ARTICLE

doi:10.1038/nature10336

The landscape of recombination in African Americans

A list of authors and their affiliations appears at the end of the paper

Recombination, together with mutation, gives rise to genetic variation in populations. Here we leverage the recent mixture of people of African and European ancestry in the Americas to build a genetic map measuring the probability of crossing over at each position in the genome, based on about 2.1 million crossovers in 30,000 unrelated African Americans. At intervals of more than three megabases it is nearly identical to a map built in Europeans. At finer scales it differs significantly, and we identify about 2,500 recombination hotspots that are active in people of West African ancestry but nearly inactive in Europeans. The probability of a crossover at these hotspots is almost fully controlled by the alleles an individual carries at *PRDM9* (P value $< 10^{-245}$). We identify a 17-base-pair DNA sequence motif that is enriched in these hotspots, and is an excellent match to the predicted binding target of *PRDM9* alleles common in West Africans and rare in Europeans. Sites of this motif are predicted to be risk loci for disease-causing genomic rearrangements in individuals carrying these alleles. More generally, this map provides a resource for research in human genetic variation and evolution.

In humans and many other species, recombination is not evenly distributed across the genome, but instead occurs in 'hotspots': 2-kilobase (kb) segments where the crossover rate is far higher than in the flanking DNA sequence[1–3]. The highest-resolution genetic map in contemporary humans so far—the deCODE map—is based on about 500,000 crossovers identified in 15,000 Icelandic meioses[4]. However, a limitation of maps built in people of European descent[4–6] is that they may not apply equally well in other populations, as suggested by comparisons of maps across ethnic groups[4,7–9] and patterns of linkage disequilibrium breakdown, which indicate that more of the genome may be recombinationally active in West Africans[10]. It is known that a major determinant of the positions of recombination hotspots is PRDM9, a meiosis-specific histone H3 methyltransferase whose zinc finger (ZF) domain binds DNA sequence motifs[11–13]. In Europeans, PRDM9 ZF arrays are predominantly of two similar types, A and B, both of which bind the 13-bp motif CCNCCNTNNCCNC[11]. In contrast, 36% of West African alleles are not of the A or B type[9,13]. Sperm typing of males who carry neither the A nor the B allele has shown no evidence of crossover activity at recombination hotspots associated with the 13-bp motif[9].

Building an African-American genetic map

To investigate differences in the crossover landscape across human populations, we built a genetic map in African Americans, who have an average of about 80% West African and 20% European ancestry, leading to genomes comprised of multi-megabase stretches of either West African or European ancestry[14]. Computational approaches, including HAPMIX[15], have been developed to infer the probability of 0, 1 or 2 European or African alleles at each locus in individuals genotyped at hundreds of thousands of single nucleotide polymorphisms (SNPs)[15–17]. Positions where the inferred number of European or African alleles changes reflect crossover events that have occurred since admixture began (on average six generations ago[15]). Change in the probability of European ancestry between adjacent SNPs can be interpreted as the probability of such a crossover between them. We inferred crossover events in 29,589 apparently unrelated African Americans who had been genotyped on SNP arrays in genetic association studies (Methods; Fig. 1a). To minimize false-positive crossovers, we restricted to crossovers that HAPMIX inferred with a probability of >95%, and that were flanked by a minimum of 2-centimorgan (cM) stretches where the ancestry was inferred to be unchanging (Supplementary Note 1). This produced 2,113,293 high-confidence crossovers, with a typical switch point resolved within 70 kb with probability 50% (Supplementary Note 1).

To build a high-resolution African-American genetic map (AA map), we leveraged the fact that most crossovers occur in hotspots shared across individuals[2] (Methods). Intuitively, although any crossover can only be roughly localized, inter-SNP intervals that are inferred to have an appreciable probability of crossover in multiple individuals are likely to contain recombination hotspots, allowing much better localization (Supplementary Fig. 1). To implement this idea, we modelled the recombination rate for each inter-SNP interval as shared across individuals and used Markov chain Monte Carlo (MCMC) to sample rates consistent with the data (Methods). This provides well-calibrated estimates of the crossing-over rate between all pairs of markers as well as estimates of rate uncertainty (Supplementary Note 1 and Supplementary Fig. 2). We find that the interval size at which the average recombination rate is equal to the standard error is 6 kb, which is the same accuracy that would be expected from a map based on 500,000 crossovers whose boundaries were precisely resolved (Supplementary Note 1). Despite this high resolution, there are also some limitations. First, the AA map does not separately infer male and female recombination rates (it is a sex-averaged map) and requires normalization by the total map length (like linkage disequilibrium maps[3,18]). Second, the map has less resolution and may miss a higher fraction of true crossovers at loci where it is more difficult to detect and resolve crossovers owing to low SNP density or low differentiation between West Africans and Europeans. Third, the map may be biased where ancestry deviates from the average, for example at chromosome 8q24, where the 10% of the people in this study who have prostate cancer have an increased proportion of African ancestry[19]. Fourth, the map assumes that all individuals are unrelated, whereas in fact there is probably some shared ancestry, resulting in multiple counting of some crossovers and an overestimation of map precision.

To assess the accuracy of the AA map, we generated an independent African-American pedigree map by analysing 222 nuclear families

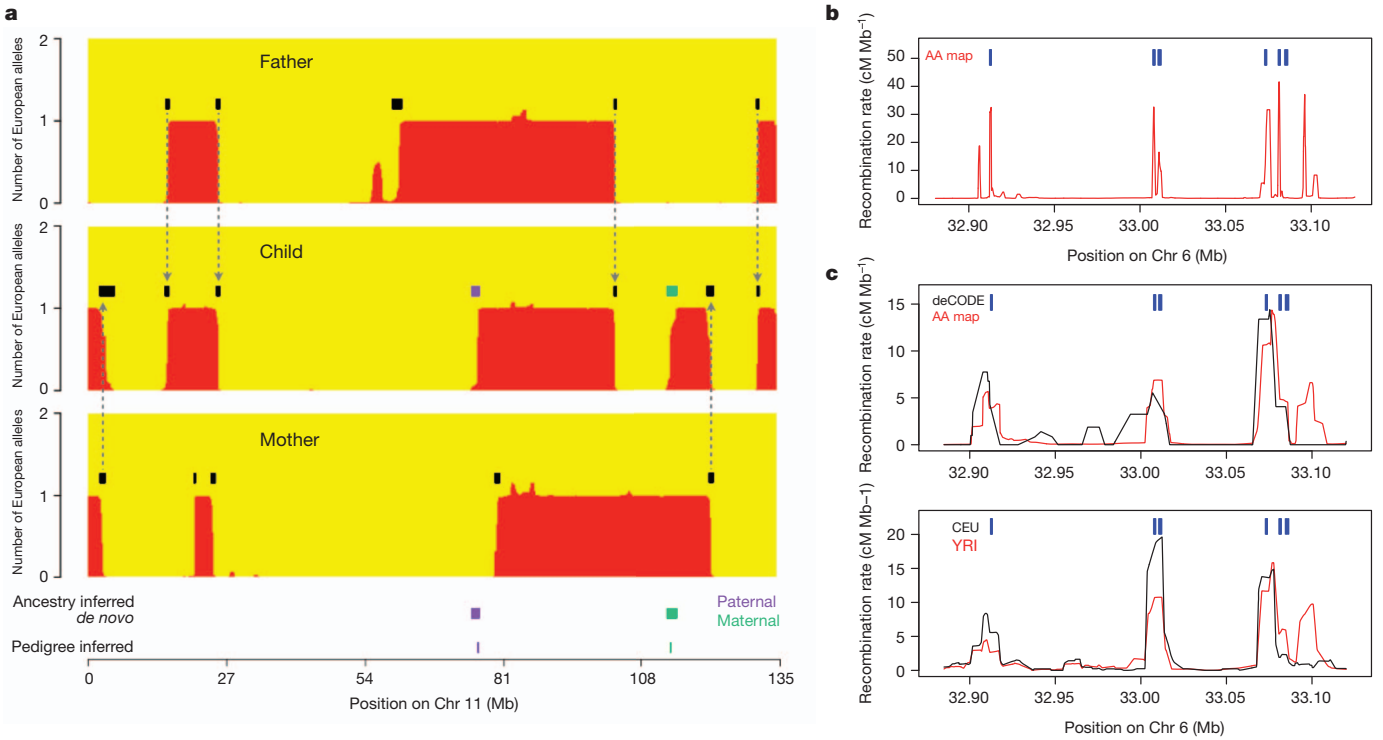

Figure 1 | Building an African-American genetic map. **a**, HAPMIX detection of crossovers between segments of inferred ancestry is illustrated in a father–mother–child trio. Black segments show inferred crossovers; arrows show transmission of ancestral crossovers from parent to child; purple/green segments show *de novo* events (paternal/maternal origin, respectively) corresponding to events identified directly using two additional children (bottom, 'pedigree inferred'). **b**, The AA map localizes five hotspots in a region of the MHC whose positions (blue) were previously mapped by sperm typing[1]. **c**, Comparison of maps shows a hotspot at 33.1 Mb in the African-derived AA and YRI maps, but not the deCODE and CEU maps (all maps smoothed to 10 kb).

that included 1,056 meioses in which we could directly detect crossovers between parent and child (Methods; Fig. 1a). Examination of the AA map rate around directly detected crossovers confirms the high resolution: the rate around such crossovers shows at least as strong a peak as that observed in maps based on linkage disequilibrium[2,3,18] (Supplementary Fig. 3). We next computed correlation coefficients for both the AA map and the deCODE map[4] to maps derived from the breakdown of linkage disequilibrium in Europeans (CEU) and West Africans (YRI)[18]. At broad scales (>3 Mb) they are almost identical ($\rho > 0.97$; Table 1). At fine scales, the AA map is more accurate (Table 1 and Supplementary Table 1), as reflected in a modest improvement in correlation to the CEU map at a 3-kb scale ($\rho_{AA,CEU} = 0.66$ versus $\rho_{deCODE,CEU} = 0.58$), and a major improvement for the YRI map, also at a 3-kb scale ($\rho_{AA,YRI} = 0.71$ versus $\rho_{deCODE,YRI} = 0.53$). The deCODE map is more correlated to the CEU map than to the YRI map at scales <1 Mb, suggesting that this map, built in Icelanders, reflects more European recombination rates. The AA map shows the opposite pattern, suggesting that it reflects more West African recombination patterns.

Population differences in hotspot locations

We compared the rate estimates for all four maps (AA, deCODE, CEU and YRI) over a 200-kb region within the major histocompatibility complex (MHC) locus where recombination rates in European males have been characterized through sperm typing[1] (Fig. 1b). The AA map detects five of six known hotspots, and localizes them to within 1 kb (the sixth hotspot is weak, with a peak male rate below the genome average[1]). Notably, the two maps based on samples with African ancestry (AA and YRI) found a hotspot not present in either map based on samples of European ancestry (deCODE and CEU) (Fig. 1c; Supplementary Fig. 4 gives a second example). We confirmed that such 'African-enriched' hotspots also occur genome-wide, by examining 2,375 loci with recombination rate peaks in the YRI map (>5 cM Mb^{-1}) but not the CEU map (<1 cM Mb^{-1}), and finding a rate rise in the independently generated AA map, but not in the deCODE map (Supplementary Fig. 5A). In the reciprocal experiment searching for European-specific hotspots, we find no such evidence for genuine ancestry specificity; at loci with recombination rate peaks in the CEU map but not the YRI map, there are weak peaks in both the deCODE and AA maps

Table 1 | Genetic map assessments at different size scales

Scale (interval size)	Pearson correlation (ρ) of the AA map (deCODE map) to the specified LD map			Estimated correlation of AA map to the true map (inferred by MCMC)†	Estimated coefficient of variation of AA map (s.e. divided by crossover rate expected for interval size)†
	Combined LD*	CEU	YRI		
3 kb	0.75 (0.63)	0.66 (0.58)	0.71 (0.53)	0.93	1.41
10 kb	0.82 (0.74)	0.73 (0.70)	0.78 (0.65)	0.96	0.73
30 kb	0.86 (0.83)	0.78 (0.78)	0.83 (0.74)	0.98	0.36
100 kb	0.91 (0.89)	0.84 (0.85)	0.87 (0.81)	0.99	0.17
300 kb	0.94 (0.93)	0.89 (0.90)	0.92 (0.88)	1.00	0.08
1 Mb	0.97 (0.96)	0.94 (0.94)	0.95 (0.95)	1.00	0.04
3 Mb	0.98 (0.98)	0.97 (0.97)	0.98 (0.97)	1.00	0.02

The numbers in this table are restricted to the autosomes and genomic segments more than 5 Mb from the telomeres. LD, linkage disequilibrium; s.e., standard error.
* The combined map is the HapMap2 population-averaged linkage-disequilibrium-based map[18].
† The s.e. of the map at each size scale is determined by the posterior probability distribution from the MCMC.

(Methods and Supplementary Fig. 5B). Thus, hotspots active in Europeans are consistently 'shared' with YRI and African Americans, whereas populations with African ancestry harbour additional, non-shared hotspots that we call 'African-enriched'.

Mapping variants underlying population differences

To understand the features of recombination in West Africans that differ from Europeans, we estimated the degree to which each African-American person's crossovers occur in African-enriched hotspots, compared with shared hotpots, a phenotype we refer to as their African enrichment (AE). We view each individual's crossovers as sampled from a mixture of two genetic maps—an 'S map' of shared hotspots based on the deCODE map, and an 'AE map' of African-enriched hotspots that is learned from comparing the deCODE and AA maps—so that the proportion of crossovers assigned to the AE map is a person's AE phenotype (Supplementary Note 4). We tested approximately 3 million SNPs (genotyped and imputed) for association with three phenotypes: AE, usage of linkage-disequilibrium-based hotspots known to be enriched for the 13-bp motif CCNCCNTNNCCNC[20] and genome-wide crossover rate (in pedigrees) (Methods and Supplementary Note 4). In crossovers detected in unrelated African Americans, the alleles a person carries are only sometimes descended from the ancestor in whom the crossover occurred, thus adding noise to the association signal (nevertheless there is useful signal given the large sample size; Supplementary Note 4). In the pedigree map, association between alleles and AE can be tested directly because we have genotypes in the parents.

The SNP showing the strongest association with AE is rs6889665 ($P = 1.5 \times 10^{-246}$; Fig. 2a and Supplementary Fig. 6), which has a derived allele frequency of 29% in YRI and 2% in CEU, and is within 4 kb of the ZF array of *PRDM9* (refs 4, 9, 11–13). This SNP is associated with AE in both the pedigree individuals and the unrelated individuals (Supplementary Note 4), and is also the SNP most strongly associated with usage of linkage-disequilibrium-based hotspots ($P = 1.8 \times 10^{-52}$) (Supplementary Table 2). No locus outside *PRDM9* is significant ($P < 0.01$ after Bonferroni correction; Supplementary Table 2). To understand better the association at rs6889665, we inferred the alleles in the *PRDM9* ZF array carried

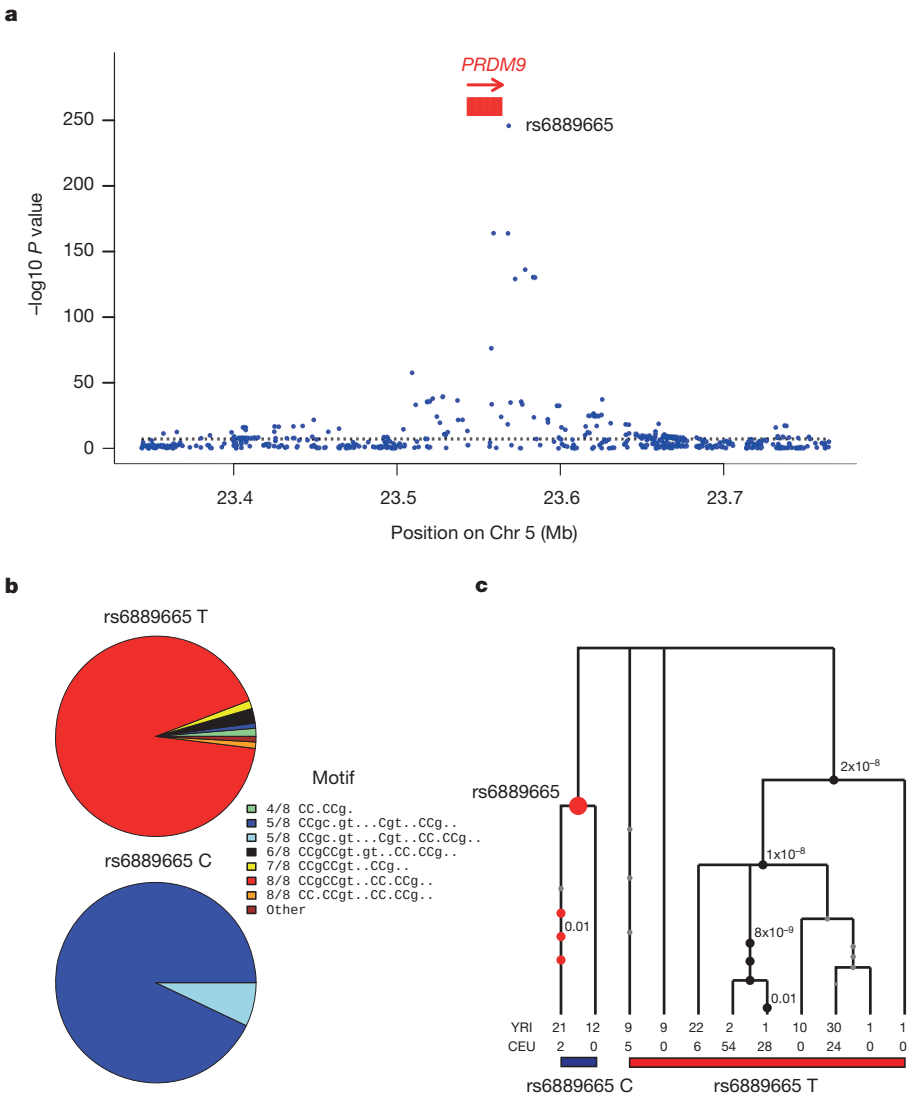

Figure 2 | Association of *PRDM9* genetic variation with hotspot activity. **a**, A genome-wide association study measuring association of the AE phenotype shows a single genome-wide significant peak at *PRDM9*, with rs6889665 the best-associated SNP. **b**, Relationship between alleles of rs6889665 and predicted binding target of the *PRDM9* ZF array[9] for West African and European samples. The binding predictions are grouped into 8 clusters according to their best-matching region to the 13-bp motif, and annotated by the number of bases matching the motif. The African-enriched rs6889665 C allele always co-occurs with motifs with a poor (5/8) match to the 13-bp motif. **c**, Gene tree[25] of the linkage disequilibrium block containing the *PRDM9* ZF array (Methods); numbered circles show SNPs and significant *P* values for association, after conditioning on rs6889665.

by 139 individuals based on sequencing data from the 1000 Genomes Project[10], using the reads to infer each individual's *PRDM9* alleles among 29 alleles whose full sequences were previously determined[9] (Supplementary Note 5). Grouping *PRDM9* alleles on the basis of how closely their binding target predictions match the 8 non-degenerate bases of the 13-bp motif, following a previously described approach[9], we find that the ancestral 'T' variant at rs6889665 is strongly correlated to alleles with an exact (8/8) match to the 13-bp motif (including the A and B alleles), whereas the derived 'C' variant is almost perfectly correlated to a group of alleles, all predicted to bind a common, different 17-bp motif—CCgCNgtNNNCgtNNCC[9]—which matches the 13-bp motif at only 5 bases (5/8 match; less strongly signalled bases in the motif are in lowercase and 'N' may be any base). This implies a common historical origin for alleles matching this 17-bp motif (Fig. 2b, Supplementary Fig. 7 and Supplementary Note 5). We also experimentally measured the number of ZF domains in *PRDM9* in 354 individuals including 166 African Americans from the pedigree study (Methods). This showed, again, that rs6889665 differentiates *PRDM9* alleles into two different classes, with 96% of haplotypes carrying the ancestral allele having <14 ZFs, and 93% of haplotypes carrying the derived allele having ≥14 ZFs (Supplementary Fig. 7). After conditioning on rs6889665, there is no evidence that ZF array length is associated with the AE phenotype. Several SNPs near the *PRDM9* ZF array show a conditional association signal that is much weaker than rs6889665, but still significant (Fig. 2c, Supplementary Fig. 6 and Supplementary Note 4), with the strongest at rs10043097 ($P = 8.3 \times 10^{-14}$), upstream of the *PRDM9* transcription start site. These SNPs may tag additional variation in the *PRDM9* ZF array, or potentially expression levels.

Finding a motif for African-enriched hotspots

To identify directly candidate African-enriched hotspot motifs, we selected 2,454 loci with a high crossover rate in the AE map and YRI map (>2 cM Mb^{-1} over 2 kb), and no more than half this rate in the S map and CEU map (this set is more powerfully enriched for higher recombination in people of African ancestry than the 2,375 above, as it includes information from the contemporary maps). We compared these to a 'control set' of 7,328 candidate hotspots more active in the European- than the African-derived maps (Methods and Supplementary Note 6). To identify sequence motifs associated with the African-enriched hotspots[3,21], we identified short motifs that occurred at increased frequency in the African-enriched hotspot set (Supplementary Note 6). Testing all motifs with lengths of 5–9 bases revealed a 9-nucleotide motif CCCCAGTGA (odds ratio (OR) = 1.79, $P = 2.24 \times 10^{-8}$, Bonferroni corrected $P = 0.004$), which exhibited a kilobase-scale rate peak near occurrences of this motif in African-derived maps, but in neither of the European-derived maps (Supplementary Fig. 8). Further analysis revealed a strong influence of downstream flanking bases (Supplementary Fig. 9) and degeneracy, yielding a 17-bp consensus sequence, CCCCaGTGAGCGTtgCc (Fig. 3a; more strongly signalled bases are in uppercase), with the same consensus obtained when we considered flanking sequences for only odd or even chromosomes, and whether we based the analysis on AE–S or YRI–CEU map comparisons (Supplementary Note 6). The 500 best matches to this motif have a ~3-fold increase in average rate in the AA and YRI relative to the deCODE and CEU maps (Fig. 3b and Supplementary Fig. 8G). Hotspots associated with the motif occur in both unique and repetitive DNA (for example, L1PA10/13 LINE elements; Supplementary Fig. 10 and Supplementary Note 6). We also compared the 17-bp consensus to the binding motif predicted for 5/8 match alleles, and found that they match almost precisely (Fig. 3a; 10 of 11 bases, $P = 8.1 \times 10^{-6}$).

Assessing the impact of *PRDM9* on recombination

How much of the African-enriched recombination pattern can be explained by *PRDM9*? We estimated the fraction of variation in the AE phenotype explained by rs6889665 in our pedigree data after accounting for noise in the phenotype estimation (Supplementary Note 4). Over 82% of map usage variability is explained by the rs6889665 genotype alone. Given that there are further influential *PRDM9* variants (Fig. 2c), this gene may thus explain almost all differences in local rate between the West African and European populations. We next examined rates around 82 narrowly defined (<10 kb) crossover sites in 7 individuals homozygous for the derived allele at rs6889665. There is no evidence of hotspots at these loci in either the deCODE or CEU maps (Fig. 3c), in contrast to crossovers in individuals carrying the ancestral allele at rs6889665 (Supplementary Fig. 11). Thus, crossover positions in individuals who are homozygous for the derived allele at rs6889665 are consistent with an entirely different recombination hotspot landscape, which would imply *PRDM9* control of all hotspots[9]. Despite the strong correlation between maps at megabase scales, there is mounting evidence that *PRDM9*'s influence

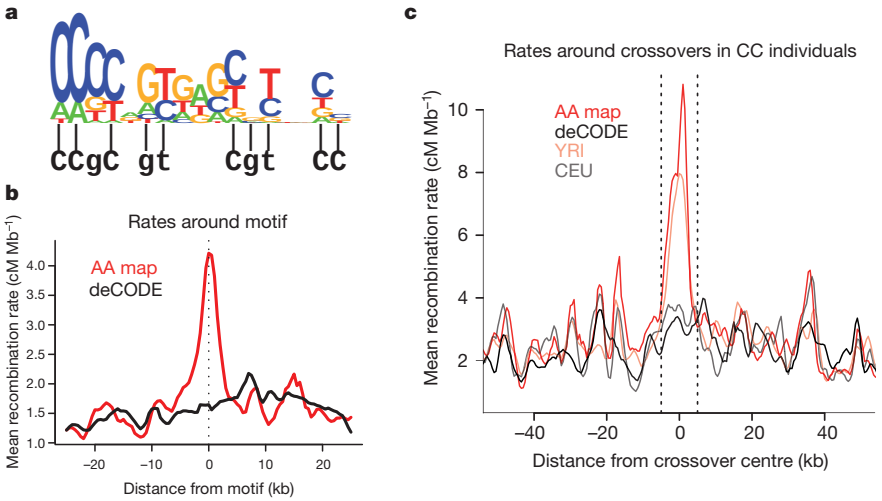

Figure 3 | A sequence motif specifying the positions of African-enriched hotspots. **a**, Logo plot showing a degenerate 17-bp hotspot motif, with stack height proportional to −log *P* value, and relative letter height proportional to the mean crossover rate increase given each base. Below is the bioinformatic *PRDM9* binding prediction for the alleles associated with rs6889665 allele C (from Fig. 2b), matching this motif at 10/11 bases (lines). **b**, Average crossover rate (in 2-kb sliding windows) in the AA (red line) and deCODE (black line) maps surrounding the 500 strongest motif matches. **c**, In seven rs6889665 CC individuals from the pedigree study, we localized 82 crossovers to within 10 kb, and plot average AA, YRI, deCODE and CEU map rates. There is no strong peak above local background in the deCODE or CEU maps.

on crossing over may not be limited to fine scales[4,11]: we observe a weakly significant association of rs6889665 with the total number of crossovers genome-wide in pedigrees ($P = 0.04$), corresponding to an average 1.3 crossovers more per meiosis per derived allele, exceeding the strongest previously known association[22] at *RNF212*.

Conclusions

We have shown that *PRDM9* alleles that bind a novel 17-bp motif and occur at greatly increased frequency in people of West African ancestry have led to a shift in the recombination landscape compared with people of non-African ancestry. The larger number of hotspots available to West Africans implies that at the population level, crossovers are more evenly distributed than in Europeans[10], and thus the shorter extent of West African linkage disequilibrium is not due to differences in demographic history alone (such as the lack of an out-of-Africa founder event)[23]. Our findings also have medical implications, as recombination errors leading to insertions or deletions are known to be associated with recombination hotspots[9,21,24]. Our results predict that the congenital abnormalities that have been associated with the recombination hotpots bound by *PRDM9* A and B alleles will occur at a decreased rate in people of West African ancestry, whereas new diseases will arise due to recombination errors near African-enriched hotspots.

METHODS SUMMARY

We assembled SNP array data from 29,589 unrelated people and 222 nuclear families genotyped at 490,000–910,000 SNPs from the Candidate Gene Association Resource (CARe), studies at the Children's Hospital of Philadelphia (CHOP), the African American Breast Cancer Consortium, the African American Prostate Cancer Consortium and the African American Lung Cancer Consortium. To build a recombination map, we used HAPMIX to localize candidate crossover positions[15], and implemented a MCMC that used the probability distributions for the positions of the filtered crossovers to infer recombination rates for each of 1.3 million inter-SNP intervals. We also implemented a second MCMC that models each individual's set of crossovers as a mixture of an S map, similar to the European deCODE map, and an AE map, and then assigned each individual an 'AE phenotype' corresponding to the proportion of their newly detected crossovers assigned to the AE map. We imputed genotypes at up to three million HapMap2 SNPs[18] and then tested each of these SNPs for association with the AE phenotype and other recombination-related phenotypes. We identified 2,454 candidate African-enriched hotspots with increased recombination rates in the YRI versus CEU maps, and in the AE versus S maps, and searched for motifs enriched at these loci, thus identifying a degenerate 17-bp motif. To study the structure of *PRDM9*, we measured the length of the *PRDM9* ZF array and genotyped rs6889665 in YRI, CEU and the CARe nuclear families; we also carried out imputation based on 1000 Genomes Project short read data[10] to infer the alleles individuals carry, among 29 previously characterized in a sequencing study of *PRDM9* (ref. 9).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 2 February; accepted 27 June 2011.
Published online 20 July 2011.

1. Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29,** 217–222 (2001).
2. McVean, G. A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304,** 581–584 (2004).
3. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310,** 321–324 (2005).
4. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467,** 1099–1103 (2010).
5. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31,** 241–247 (2002).
6. Matise, T. C. *et al.* A second-generation combined linkage-physical map of the human genome. *Genome Res.* **17,** 1783–1786 (2007).
7. Weitkamp, L. R. Proceedings: population differences in meiotic recombination frequency between loci on chromosome 1. *Cytogenet. Cell Genet.* **13,** 179–182 (1974).
8. Jorgenson, E. *et al.* Ethnicity and human genetic linkage maps. *Am. J. Hum. Genet.* **76,** 276–290 (2005).
9. Berg, I. L. *et al. PRDM9* variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genet.* **42,** 859–863 (2010).
10. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
11. Baudat, F. *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327,** 836–840 (2010).
12. Myers, S. *et al.* Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. *Science* **327,** 876–879 (2010).
13. Parvanov, E. D., Petkov, P. M. & Paigen, K. *Prdm9* controls activation of mammalian recombination hotspots. *Science* **327,** 835 (2010).
14. Smith, M. W. *et al.* A high-density admixture map for disease gene discovery in African Americans. *Am. J. Hum. Genet.* **74,** 1001–1013 (2004).
15. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5,** e1000519 (2009).
16. Sankararaman, S., Sridhar, S., Kimmel, G. & Halperin, E. Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.* **82,** 290–303 (2008).
17. Patterson, N. *et al.* Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74,** 979–1000 (2004).
18. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
19. Freedman, M. L. *et al.* Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl Acad. Sci. USA* **103,** 14068–14073 (2006).
20. Coop, G., Wen, X., Ober, C., Pritchard, J. K. & Przeworski, M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* **319,** 1395–1398 (2008).
21. Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hotspots and genome instability in humans. *Nature Genet.* **40,** 1124–1129 (2008).
22. Kong, A. *et al.* Sequence variants in the *RNF212* gene associate with genome-wide recombination rate. *Science* **319,** 1398–1401 (2008).
23. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411,** 199–204 (2001).
24. Raedt, T. D. *et al.* Conservation of hotspots for recombination in low-copy repeats associated with the *NF1* microdeletion. *Nature Genet.* **38,** 1419–1423 (2006).
25. Griffiths, R. C. & Tavaré, S. Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosci.* **127,** 77–98 (1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to the participants who donated DNA samples, to D. Altshuler, J. Buard, K. Bryc, J. Kovacs, B. de Massy, G. McVean, B. Pasaniuc and S. Sankararaman for conversations and critiques, and to A. Auton for facilitating analysis of the 1000 Genomes Project data. Analysis was supported by the Wellcome Trust and NIH grants HL084107 and GM091332. CARe was supported by a contract from the National Heart, Lung and Blood Institute (HHSN268200960009C) to create a phenotype and genotype database for dissemination to the biomedical research community. Eight parent studies contributed phenotypic data and DNA samples through the Broad Institute (N01-HC-65226): 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), the Multi-Ethnic Study of Atherosclerosis (MESA) study, the Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS) and the Sleep Heart Health Study (SHHS). Support for CARe also came from the individual research institutions, investigators, field staff and study participants. Individual funding information is available at http://public.nhlbi.nih.gov/GeneticsGenomics/home/care.aspx. All genome-wide genotyping of samples from the Children's Hospital of Pennsylvania (CHOP) was supported by an Institutional Development Award to the Center for Applied Genomics from the Children's Hospital of Philadelphia, a research award from the Landenberger Foundation and the Cotswold Foundation. We thank all study participants and the staff at the Center for Applied Genomics for performing the genotyping. The African American Breast Cancer Consortium (AABCC) was supported by a DoD Breast Cancer Research Program Era of Hope Scholar Award to C.A.H. and the Norris Foundation, and by grants to the component studies: MEC (CA63464, CA54281); CARE (HD33175); WCHS (CA100598, DAMD 170100334, Breast Cancer Research Foundation); SFBC (CA77305, DAMD 17966071); CBCS (CA58223, ES10126), PLCO (NCI Intramural Research Program); NHBS (CA100374); WFBC (R01-CA73629); and CPS-II (the American Cancer Society). The African American Prostate Cancer Consortium (AAPCC) was supported by grants CA63464, CA54281, CA1326792, CA148085 and HG004726, and by grants to the component studies: PLCO (NCI Intramural Research Program), LAAPC (Cancer Research Fund 99-00524V-10258), both MEC and LAAPC (PC35139, DP000807); MDA (CA68578, CA140388, ES007784, DAMD W81XWH0710645); GECAP (ES011126); CaP Genes (CA88164); IPCG (W81XWH0710122); DCPC (GM08016, DAMD W81XWH0710203, DAMD W81XWH0610066); and SCCS (CA092447, CA68485). The African American Lung Cancer Consortium (AALCC) was supported by grants CA060691, CA87895, PC35145 and CA22453, CA68578, CA140388, ES007784, ES06717, CA55769, CA127219, CA1116460S1, CA1116460, CA121197, CA141716, CA121197S2, CPRIT RP100443, CA148127, DAMD W81XWH0710645, University Cancer Foundation, Duncan Family institute, Center for Community, Implementation and Dissemination Research Core, and by grants to the component studies: PLCO and the Maryland Studies (NCI Intramural Research Program), LAAPC (Cancer Research Fund 99-00524V-10258), and both MEC and LAAPC (PC35139, DP000807).

Author Contributions D.R. and S.R.M. conceived the study. A.G.H., A.T., N.P., Y.S., N.R., C.D.P., G.K.C., K.W., S.G.B., D.R. and S.R.M. performed analyses. N.R. performed the experimental work (genotyping of polymorphisms at *PRDM9*). A.G.H., N.P., J.N.H.,

B.E.H., H.A.T. Jr, A.L.P., H.H., S.J.C., C.A.H., J.G.W., D.R. and S.R.M. coordinated the study. A.G.H., D.R. and S.R.M. wrote the paper. N.R., C.D.P., G.K.C., K.W., S.G.B., S.R., J.N.H., B.E.H., H.A.T. Jr, H.H., S.J.C., C.A.H., J.G.W., D.R. and all the alphabetically listed authors contributed to sample collection and generation of SNP array data. All authors contributed to revision and review of the manuscript.

Author Information Crossover rate estimates for the AA map can be found at http://www.well.ox.ac.uk/~anjali/AAmap/. We also provide estimates of uncertainty for the map based on samples from the MCMC. Association testing results for each SNP are available from the authors on request. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.R. (reich@genetics.med.harvard.edu) or S.R.M. (myers@stats.ox.ac.uk).

Anjali G. Hinch[1], Arti Tandon[2,3], Nick Patterson[2], Yunli Song[4], Nadin Rohland[2,3], Cameron D. Palmer[5,6], Gary K. Chen[7], Kai Wang[8,9], Sarah G. Buxbaum[10], Ermeg L. Akylbekova[10,11], Melinda C. Aldrich[12,13], Christine B. Ambrosone[14], Christopher Amos[15], Elisa V. Bandera[16], Sonja I. Berndt[17], Leslie Bernstein[18], William J. Blot[13,19], Cathryn H. Bock[20], Eric Boerwinkle[21], Qiuyin Cai[13], Neil Caporaso[17], Graham Casey[7], L. Adrienne Cupples[22], Sandra L. Deming[13], W. Ryan Diver[23], Jasmin Divers[24], Myriam Fornage[25], Elizabeth M. Gillanders[26], Joseph Glessner[9], Curtis C. Harris[27], Jennifer J. Hu[28], Sue A. Ingles[7], William Isaacs[29], Esther M. John[30], W. H. Linda Kao[31], Brendan Keating[9], Rick A. Kittles[32], Laurence N. Kolonel[33], Emma Larkin[34], Loic Le Marchand[33], Lorna H. McNeill[35], Robert C. Millikan[36], Adam Murphy[37], Solomon Musani[11], Christine Neslund-Dudas[38], Sarah Nyante[36], George J. Papanicolaou[39], Michael F. Press[7], Bruce M. Psaty[40], Alex P. Reiner[41], Stephen S. Rich[42], Jorge L. Rodriguez-Gil[28], Jerome I. Rotter[43], Benjamin A. Rybicki[38], Ann G. Schwartz[20], Lisa B. Signorello[13,19], Margaret Spitz[15], Sara S. Strom[44], Michael J. Thun[23], Margaret A. Tucker[17], Zhaoming Wang[45], John K. Wiencke[46], John S. Witte[47], Margaret Wrensch[46], Xifeng Wu[15], Yuko Yamamura[44], Krista A. Zanetti[26,27], Wei Zheng[13], Regina G. Ziegler[17], Xiaofeng Zhu[48], Susan Redline[49], Joel N. Hirschhorn[5,6,50], Brian E. Henderson[7], Herman A. Taylor Jr[11,51,52], Alkes L. Price[53], Hakon Hakonarson[9,54], Stephen J. Chanock[17],Christopher A. Haiman[7], James G. Wilson[55], David Reich[2,3]* & Simon R. Myers[1,4]*

[1]Wellcome Trust Centre for Human Genetics, Oxford University, Roosevelt Drive, Oxford OX3 7BN, UK. [2]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [3]Deptartment of Genetics, Harvard Medical School, New Research Building, 77 Ave. Louis Pasteur, Boston, Massachusetts 02115, USA. [4]Department of Statistics, Oxford University, 1 South Parks Road, Oxford OX1 3TG, UK. [5]Program in Medical and Population Genetics, Broad Institute, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [6]Divisions of Endocrinology and Genetics and Program in Genomics, Children's Hospital Boston, Massachusetts 02115, USA. [7]Department of Preventive Medicine and Department of Pathology, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, California 90033, USA. [8]Zilkha Neurogenetic Institute, University of Southern California, Los Angeles, California 90089, USA. [9]Center for Applied Genomics, The Childrens Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. [10]Jackson Heart Study Coordinating Center, Jackson State University, 350 W. Woodrow Wilson Ave., Suite 701, Jackson, Mississippi 39213, USA. [11]Department of Medicine, University of Mississippi Medical Center, 2500 N. State St., Jackson, Mississippi 39216, USA. [12]Department of Thoracic Surgery, Vanderbilt University School of Medicine, Nashville, Tennessee 37203, USA. [13]Division of Epidemiology in the Department of Medicine, Vanderbilt Epidemiology Center; and the Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee 37203, USA. [14]Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, New York 14263, USA. [15]Department of Epidemiology, Division of Cancer Prevention and Population Sciences, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [16]The Cancer Institute of New Jersey, New Brunswick, New Jersey 08903, USA. [17]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892, USA. [18]Division of Cancer Etiology, Department of Population Science, Beckman Research Institute, City of Hope, California 91010, USA. [19]International Epidemiology Institute, Rockville, Maryland 20850, USA. [20]Karmanos Cancer Institute and Department of Oncology, Wayne State University of Medicine, Detroit, Michigan 48201, USA. [21]Human Genetics Center and Division of Epidemiology, University of Texas at Houston, 1200 Herman Pressler St., Houston, Texas 77030, USA. [22]Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Avenue, Boston, Massachusetts 02118 and Framingham Heart Study, Framingham, Massachusetts 01702, USA. [23]Epidemiology Research Program, American Cancer Society, Atlanta, Georgia 30303, USA. [24]Department of Biostatistical Sciences, Wake Forest University School of Medicine WC-2326, Medical Center Blvd., Winston Salem, North Carolina 27157, USA. [25]Institute of Molecular Medicine and Division of Epidemiology, School of Public Health, University of Texas Health Sciences Center at Houston, 1825 Pressler Street, Houston, Texas 77030, USA. [26]Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Maryland 20892, USA. [27]Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland 20892, USA. [28]Sylvester Comprehensive Cancer Center and Department of Epidemiology and Public Health, University of Miami Miller School of Medicine, Miami, Florida 33136, USA. [29]James Buchanan Brady Urological Institute, Johns Hopkins Hospital and Medical Institutions, Baltimore, Maryland 21287, USA. [30]Cancer Prevention Institute of California, Fremont, California 94538; and Stanford University School of Medicine and Stanford Cancer Center, Stanford, California 94305, USA. [31]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, Maryland 21205, USA. [32]Department of Medicine, University of Illinois at Chicago, Chicago, Illinois 60607, USA. [33]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii 96813, USA. [34]Department of Medicine, Division of Allergy, Pulmonary and Critical Care, 6100 Medical Center East, Vanderbilt University Medical Center, Nashville, Tennessee 37232-8300, USA. [35]Department of Health Disparities Research, Division of OVP, Cancer Prevention and Population Sciences, and Center for Community Implementation and Dissemination Research, Duncan Family Institute, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [36]Department of Epidemiology, Gillings School of Global Public Health, and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina 27599, USA. [37]Department of Urology, Northwestern University, Chicago, Illinois 60611, USA. [38]Department of Public Health Sciences, Henry Ford Hospital, Detroit, Michigan 48202, USA. [39]Division of Cardiovascular Sciences, National Heart, Lung and Blood Institute, 6701 Rockledge Drive, Bethesda, Maryland 20892, USA. [40]Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology & Health Services, University of Washington; Group Health Research Institute; Group Health Cooperative; 1730 Minor Ave., Seattle, Washington 98101, USA. [41]Department of Epidemiology, University of Washington, Box 357236 Seattle, Washington 98195, USA. [42]Center for Public Health Genomics, University of Virginia, West Complex Room 6111, Charlottesville, Virginia 22908, USA. [43]Medical Genetics Institute, Cedars-Sinai Medical Center, 8700 Beverly Blvd, Los Angeles, California 90048, USA. [44]Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, USA. [45]Core Genotype Facility, SAIC-Frederick, Inc., National Cancer Institute-Frederick, Frederick, Maryland 20877, USA. [46]University of California San Francisco, San Francisco, California 94158, USA. [47]Institute for Human Genetics, Departments of Epidemiology and Biostatistics and Urology, University of California, San Francisco, San Francisco, California 94158, USA. [48]Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Wolstein Research Building, Cleveland, Ohio 44106, USA. [49]Brigham and Women's Hospital, Department of Medicine, Division of Sleep Medicine, 75 Francis Street, Boston, Massachusetts 02115, USA. [50]Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. [51]Jackson State University, 1400 Lynch Street, Jackson, Mississippi 39217, USA. [52]Tougaloo College, 500 West County Line Road, Tougaloo, Mississippi 39174, USA. [53]Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [54]Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA. [55]Department of Physiology and Biophysics, University of Mississippi Medical Center, 2500 N. State St., Jackson, Mississippi 39216, USA.
*These authors contributed equally to this work.

METHODS

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 analysed 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 paediatric patients. All samples analysed here were genotyped on either the Illumina 610-Quad or Illumina HumanHap550 array. After removing individuals known to be related, identifying American Americans by multidimensional 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[26] (>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), 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 (AALCC), 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 analyse) 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. Supplementary Note 2 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 YRI and CEU data from the HapMap3 data set[27]. 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 as a 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[6] (Supplementary Note 1).

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 linkage disequilibrium[3,18] (Supplementary Note 1). For comparisons to the deCODE map and linkage-disequilibrium-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 analysed across all five sample sets. Our goal was to estimate a crossover rate for each of these intervals. We modelled 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 (Supplementary Note 1). 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 Supplementary Note 1 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 used 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 $>2\,\mathrm{cM\,Mb^{-1}}$ and at least double that in the S map, and in addition the YRI map rate was $>2\,\mathrm{cM\,Mb^{-1}}$ 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 $>2\,\mathrm{cM\,Mb^{-1}}$, 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[28] was used to impute up to 3,058,149 SNP genotypes from HapMap2 (ref. 18) into all African Americans we analysed, using the unrelated YRI and CEU samples as combined reference panels. We tested for association at all SNPs with minor allele frequency > 1%. 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 (Supplementary Note 4). We observe slight inflation of the association statistics genome-wide compared with the expectation (the Genomic Control inflation factor[29] is 1.046 for the AE phenotype and 1.038 for the hotspot usage phenotype), which we propose may reflect cryptic relatedness among samples (Supplementary Note 4). We report *P* values after correction using Genomic Control[29].

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 (ref. 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*[25], 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 YRI data only and used this to construct Fig. 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: (1) they were statistically significant after Bonferroni correction for the number of motifs tested; (2) they were overrepresented in the African-enriched set; (3) they were statistically significant on both the repeat and non-repeat backgrounds ($P < 0.01$) independently; and (4) 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-nucleotide oligomer CCCCAGTGA. We explored whether flanking DNA around exact matches to this motif also had 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 (Supplementary Note 6 has full details). To identify close matches to this 17-bp motif among all matches to the 9-bp motif in the genome, for every occurrence of the 9-bp motif, 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 (Fig. 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 (Supplementary Fig. 8).

***PRDM9* ZF length typing and genotyping of rs6889665.** To determine the number of ZF 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′-GGGGAATATAAGGGG TCAGC-3′). Product lengths ranged between 7 and 20 repeats (801–1,893 bp). Four of the 166 African-American samples did not show an amplification product, presumably because of insufficient DNA quality. We also genotyped 90 YRI and 90 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′, Locked Nucleic Acid (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.

26. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
27. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
28. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34,** 816–834 (2010).
29. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).