METHODS (TO REPLACE THE METHODS SUMMARY IN THE ONLINE VERSION)

Samples used for building the AA Map: The 29,589 unrelated African American samples derive from five sources. Informed consent was provided by all the individuals participating in the study, and was approved by all of the institutions responsible for sample collection.

The first source is the Candidate Gene Association Resource (CARe) study, a consortium of cohorts. We analyzed CARe samples genotyped on the Affymetrix 6.0 array from the Atherosclerosis Risk in Communities study (ARIC), the Cleveland Family Study (CFS), the Coronary Artery Risk Development in Young Adults study (CARDIA), the Jackson Heart Study (JHS), and the Multi-Ethnic Study of Atherosclerosis (MESA). After removing individuals known to be related, and restricting to SNPs with good completeness in all cohorts, we had data from 6,209 individuals typed at 580,000 SNPs.

The second source consists of diverse studies carried out at the Children's Hospital of Philadelphia (CHOP), which has established a biobank for Philadelphia children to facilitate large genotype-phenotype association analysis. The cohort was recruited by CHOP clinicians, nursing and medical assistant staff within the CHOP Health Care Network, including primary care clinics and outpatient practices, from the hospital's patient base of over one million pediatric patients. All samples analyzed here were genotyped on either the Illumina 610-Quad or Illumina HumanHap550 array. After removing individuals known to be related, identifying American Americans by multi-dimensional scaling on genotype data, and restricting to SNPs with a high level of completeness across samples, we had data from 7,503 samples typed at 491,572 SNPs.

The third source is the African American Breast Cancer Consortium (AABCC), consisting of the Multiethnic Cohort study (MEC), the Los Angeles component of the Women's Contraceptive and Reproductive Experiences study (CARE), the Women's Circle of Health Study (WCHS), the San Francisco Bay Area Breast Cancer study (SFBC), the Carolina Breast Cancer Study (CBCS), the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Cohort (PLCO), the Nashville Breast Health Study (NBHS) and the Wake Forest University Breast Cancer Study (WFBC), all genotyped on an Illumina 1M array. After data curation, including removal of samples with genetic evidence of being second degree relatives or closer using the *smartrel* package of EIGENSOFT[27] (>0.2 correlation of genotype state), we had data from 5,203 women (about half cases and half controls) typed at 894,717 SNPs.

The fourth source is the African American Prostate Cancer Consortium (AAPCC), consisting of the MEC, the Southern Community Cohort Study (SCCS), PLCO, the Cancer Prevention Study II Nutrition Cohort (CPS-II), the Prostate Cancer Case-Control Studies at MD Anderson (MDA), the Identifying Prostate Cancer Genes study (IPCG), the Los Angeles Study of Aggressive Prostate Cancer (LAAPC) study, the Prostate Cancer Genetics Study (CaP Genes), the Case-Control Study of Prostate Cancer among African Americans in Washington DC (DCPC), the Gene-Environment Interaction in Prostate Cancer Study (GECAP) and the Cancer Prevention Study II (CPS-II), all typed on an Illumina

1M array. After the same data curation as the breast cancer study, we had data from 6,540 men (about half cases and half controls) typed at 896,036 SNPs.

The fifth source is individuals from the African American Lung Cancer Consortium, including cases and controls from the MEC, the SCCS, PLCO, the MD Anderson (MDA) African American Lung Cancer Study, the NCI-Maryland Lung Cancer Case-Control Study, the University of California at San Francisco African American Lung Cancer Study and the Wayne State African American Lung Cancer Study, all genotyped on the Illumina 1M array. After data curation, we had data from 4,134 individuals typed at 906,687 SNPs.

Samples used for building the Pedigree Map: The Pedigree Map was built using data from 135 African American nuclear families from CARe and 87 African American families from CHOP for which genotyping data were available from at least two full siblings and at least one parent. The CARe studies that contributed samples were JHS (70 families, including 58 samples that we newly genotyped on the Affymetrix 6.0 array to increase the number of crossovers we could analyze) and CFS (65 families). For the families with a missing parent, we developed a Hidden Markov Model (HMM) approach to jointly estimate the genotype of the missing parent as well as to infer the position of crossover events in the offspring. The observed variables in the HMM were the genotypes of the available family members and the states of the HMM were the genotypes of the parents and the identity by descent (IBD) status of the children. A change in IBD status in an offspring is interpreted as a crossover event. Note S2 provides details of the HMM used to infer positions of these pedigree crossover events.

Local ancestry inference and identification of crossover events: We merged the data for each cohort with phased West African (YRI) and European American (CEU) data from the HapMap Phase 3 data set[28]. We filtered SNPs that had a frequency inconsistent with an 80%-20% linear combination of YRI and CEU frequencies (t-statistic with an absolute value of greater than 3), potentially reflecting genotyping error in either the HapMap3 or the cohort data.

We ran HAPMIX on these data using a prior hypothesis of 20% European ancestry and 6 generations since mixture for each individual[15]. HAPMIX requires users to input a recombination map prior distribution, and we assumed that rates were constant across each chromosome arm with a total rate across each arm determined by the Rutgers genetic map[29] (Table S4).

Filtering of crossover events had three stages. First, we removed crossover events where the probability of occurrence was estimated to be less than 95% by HAPMIX. Second, we removed candidate crossover events that were non-monotonic, that is, where the probability of an overlapping crossover event with an ancestry switch in a different direction was ≥1% within any inter-SNP interval. Third, we removed crossover events where either of the two flanking ancestry blocks was smaller than 2 cM in size as measured with respect to a published map based on LD[2,18] (Note S1). For comparisons to the deCODE

Map and LD-based maps, we also removed segments of the genome within 5 Mb of the telomeres (to be consistent with the comparisons presented in the deCODE study where the same restriction was applied[4]).

Construction of the AA Map: All 22 autosomes and chromosome X were split into approximately 1.3 million inter-SNP intervals based on the union of SNPs analyzed across all five sample sets. Our goal was to estimate a crossover rate for each of these intervals. We modeled crossover rates such that the rate for each SNP interval is independent of every other SNP interval, motivated by a hotspot model. We used a gamma prior on rates with the mean estimated from the filtered HAPMIX output (Note S1). We used a Gibbs sampler to sample rates in every SNP interval and to determine the location of a crossover event within the 95% range estimated by the HAPMIX output. In each round of the Gibbs sampler, we used the set of sampled rates in the previous round to construct a probability mass function for the SNP interval in which each crossover occurred, using an approach described in Note S1 to approximate the probability mass function that HAPMIX would have produced conditional on the previous set of sampled rates. After sampling the location of the crossover events, we counted how many crossovers occurred in every SNP interval. We used these counts to construct a posterior distribution for the crossover rate in each SNP interval, taking advantage of the conjugacy of a Poisson likelihood and a gamma prior. We then sampled a crossover rate for each SNP interval from its respective gamma posterior distribution.

Candidate African-enriched hotspots: To identify candidate African-enriched hotspots, we utilized two pairs of maps: the previously available YRI Map and CEU Map, and the AE Map and the S Map. We combined information from both map pairs to enrich for regions with genuine differences between the West African and European populations. Specifically, we identified candidate hotspots as 2 kb intervals representing a peak in the AE Map rate, where the estimated rate in the AE Map was >2 cM/Mb and at least double that in the S Map, and in addition the YRI Map rate was >2cM/Mb and at least double the CEU Map rate. We took the resulting candidate hotspot set and defined hotspot boundaries by identifying the region flanking the 2 kb rate peak that had rates at least 50% of the peak value in the AE Map. Regions larger than 5 kb were discarded. We similarly constructed a set of "shared" hotspots but modified the initial criteria given the lack of obvious hotspots present only in people of European ancestry. Specifically, we identified 2 kb S Map rate peak locations where both the S and CEU estimated rates were >2 cM/Mb, while the AE and YRI Map rates were below those in these respective European populations. We then narrowed the regions and filtered using the same procedure we had developed for the candidate African-enriched hotspots.

Association testing: MaCH[30] was used to impute up to 3,058,149 SNP genotypes from HapMap2[18] into all African Americans we analyzed, using the unrelated YRI and CEU samples as combined reference panels. We tested for association at all SNPs. To restrict our analysis to individuals in whom the phenotype was measured accurately, we performed the association analysis with the AE and Hotspot Usage phenotypes only in individuals with at least 35 inferred crossovers. Association testing was carried out using linear regression, after controlling for gender, genome-wide European ancestry proportion (inferred by HAPMIX) and study (Note S4). We observe slight inflation of the association statistics genome-wide compared with the expectation (the Genomic Control inflation factor[31] is 1.046 for AE phenotype and 1.038 for the hotspot usage phenotype which we hypothesize may reflect cryptic relatedness among samples (Note S4). We report P-values after correction using Genomic Control[31].

Construction of *PRDM9* tree: To examine the history of the *PRDM9* ZF array and to place SNPs showing association with AE Map usage within the framework of this history, we identified 19 SNPs from HapMap2[18] that surrounded the ZF array and that form a maximal block of SNPs where there is almost no evidence of recombination: |D'| =1 for all pairs of SNPs in the data after removing 2 of 120 YRI and 1 of 120 CEU haplotypes (the chimpanzee genome was used to define the ancestral alleles). A unique "gene tree" was then built, and we used *genetree*[32], which assumes a coalescent prior on genealogies, to approximately infer ages for these mutations conditional on the data (a caveat is that the tree building does not account for the HapMap SNP ascertainment scheme). Because *genetree* assumes a randomly mating population, and the YRI represent almost all the HapMap haplotype diversity in this region, we ran the software (2,000,000 importance samples, otherwise default parameters) on the data only using YRI and used this to construct Figure 2C. Each node of the tree corresponds to a unique haplotype at these 19 SNPs, whose frequency in both CEU and YRI is shown at the base of the figure.

Motif searching: We tested all candidate motifs of 5 to 9 base pairs for enrichment in our African-enriched hotspot set relative to our shared hotspot set. We counted occurrences of all tested motifs in repeat and non-repeat backgrounds separately, and computed a separate P-value for each genomic background with a chi-squared test, based on a contingency table that compares the counts of a particular motif to the counts of all motifs of that size. We converted each P-value to a Z-score, added the scores on each background, and then obtained a corresponding combined P-value. Motifs were considered statistically significant only if they passed four stringent criteria: (i) they were statistically significant after Bonferroni correction for the number of motifs tested; (ii) they were overrepresented in the African-enriched set; (iii) they were statistically significant on both the repeat and non-repeat backgrounds ($P<0.01$) independently; and (iv) they were statistically significant when the joint P-value was calculated only by comparing the frequency of the motif to other motifs of identical G/C content (to eliminate false positives due to any difference in G/C content between the hotspot sets). This testing revealed a unique significant motif, the 9-mer CCCCAGTGA. We explored whether flanking DNA around exact matches to this motif also played a role by testing whether bases at a given site relative to the motif were associated with the difference in rates between African- and European-ancestry populations (Kruskal-Wallis test).

Rates were evaluated in the 2 kb surrounding each motif occurrence. We separately evaluated flanking sequence using both the difference between YRI/CEU Map rates, and the difference between the AE/S Map rates, leading to the identification of the 17-bp consensus African-enriched motif (Note S6 has full details). To identify close matches to this 17-bp motif among all matches to the 9-mer in the genome, for every occurrence of the 9-mer, we scored the flanking sequence bases proportionately to the relative increase in average crossover rate difference associated with each base, then multiplied across bases in the 17-mer region to provide an overall score. We ranked occurrences according to this score, and plotted rates around the top 500 (Figure 3B). We verified these findings by measuring average crossover differences for each base using only odd chromosomes and used these to score motif occurrences on the (non-overlapping) set of even chromosomes, and vice-versa (Figure S8).

PRDM9 zinc finger length-typing and genotyping of rs6889665: To determine the number of zinc finger motifs of *PRDM9* in a subset of the samples used to build the map, published primer pairs[4] were used to amplify this region (forward: 5'-GGCCAGAAAGTGAATCCAGG-3', reverse: 5'-GGGGAATATAAGGGGTCAGC-3'). Product lengths ranged between 7 and 20 repeats (801-1893 bp). Four of the 166 African American samples did not show an amplification product, presumably because of insufficient DNA quality. We also genotyped 96 YRI and 92 CEU HapMap samples.

The SNP rs6889665 was genotyped in the same samples using an allelic discrimination assay (forward primer: 5'-aaacttggaacatccatagggt-3', reverse primer: 5'-cgaaaggagaaaagcataatcc-3', LNA-probe 'C': 5'-/6-FAM/aGGGatAaatgaag/BHQ/-3', LNA-probe'T': 5'-/HEX/ AGAGatAaatGaagg/BHQ/-3'; LNA-bases are given in capital letters; reporter dyes: 6-FAM: 6-Carboxyfluorescein, HEX: Hexachlorofluorescein; Quencher: BHQ: Black Hole Quencher® 1). Only one out of the 166 African American samples failed in this assay. The same YRI and CEU samples as above were also genotyped.

ADDITIONAL REFERENCES (FOR ONLINE VERSION WITH FULL METHODS

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

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

[29] Matise, T.C. *et al.* A second-generation combined linkage physical map of the human genome. *Genome Res.* **17**, 1783-1786 (2007).

[30] Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34,** 816-834 (2010).

[31] Devlin, B., Roeder, K. Genomic control for association studies. *Biometrics* 55, 997-1004 (1999).

[32] Griffiths, R.C. Tavaré. S. Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosci.,* **127,** 77-98 (1995).