cell trait HbAS (HbAS T) are protected against severe malaria[10]. We note that a study of unusual population differentiation between Han Chinese and Tibetans[14] also showed evidence of selection at the *HBB* locus. However, the most significantly differentiated marker in that analysis, rs10768683, and the most significantly differentiated marker in our analysis, rs2213169, are not polymorphic in any of the same HapMap populations. While we cannot rule out separate selective sweeps on the same variant, the absence of HbAS T allele in East Asia leads us to believe that separate selective events on separate causal variants is most consistent with this finding.

Genome-wide significant evidence of selection (see Figure 4) exists for *HLA* on chromosome 6, known to be heavily involved in human immunity and a well studied example of natural selection[7, 34, 35]. Peaks at *HLA* are observed on all three branches of the tree. However, our analysis of selection at *HLA* shows distinct sets of SNPs with significant evidence of selection on the Gambian, Nigerian and African-American branches of the tree. Specifically, there are unlinked SNPs which show strongest evidence of selection in different populations. The most significantly differentiated SNP

along the Gambian branch, rs28366191 ($P=6.3x10^{-16}$), is differentiated to a much lesser degree on either of the Nigerian or African-American branches ($P=2.5x10^{-4}$ and $P= 0.59$, respectively) or in a pairwise comparison of these populations ($P = 0.02$). Additionally, the Nigerian and African-American branches show significant evidence of selection at SNPs in the *HLA* region, for example rs2179915 ($P=1.48x10^{-9}$ and $P=2.45x10^{-10}$, respectively), which are not significant on the Gambian branch ($P=0.53$). This SNP was not significantly differentiated in a pairwise analysis of Gambians and African Americans ($P = 0.47$) indicating that selection likely took place on the Nigerian branch. This leaves multiple selective events as a parsimonious explanation of our findings at *HLA*.

We also observe a signal in the *HLA* at rs6901541 which is highly differentiated on all branches of the tree, $P=3.61x10^{-5}$, $6.37x10^{-10}$, and $1.71x10^{-6}$, for African American, Nigerian and Gambian branches, respectively. This SNP is also highly differentiated in all three pairwise analyses. We note that this is consistent with selection on multiple branches of the tree and further indicates the widespread nature of selection at the *HLA*.

We observe a suggestive signal, rs2920283 ($P = 1.1x10^{-7}$), on chromosome 8 within the protein-coding gene Prostate Stem Cell Antigen (*PSCA*). Further evidence of selection at this locus was obtained by analyzing additional populations (see below). A nonsynonymous SNP in *PSCA*, rs2294008, causing a 9 amino acid truncation of the protein, has been shown to be associated to both gastric and bladder cancers with $P = 8x10^{-11}$ and $P = 2.14x10^{-10}$, respectively[36, 37]. The marker with the most significant evidence of selection on the African-American branch, rs2920283, is in very high LD with the disease associated SNP ($r^2 > 0.85$). We note that rs2920283 is polymorphic in all of the populations studied here (see Table 3c) and those included in the Human Genome Diversity Project (see Figure 5). This indicates that the classical "selective sweep", in which a novel variant rises to high frequency under selection, is unlikely to apply. Instead, we posit that selection at *PSCA* is a case of selection on "standing variation" and an ideal candidate for a test based on population differentiation. We note that no Extended Haplotype Homozygosity[12] or integrated Haplotype Score[51] signal has been previously reported at this locus[1, 5].

For comparison purposes we implemented the LSBL statistic[29], which has been used to discover or validate loci under selection with associations to altitude response[13, 14], cystic fibrosis[52], skin pigmentation[53-55], and hair straightness[56], and ran it on our data (see Table S3 of Bhatia et al.[62]). The *HLA*, *HBB*, and *CD36* loci have statistics that rank in the top 0.01% (see Figure S2 of Bhatia et al.[62]). The *PSCA* locus has a statistic in the top 1%. However, many SNPs (nearly 10,000) rank in the top 1% and it is unclear which of these, if any, present significant evidence of selection.

We note that all reported loci are constrained to contain multiple highly differentiated markers, ruling out the possibility of spurious signals due to assay. While 2 markers 16 Mb apart on chromosome 16 achieved genome-wide significance, they were not reported because they did not satisfy this criteria.

Examining Additional Populations

In order to further explore evidence of selection at our implicated loci, we examined pairs of populations from HapMap3 that were closely related ($F_{ST} < 0.01$). We compared YRI to LWK ($F_{ST} = 0.0080$), TSI to CEU ($F_{ST} = 0.0039$), and JPT to the combined individuals

from CHB and CHD (F_{ST} = 0.0075)[28]. In this analysis we corroborated several published examples of natural selection including *LCT*[67] [MIM 603202] and *OCA2*[51] [MIM 611409] in Europeans, *KITLG*[58] [MIM 184745] in East Asians, and *CD36*[24] in Africans (see Table 4). Unsurprisingly, we observe that markers in *HLA*[34] are highly differentiated in all three pairwise analyses consistent with the role of *HLA* in immunity. While our comparison of African populations (LWK-YRI) does show a high degree of differentiation at the *HLA* and *CD36* loci (Table 4) we do not observe a signal at the *HBB* locus. This may be due to insufficient sample size or similar selection pressures in both populations. We note that this comparison is *not* independent of our tree-based analysis as both involve Yoruba populations.

We note the surprising finding of a high degree of differentiation between JPT and CHB+CHD at the *PSCA* locus (rs2928023, χ_1^2 =21.03, P=4.58x10^{-6}) and (rs2976397, χ_1^2 =24.95, P=5.88x10^{-7}). This is one of the strongest signals of selection in our analysis and corresponds to a 34% allele frequency difference between JPT and CHB+CHD. We note that this comparison is independent of the tree-based analysis because no population in East Asia was used in the tree (i.e. YRI) or to correct for European-related admixture (i.e. CEU). Independence allows us to sum the statistic for differentiation in East Asia with that obtained from the tree at any SNP and produce a χ_2^2 2 degree of freedom statistic. Doing so yields (rs2920283, χ_2^2 =48.12, P=3.56x10^{-11}), which remains genome-wide significant (P<5.56x10^{-9}) after correction for multiple hypotheses tested (see Methods).

We have also plotted allele frequencies at SNP rs2294008 in all of the populations included in HGDP[40] (Figure 5). There exist large differences in allele frequency throughout East Asia as well as Europe and South America. While the small sample sizes taken from each population make studies of differentiation underpowered, further studies may elucidate the underlying cause of the selective pressure by analyzing global allele frequency differences.

Population Differences By SNP Class

We analyzed coding and nonsynonymous SNPs for excessive differentiation similar to previous work[46]. We examined SNPs which were differentiated with P < 0.0001 on any branch of the tree and compared the number of nonsynonymous coding SNPs and genic SNPs to the number expected under neutrality. We observed 22 nonsynonymous coding SNPs differentiated to this degree, a 3.7-fold enrichment compared with expectations under neutrality (χ_1^2 =42.08, P=8.77x10^{-11}). However, several of these nonsynonymous variants were highly collocated—many occurring in the *HLA* region—and are unlikely to have been subject to independent selective events. Once we restricted to a single variant per locus, only 8 highly differentiated, nonsynonymous SNPs remained (χ_1^2 =0.7, P=0.40). We did not observe a statistically significant enrichment of genic SNPs (χ_1^2 =0.21, P=0.65).

A recent study of natural selection in sequence data[59] found that nonsynonymous coding sites were not enriched for excessive differentiation relative to synonymous sites. This is consistent with our findings. The authors of this study suggest that the "selective sweep" is an uncommon model of human evolution and that methods based on population

differentiation between closely related populations may be more powerful for detecting selection. We provide such a method.

Variation in functional status and strength of background selection has been shown to influence the effective population size and, therefore, genetic drift at a locus-specific level[60]. Specifically, background selection, often observed in known functional regions, tends to increase the rate of drift and increase the average differentiation at the locus. In our data we observed a difference in F_{ST} estimates (Table 5a) when computed using markers classified as genic or non-genic[46]. This trend was also apparent when we classified markers by the strength of background selection[47] at the locus (Table 5b) and was especially prevalent when we examined loci with the strongest evidence of background selection.

In order to verify that our results were not spurious signals due to variation in genetic drift across the genome[47], we repeated our analysis in separate bins according to the strength of background selection. Our results prior to (Table 3a) and after (Table 3b) correction for the strength of background selection at each locus are very similar. This would indicate that our results including the signal at *PSCA* are robust to this correction.

Discussion

We have examined population differentiation in a genome-wide fashion in three closely related African populations. Similar studies of population differentiation have been previously performed with some success[2, 17-19, 22-24], however, many of these have focused on continental populations with much larger genetic distance. While studies have examined closely related populations within Europe or Asia, such studies require the availability of data from large numbers of individuals. Now, as such data has become available we are able to apply this approach to closely related African populations. In addition to performing pairwise comparisons between closely related populations, we have developed a method of analysis based upon differentiation in a tree of populations.

The tree-based analysis that we use is somewhat comparable to the Population Branch Statistic (PBS) described by Yi et. al.[14] and the Locus Specific Branch Length (LSBL) [13, 29, 52-56]. The PBS seeks to estimate the time since divergence from a central node using SNP-specific F_{ST} and has been shown to have power to detect recent population-specific natural selection. One challenge associated with using the PBS/LSBL is that the null distribution of these statistics is not well-defined. Thus, significance can be assessed using extensive simulations according to a specific demographic history or a simple ranking of results. When implemented on our dataset the LSBL replicated clear peaks at *HLA*, *HBB*, and *CD36*; however, no other significant peaks were observed.

Our results provided genome-wide significant or suggestive corroboration of several known loci including *HLA*, *HBB* and *CD36*. We identified a new genome-wide significant locus in *PSCA*. Our most significantly differentiated marker is tightly linked to a marker with prior, genome-wide significant associations to both gastric and bladder cancer. Additionally, our evidence suggests that multiple, independent selective events have occurred in the *HLA* region.

Several questions of interest arise from this work. Notably, imputation of the *HLA* genotypes of individuals in our datasets would allow us to pinpoint specific alleles under selection. By analyzing the various *HLA* alleles individually for population differentiation, it may be possible to infer which *HLA* alleles are being pushed to high frequency. Understanding this may give further insight into infectious disease resistance. Similarly, understanding the selective pressure acting at *PSCA* is a question of interest. Analysis of data specific to infectious disease and other possible drivers of selection[61] may yield insight into the environmental pressure responsible for selection at this locus.

Acknowledgements

This work was funded by NIH grant R01 HG005224 (B.P., S.P., A.L.P.) , by NIH grant RC1 GM091332 (N.P., D.R., J.G.W.) and by grant T32 HG002295 from the National Human Genome Research Institute (NHGRI) (G.B.) and NIH fellowship 5T32ES007142-27 (N.Z.), using data from NHLBI's Candidate Gene Association Resource (CARe) project. C.H., D.R., N.P., and A.T. were supported by NIH/NHGRI grant U01 HG004726-01.

We acknowledge the contributions of the participants and investigators of NHLBI's CARe consortium (contract number HHSN268200960009C). Funding information for CARe and its parent cohorts can be found at http://public.nhlbi.nih.gov/GeneticsGenomics/home/care.aspx.

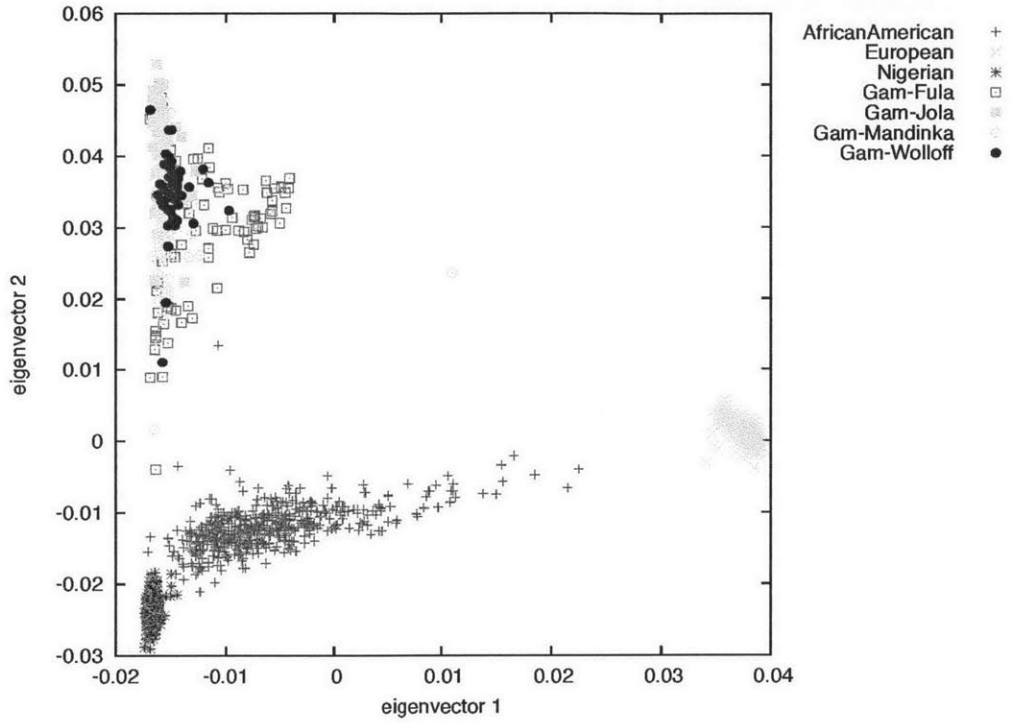
This study makes use of data generated by MalariaGEN. A full list of the investigators who contributed to the generation of the data is available from www.MalariaGEN.net. Funding for this project was provided by the Foundation for the National Institutes of Health and the Wellcome Trust. The funding for this project comes through the Grand Challenges in Global Health Initiative.

Web Resources

http://www.hsph.harvard.edu/faculty/alkes-price/software/ (TreeSelect software)
http://www.hsph.harvard.edu/faculty/alkes-price/software/ (EIGENSOFT software)
http://www.sph.umich.edu/csg/abecasis/MACH/ (MaCH software)
http://www.omim.org (Online Mendelian Inheritance in Man)

Figures

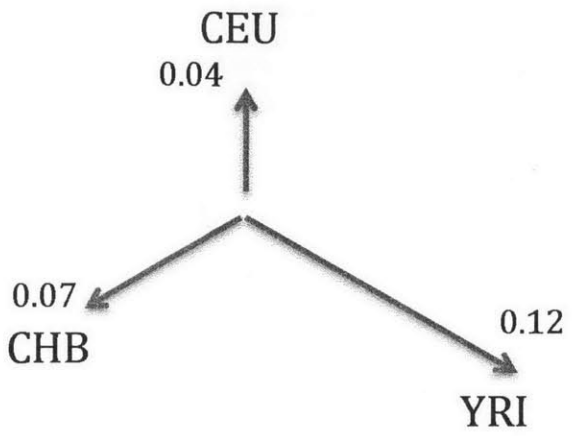
Figure 1



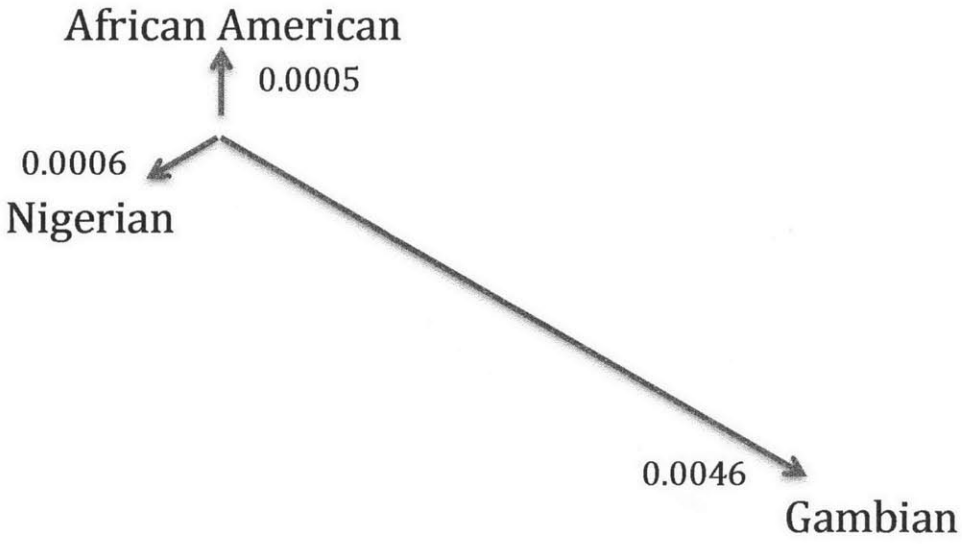
PCA Analysis of Population Structure. This analysis of population structure in our main data sets shows Europeans and Nigerians forming separate tight clusters. African Americans form a cline between the Nigerian and European clusters indicative of varying degrees of European ancestry. The Gambian samples are separated from the Nigerians on PC2, form separate but overlapping clusters, and show evidence of European-like admixture within the Fula subpopulation.

Figure 2

a)



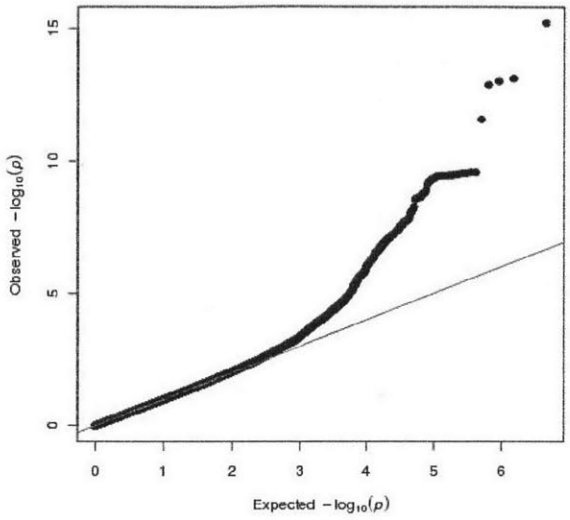
b)



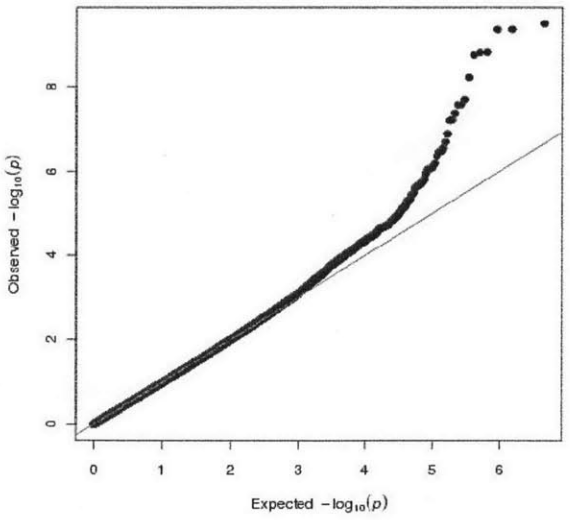
Tree Estimates From Sample Data. a) This tree was estimated using unrelated individuals from the YRI, CEU and CHB populations sampled as part of the International HapMap Project Phase III. The branch lengths show strong concordance with estimated pairwise values for F_{st}. b) This tree was estimated using our main data sets of African American, Nigerian and Gambian samples after accounting for significant European-like admixture in the African-American and Gambian datasets. We note that the second tree is scaled approximately by a factor of 100 with respect to the first. The values quoted are based on genome-wide average estimates of F_{st}.

Figure 3

a)

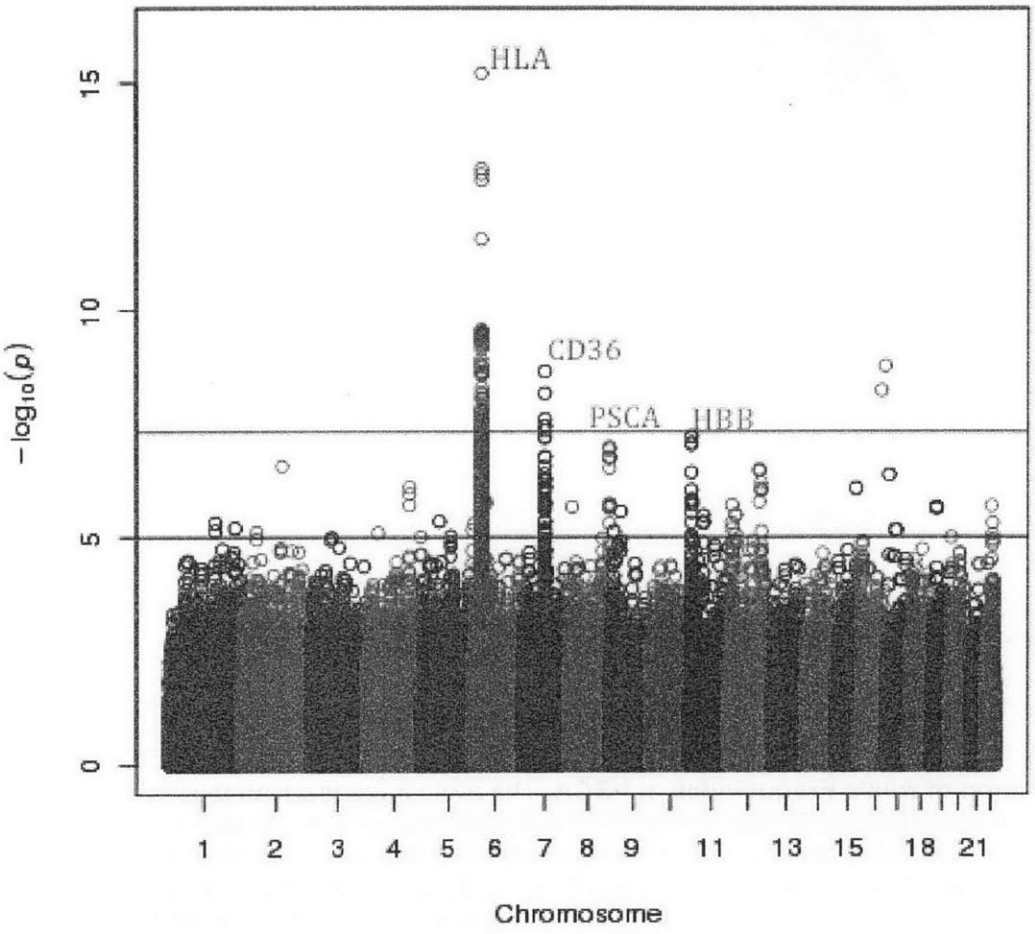


b)



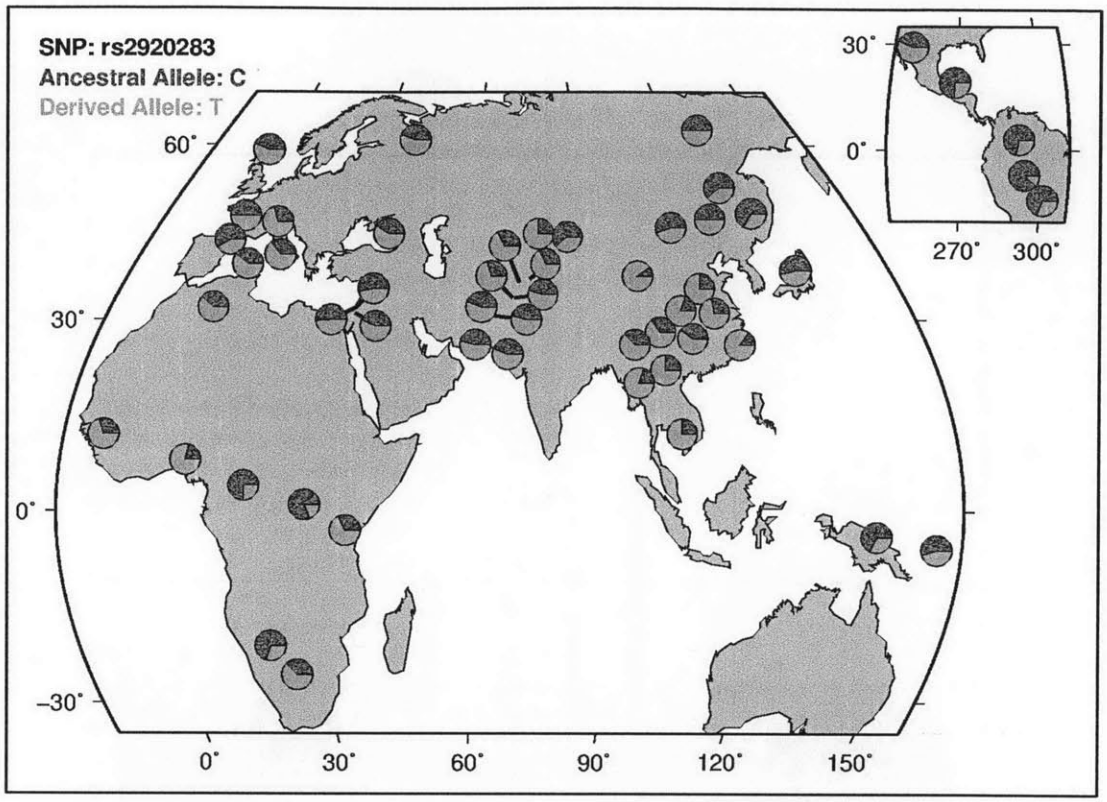
Q-Q Plots of Population Differentiation in Africans. a) We compare the actual and expected distribution of selection statistics. The red line represents expectation under neutrality. It is clear that a "fat-tail" of highly differentiated markers exists, consistent with multiple selective events. b) We repeated the analysis after removing the 5 Mb regions containing each of our most significant SNPs and still observe a fat-tail of highly differentiated markers.

Figure 4



Genome-Wide Population Differentiation in Africans. All values are reported after correcting for variation in F_{st} according to quantity of background selection. We note genome-specific peaks in the HLA locus on chromosome 6, and CD36 on chromosome 7. HLA has a major role in immunity with multiple prior disease associations, and CD36 is known for its role in malaria resistance. We also observe a highly suggestive peak at PSCA (chromosome 8) tightly linked to a protein-altering variant with prior associations to gastric and bladder cancers. The highly suggestive signal at HBB is unsurprising given its role in malaria resistance. HLA, HBB and CD36 have been previously reported targets of selection.

Figure 5



Distribution of Allele Frequencies at PSCA. The allele frequencies of the most differentiated SNP at PSCA are plotted in 52 distinct ethnic groups genotyped as part of the Human Genome Diversity Project. We note the high degree of differentiation in East Asia, Africa and South America (insert, upper right). While small samples sizes of these populations hinder analysis of selection, analysis of selection pressures in each of these populations may elucidate the cause of the large allele frequency differences at PSCA.

Tables

Table 1

a)

	African American	Nigerian	Gambian
African American		0.0074 ± 5.2E-5	0.0072 ± 5.1E-5
Nigerian			0.0059 ± 4.6E-5

b)

	African American	Nigerian	Gambian
African American		0.0011 ± 1.1E-5	0.0045 ± 3.3E-5
Nigerian			0.0058 ± 4.6E-5

Pairwise F_{st} Between African Populations. Here we combined all of the Gambian samples and compared these with the African American and Nigerian samples. We list both the estimate and standard error of the estimate for F_{st}. In (a) we have not accounted for significant European-like admixture in the Gambians and African Americans. In (b) we show the values after accounting for admixture by subtracting European allele frequencies weighted by admixture proportion. The large decrease in F_{st} between African Americans and both Nigerians and Gambians is expected to increase our power to detect signals of selection. While the drop in F_{st} between Nigerians and Gambians is small, this is expected due to the small admixture proportion estimated.

Table 2

	Mandinka	Jola	Fula	Wolloff
Mandinka		0.0012	0.0030	0.0005
Jola			0.0051	0.0020
Fula				0.0027

Pairwise F_{st} Between Gambian Subpopulations. We note that the values for F_{st} do not account for significant European-like admixture within the Fula subpopulation and these values could potentially be reduced further. With these low values for F_{st} exploring population differentiation within the Gambia may be a fruitful endeavor. However, such a study may require larger samples than we have available.

Table 3

a)

Chr	Gene or Region	SNP	Position	P-values		
				African American	Nigerian	Gambian
6	HLA	rs28366191	32472168	0.62	3.62E-04	1.89E-15
6	HLA	rs6901541	32550239	4.28E-05	1.29E-09	2.75E-06
6	HLA	rs2179915	33173712	2.45E-10	1.48E-09	0.53
7	CD36	rs12721454	79678275	6.82E-09	1.76E-07	0.97
7	CD36	rs513740	79872884	5.64E-08	4.03E-09	0.05
8	PSCA	rs2920283	143754039	1.66E-07	2.60E-06	0.95
11	HBB	rs7936387	5256204	3.15E-05	5.99E-08	1.05E-03

b)

Chr	Gene or Region	SNP	Position	P-values		
				African American	Nigerian	Gambian
6	HLA	rs28366191	32472168	0.59	2.51E-04	6.25E-16
6	HLA	rs6901541	32550239	3.61E-05	6.37E-10	1.71E-06
6	HLA	rs2179915	33173712	3.16E-10	1.78E-09	0.52
7	CD36	rs12721454	79678275	7.05E-09	1.76E-07	0.96
7	CD36	rs513740	79872884	3.78E-08	2.32E-09	0.05
8	PSCA	rs2920283	143754039	1.06E-07	1.88E-06	0.96
11	HBB	rs7936387	5256204	4.06E-05	6.15E-08	9.53E-04

c)

Chr	Gene or Region	SNP	Position	Allele Frequencies		
				African American	Nigerian	Gambian
6	HLA	rs28366191	32472168	0.08	0.05	0.28
6	HLA	rs6901541	32550239	0.31	0.45	0.14
6	HLA	rs2179915	33173712	0.42	0.59	0.46
7	CD36	rs12721454	79678275	0.25	0.39	0.31
7	CD36	rs513740	79872884	0.27	0.41	0.23
8	PSCA	rs2920283	143754039	0.37	0.24	0.32
11	HBB	rs7936387	5256204	0.17	0.28	0.08

Loci with Evidence of Selection in African Populations. We report the most significant SNPs in loci that showed genome-wide significant or suggestive evidence of natural selection. All SNPs are imputed. Table 4(a) shows the P-values for each SNP without correcting for background selection at the locus and Table 4(b) shows the results after the correction. We note the relative insensitivity of our results to correcting for evidence of background selection. Table 4(c) lists the allele frequencies of the highly differentiated SNPs.

Table 4

a)

Chr	Gene or Region	SNP	Position	P-values		
				JPT-CH	LWK-YRI	CEU-TSI
2	LCT	rs6754311	136424452	N/A	0.60	2.03E-15
3	SLC9A9/Corf58	rs7649861	145653390	0.65	2.68E-07	0.04
6	HLA	rs7745413	30023448	1.35E-07	0.15	0.15
6	HLA	rs28366191	32472168	0.08	0.23	0.69
6	HLA	rs6901541	32550239	0.13	0.19	0.64
6	HLA	rs2179915	33173712	N/A	1.08E-03	0.87
7	CD36	rs12721454	79678275	N/A	2.63E-05	0.40
7	CD36	rs513740	79872884	0.11	9.74E-04	0.65
7	CD36	rs6944302	79942827	N/A	7.47E-07	0.09
8	PSCA	rs2976397	143761615	5.87E-07	0.01	0.60
8	PSCA	rs2920283	143754039	4.58E-06	0.01	0.75
11	HBB+HBG2	rs7936387	5256204	N/A	0.66	N/A
11	OPCML	rs11223548	133036865	8.90E-07	0.90	N/A
12	KITLG	rs11104947	87467111	4.88E-07	N/A	0.43
15	OCA2	rs12913832	26039213	N/A	N/A	1.42E-08

b)

Gene or Region	JPT-CH		LWK-YRI		CEU-TSI	
	P-value	SNP	P-value	SNP	P-value	SNP
HLA	1.35E-07	rs7745413	9.30E-05	rs7905	3.39E-06	rs2256175
CD36	-	-	7.47E-07	rs6944302	-	-
HBB+HBG2	-	-	-	-	-	-
PSCA	5.87E-07	rs2976397	-	-	-	-

Loci with Evidence of Selection in Other Comparisons. a) We report all highly differentiated SNPs with strong or suggestive evidence for selection ($P < 10^{-6}$). We see several well studied examples of selection such as LCT, and OCA2 in Europeans, KITLG in East Asians and CD36 in Africans. However, several markers significant in our original analysis of African populations do not appear significant in this analysis. This may be because of the small sample size taken from each of the HapMap3 populations. b) To test concordance with the signals observed in our analysis of Africans, the regions surrounding (2.5 Mb on either side) the most highly differentiated markers in our analysis are analyzed here. We report the most significant P-value in the region provided that $P < 10^{-4}$. Surprisingly no SNP appears differentiated with $P << 10^{-4}$ in our analysis of the HBB region in Yoruba (YRI) and Luhya (LWK). This may be due to small sample size or an absence of different malaria pressure between these populations.

Table 5

a)

	AA-Nigerian	AA-Gambian	Nigerian-Gambian
Genic	0.0011	0.0045	0.0061
Nongenic	0.0011	0.0044	0.0060

b)

B	AA-Nigerian	AA-Gambian	Nigerian-Gambian
0.0-0.1	0.0017	0.0069	0.0100
0.1-0.2	0.0011	0.0051	0.0070
0.2-0.3	0.0011	0.0052	0.0065
0.3-0.4	0.0011	0.0050	0.0066
0.4-0.5	0.0012	0.0047	0.0065
0.5-0.6	0.0011	0.0046	0.0064
0.6-0.7	0.0011	0.0046	0.0063
0.7-0.8	0.0011	0.0044	0.0060
0.8-0.9	0.0011	0.0043	0.0059
0.9-1.0	0.0010	0.0042	0.0057

Pairwise F_{st} Estimated Using Partitioned Sets of SNPs. a) Pairwise estimates of F_{st} calculated using genic and nongenic SNPs. b) Pairwise estimates of F_{st} calculated after binning SNPs according to the strength of background selection at the locus as quantified by the B statistic of McVicker and colleagues. The trend observed in (a) is magnified when looking at B values between 0 and 0.1 with respect to the remainder of the genome. Because of this difference, we performed all subsequent analysis separately for each bin of *B*.

References

1. Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J. et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832-837.

2. The International HapMap Project. (2005). A haplotype map of the human genome. Nature 437, 1299-1320.

3. Ko, W., Kaercher, K.A., Giombini, E., Marcatili, P., Froment, A., Ibrahim, M., Lema, G., Nyambo, T.B., Omar, S.A., Wambebe, C. et al. (2011). Effects of Natural Selection and Gene Conversion on the Evolution of Human Glycophorins Coding for MNS Blood Polymorphisms in Malaria-Endemic African Populations. Am. J. Hum. Genet. 88, 741-754.

4. Hamblin, M.T., Di Rienzo, A. (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. 66, 1669-1679.

5. 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. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature 449, 913-918.

6. Genovese, G., Friedman, D.J., Ross, M.D., Lecordier, L., Uzureau, P., Freedman, B.I., Bowden, D.W., Langefeld, C.D., Oleksyk, T.K., Uscinski Knob, A.L. et al. (2010). Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans. Science 329, 841-845.

7. de Bakker, P.I., McVean, G., Sabeti, P.C., Miretti, M.M., Green, T., Marchini, J., Ke, X., Monsuur, A.J., Whittaker, P., Delgado, M. et al. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat. Genet. 38, 1166-1172.

8. Lewinsohn, D.A., Winata, E., Swarbrick, G.M., Tanner, K.E., Cook, M.S., Null, M.D., Cansler, M.E., Sette, A., Sidney, J., Lewinsohn, D.M. (2007). Immunodominant tuberculosis CD8 antigens preferentially restricted by HLA-B. PLoS Pathog. 3, 1240-1249.

9. Fellay, J., Shianna, K.V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A. et al. (2007). A whole-genome association study of major determinants for host control of HIV-1. Science 317, 944-947.

10. Kwiatkowski, D.P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. Am. J. Hum. Genet. 77, 171-192.

11. Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., Clark, A.G. (2007). Recent and ongoing selection in the human genome. Nat. Rev. Genet. 8, 857-868.

12. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., Lander, E.S. (2006). Positive natural selection in the human lineage. Science 312, 1614-1620.

13. Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J.M., Mei, R., Scherer, S.W., Julian, C.G., Wilson, M.J., Lopez Herraez, D. et al. (2010). Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet. 6.

14. Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S. et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329, 75-78.

15. Simonson, T.S., Yang, Y., Huff, C.D., Yun, H., Qin, G., Witherspoon, D.J., Bai, Z., Lorenzo, F.R., Xing, J., Jorde, L.B. et al. (2010). Genetic evidence for high-altitude adaptation in Tibet. Science 329, 72-75.

16. Novembre, J., Di Rienzo, A. (2009). Spatial patterns of variation due to natural selection in humans. Nat. Rev. Genet. 10, 745-755.

17. The International HapMap Project. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851-861.

18. Campbell, C., Sampas, N., Tsalenko, A., Sudmant, P., Kidd, J., Malig, M., Vu, T., Vives, L., Tsang, P., Bruhn, L. et al. (2011). Population-Genetic Properties of Differentiated Human Copy-Number Polymorphisms. The American Journal of Human Genetics 88, 317-332.

19. Akey, J.M., Zhang, G., Zhang, K., Jin, L., Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 12, 1805-1814.

20. Holsinger, K.E., Weir, B.S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. Nat. Rev. Genet. 10, 639-650.

21. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O. et al. (2009). The genetic structure and history of Africans and African Americans. Science 324, 1035-1044.

22. Price, A.L., Helgason, A., Palsson, S., Stefansson, H., St Clair, D., Andreassen, O.A., Reich, D., Kong, A., Stefansson, K. (2009). The impact of divergence time on the nature of population structure: an example from Iceland. PLoS Genet. 5, e1000505.

23. Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., Gong, X., Wang, H., Shen, Y., Pan, X. et al. (2009). Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am. J. Hum. Genet. 85, 762-774.

24. Ayodo, G., Price, A.L., Keinan, A., Ajwang, A., Otieno, M.F., Orago, A.S., Patterson, N., 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. 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. The American Journal of Human Genetics 74, 979.

26. Jallow, M., Teo, Y.Y., Small, K.S., Rockett, K.A., Deloukas, P., Clark, T.G., Kivinen, K., Bojang, K.A., Conway, D.J., Pinder, M. et al. (2009). Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat. Genet. 41, 657-665.

27. Thye, T., Vannberg, F.O., Wong, S.H., Owusu-Dabo, E., Osei, I., Gyapong, J., Sirugo, G., Sisay-Joof, F., Enimil, A., Chinbuah, M.A. et al. (2010). Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. Nat. Genet. 42, 739-741.

28. The International HapMap 3 Project. (2010). Integrating common and rare genetic variation in diverse human populations. Nature 467, 52-58.

29. Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akey, J.M., Jones, K.W. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum.Genomics 1, 274-286.

30. Akey, J.M. (2009). Constructing genomic maps of positive selection in humans: Where do we go from here? Genome Research 19, 711-722.

31. Fry, A.E., Ghansa, A., Small, K.S., Palma, A., Auburn, S., Diakite, M., Green, A., Campino, S., Teo, Y.Y., Clark, T.G. et al. (2009). Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. Human Molecular Genetics 18, 2683-2692.

32. Currat, M., Trabuchet, G., Rees, D., Perrin, P., Harding, R.M., Clegg, J.B., Langaney, A., Excoffier, L. (2002). Molecular Analysis of the [beta]-Globin Gene Cluster in the Niokholo Mandenka Population Reveals a Recent Origin of the [beta]S Senegal Mutation. The American Journal of Human Genetics 70, 207-223.

33. Hedrick, P.W. (2011). Population genetics of malaria resistance in humans. Heredity.

34. Hedrick, P.W., Thomson, G. (1983). Evidence for Balancing Selection at HLA. Genetics 104, 449-456.

35. Cao, K., Moormann, A.M., Lyke, K.E., Masaberg, C., Sumba, O.P., Doumbo, O.K., Koech, D., Lancaster, A., Nelson, M., Meyer, D. et al. (2004). Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. Tissue Antigens 63, 293-325.

36. Sakamoto, H., Yoshimura, K., Saeki, N., Katai, H., Shimoda, T., Matsuno, Y., Saito, D., Sugimura, H., Tanioka, F., Kato, S. et al. (2008). Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. Nat. Genet. 40, 730-740.

37. Wu, X., Ye, Y., Kiemeney, L.A., Sulem, P., Rafnar, T., Matullo, G., Seminara, D., Yoshida, T., Saeki, N., Andrew, A.S. et al. (2009). Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. Nat. Genet. 41, 991-995.

38. Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G.K., Tandon, A., Kao, W.H.L., Ruczinski, I., Fornage, M., Siscovick, D.S., Zhu, X. et al. (2011). Enhanced Statistical Tests for

GWAS in Admixed Populations: Assessment using African Americans from CARe and a Breast Cancer Consortium. PLoS Genet 7, e1001371.

39. Lettre, G., Palmer, C.D., Young, T., Ejebe, K.G., Allayee, H., Benjamin, E.J., Bennett, F., Bowden, D.W., Chakravarti, A., Dreisbach, A. et al. (2011). Genome-Wide Association Study of Coronary Heart Disease and Its Risk Factors in 8,090 African Americans: The NHLBI CARe Project. PLoS Genet 7, e1001300.

40. Cann, H.M., De Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G.B., Friedlaender, J.S., Groot, H., Gurwitz, D., Jenkins, T., Herrera, R.J., Huang, X., Kidd, J., Kidd, K.K., Langaney, A., Lin, A.A., Mehdi, S.Q., Parham, P., Piazza, A., Pistillo, M.P., Qian, Y., Shu, Q., Xu, J., Zhu, S., Weber, J.L., Greely, H.T., Feldman, M.W., Thomas, G., Dausset, J., Cavalli-Sforza,L.L. (2002). A human genome diversity cell line panel Science 296, 261-262.

41. Clayton, D.G., Walker, N.M., Smyth, D.J., Pask, R., Cooper, J.D., Maier, L.M., Smink, L.J., Lam, A.C., Ovington, N.R., Stevens, H.E. et al. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat. Genet. 37, 1243-1246.

42. Nicholson, G., Smith, A.V., Jonsson, F., Gustafsson, A., Stefansson, K., Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64, 695-715.

43. The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661-678.

44. Cavalli-Sforza, L.L., Edwards, A.W. (1967). Phylogenetic analysis. Models and estimation procedures. Am. J. Hum. Genet. 19, 233-257.

45. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 5, e1000519.

46. Barreiro, L.B., Laval, G., Quach, H., Patin, E., Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. Nat. Genet. 40, 340-345.

47. McVicker, G., Gordon, D., Davis, C., Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 5, e1000471.

48. Li, Y., Willer, C.J., Ding, J., Scheet, P., Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34, 816-834.

49. Patterson, N., Price, A.L., Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. 2, e190.

50. Wainscoat, J.S., Hill, A.V.S., Boyce, A.L., Flint, J., Hernandez, M., Thein, S.L., Old, J.M., Lynch, J.R., Falusi, A.G., Weatherall, D.J. et al. (1986). Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. Nature 319, 491-493.

51. Voight, B.F., Kudaravalli, S., Wen, X., Pritchard, J.K. (2006). A Map of Recent Positive Selection in the Human Genome. PLoS Biol 4, e72.

52. Mattiangeli, V., Ryan, A., McManus, R., Bradley, D. (2006). A genome-wide approach to identify genetic loci with a signature of natural selection in the Irish population. Genome Biol. 7, R74.

53. Norton, H.L., Kittles, R.A., Parra, E., McKeigue, P., Mao, X., Cheng, K., Canfield, V.A., Bradley, D.G., McEvoy, B., Shriver, M.D. (2007). Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians. Molecular Biology and Evolution 24, 710-722.

54. McEvoy, B., Beleza, S., Shriver, M.D. (2006). The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model. Human Molecular Genetics 15, R176-R181.

55. Edwards, Melissa AND Bigham, Abigail AND Tan, Jinze AND Li, Shilin AND Gozdzik, Agnes AND Ross, Kendra AND Jin, Li AND Parra,Esteban J. (2010). Association of the OCA2 Polymorphism His615Arg with Melanin Content in East Asian Populations: Further Evidence of Convergent Evolution of Skin Pigmentation. PLoS Genet 6, e1000867.

56. Medland, S.E., Nyholt, D.R., Painter, J.N., McEvoy, B.P., McRae, A.F., Zhu, G., Gordon, S.D., Ferreira, M.A.R., Wright, M.J., Henders, A.K. et al. (2009). Common Variants in the Trichohyalin Gene Are Associated with Straight Hair in Europeans. The American Journal of Human Genetics 85, 750.

57. Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., Hirschhorn, J.N. (2004). Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. The American Journal of Human Genetics 74, 1111-1120.

58. Williamson, S.H., Hubisz, M.J., Clark, A.G., Payseur, B.A., Bustamante, C.D., Nielsen, R. (2007). Localizing Recent Adaptive Evolution in the Human Genome. PLoS Genet 3, e90.

59. Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., 1000 Genomes Project, Sella, G., Przeworski, M. (2011). Classic Selective Sweeps Were Rare in Recent Human Evolution. Science 331, 920-924.

60. Weir, B.S., Cardon, L.R., Anderson, A.D., Nielsen, D.M., Hill, W.G. (2005). Measures of human population structure show heterogeneity among genomic regions. Genome Res. 15, 1468-1476.

61. Hancock, A.M., Witonsky, D.B., Gordon, A.S., Eshel, G., Pritchard, J.K., Coop, G., Di Rienzo, A. (2008). Adaptations to Climate in Candidate Genes for Common Metabolic Disorders. PLoS Genet 4, e32.

62. 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.

Appendix

Neutral Simulations

We simulated allele frequencies from a pair of populations to verify that this statistic follows the correct null distribution. In order to do this, we chose a variety of starting allele frequencies, f_s, and values for F_{ST}. For each f_s and F_{ST} we sampled pairs of allele frequencies from a normal distribution with mean f_s and variance given by $2F_{ST}$. We then estimated F_{st} from the generated samples and computed the statistic for each pair of sample allele frequencies. In doing this we notice inflation of the χ^2 statistic for small values of f_s. However, we note that this inflation is very small with respect to the fat tail observed on real data and is negligible for the allele frequencies of the SNPs that we report to be showing a signal of selection (See Table S1 of Bhatia et al.[62]).

Locus Specific Branch Length

The Locus Specific Branch Length generates a statistic for population differentiation on each of the branches of a tree of three populations. This method assumes that F_{ST} statistics are additive and assesses the branch specific F_{ST} for each population. Specifically, given 3 populations, three pairwise F_{ST} statistics ($F_{ST}^{A,B}, F_{ST}^{B,C}, F_{ST}^{A,C}$) can be computed for each marker. Then, each of the branch-specific F_{ST} statistics can be calculated by solving a system of equations giving

$$F_{ST}^{A} = \frac{F_{ST}^{A,B} + F_{ST}^{A,C} - F_{ST}^{B,C}}{2}$$

$$F_{ST}^{B} = \frac{F_{ST}^{A,B} + F_{ST}^{B,C} - F_{ST}^{A,C}}{2}$$

$$F_{ST}^{C} = \frac{F_{ST}^{B,C} + F_{ST}^{A,C} - F_{ST}^{A,B}}{2}$$

However, this method is applicable specifically to the case of three populations. Once these statistics are computed significance is assessed by ranking. Thus, LSBL can not provide evidence of genome-wide significance.

Chapter 4: Genome-wide scan of 29,141 African Americans finds no evidence of selection since admixture

We scanned through the genomes of 29,141 African Americans, searching for loci where the average proportion of African ancestry deviates significantly from the genome-wide average. We failed to find any genome-wide significant deviations, and conclude that any selection in African Americans since admixture is sufficiently weak that it falls below the threshold of our power to detect it using a large sample size. These results stand in contrast to the findings of a recent study with 15 times fewer samples that reported six loci with significant deviations; we show that the discrepancy is likely due to insufficient correction for multiple hypothesis testing in the previous study. We also tested for polymorphic sites in the genome that exhibit greater population differentiation between African Americans and Nigerian Yoruba than would be expected in the absence of natural selection. Four such loci were previously shown to be genome-wide significant and likely to be affected by selection, but we show that most of the 10 additional loci reported in a recent study are likely to be false positives. Additionally, the most parsimonious explanation for the loci that have significant evidence of unusual differentiation in frequency between Nigerians and Africans Americans is selection in Africa prior to their forced migration to the Americas.

Introduction

Admixed populations such as African Americans and Latinos are formed by the mixing of populations from different continents. Alleles that are highly differentiated between the ancestral populations and advantageous in the admixed population are expected to rise in frequency after admixture, causing a deviation in local ancestry compared with the genome-wide average [1]. This signal can be used to test for selection since admixture.

A recent study applied this approach to 1,890 African Americans [2]. The study reported six loci as likely targets of natural selection since admixture. However, that study used a genome-wide significance threshold of $P < 2.7 \times 10^{-3}$, correcting for ~20 hypotheses tested. Based on the scale of admixture linkage disequilibrium in African Americans, a more appropriate threshold would be $P < 10^{-5}$, correcting for 5,000 hypotheses tested as recommended by Seldin et al. (2011).

To revisit the issue of whether there is evidence of natural selection since admixture in African Americans, we scanned through the genomes of 29,141 African Americans, using exactly the same genotyping data set that had previously been used to study the landscape of recombination (meiotic crossover) in African Americans [3]. This is the largest sample size analyzed to date for this type of study. Using a genome-wide significance threshold of $P < 10^{-5}$, we find no genome-wide significant signals of selection since admixture. The six previously reported loci do not attain nominal significance ($P < 0.05$), suggesting that they are false positives due to insufficient correction for multiple tests in the previous study.

We also evaluated the 14 signals of unusual population differentiation between African Americans and Yoruba reported by Jin et al. (2012). Four of these loci were previously shown to be genome-wide significant [4,5]. However, we show that most of the 10 remaining loci are likely to be false positives due to biases that arise when using the Weir and Cockerham (1984) estimator of F_{ST} to compare two populations of very unequal sample size, or due to an insufficient correction for multiple testing. Additionally, at loci with robust signals of selection, the selection is most likely to have occurred within Africa, prior to the arrival of Africans in the Americas. Thus, any conclusions of selection since the arrival of Africans in the Americas should be viewed with caution, and indeed, at present no unambiguous examples of such selection have been empirically documented.

Results

Genome-wide scan of 29,141 African Americans

We performed an admixture scan for unusual deviations in local ancestry in 29,141 African Americans from five cohorts, genotyped on three different platforms (see Methods). We used the HAPMIX software [6] to infer local ancestry in each individual and averaged local ancestry estimates across individuals (see Methods). To search for signals of selection since admixture, we computed the difference between the average local ancestry estimate at each locus and the genome-wide average, divided by the empirical standard deviation in local ancestry estimates across SNPs. It is important to divide by the empirical standard deviation, rather than by the theoretical standard deviation expected if all individuals are independent, as in practice there may be cryptic relatedness among samples—as well as systematic error in the ancestry inference—that

will inflate the variance across loci compared with what is theoretically expected (see Methods).

The average genome-wide estimate of European ancestry over all samples in the dataset is 0.204 with a standard deviation of 0.117 across individuals and 0.0036 across SNPs. We considered any deviation in local ancestry greater than 4.42 s.d (i.e. greater than 0.0154) to be genome-wide significant, corresponding to a threshold of $P < 10^{-5}$ (correcting for 5,000 hypotheses tested) as recommended by Seldin et al. (2011). No locus achieved this threshold of genome-wide significance (see Figure 1).

A previous study of selection in 1,890 African Americans [2] reported six loci that passed a (less stringent) significance threshold of $P < 2.7 \times 10^{-3}$ (equivalent to a Bonferroni correction for ~20 hypotheses tested). The six loci did not replicate at nominal significance ($P < 0.05$) in our much larger dataset (see Table 1 and Figure 1), and are likely to be false positives due to an insufficient correction for multiple tests in the previous study. For 5 of the 6 loci in Table 1, the deviation that we observe has the same sign as the deviation reported by Jin et al. This could be due either to statistical chance (P=0.11; 1-sided Fisher's exact test) or small systematic deviations in local ancestry inference that are correlated between the two analyses (see Supplementary Note of Bhatia et al.[25]). In either case, our results show that the proportion of African ancestry at these six loci is not likely to have been strongly affected by natural selection since admixture.

Inferring selection using allele frequency differences

Studies of selection often rank single SNP estimates of F_{ST} and report the most highly differentiated SNPs as signals of selection [2,7-11]. These estimates are most often produced using the Weir and Cockerham (1984) (WC) F_{ST} estimator (see Methods). However, a concern with the use of the WC estimator for this application is that estimates can depend on the sample sizes used, potentially resulting in overestimates of the degree of differentiation at single SNPs [12]. In the situation most prone to overestimation, which would be a study of rare variants with large differences in sample sizes between populations, greater than 99% of the highest single SNP F_{ST} estimates would be expected to be the result of inflation due to unequal sample sizes (Figure S1 of Bhatia et al. (2013)). On the other hand, the Hudson estimator [12,13], which is a simple average of the population-specific estimators of Weir and Hill (2002), does not have this bias.

We tested the magnitude of inflation of WC estimates in real data by reanalyzing the most highly differentiated SNPs reported in a recent analysis of this type [2] (see Table 2). This study compared African segments of 1,890 African Americans (AAF) and 113 Yoruba (YRI) at SNPs with MAF > 5%, and reported 40 SNPs—the 99.99th percentile of 401,559 SNPs tested—that have F_{ST} greater than 0.0452. These 40 SNPs are clustered into 14 regions, of which 10 are previously unreported targets of natural selection and 4 were reported as genome-wide significant in the parallel study of [4] (or nearly genome-wide significant in the case of HBB, a previously identified target of selection [5]). Of the 10 novel signals, 9 produce lower estimates when we used the Hudson estimator and 3 fall below the Jin et al. threshold ($F_{ST} > 0.0452$) (see Table 2). We note that the 99.99th percentile of F_{ST} could change when switching from the WC estimator to the Hudson estimator. In our analysis, the magnitude of this change was smaller than the decreases

observed (see Methods), suggesting that inflated WC F_{ST} estimates may lead to false positive signals of selection.

In addition to issues of F_{ST} estimation, studies that simply rank the most highly differentiated SNPs between populations are unable to evaluate genome-wide significance of reported signals. Model-based approaches, on the other hand, [4,5,14,15] can formally assess genome-wide significance and are robust to the biases of the WC F_{ST} estimator at single SNPs. In general, studies that use a model-based approach are well powered if sample sizes are much larger than $1/F_{ST}$ [4], as both F_{ST} and sampling noise contribute to normal variation in allele frequency differences. In the Jin et al. comparison, the sample size of Yoruba (n=113), is much smaller than the reciprocal of F_{ST} between AAF and YRI ($1/F_{ST}$ = 1429). Thus, sampling variation is expected to dominate the variance of the distribution of allele frequency differences. This may contribute to the fact that, when re-evaluated using a model-based approach [4,5], none of the reported SNPs achieves genome-wide significance ($P < 5x10^{-8}$) (see Table 2).

We re-examined the statistical significance of the 10 novel loci reported by Jin et al. in the Bhatia et al. (2011) dataset, which included 6,209 African Americans and 756 Yoruba. (Extending the analysis to all 29,141 African Americans in the current study yields very similar results, as the Yoruba sample size is the limiting factor.) The Bhatia et al. data includes 9 of these 10 loci, and only 4 of the 9 loci were nominally significant ($P<0.05$ without correcting for multiple hypothesis testing) (see Table 2). We caution however that these 4 loci should not be viewed as an independent replication, because the two analyses are not statistically independent due to genetic drift between AAF and YRI populations that is common to both analyses, so 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. data suggests that most of the reported novel loci are false positives, although a subset may be genuine.

It is important to recognize that even robust, genome-wide significant evidence of unusual population differentiation, e.g. at the 4 loci identified by both Bhatia et al. (2011) and Jin et al. (2012), does not imply selection following the forced migration from Africa. The observed population differences at these loci are more parsimoniously explained by selection within Africa, in the ancestors of Yoruba and/or in the African ancestors of African Americans (prior to enslavement). This is because the majority of time since these populations diverged was spent in Africa, giving more time for selection to produce an allele frequency difference.

As a counterexample that proves the rule, we consider the well-studied sickle-cell variant rs334 at the HBB locus, where biological evidence suggests 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 [16]. The minor allele frequency at rs334 in African Americans is 0.050 [17], 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—that heterozygotes have no advantage (due to much lower rates of malaria in the Americas) and that no people with sickle-cell anemia have children—the allele frequency in the African ancestors of African Americans 7 generations ago [6] would have been 0.096 (see Methods). This corresponds to a maximum possible allele frequency difference of 0.033 due to selection

since the migration from Africa. Allele frequency differences at HBB of >10% have been reported between African populations [4,5], showing that selection in Africa cannot be ruled out as explaining most of the observed frequency difference between African Americans and Yoruba. We have no doubt that the frequency of the minor allele decreased in African Americans since arrival in the Americas. However, the empirical data do not allow us to conclude that the allele frequency difference between African Americans and Yoruba at the sickle cell variant is primarily explained by selection since the arrival of Africans in the Americas.

We note in passing that the lack of a genome-wide significant deviation in average local ancestry at the HBB locus (γ=0.206, vs. a genome-wide average of 0.204) is not unexpected, even given the decrease in frequency of this allele that must surely have occurred. Even though the per-allele selection coefficient is strong ($s_{allele} = 0.077$), the selection coefficient per copy of local ancestry is still quite low ($s_{ancestry} = 0.0074$), and below the threshold of $s_{ancestry}$ = 0.019 that our local ancestry analysis is powered to detect (see Methods). According to our model of selection we expect an average local ancestry of 0.210 at the HBB locus (see Methods), consistent with our observed result of γ=0.206 (1.11 s.d. apart).

Discussion

We performed a scan for unusual deviations in local ancestry in African Americans compared with the genome-wide average, which found no genome-wide significant loci and did not replicate 6 previously reported loci with unusual deviations in local ancestry. We also reanalyzed 14 unusually differentiated loci from a previous study using a different F_{ST} estimator, showing that many single-SNP F_{ST} estimates were inflated. Even after correcting for this inflation, most of the reported loci are not genome-wide significant. Furthermore, even for loci with robust, genome-wide significant evidence of selection based on population differentiation, selection within Africa provides a parsimonious explanation for most of the empirically observed differentiation.

We caution that although there is little evidence of selection since the forced migration out of Africa in the data analyzed by Jin et al. (2012) or in the current study, we cannot exclude the possibility that some selection of this type has occurred. Indeed, selection is an ongoing process, and has surely occurred to some degree in African Americans since migration out of Africa. The key point here is statistical power. Our genome-wide scan of 29,141 samples study is well-powered (>95%) to detect signals of selection with a selection coefficient for local ancestry ($s_{ancestry}$) greater than 0.019 (see Methods). Although selection of this magnitude or greater is ruled out by our data, weaker selection may have occurred. For example, selection after the forced migration from Africa is likely to have occurred at the HBB locus due to much lower rates of malaria and a corresponding reduction in positive selection. However, our estimate of the likely selection coefficient for local ancestry ($s_{ancestry}$ = 0.0074) is too small for us to produce genome-wide significant evidence of selection even in the context of the large sample size we analyzed.

We conclude with three recommendations for future studies. First, studies of selection since admixture based on 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 $P=1x10^{-5}$ [1], and should be cognizant of the possibility that systematic errors in local ancestry inference can lead to false-positive signals. Second, studies of selection based on population differentiation that involve unequal sample sizes should not use the F_{ST} estimator of Weir and Cockerham (1984), which is susceptible to bias in this case, and instead should use the Hudson estimator [12,13,18]. Third, genome-wide significance should never be reported based on a simple ranking, and instead should be reported via model-based approaches [4,5,14,15,19].

Methods

Samples

We studied the local ancestry distribution of African Americans from five different cohorts (N = 29,141 samples combined across cohorts), using a previously published data set where a nearly identical local ancestry inference procedure was used to study the rate of recombination [3].

In detail, the dataset was derived from five genome-wide association studies conducted on African Americans, all of which differ in population size and characteristics. A coherent summary of the generation of these five datasets and the filters we used to harmonize the genotyping data is presented in [3], and hence we do not present it again here.

A complication in the analysis of these data is that the data were produced on three different genotyping platforms. Three of these data sets are genotyped on the Illumina 1M array: samples from the African American Lung Cancer Consortium (AALCC), the African American Breast Cancer Consortium (AABCC), and t the African American Prostate Cancer Consortium (AAPCC). The fourth data set was genotyped on the Illumina Human Hap550 array and is from the Children's Hospital of Philadelphia (CHOP) [3]. The fifth data set is the Candidate Gene Association Resource (CARe) study, a consortium of cohorts. This data set consists of the ARIC, CFS, CARDIA, JHS and MESA cohorts and is genotyped on the Affymetrix 6.0 chip. We note that the AALCC, AABCC and AAPCC data sets consist of disease cases and controls, but phenotype information was not available in the current study. The inclusion of disease cases could produce false-positive signals of selection due to admixture associations to disease (no such signals were observed), but are unlikely to produce false-negative signals of selection.

For our local ancestry inferences of the CARe dataset, we simply used the already published results of Pasaniuc et al. (2011). All of the remaining datasets were curated to retain only markers and samples with high genotyping completeness (>90%). We removed samples with genetic evidence of being second-degree relatives or closer using either PLINK [20] or EIGENSOFT [21] (*smartrel*). Samples with genome-wide European ancestry proportion less than 2.5% or greater than 75% were removed in all cohorts. All these datasets were separately combined with the phased Hapmap3 data of 88 European (CEU) and 100 West African (YRI) samples. Markers were removed if their allele frequency was inconsistent with a linear combination of 0.8 African and 0.2 European allele frequencies according to a t-statistic greater than 15 (or less than -15). This filter was applied to each cohort individually. A total of 626 markers were removed, none of which were located inside the 6 regions that were previously reported as being loci affected by natural selection since admixture [2]. The markers that were removed had values of $(p_{AA}-p_{EUR})/(p_{AFR}-p_{EUR})$ that were either greater than 0.25 or less than 2.86. These extreme deviations from the expected admixture proportion of African and European ancestral allele frequencies are likely due to genotyping artifacts.

Following QC, the remaining samples were 4,094 AALCC samples genotyped at 877,881autosomal markers; 5,131 AABCC samples genotyped at 866,269 autosomal markers; 6,339 AAPCC samples genotyped at 867,658 autosomal markers; 7,368 CHOP samples genotyped at 480,029 autosomal markers; and 6,209 CARE samples

genotyped at 738,831 autosomal markers. These five datasets have 121,511 autosomal markers in common, which we used to generate Figure S1 of Bhatia et al.[25].

Inferring Local Ancestry

For the CARe cohort, we used the exact same local ancestry inference reported in Pasaniuc et al. (2011). For the remaining datasets, we ran the HAPMIX [6] software separately in each cohort to infer local ancestry estimates for all the samples at each of the autosomal loci. The software was run using the Hapmap3 CEU and YRI haplotypes as the ancestral populations, assuming that the number of generations since admixture (λ) was 6, using an individual specific average European ancestry proportion (θ) prior, and using the Oxford recombination map [22]. The individual-specific θ values were calculated by running HAPMIX on these same samples using a uniform recombination map. The software was run in a mode in which it outputs an integer estimate of local ancestry by sampling from the probabilities for 0, 1 or 2 European chromosomes at each locus. The genome wide ancestry for each sample was calculated by averaging over local ancestry estimates genome-wide, after scaling these estimates by half. The average local ancestry at each locus was calculated as an average of the local ancestry estimates across all the samples. The entire analysis was conducted separately for each cohort and then combined across cohorts. 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 3 loci in these regions (which do not overlap with the Jin et al. loci) did show significant deviations in local ancestry, but these are very likely to be artifacts (see Supplementary Note of Bhatia et al.[25]). This filtering left a total of 118,006 SNPs in our analysis of local ancestry, which we used to generate Figure 1.

The mean European genome-wide ancestry proportion was 0.210 (S.D across individuals 0.123) in the AALCC sample, 0.218 (0.134) in the AABCC sample, 0.215 (0.131) in the AAPCC sample, and 0.193 (0.086) in the CHOP sample. These estimates are similar to previous studies [23,24].
The standard deviation in average local ancestry estimates across SNPs was 0.0036 for the full set of 29,141 samples.

Theoretical Standard Deviation in Local Ancestry

We calculated the theoretical standard deviation in average local ancestry as follows:

$$SD^*(\gamma) = \frac{\sqrt{\sum_i 2\overline{\gamma}_i(1-\overline{\gamma}_i)}}{2N}$$

Where $\overline{\gamma}_i$ is the average genome-wide ancestry for individual *i*, and *N* is the total number of samples. Using this calculation we obtain a theoretical standard deviation of 0.0016, less than half the empirical standard deviation of 0.0036.

Assessing Population Differentiation with the WC Estimator

Consider a biallelic SNP in a sample of individuals from 2 populations. The WC estimator is:

(1)

$$\hat{F}_{ST}^{WC} = 1 - \frac{2\frac{n_1 n_2}{n_1+n_2}\frac{1}{n_1+n_2-2}[n_1\tilde{p}_1(1-\tilde{p}_1)+n_2\tilde{p}_2(1-\tilde{p}_2)]}{\frac{n_1 n_2}{n_1+n_2}(\tilde{p}_1-\tilde{p}_2)^2+(2\frac{n_1 n_2}{n_1+n_2}-1)\frac{1}{n_1+n_2-2}[n_1\tilde{p}_1(1-\tilde{p}_1)+n_2\tilde{p}_2(1-\tilde{p}_2)]}$$

where n_i is the sample size and $\tilde{p}_i$ is the sample allele frequency in population *i* for $i \in \{1,2\}$. Then, in the limit of large sample sizes ($n_i - 1 \approx n_i$), we can assume that sample allele frequencies become close to population allele frequencies ($\tilde{p}_i \to p_i$). We analyze the estimator as the sample sizes increase, but their ratio goes to a constant *M*. In the limit of infinite, but not necessarily equal sample sizes the estimator is:

$$\lim_{\substack{n_1,n_2\to\infty \\ n_1/n_2\to M}} \hat{F}_{ST}^{WC} = \frac{(p_1-p_2)^2}{(p_1-p_2)^2+2\frac{1}{(M+1)}[Mp_1(1-p_1)+p_2(1-p_2)]} \quad (2)$$

Consider a SNP that is rare in one population and has allele frequency zero in the other population: $p_1 = 0, p_2 = \varepsilon$. If sample sizes are equal, the single SNP estimate of F_{ST} from the WC estimator is approximately ε. Now, consider what happens as we increase n_1 arbitrarily. It is clear that both numerator and denominator tend toward the same quantity and F_{ST} approaches 1.

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. However, it is possible that the 99.99th percentile threshold 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. (2012) we sub-sampled 2,500 African American individuals from our data, subtracted European allele frequencies from CEU [4], and compared the result to YRI 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. report a threshold of 0.0452. Even if this decreases by 0.002 due to use of the Hudson estimator, 2 of the 14 reported loci would no longer be in the 99.99th percentile.

Selection at HBB after the migration of African American ancestors from Africa

In order to assess the maximum effect of selection at HBB, we consider the following situation. The minor allele at rs334—which is known to cause sickle cell anemia—in African Americans today has a frequency of 0.050 [17], corresponding to a MAF of 0.0625 (0.05/0.8) on African segments. From this information, we can work backwards in time with the following equation:

$$p_{g+1} = \frac{p_g}{1-p_g} \quad (3)$$

Assuming that $p_0 = 0.0625$, and 7 generations since the admixture of the African and European ancestors of African Americans [6], we have $p_7 = 0.0962$. According to these estimates, the maximum allele frequency difference since admixture is 0.0337.

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 ($s^g_{allele} = \gamma p_g$), where γ is the proportion of African ancestry at the HBB locus during the current generation. Assuming that the proportion of local ancestry at each locus 7 generations ago is equivalent to the current genome-wide average, the maximum value of this coefficient is $s_{allele} = 0.796 p_7 = 0.077$. The selection coefficient per copy of African local ancestry is given by $s_{ancestry} = \gamma(p)^2$. That is, given that an individual carries one African chromosome at the HBB locus he must also carry (1) the sickle 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 $s_{ancestry} = 0.796(p_7)^2 = 0.0074$.

Estimating Local Ancestry Proportions After Selection

To perform power calculations and estimate the expected deviation in local ancestry at HBB (see above) we need to be able to assess the effect of a particular selection coefficient on local ancestry proportion. We did this iteratively, using the equation

$$\gamma_{g+1} = \frac{\gamma_g(1 - s_{ancestry})}{1 - \gamma_g s_{ancestry}} \qquad (4)$$

We performed this iteration *g* times to assess the effect of selection at the locus. To perform power calculations, we used a static value of $s_{ancestry}$, to assess the expected deviation in local ancestry at HBB we substituted $s_{ancestry} = \gamma_g(p_g)^2$.

Estimating the Minimum Detectable Selection Coefficient

In order to call a genome-wide significant deviation in local ancestry, we must have $|\hat{\gamma}_L - \overline{\gamma}| > 4.4\hat{\sigma}_{\gamma_L}$. That is, the observed average ancestry at a locus $\hat{\gamma}_L$, must deviate from the genome wide average $\overline{\gamma}$ by more than 4.42 standard deviations (estimated from the data). We can assume that the sampling distribution of observed average local ancestry is normal and centered around the true population average ancestry at the locus. That is $\hat{\gamma}_L \sim N(\gamma_L, \frac{\gamma_L(1-\gamma_L)}{n})$ where γ_L is the true population average ancestry at the locus and n is sample size of the study. We can then solve for γ_L, so that $\Pr(|\hat{\gamma}_L - \overline{\gamma}| > 4.4\hat{\sigma}_{\gamma_L} | \gamma_L) = 0.95$. In our case, assuming $\overline{\gamma} = 0.204$, and $\hat{\sigma}_{\gamma_L} = 0.0036$ we obtain γ_L = 0.183 or 0.225. Then, assuming 7 generations since admixture we perform a grid search over possible values of the selection coefficient for local ancestry that would produce these values of γ_L (see below) and obtain an estimate of 0.019.

Using a model-based approach to detect selection on Jin et al. (2012) data

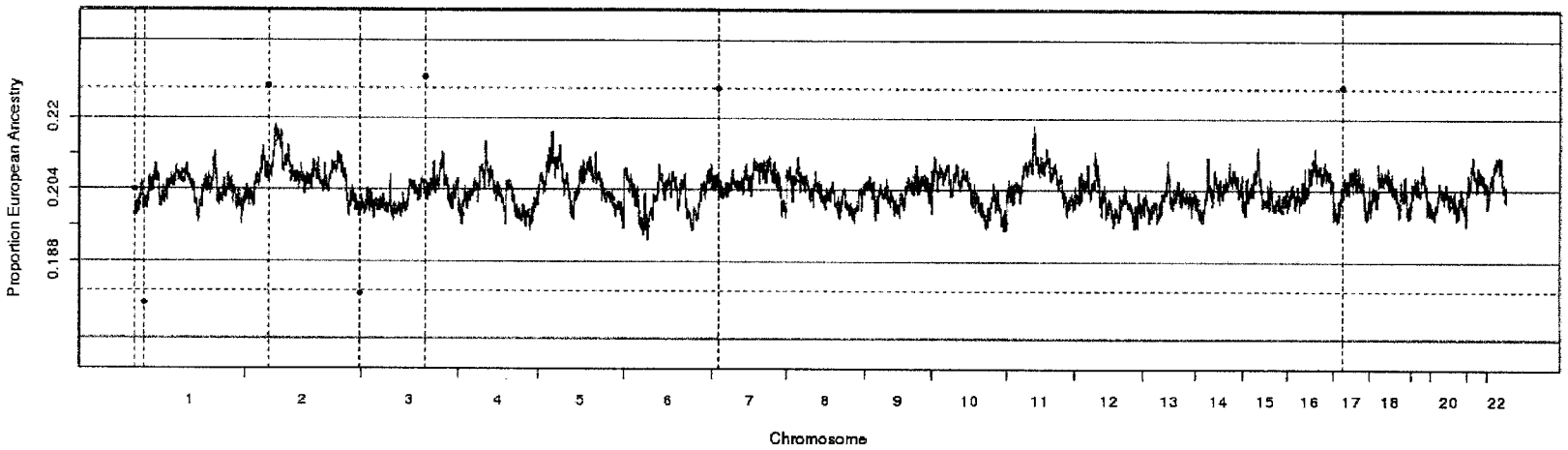
When using a model-based approach to reanalyze Jin et al. results, we were unable to estimate F_{ST} in their sample (since we did not have the raw genotypes) and thus used their reported F_{ST} of 0.0007 in the model [4,5,14,15]. We note that this F_{ST} was calculated using the WC estimator, which may be susceptible to biases when very different sample sizes are analyzed. The reported F_{ST} of 0.0007 was less than the F_{ST} of 0.0011 used in Bhatia et al. (2011). This difference has a minimal impact on the resulting statistics, as the variance is primarily due to the small sample size from Yoruba. However, using an F_{ST} of 0.0011 would lead to even less statistically significant results than those reported in Table 2, so that all model-based P-values using Jin et al. data would remain non-genome-wide significant.

Acknowledgements

We thank N. Patterson, J. Wilson and R. Kittles for helpful discussions and comments on the manuscript. This research was funded by NIH grants R01 HG006399 and R03 HG006170.

Figures

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 European ancestry. The red and blue lines indicate the threshold for genome-wide significance ($P < 10^{-5}$) in our study, and the Jin et al. study, respectively. The dashed blue line indicates the threshold for significance ($P < 2.7 \times 10^{-3}$) that was actually used in the Jin et al. study. The standard deviation was computed empirically over all SNPs. It is clear that no region attains genome-wide significance in our scan. Dashed vertical lines indicate the location and blue points the deviation in local ancestry of the six loci reported under selection in Jin et al. These deviations are reported relative to the genome-wide average ancestry proportion in our study. None of the six reported loci exceed the $P < 10^{-5}$ genome-wide significance threshold for the Jin et al. study (blue lines).

Tables

Table 1

Region	Jin et al.		Current study	
	Deviation	Nominal P-Value	Deviation	Nominal P-Value
chr1:17409539..21604321	-0.025	7.43E-04	-0.004	0.55
chr2:241750403..242568618	-0.023	2.07E-03	-0.006	0.44
chr2:37451925..37508581	0.023	2.16E-03	0.005	0.51
chr3:116930811..118313302	0.025	8.58E-04	-0.002	0.83
chr6:163653158..163653428	0.023	2.70E-03	0.004	0.60
chr16:61214438..61242497	0.023	2.26E-03	0.006	0.41

We list the 6 regions with unusual deviations in local ancestry reported by Jin et al. and compare these to our scan. None of the 6 regions replicated at nominal significance ($P < 0.05$) in our analysis.

Table 2

SNP id	Region	Gene	WC F_{ST} Jin Data	Hudson F_{ST} Jin Data	Model-based P-value Jin Data	Model-based P-value Bhatia Data
rs1541044	chr1:100125058..100183875		0.0562	**0.0439**	4.7×10^{-5}	0.04
rs4460629	chr1:153401959..153464086		0.0692	0.065	6.8×10^{-7}	2.1×10^{-4}
rs12094201	chr1:236509336		0.0561	0.0489	1.7×10^{-5}	0.86
rs7642575	chr3: 31400165		0.0453	**0.0393**	1.1×10^{-4}	0.41
rs652888	chr6:26554684..33961049	HLA	0.0711	0.0627	1.1×10^{-6}	1.8×10^{-11}
rs9478984	chr6:151555551..151569258		0.0545	0.0596	2.1×10^{-6}	0.02
rs10499542	chr7: 22235870		0.0461	0.0453	3.6×10^{-5}	0.35
rs304735	chr7:79768487..80482597	CD36	0.0946	0.069	3.0×10^{-7}	3.7×10^{-13}
rs2920283	chr8:143754039..143758933	PSCA	0.0468	0.0532	7.6×10^{-6}	6.4×10^{-7}
rs1498487	chr11:5034229..5421456	HBB	0.0617	0.0464	2.4×10^{-5}	1.7×10^{-7}
rs4883422	chr12:7189594		0.0472	0.0461	3.0×10^{-5}	1.3×10^{-3}
rs6491096	chr13:25488362		0.0472	**0.0373**	1.5×10^{-4}	0.4
rs1075875	chr16: 47595721		0.0766	0.0608	1.3×10^{-6}	N/A
rs6015945	chr20:59319574		0.0627	0.055	4.3×10^{-6}	0.5

We recreate Table 2 of Jin et. al (2012) analyzing the same data with the Hudson instead of the WC estimator. The bolded cells indicate loci that fall below the 99.99th percentile threshold of 0.0452 when the Hudson estimator is used. We also estimated the P-value at each SNP using the reported F_{ST} = 0.0007 of Jin et al. (2012) (see Methods), and a model based approach [5]. Finally, we report the model-based P-value of the most significant SNP in the region reported in the parallel study of Bhatia et al. (2011). We note that results reported in that paper were more significant than those reported here due to analysis of additional populations. The chr16 locus is reported as N/A due to a lack of data at this locus in the Bhatia et al. data.

References

1. Seldin, M.F., Pasaniuc, B., and Price, A.L. (2011). New approaches to disease mapping in admixed populations. Nature Reviews. Genetics *12*, 523–528.

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

7. Akey, J.M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. *19*, 711–722.

8. 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.

9. 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., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.

10. 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., et al. (2009). Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res. *19*, 2154–2162.

11. 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.

12. Bhatia, G., Patterson, N., Sankararaman, S., and Price, A.L. (2013). Estimating and interpreting FST: The impact of rare variants. Genome Res.

13. Hudson, R.R., Slatkin, M., and Maddison, W.P. (1992). Estimation of levels of gene flow from DNA sequence data. Genetics *132*, 583–589.

14. 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.

15. 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.

16. Aidoo, M., Terlouw, D.J., Kolczak, M.S., McElroy, P.D., Kuile, ter, 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.

17. 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. The American Journal of Human Genetics *91*, 794–808.

18. Weir, B.S., and Hill, W.G. (2002). Estimating F-statistics. Annu. Rev. Genet. *36*, 721–750.

19. Grossman, S.R., Shlyakhter, I., Shylakhter, I., Karlsson, E.K., Byrne, E.H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. Science *327*, 883–886.

20. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

21. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet *2*, e190.

22. Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. Science.

23. Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G.K., Tandon, A., Kao, W.H.L., Ruczinski, I., Fornage, M., Siscovick, D.S., Zhu, X., et al. (2011). Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. PLoS Genet *7*, e1001371.

24. 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.

25. Bhatia, G., Tandon, A., Aldrich, M.C., Ambrosone, C.B., Amos, C., Bandera, E.V., Berndt, S.I., Bernstein, L., Blot, W.J., Bock, C.H., et al. (2013). Genome-wide scan of 29,141 African Americans finds no evidence of selection since admixture. arXiv.

Chapter 5: Conclusions

In this thesis we presented methods to measure genetic distances between human populations, and detect natural selection. We applied these methods to genotyping and sequencing data-sets from multiple populations.

In Chapter 2, we gave a definition for F_{ST} based on allele frequencies in the currently observed population and the most recent common ancestral population. We showed that commonly used estimators for F_{ST} [1,2] generate biased F_{ST} estimates that may depend on the sample sizes studied. We formalized an estimator [3,4] of F_{ST} that gives results consistent with population genetic theory across sampling regimes. While the numerator of this estimator is an unbiased estimator of the variance between populations, and the denominator of this estimator is an unbiased estimator of the total variance, the ratio of these quantities is not an unbiased estimator of F_{ST}. As a result, we recommended that a ratio of averages be used to generate asymptotically consistent estimates. Finally, we demonstrated that outgroup ascertainment is desirable if the established relationship between F_{ST} and divergence time[5] is to hold. More generally, differing ascertainment schemes can be used to test a variety of hypotheses about human population history and selection and we encourage authors to publish the ascertainment scheme used to produce published estimates of F_{ST}.

We used this protocol for F_{ST} estimation to show that F_{ST} estimates from sequence data [6] are highly concordant with estimates from genotype data [7], contrary to prior reports [8]. While these earlier reports might have lead to the conclusion that F_{ST} estimated from rare variant data is substantially lower than F_{ST} estimated using common variants, our results suggest the opposite. Under certain ascertainment schemes, F_{ST} estimated from rare variants is actually *higher* than F_{ST} estimated from common variants. This is consistent with the strong effects of bottlenecks in human population history on F_{ST} at rare variants and is more pronounced for strongly bottlenecked populations[4]. We note that our approach assumes that the populations under study are outbred. A method to deal with inbred populations has been described in the supplementary note of [9].

In Chapter 3, we utilized a model-based approach[10-13] to detect selection between West African populations. An alternate approach is to rank loci by a local estimate of selection (i.e. F_{ST}, XP-EHH) [14,15]. While this approach is commonly used, an advantage of a model-based approach is the ability to generate robust genome-wide significant evidence for selection at a locus. In addition to using the model-based approach to compare two populations, we extended this approach to three populations by using an unrooted tree with a single central node. We caution that this model-based approach may be applicable to a smaller range of usage scenarios than ranking based approaches as the underlying null model—that the distribution of allele frequency differences is normal—only applies for common mutations and a small amount of genetic drift.

We used this model-based approach to validate previously published targets of selection[16-19], to provide evidence of multiple, population-specific selective events at HLA, and to discover a novel target of selection at PSCA. Of note, the mutations in this locus that show maximal differentiation between the West African populations are also highly differentiated between Japanese and Chinese populations[7]. This may serve as a motivating example for future work investigating concordance of highly differentiated loci between closely related populations. Finally, a potential criticism of our approach is that matching the empirical distribution, in expectation, to a χ^2 (1 d.f.) distribution requires that we use an average of ratios estimate of F_{ST}, which, was not our recommendation in Chapter 2. We note that under reasonable usage scenarios for model-based approach—small F_{ST} and common variants—we expect the ratio of averages and the average of

ratios to be largely concordant.

In Chapter 4, we used our work in Chapter 2 and additional analysis of local ancestry to reevaluate reports of natural selection in African Americans since the arrival of their ancestors in the Americas. Considering the very small number of generations[20] since this event, any detected selection would have to be very strong. If present, this could result in an excess (or loss) of African ancestry at the locus under selection due to hitchhiking with the selected allele. Jin and colleagues provide evidence for this deviation in average local ancestry at 6 loci. These are all loci that achieved a deviation of greater than three standard deviations from the genome-wide average. However, we point out that a threshold of 3 standard deviations corresponds to a Bonferroni correction for less than 20 independent hypotheses. As there are more than 20 chromosomes in the genome, we believe that this is likely to be insufficient and the more stringent threshold of 4.4 standard deviations (corresponding to 5000 independent hypotheses) is more appropriate[21]. Using this more stringent threshold, none of the loci reported by Jin and colleagues achieves genome-wide significance. Moreover, in a reanalysis of nearly 30 thousand African American individuals, no locus showed genome-wide significant deviations in average local ancestry. This suggests that we can rule out selection since admixture stronger than $s_{ancestry} > 0.019$. In general, we caution that detecting selection via admixture analysis may be unreliable due to biases in local ancestry inference and insufficient correction for multiple hypotheses.

The second line of evidence used by Jin and colleagues was population differences between Yoruba [7] and allele frequencies in the inferred African segments from the genomes of African American individuals. This analysis was done using an F_{ST} estimator [1] that may produce false positive results when comparing samples with very different sizes. We demonstrated that nine of the ten loci reported by Jin et al. had F_{ST} estimates inflated by this estimator, with three of these dropping below their nominal significance threshold. While Jin and colleagues do corroborate previously established targets of selection at HLA, CD36 and HBB as well as the result we described in chapter 3 at PSCA, we believe these are much more likely to indicate selection in African than selection in the Americas. As an example, we considered strong negative selection on the sickle cell allele in African Americans. Considering the current frequency in African Americans [22], and working backwards over 7 generations, we show that a difference of only ~3% can be explained by selection in the Americas. At this locus, allele frequency differences of greater than 10% have been reported in Africa[12], suggesting that natural selection in Africa is a more parsimonious explanation of observed signals.

In this thesis, we have examined questions of estimating genetic distance and detecting natural selection with regard to pairs of populations. In Chapter 3, we did generalize this to three populations using a tree of populations, but we believe that further extensions will be difficult due to a combinatorial explosion in tree topologies[23]. Additionally, while simplifying methods may allow a reasonable first approximation to a tree of human populations [24], recent studies suggest that admixture has been the rule, rather than the exception, in human population history [9,25,26]. As a result, tree based approaches, even those that allow for admixture events through use of directed acyclic graphs [27], are unlikely to model the full complexity of relationships between large numbers of human populations. Our belief is that a population covariance matrix [28] may present a more reasonable approximation of the true relationships between human populations. Additionally, a weakness of the literature on natural selection is the limited capacity with which researchers can identify the phenotype that is under selection. Methods incorporating a large number of human populations, perhaps through such covariance matrices, may be better powered to detect selection correlated with specific environmental pressures, potentially improving our understanding of human population

history, biology, and disease.

References

1. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. Evolution *38*, 1358–1370.

2. Nei, M. (1986). Definition and Estimation of Fixation Indices. Evolution *40*, 643–645.

3. Hudson, R.R., Slatkin, M., and Maddison, W.P. (1992). Estimation of levels of gene flow from DNA sequence data. Genetics *132*, 583–589.

4. Keinan, A., Mullikin, J.C., Patterson, N., and Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nature Genetics *39*, 1251–1255.

5. Cavalli-Sforza, L.L., and Bodmer, W.F. (1971). The genetics of human populations (San Francisco: W.H. Freeman).

6. 1000 Genomes Project Consortium, 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. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

7. International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

8. 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

9. Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. (2009). Reconstructing Indian population history. Nature *461*, 489–494.

10. 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.

11. Nicholson, G., Smith, A.V., Jonsson, F., Gústafsson, Ó., Stefansson, K., and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *64*, 695–715.

12. 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.

13. 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.

14. 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.

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., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res. *19*, 826–837.

16. Hedrick, P.W., and Thomson, G. (1983). Evidence for balancing selection at HLA. Genetics *104*, 449–456.

17. Cao, K., Moormann, A.M., Lyke, K.E., Masaberg, C., Sumba, O.P., Doumbo, O.K., Koech, D., Lancaster, A., Nelson, M., Meyer, D., et al. (2004). Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. Tissue Antigens *63*, 293–325.

18. Fry, A.E., Ghansa, A., Small, K.S., Palma, A., Auburn, S., Diakite, M., Green, A., Campino, S., Teo, Y.Y., Clark, T.G., et al. (2009). Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. Human Molecular Genetics *18*, 2683–2692.

19. Kwiatkowski, D.P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. Am. J. Hum. Genet. *77*, 171–192.

20. 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.

21. Seldin, M.F., Pasaniuc, B., and Price, A.L. (2011). New approaches to disease mapping in admixed populations. Nature Reviews. Genetics *12*, 523–528.

22. 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. The American Journal of Human Genetics *91*, 794–808.

23. Cavalli-Sforza, L.L., and Edwards, A.W.F. (1967). Phylogenetic Analysis: Models and Estimation Procedures. Evolution *21*, 550–570.

24. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.

25. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., Mesa, N., et al. (2012). Reconstructing Native American population history. Nature *488*, 370–374.

26. 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.

27. Pickrell, J.K., and Pritchard, J.K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet *8*, e1002967.

28. Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1994). The History and Geography of Human Genes.