doi:10.1038/nature10336

Table of contents

Supplementary notes

1) Building the African American Map 2-9
2) Building the Pedigree Map 10-13
3) Building the African-enriched Map 14-17
4) Association testing to recombination-related phenotypes 18-22
5) Determination of *PRDM9* zinc finger allele using 1000 Genomes Project data 23-27
6) Discovery of a sequence motif associated with African American recombination 28-37

Supplementary tables

1) Comparison of the AA Map and deCODE Map using Pearson and Spearman correlations 38
2) Results of genome-wide association testing for the AE and hotspot usage phenotypes 39

Supplementary figures

1) Localization of hotspots by the MCMC 40
2) Comparing the calibration of the AA and Basic maps (based on the Pedigree Map) 41
3) Recombination rates near crossovers detected in African American pedigrees 42
4) Detailed comparisons of genetic maps over selected 200kb regions 43
5) The AA and YRI maps show a shared signal for African-enriched hotspots… 44
6) Variation at *PRDM9* is associated with the African Enrichment phenotype 45
7) Relationship between alleles at rs6889665 and *PRDM9* variants 46
8) Hotspot activity around motif CCCCAGTGA and flanking sequencing 47
9) Influence of flanking bases around 9-mer motif CCCCAGTGA 48
10) Rates around occurrences of CCCCAGTGA within L1PA10 and L1PA13 repeats 49
11) Rates in different maps for pedigree individuals who are CT or TT at SNP rs6889665 50

Note S1
Building the African American Map

(i) Strategy for building a "Basic Map"

We ran HAPMIX on the unphased genotypes of every African American in the study separately. HAPMIX requires a genetic map as a prior distribution, and since we did not want our map to be influenced by previous recombination maps, we provided it with a map with a uniform recombination rate over every chromosome arm, using the rates from the table below:

Chrom-osome	p-arm end		q-arm start		q-arm end	
	physical (Build 36)	genetic (cM)	physical (Build 36)	genetic (cM)	physical (Build 36)	genetic (cM)
1	120,330,455	150.83	144,243,274	151.41	266,127,003	308.19
2	88,457,695	111.79	94,912,172	111.90	262,575,050	284.34
3	90,335,510	110.04	95,301,156	110.16	219,287,577	246.35
4	47,052,440	67.38	52,814,094	68.47	211,158,857	235.77
5	45,603,505	67.28	49,659,296	67.50	200,420,866	231.25
6	58,440,820	84.17	62,118,323	84.37	190,747,902	214.57
7	57,884,453	77.26	62,171,909	77.55	178,710,965	210.26
8	42,668,804	64.01	47,154,379	64.38	166,114,869	191.11
9	45,617,414	66.52	66,481,119	66.62	160,137,385	195.02
10	38,715,905	63.33	42,274,907	63.47	155,053,665	198.19
11	51,303,111	69.25	54,945,985	69.30	154,402,514	184.59
12	34,142,287	58.67	36,887,280	58.81	152,287,718	200.90
13	17,394,434	0.00	18,394,434	1.38	134,031,187	160.81
14	18,515,210	0.00	19,515,210	1.44	126,357,542	155.54
15	19,367,093	0.00	20,367,093	1.66	120,181,650	167.32
16	34,715,504	59.66	45,272,030	59.77	108,640,449	168.01
17	21,864,298	52.69	22,336,210	52.92	98,609,338	167.48
18	15,072,789	47.54	16,909,449	47.67	95,965,205	157.88
19	24,163,969	50.26	32,842,516	50.46	83,785,296	153.90
20	26,114,909	53.48	29,310,063	53.69	82,381,835	151.81
21	12,678,219	0.00	13,678,219	2.11	66,902,240	114.41
22	14,440,166	0.00	15,440,166	2.32	69,508,925	127.92
X *	57,820,632	74.60	63,670,690	74.75	174,803,587	164.77

Note: This table was used to determine the recombination rate priors for running HAPMIX (uniform on each chromosome arm as described in the main text). The mappings between physical and genetic positions shown were adapted from the Rutgers smoothed genetic map, which we download from http://compgen.rutgers.edu/RutgersMap/Default.aspx on September 30, 2009. We then modified the map in two ways. First, we added a constant on each chromosome so that the extrapolated genetic position at physical position 0 bp was 0 cM. Second, we looked up the positions of the centromeres on the UCSC genome browser, and manually defined a ~~"p-~~arm end" and ~~"q-~~arm" start on either boundary. The recombination rate prior used in the HAPMIX runs was piecewise uniform over three segments: (i) the ~~"p-~~arm" between position 0 of both the physical and genetic maps and the p-arm end, (ii) the ~~"ce~~ntromere" between the p-arm end and q-arm start, and (ii) the ~~"q-~~arm" between the q-arm start and the q-arm end. It was important to separately define the centromeric region as otherwise we had extremely low recombination rate priors for inter-SNP intervals near the boundary, inappropriately biasing our estimates.

To infer the positions of putative recombination hotspots, we use the fact that for every individual, HAPMIX infers the posterior probability (integrating over paths in the Hidden Markov Model) that the individual has 0, 1 or 2 European or African alleles at each genotyped SNP. We build our map using only these probabilities, in each individual and across the genome. First, we use these probabilities to compute the expected number of European alleles at each SNP. SNPs with an expected value between 0 and 1, or between 1 and 2, are in a state where HAPMIX is uncertain about ancestry. These may lie within a recombination event. We base our map on the idea that within regions across which ancestry changes by an integer value (e.g. from 0 to 1 African alleles), the change in the number of African alleles between two SNPs almost exactly equals the probability of the ancestry switch occurring between those SNPs, provided the likelihood of more than one ancestry switch in the same region is small. We accordingly filter out events where there is any weight placed on multiple events.

(ii) Restricting analysis to the most confidently inferred crossover events

We found that using the raw output of HAPMIX to infer the positions of crossovers resulted in a less accurate map (as assessed by correlation to previously published maps based on linkage disequilibrium) than we obtained when we applied rigorous filters to remove potentially false-positive crossover events. The filters we applied were: (a) we restricted to the crossover events that HAPMIX was most confident had occurred; (b) we restricted to crossover events that were accurately localized; and (c) we restricted to crossover events that were flanked by large blocks of contiguous European or African ancestry. We discuss each of these filters in turn.

(a) Restricting to the crossover events in which HAPMIX is most confident

We observed that many loci have the same inferred number of European alleles within a short distance on either side of a putative crossover event. For example, an individual may have an ancestry sequence of {0, 0, 0.1, 0.2, 0.1, 0, 0} European alleles, suggesting a 20% probability of European ancestry in the middle of a block of African ancestry. This could happen, for example, when two overlapping events have occurred on each of the two homologous chromosomes, thereby partially erasing the signal. This could also happen if HAPMIX has made an error. Either way, we cannot be confident that a unique event has occurred in that interval. We filter all transitions such that we are at least 95% sure that they have occurred. To do this, we require that there has been a transition from less than 0.025 expected European alleles to more than 0.975 expected European alleles (or vice versa) or from 1.025 expected European alleles to 1.975 (or vice versa).

(b) Restricting to crossovers that are accurately localized

Some transitions are non-monotonic, that is, the expected number of European alleles rises and falls even though the end-points indicate a complete transition. For example, an individual may have an ancestry sequence of {0, 0, 0.4, 0.6, 0.4, 0.8, 1, 1}. This could happen when two overlapping crossover events have occurred on each of the individual‘s chromosomes and the signals have become merged. We cannot, then, work out the correct end-points of the transition. We filter out events with a non-monotonic ancestry change that is greater than 1%, where non-monotonicity is defined as an ancestry change opposite to the rest of the transition.

(c) Restricting to crossovers flanked by large blocks of contiguous African or European ancestry

We observed an excess of very short blocks of contiguous ancestry compared with an exponential distribution. Since in theory, the location of recombination events should be Poisson-randomly distributed[1], this suggests the presence of false-positive crossover events in

our dataset, which are especially common near short blocks of contiguous ancestry. On these grounds we filtered out putative crossover events with flanking ancestry chunks shorter than 2 cM (using the HapMap population-averaged LD-based map to assess length[2]).

We identified 2,805,677 events in total using criterion (a) above. A total of 221,208 events, or approximately 7.9% were filtered out using the monotonicity criterion. A further 471,176 events, or approximately 16.8% of all events, were removed for having short flanking ancestry blocks. This left 2,113,293 candidate crossover events, or approximately 71 events per individual.

We next built a probability distribution for each event as follows:

(1) We assumed that the crossover event occurred where the expected number of European chromosomes inferred by HAPMIX is between 0.025-0.975 or 1.025-1.975.
(2) Within this region, we interpreted the change in the number of European alleles in each inter-SNP interval as proportional to the probability a recombination event occurred.
(3) We then chose the constant of proportionality following (2), so that for each event, the probabilities over all inter-SNP intervals overlapping the event sum to one.

To assess how precisely this procedure infers the physical positions of crossovers, we examined the probability of recombination across all events in all SNP intervals and sorted by probability density (probability of recombination divided by length of the interval) in unrelated individuals (restricting this analysis to the 6,209 unrelated individuals from the CARe study). Upon adding the probability of recombination in intervals in order of decreasing probability density and normalizing by the number of events, we found that 50% of the recombination probability fell within 70 kb. Informally, this means that, on average, the ‗best' half of information about recombination in our data is contained with 70 kb.

We conclude this section on the identification of putative crossovers by noting the potential biases in rate estimates that might arise from our filters. We considered two in particular:

(i) Our resolution-based filters and monotonicity filters (a) and (b) might induce some bias if they remove a larger fraction of events in particular regions that are difficult to analyze.
(ii) In principle, restricting to loci flanked by large stretches of contiguous ancestry ought not to bias the map since block sizes have a distribution independent of the rate (when block sizes are measured in units of Morgans or equivalent). However, we might have some bias in regions where broad-scale LD-based recombination rates have large errors. We do not believe these to be common problems based on strong genome-wide correlations between maps, particularly at broad scales.

To determine whether these filters introduced bias in the map, we compared them to a directly inferred Pedigree Map and to previous published maps, as discussed in more detail below.

(iii)Building a "Basic" Map

The individuals in our dataset were genotyped on several different SNP arrays. The Affymetrix 6.0 array was genotyped in 6,209 individuals from CARe (580,000 SNPs after filtering). The Illumina 1M array was genotyped in 6,540 men from the African American Prostate Cancer Consortium (896,036 SNPs after filtering), 5,203 women from the African American Breast Cancer Consortium (894,717 SNPs after filtering) and 4,134 individuals from the African American Lung Cancer Consortium (906,687 SNPs after filtering). One of the Illumina 610-Quad or Illumina HumanHap550 arrays was genotyped in 7,503 individuals enrolled in studies at

the Children's Hospital of Pennsylvania. There were 1,271,992 SNPs in the union of these datasets. We performed linear interpolation to assign probabilities of recombination to sub-intervals created by SNPs that were not present in the HAPMIX output for any one individual.

To build the Basic Map based on the inferred probability distributions for each crossover event, we added up the probability distributions across all 2.1 million candidate crossovers to obtain a function proportional to the probability of crossover at each point in the genome.

We do not know the total number of generations of recombination represented in the ancestry of each individual since European and African admixture began. As a result, this is a relative genetic map rather than an absolute map. We therefore renormalized the total length of the genetic map to force it to equal that of the Hapmap2 population-averaged COMBINED LD-based map[2].

The crossovers we detect in an individual today may have occurred in maternal or paternal ancestors. Therefore, this map is a sex-averaged map.

(iv) Building an "AA Map" by leveraging shared recombination rates across samples

An important fact that is not taken into account in the construction of the Basic Map is that recombination rates are shared across individuals within a population (indeed, identifying the common rates is the goal of building a genetic map). Another fact that is not taken into account is that in humans, the majority of recombination occurs in discrete hotspots that are 1-2 kb in size[3,4,5]. To capture these observations, we implemented a Bayesian Markov Chain Monte Carlo (MCMC) procedure that takes advantage of the knowledge that rates are shared across samples and that recombination occurs in hotspots.

In detail, we modeled the genome as a sequence of SNP intervals with independent recombination rates shared by all individuals (motivated by the fact that most hotspots are small enough to be contained within a unique inter-SNP interval). We imposed a gamma prior on the recombination rate in each interval so that the model is unaffected by SNP density or a particular SNP set that is chosen. A challenge in building an MCMC in this context is that there is known to be variation in the density and intensity of hotspots along the genome (for example, the telomeres are different from the centromeres, and GC rich regions have systematically different rates over large scales). To account for this long-range autocorrelation in rates without making the model intractable, we use the rates in the "Basic Map" as the mean of the gamma prior for each interval. As described later in Note S1, the resolution of the Basic Map—defined as the interval size in which the mean standard error in the rate estimate in the map is the same as the average rate—is 35kb, and thus is much larger than a hotspot.

To use these priors to build the AA Map, we performed Gibbs sampling over the location of each event and the recombination rate in each interval. In detail, the MCMC proceeds as follows:

(1) *Initialization of the MCMC*
We initialize the Gibbs sampler with recombination rates sampled from the Basic Map prior.

(2) Step 1: Sampling the locations of each of the 2.1 million crossover events

For each crossover, we sample the SNP interval in which it occurred conditional on the most recent set of sampled rates. It would be ideal to re-run HAPMIX again at each iteration of the MCMC, but this is computationally prohibitive. However, we can use the output of HAPMIX when run with a uniform recombination rate prior (Note S1, part (ii)), and mathematically rescale these probabilities to be appropriate for our purpose as described below.

In this step of the Gibbs sampler we sample the locations of all events E conditional on the vector $\overline{r}$ of recombination rates in each interval. In a later step, we will sample the rates conditional on events. In order to do the former, we need to compute $P(E \mid \overline{r})$.

Let an observed recombination event $e \in E$ span SNP intervals from 1 to n. Without loss of generality we assume that the switch in e happens from African to European ancestry. If a_i is ancestry and s_i is the allele type at locus i on the chromosome undergoing recombination, then e is the event that $(a_1 = Afr)$ and $(a_{n+1} = Eur)$ and for all SNPs $i \in \{1,2,\cdots,n\}$, the allele observed in the individual harboring the recombination event is s_i. Let r_i represent the rate for interval $i \in \{1,2,\cdots,n\}$. Let rec_i represent the event that the crossover in e occurs in interval i. We assume that exactly one ancestry switch happens in any observed event. Therefore,

$$P(e \mid \overline{r}) = \sum_{i=1}^{n} P(e, rec_i \mid \overline{r}).$$

For each interval i, we define $h_{1,i}$ as the haplotype to the left of i, i.e.,$\{s_1,\cdots,s_i\}$and similarly $h_{i+1,n+1}$ as the haplotype to the right of i, i.e., $\{s_{i+1},\cdots,s_{n+1}\}$. Looking at recombination in interval i, we have, therefore,

$$P(e, rec_i \mid \overline{r}) = P\big(h_{1,i}, a_1 = Afr \mid rec_i, \overline{r}\big)P\big(h_{i+1,n+1}, a_{n+1} = Eur \mid rec_i, \overline{r}\big)P(rec_i \mid \overline{r}).$$

Conditional on recombination in interval i, the probabilities of the left and right haplotypes are independent. In addition, suppose the prior probability of an African allele is θ and a European allele is $(1-\theta)$ (based on the genome-wide ancestry distribution across a population). The probability of which ancestry is found on the donor chromosome after a double strand break on the initiating chromosome is obviously independent of recombination rates. We make the further assumption that the probability of haplotype $h_{1,i}$ in Africans and $h_{i+1,n+1}$ in Europeans is independent of the probability of recombination in interval i. Therefore,

$$P(\mathrm{e}, rec_i \mid \overline{r}) = P\big(h_{1,i} \mid a_1 = Afr, \overline{r}\big)P\big(h_{i+1,n+1} \mid a_{n+1} = Eur, \overline{r}\big)r_i\theta(1-\theta).$$

We make the assumption that the probability of observing the haplotypes $h_{1,i}$ in Africans and $h_{i+1,n+1}$ in Europeans is not influenced by the probability distribution of ancestry switches and recombination rates $\overline{r}$. This is not strictly true. The reason for this approximation is that we do not want to use LD patterns to estimate rates, and thus we ignore the effect of rates on LD patterns within a population. We run HAPMIX with uniform rates $\overline{r}_{flat}$, and make the approximation that the probability of the haplotypes is the same if the rates are $\overline{r}$ or $\overline{r}_{flat}$:

$$P\big(h_{1,i} \mid a_1 = Afr, \overline{r}\big)P\big(h_{i+1,n+1} \mid a_{n+1} = Eur, \overline{r}\big)\theta(1-\theta) \approx P(\mathrm{e}, rec_i \mid \overline{r}_{flat})/r_{flat_i}.$$

We add over the possibility of recombination in all intervals to get

$$P(e \mid \bar{r}) \approx \sum_{i=1}^{n} \frac{r_i}{r_{flat_i}} P(rec_i \mid e, \bar{r}_{flat}) P(e \mid \bar{r}_{flat})$$
$$\propto \sum_{i=1}^{n} \frac{r_i}{r_{flat_i}} P(rec_i \mid e, \bar{r}_{flat}).$$

Therefore, assuming independence of events, we multiply over all events $e_j \in E$:

$$P(E \mid \bar{r}) \propto \prod_j \sum_i \frac{r_i}{r_{flat_i}} P(rec_i \mid e_j, \bar{r}_{flat}).$$

We obtain the probability of recombination in any interval in an event given uniform rates from the HAPMIX output, as discussed in Note S1 (ii). Therefore, for every event we have obtained the likelihood of the occurrence of recombination in each interval overlapping the event. We normalize these likelihoods to obtain a probability distribution for each event.

We use this distribution for every event to sample the SNP interval in which an event occurred. We do this for all events, and then count how many events were placed into each SNP interval *i* and call it n_i.

(3) *Step 2: Sampling the recombination rates in each of the inter-SNP intervals*
For each of the 1.3 million inter-SNP intervals, we calculate the distribution of rates using the number of events sampled to have occurred in that interval in the previous step.

Specifically, in this step of the Gibbs Sampler, we sample rates in each interval $\bar{r}$ conditional on the recombination events E by computing $P(\bar{r}|E)$. Using Bayes rule, we have

$$P(\bar{r}|E) \propto P(E \mid \bar{r}) P(\bar{r})$$

Since we know the location of each event (sampled at the end of the previous step), any further information about events such as ancestry is irrelevant. At the end of the previous step, we calculated the total number of events n_i in each interval *i*. We assume that, given the rate r_i in interval *i*, the number of events n_i in that interval is independent of the rate in any other SNP interval. Therefore, if the total number of SNP intervals is *N*,

$$P(\bar{r}|E) \propto P(\bar{r}) \prod_{i=1}^{N} P(n_i \mid r_i).$$

We model the event counts in each interval *i* on each chromosome as having a Poisson distribution with rate r_i. If the total number of individuals is *I*,

$$n_i \sim \text{Poisson}(2Ir_i).$$

We use an independent Gamma prior on each rate r_i. The implicit assumption here is that hotspots are small enough to be contained within a SNP interval though this may not be

strictly true and is an approximation. The parameters of the gamma distribution for each r_i were calculated such that the mean is the same as that of the Basic Map. The variance is estimated using the genome-wide distribution of rates in the population averaged LD-based map. The unit variance is constant genome-wide, and the variance in any interval is proportional to the size of the interval, consistent with a model for independent rates.

$$r_i \sim \text{Gamma}(\alpha_i, \beta_i)$$

Since we use a gamma prior, we are able to take advantage of the conjugacy of Poisson and gamma distributions to obtain a gamma posterior.

$$r_i | \, E \sim \text{Gamma}(\alpha_i + n_i, \beta_i + 2I)$$

We can now sample a rate for each SNP interval from its respective posterior distribution to complete one iteration of the Gibbs Sampler.

(4) *Iteration of MCMC*
To obtain the AA Map, we ran at least five MCMC chains with at least 1 million iterations each for every chromosome. We removed the first 300,000 samples as burn-in and computed the mean recombination rate in each inter-SNP interval using every 500th sample. We normalized the AA Map rates to ensure that the total map length is equal to that of the COMBINED LD-based map.[2]

(iv) Resolution of AA map and uncertainty in rate estimates

One important advantage of using the MCMC procedure is that it can be used to generate an uncertainty estimate for the recombination rate in each inter-SNP interval. The reason for this is that the MCMC resamples the positions of ancestry switch-points in each iteration and thus builds up a probability distribution for the true position. Thus, we can use samples collected during the procedure to estimate the resolution of the AA and Basic maps. For every 10,000th sample from the MCMC (after discarding burn-in), we calculated the squared error of the sample relative to the map at different scales. If the posterior samples the correct distribution of potential true maps, the average of these values estimates the mean standard error of the map.

One possible definition of map resolution is the interval size in which, on average, there is one perfectly resolved recombination event. Crossovers in the AA map are *not* perfectly resolved, so we look for an analogous way of estimating resolution. If crossovers were perfectly resolved, the number of crossovers in a short interval can be modeled as having a Poisson distribution with mean $\lambda > 0$, and so standard deviation $\sqrt{\lambda}$. The possible definition above corresponds to $\lambda = 1$; in this case and only this case, the coefficient of variation, i.e., the ratio of the expected number of crossovers and its standard deviation, is 1. The interval size corresponding to the coefficient of variation being equal to 1 can therefore be used, in general, as a map resolution estimate. Using this definition, we find that the resolution of the AA Map is 6 kb which is greatly improved relative to 35 kb for the Basic Map. Given that the size of the human genome is approximately 3 Gb, a resolution of 6kb corresponds to roughly 500,000 perfectly resolved crossovers.

To assess whether uncertainty estimates are well calibrated, we examined families where we had

genotype data for both parents and at least three children. We ran HAPMIX on the parents and each child in those families and identified a set of 370 candidate crossovers events that were present in the children but not the parents (Figure 1A). We matched them to crossover events that were independently identified using a pedigree-based approach in the same individuals (Note S2). Figure S2 compares the probability of recombination in SNP intervals within the HAPMIX version of those events when using the AA map as prior with the probability of recombination in those intervals estimated using pedigrees. (Since pedigree events are better resolved than HAPMIX events, the latter will frequently be zero.) The SNP intervals are then sorted in order of decreasing recombination probability density per base of sequence (i.e., heat) in the HAPMIX version of events. The x-axis has 10% bins of recombination probability, with the leftmost containing the hottest SNP intervals and rightmost bin the coolest. The y-axis shows the corresponding actual probability in those SNP intervals, estimated using pedigrees. The decreasing curve of the Basic Map suggests that it underestimates the heat of the hottest SNP intervals (and, by extension, the hotspots) while overestimating the heat of cooler regions. The curve of the AA map, on the other hand, is almost flat suggesting accurate estimation of heat in the AA map, at least up to the resolution of the pedigree based events.

(iv) Correlation Analysis of AA and deCODE maps

In Table 1, we compare the Pearson correlations of the AA and deCODE maps with LD-based genetic maps. This method calculates the squared difference in rates between the two maps. Squared difference in rates is naturally dominated by the error in the hotter regions of the genome. To give a more equitable weight to error in the less hot regions of the genome, we can calculate the Spearman correlation. This calculation assigns a rank to each rate and uses the ranks to calculate the squared difference instead of the rates themselves. Spearman correlations (Table S1) show that the AA map has a far higher correlation than the deCODE map with LD-based maps in both Europeans and Yoruba at fine scales. The Pearson and Spearman correlations together suggest that the AA map has a fundamentally higher resolution and is more accurate than the deCODE map for all hotspot intensities.

References for Note S1

[1] Falush D., Stephens M. & Pritchard J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164,** 1567-1587 (2003).
[2] The International Haplotype Map Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851-861.
[3] Jeffreys, A.J., Kauppi, L., Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29,** 217-212 (2001).
[4] McVean, G.A. *et al.* The fine-scale structure of recombination rate variation in the human genome. Science **304,** 581-584.
[5] 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).

Note S2
Building the Pedigree Map

Our African American Pedigree Map is based on families genotyped on three different SNP arrays. First, it is based on 135 nuclear African American families from the Jackson Heart Study and the Cleveland Family Study in which at least two children and at least one parent were genotyped on the Affymetrix 6.0 array. Second, it is based on 87 nuclear families from a study at the Children's Hospital of Philadelphia (CHOP), also with at least two children and one parent, which were genotyped on either an Illumina 610-Quad or an Illumina HumanHap550 array. Of these families, 27 had genotype data for both parents and 195 had genotype data for one parent. A total of 1,056 meioses were represented in the data set, which is the first large pedigree map reported in a population of West African ancestry.

To detect recombination events in the presence of missing parental genotype data, we implemented a Hidden Markov Model (HMM) to jointly infer the parental genotype and Identity-by-Descent (IBD) status in children. The idea is that in a family with 2 children we expect to see both alleles from the missing parent half the time. That fraction is three-quarters in a family with 3 children, seven-eighths in a family with 4 children and so on. When both alleles from a parent can be inferred, we can identify the IBD status from the missing parent to each of the children and identify recombination events whenever there is an informative marker.

To perform this inference, we thinned SNPs to a set of unlinked SNPs with minor allele frequency ≥ 0.4. The hidden states in the HMM are the phased parental chromosomes and IBD status of children at each of these sites.

Let M_0 and M_1 be the phased maternal chromosomes and P_0 and P_1 be the phased paternal chromosomes. We assume all SNPs are biallelic and without loss of generality we code all alleles as being either 0 or 1. Therefore, if the total number of SNPs after filtering is N, then for each $s \in \{1,2,\cdots,N\}$, $M_0\,(s) = 0$ or $M_0\,(s) = 1$. The same is true for M_1, P_0 and P_1. There are four possibilities for the hidden state of the mother at each site, and the same for the father.

Let the number of genotyped children be n. For each child $C^j \in \{C^1,\cdots,C^n\}$, let C_M^j and C_P^j be the maternal and paternal inheritance vectors, respectively. They are defined as $C_M^j(s) = 0$ if the child inherits maternal chromosome M_0 at site s and $C_M^j(s) = 1$ if the child inherits maternal chromosome M_1 at that site. Similarly, the values are defined as $C_P^j(s) = 0$ if the child inherits paternal chromosome P_0 at site s and $C_P^j(s) = 1$ if the child inherits paternal chromosome P_1 at that site. There are, therefore, four possibilities for the hidden state of each child at each site and $4n$ states altogether for the children.

Since the markers are unlinked, the transition probability from one parental state at s to the next at site s+1 is the product of the allele frequencies at site s+1. Transitions in the mother and father are obviously independent of each other and of transitions in the children's hidden states.

Transition in the hidden state of a child represents a change in the inherited chromosome from a parent and thus represents a crossover. We assume that the probability of more than one

crossover in any SNP interval is negligible. To avoid bias in the inferred location of recombination towards any particular map, we used a uniform recombination rate. Therefore, if the recombination rate (in units of Morgans) between sites s and $s+1$ is r_s, and i and j represent the full state matrix of the inheritance of the n children then

$$
\begin{array}{ll}
p^s(i,j) = e^{-2nr_s} & \text{if } \forall k : C_M^k(s) = C_M^k(s+1) \;\&\; C_P^k(s) = C_P^k(s+1) \\
p^s(i,j) = \dfrac{1-e^{-2nr_s}}{2n} & \text{if } C_M^l(s) \neq C_M^l(s+1) \;\&\; C_P^l(s) = C_P^l(s+1) \;\& \\
 & \quad \forall k \neq l : C_M^k(s) = C_M^k(s+1) \;\&\; C_P^k(s) = C_P^k(s+1) \\
p^s(i,j) = \dfrac{1-e^{-2nr_s}}{2n} & \text{if } C_M^l(s) = C_M^l(s+1) \;\&\; C_P^l(s) \neq C_P^l(s+1) \;\& \\
 & \quad \forall k \neq l : C_M^k(s) = C_M^k(s+1) \;\&\; C_P^k(s) = C_P^k(s+1) \\
0 & \text{all other cases}
\end{array}
$$

The first case represents no crossovers. The second case represents a crossover in the maternal chromosome inherited by child l. The third case represents a crossover in the paternal chromosome inherited by child l. All other cases consist of more than one crossover in the same interval and are ignored.

To allow for genotyping error, we allow a probability p_e of 0.1% for mismatch between the observed allele and the true allele (hidden state). The emission probabilities $e_X(s)$ reflect the number of ways in which the observed genotype g_s could have been produced at site s for individual X. g_s could be 0, 1, 2 or unknown. For example, for the mother,

$$
\begin{array}{lll}
e_M(s) = p_e \cdot p_e & \text{if } \mathbf{abs}(g_s - (M_0(s) + M_1(s)) = 2 & \{\text{both alleles incorrect}\} \\
= 2 \cdot p_e \cdot (1-p_e) & \text{if } \mathbf{abs}(g_s - (M_0(s) + M_1(s)) = 1 & \{\text{one allele incorrect}\} \\
= 2 \cdot (1-p_e) \cdot (1-p_e) & \text{if } g_s = M_0(s) + M_1(s) & \{\text{both alleles correct}\} \\
\propto 1 & \text{if } g_s \text{ is unknown} & \{\text{no information, multiply} \\
 & & \text{likelihood by 1 WLOG}\}
\end{array}
$$

The corresponding emission probabilities hold for the paternal genotypes.

For the children, the haplotypes depend on the alleles transmitted from the parents. For child C^j at site s, for example, the inheritance vector from the mother is $C_M^j(s)$, and therefore the allele inherited by the child is the allele on that maternal chromosome. The emission probabilities for genotype g_s for child C^j are, therefore,

$$
\begin{array}{lll}
e_{C^j}(s) = p_e \cdot p_e & \text{if } \mathbf{abs}(g_s - (M_{C_M^j(s)}(s) + P_{C_P^j(s)}(s)) = 2 & \{\text{both alleles incorrect}\} \\
= 2 \cdot p_e \cdot (1-p_e) & \text{if } \mathbf{abs}(g_s - (M_{C_M^j(s)}(s) + P_{C_P^j(s)}(s)) = 1 & \{\text{one allele incorrect}\} \\
= 2 \cdot (1-p_e) \cdot (1-p_e) & \text{if } g_s = M_{C_M^j(s)}(s) + P_{C_P^j(s)}(s) & \{\text{both alleles correct}\} \\
\propto 1 & \text{if } g_s \text{ is unknown} & \{\text{no information,} \\
 & & \text{multiply likelihood} \\
 & & \text{by 1 WLOG}\}
\end{array}
$$

We ran this HMM for all 222 families. To determine its effectiveness in the presence of missing data, we ran it twice for several of the families where genotype data was available for both parents: once with both parents and once by hiding the genotype of one parent. All crossovers inferred with both parents were also inferred with one parent.

To improve the resolution of inferred crossovers, we performed a slightly modified analysis for each event separately. For each event, we restricted the HMM to SNPs in the region inferred by the previous HMM and a few hundred SNPs in the neighborhood as long as they are not within another event. In order to infer crossover boundaries as accurately as possible, we filtered out crossovers that overlapped with others. In addition, we used all available SNPs as opposed to only SNPs in low LD with each other. Most importantly, we conditioned on there being exactly one crossover in the whole region. The HMM was otherwise unchanged.

We had an important reason for implementing this two-step procedure. An alternative may have been to use the full set of SNPs in the first HMM analysis instead of using only highly informative unlinked SNPs. Since we do not model LD between SNPs, the HMM with the full set of SNPs has a flawed transition matrix. Consider a situation where the genotype of a parent is unknown and the parent is homozygous in a multi-SNP haplotype which is then inherited by two children from separate parental chromosomes. Depending on the allele frequencies of the SNPs in the haplotype, an HMM which fails to model LD within the haplotype may find it unlikely that two children have inherited identical SNPs from different chromosomes in the parent. Depending on the recombination rates in the region, the HMM may find it more likely that the children have inherited the same chromosome from the parent and then place a recombination event at each end of the haplotype in order to accommodate the absence of IBD in the regions flanking the haplotype. This results in a pair of false recombination events being inferred.

When we use only unlinked markers in the first step of our two-step procedure, however, this problem is removed. In the next step, we rely on two conditions to produce the correct range for the recombination event. First, by restricting to the neighborhood of the recombination event inferred in the first step and conditioning on only one event being present, we reduce the chance of picking homozygous haplotypes since such a homozygous haplotype must be flanked by two recombination events. The remaining possibility is that part of a homozygous haplotype slips through into the ends of the region under consideration. We expect that that is not a problem because the region is chosen to be centered around the event identified by the first step and several hundred SNPs are included on either side. The HMM would have to choose between placing an event at the end of the partial homozygous haplotype and an event at the end of the true stretch of IBD in the children. To choose the former, it would have to mark all the SNPs in the true stretch of IBD as homozygous in the parent. To choose the latter, it would have to mark all the SNPs in the partial homozygous haplotype as homozygous. By the design of the region under consideration in this step, we expect that a far higher number of SNPs would have to be marked homozygous (and therefore have lower likelihood) for the former to be picked.

We inferred 32,180 crossover events in total, of which 26,664 were resolved within 100 kb and 12,953 within 30 kb. The median event resolution was 38 kb.

The table below compares the total autosomal genetic map length estimated in our study with other published maps, all of which were estimated in Europeans. For this analysis, we wished to

obtain an unbiased estimate of total map length, and hence made an exception to the rule of filtering out crossover events that overlapped with other crossovers events. Likewise, we used the full set of events without filtering for association testing of individuals' genetic map length with their genotypes (Note S4).

Autosomal genome-wide rate (in Morgans) in different pedigree maps

	Female	Male
African American	42.6	29.9
deCODE (2010) [1,*]	40.7	22.9
Hutterites[2]	39.6	26.2
Rutgers[3]	44.0	28.3
deCODE (2002) [4]	42.8	25.9

[*] In the Decode (2010) map[1], genetic distances are not estimated for regions within 2.5Mb of the ends of SNP coverage on each chromosome.

References for Note S2

[1] Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-1103 (2010).

[2] 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).

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

[4] Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genetics* **31,** 241-247 (2002)

Note S3
Building the African-enriched Map

This study shows that African Americans harbor two types of recombination hotspots: a –shared" class that is also seen in Europeans, and an –African-enriched" class that appears to have been much more active in West African than in non-African history.

To estimate the degree to which the crossover events detected in each African American in this study differ in their distribution from what is expected in people of European descent, we modeled the crossover events as a mixture of the European genetic map (inferred from the deCODE Map) and an –African-enriched" map (inferred from the differences between the African American recombination landscape from the European recombination landscape as described below). We refer to the proportion of an individuals' crossovers that are ascribed to the –African-enriched" map as the –African-enrichment" (AE) of their recombination landscape.

To infer AE for each individual, we extended the MCMC of Note S1 to infer two distinct maps. We began by utilizing the deCODE Map[1] as a set of events available for one of the two maps (that is, the Shared (S) map). Since we do not have direct information on the events that were identified by deCODE, we used their published sex-averaged map and multiplied the rates therein by the number of meioses that the map represents. We used the resulting number for every SNP interval as a set of counts representing definitely European events. We then used these counts in the MCMC in addition to crossovers inferred from unrelated African Americans (the same procedure as Note S1).

We performed Gibbs Sampling over (i) the location of each crossover inferred in an African American or European, (ii) the recombination rate in each inter-SNP interval, (iii) the map to which each crossover event is assigned (either Shared or African-enriched), and (iv) the degree of African-enrichment (AE) of each individual's recombination landscape.

In detail, the MCMC proceeds as follows:

Initialization of MCMC

We start the Gibbs Sampler with two maps: a Shared map and an African-enriched map. Both are initialized with recombination rates sampled from the Basic Map prior (similar to Note S1). We initialize AE (anc_k) for each individual k by sampling from a Beta(1,1) (uniform) prior.

We denote the total number of SNP intervals as N. Further, for each SNP interval $i \in \{1, 2, \cdots, N\}$, let d_i correspond to the number of events obtained from deCODE using their scaled sex-averaged map as described above.

Step 1 in each iteration - Sampling each crossover event

In this Gibbs sampling step, for each crossover, we sample the SNP interval in which it occurred conditional on the most recent set of sampled Shared and African-enriched rates. For each SNP interval $i \in \{1, 2, \cdots, N\}$, let $\text{SharedEvents} = (n_{s,1}, n_{s,2}, \cdots, n_{s,N})$ represent Shared events that occurred in each SNP interval and $\text{AEEvents} = (n_{ae,1}, n_{ae,2}, \cdots, n_{ae,N})$ the corresponding number of African-enriched events. At the start of each iteration, the number of events in every SNP interval of each type is zero.

- We add counts corresponding to the deCODE Map to the European counts
- For each African American individual, we sample the location of their crossover events using the latest set of sampled rates for both maps, and the latest AE for each individual.

Specifically, we start with computing the likelihoods of a crossover c spanning SNP intervals $i \in \{1, 2, \cdots, n\}$ conditional on the rates in the Shared map $\overline{r}^s$ and the rates in African-enriched map $\overline{r}^{ae}$. From Note S1, we obtain

$$P(c \mid \overline{r}^s) \approx P(c \mid \overline{r}_{flat}) \sum_{i=1}^{n} \frac{r_i^s}{r_{flat_i}} P(rec_i \mid c, \overline{r}_{flat})$$

$$P(c \mid \overline{r}^{ae}) \approx P(c \mid \overline{r}_{flat}) \sum_{i=1}^{n} \frac{r_i^{ae}}{r_{flat_i}} P(rec_i \mid c, \overline{r}_{flat})$$

As in Note S1, we obtain $P(rec_i|c, \overline{r}_{flat})$ from HAPMIX. We denote this h_i for simplicity of notation. The constant of proportionality $P(c|\overline{r}_{flat})$ is the same for both maps, so we can ignore it. We thus have the total likelihood of an event conditional on each of the two maps. For each crossover, the probability of whether the Shared or African-enriched map was used to generate it is informed by its prior, which is simply the AE phenotype (anc_k) of the individual k harboring the crossover. Therefore the posterior likelihood of the crossover conditional on each map is, up to a proportionality constant:

$$L(c \mid \overline{r}^s, anc_k) = (1 - anc_k) \sum_{i=1}^{n} \frac{r_i^s h_i}{r_{flat_i}}$$

$$L(c \mid \overline{r}^{ae}, anc_k) = anc_k \sum_{i=1}^{n} \frac{r_i^{ae} h_i}{r_{flat_i}}$$

We sample whether the crossover c was Shared or African-enriched proportional to their respective likelihoods $L(c \mid \overline{r}^s, anc_k)$ and $L(c \mid \overline{r}^{ae}, anc_k)$.

- If the event c was sampled to be Shared, we then sample the SNP interval in which it occurred proportional to the likelihoods computed above for each SNP intervals using the Shared Map. If the event was sampled to be in SNP interval j, we increment the Shared count $n_{s,j}$ for that SNP interval. Similarly, if c was sampled to be African-enriched, we then sample its location using the African-enriched map and increment the corresponding entry of the vector of African-enriched counts.

At the end of Step 1, we can thus count the total number of Shared and African-enriched events that occurred in every SNP interval.

Step 2 in each iteration - Sampling the recombination rates in each inter-SNP interval

In this step of the Gibbs sampler, we sample Shared $\overline{r}^s$ and African-enriched $\overline{r}^{ae}$ rates in each interval conditional on the sampled Shared and African-enriched recombination events. By Bayes rule, we have

$$P(\bar{r}^s \mid \text{SharedEvents}) \propto P(\bar{r}^s)P(\text{SharedEvents} \mid r_i^s)$$
$$= P(\bar{r}^s)P(\left[n_{s,1}, n_{s,2}, \cdots, n_{s,N}\right] \mid r_i^s)$$

$$P(\bar{r}^{ae} \mid \text{AEEvents}) \propto P(\bar{r}^{ae})P(\text{AEEvents} \mid r_i^{ae})$$
$$= P(\bar{r}^{ae})P(\left[n_{ae,1}, n_{ae,2}, \cdots, n_{ae,N}\right] \mid r_i^{ae})$$

We model the event counts as having a multinomial distribution with rates $\bar{r}^s$ or $\bar{r}^{ae}$. Let the total number of Shared and African-enriched events be T_s and T_{ae} respectively.

$$\left(n_{s,1}, n_{s,2}, \cdots, n_{s,N}\right) \sim \text{Multinomial}(T_s, \bar{r}^s)$$
$$\left(n_{ae,1}, n_{ae,2}, \cdots, n_{ae,N}\right) \sim \text{Multinomial}(T_{ae}, \bar{r}^{ae})$$

We use a Dirichlet prior for $\bar{r}^s$ and $\bar{r}^{ae}$. We use a Dirichlet and multinomial model here, as opposed to the gamma and Poisson model in Note S1, to force the total genome-wide rates in the two maps to be the same. This forces both sets of rates to sum to be 1. Thus, they produce relative and not absolute maps.

In order to define the Dirichlet priors, we construct a new "Basic" Shared map (as in Note S1) and a new "Basic" African-enriched map using the latest set of sampled events of each type from Step 1. We do this in order to account for long-range correlation of rates as discussed in Note S1. The parameters of the Dirichlet priors $\bar{\alpha}_s$ and $\bar{\alpha}_{ae}$ for the Shared and African-enriched maps respectively were calculated such that the mean is the same as that of the respective Basic map.

$$\bar{r}^s \sim \text{Dirichlet}(\bar{\alpha}^s)$$
$$\bar{r}^{ae} \sim \text{Dirichlet}(\bar{\alpha}^{ae})$$

We use the conjugacy of the multinomial and Dirichlet distributions to construct Dirichlet posterior distributions and use these distributions to sample a recombination rate for every inter-SNP interval for the Shared and African-enriched maps.

$$\bar{r}^s \mid \text{SharedEvents} \sim \text{Dirichlet}(\alpha_1^s + n_{s,1}, \alpha_2^s + n_{s,2}, \cdots, \alpha_N^s + n_{s,N})$$
$$\bar{r}^s \mid \text{AEEvents} \sim \text{Dirichlet}(\alpha_1^{ae} + n_{ae,1}, \alpha_2^{ae} + n_{ae,2}, \cdots, \alpha_N^{ae} + n_{ae,N})$$

Step 3 in each iteration - Sampling AE for each individual

At the end of Step 1, we have sampled the ancestry of recombination of each event in every individual. For individual *k*, let us say that the total number of Shared events was ind_k^s and the number of African-enriched events was ind_k^{ae}. We model these counts as a binomial distribution conditional on the total number of events in an individual which is fixed and AE phenotype (anc_k) of the individual.

$$ind_k^s \sim \text{Binomial}(ind_k^s + ind_k^{ae}, anc_k)$$

We use a uniform Beta(1,1) prior for the AE phenotype. We can therefore use the conjugacy of the binomial and beta distributions to obtain a posterior beta distribution for it.

$$anc_k \mid ind_k^s, ind_k^{ae} \sim \text{Beta}(ind_k^s + 1, ind_k^{ae} + 1)$$

We take a sample from the respective distributions of each individual to complete one iteration of the Gibbs Sampler.

Iteration of MCMC
Unlike the previous MCMC which is run separately for each chromosome, this MCMC works with the whole genome at once to maximize the information available to estimate the ancestry of recombination of each individual. Due to the high memory requirements of this process, we performed this analysis with a subset of the data. We included autosomal data from 18,000 unrelated individuals (in contrast with the AA map which is built using 30,000 individuals and includes the X chromosome).Ten chains with at least 70,000 samples each were run, and 15,000 samples per chain were discarded as burn-in. We computed the mean recombination rate for every SNP interval using every 25th sample for both maps.

References for Note S3

[1] Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-1103 (2010).

Note S4
Association testing to recombination-related phenotypes

We performed association testing for 29,589 unrelated individuals and 444 parents from the pedigrees. Genotypes at up to 3,058,149 HapMap2 SNPs[1] were imputed into the unrelated individuals using the MaCH software with the West African YRI and European American CEU as reference panels[2]. We tested for association both at genotyped and imputed SNPs.

(i) Three phenotypes that we tested for association

(1) *"African-enrichment" (AE)*

For pedigree parents and unrelated individuals, we modeled the „true" recombination map as a mixture of the S and the AE maps inferred in Note S3. This is motivated by the idea that each unrelated individual inherits crossovers from ancestors some of whom may have recombination that looks Shared and others who may have an African-enriched recombination landscape. The best estimate for the mixture proportion, which we denote λ, is viewed as the AE phenotype for that individual. We use a Bayesian approach to estimate λ. If E represents the set of events in an individual, we have by Bayes' rule

$$P(\lambda|E) \propto P(E|\lambda)P(\lambda)$$

We impose a uniform prior on λ. We compute the likelihood of an individual's events by assuming that each event is independent.

$$\begin{aligned} P(E \mid \lambda) &= \prod_{e_j \in E} P\left(e_j \mid \lambda\right) \\ &= \prod_{e_j \in E} P\left(e_j \mid \lambda \bar{r}^{ae} + (1-\lambda)\bar{r}^{s}\right) \end{aligned}$$

From Note S1, we therefore obtain

$$P(E \mid \lambda) \propto \prod_{e_j \in E} \sum_{i \in e_j} \frac{\lambda r_i^{ae} + (1-\lambda) r_i^{s}}{r_{flat_i}} P(rec_i \mid e_j, \bar{r}_{flat}).$$

As in Note S1, we obtain $P(rec_i|e_j,\bar{r}_{flat})$ from the HAPMIX output. We are thus able to calculate the posterior $P(\lambda|E)$ for any $\lambda \in [0,1]$ up to a normalizing constant.

Events overlapping the terminal 5 Mb of each chromosome were removed from the analysis since the deCODE Map paper found that rates are less reliable near the telomeres[3], and we expect the same for our S and AE maps. For estimating the MP in pedigree parents, we further restrict to crossover events that are resolved within 100 kb. We calculate $P(\lambda|E)$ over a dense grid in [0,1] and numerically integrate for each individual to obtain the expected value of λ under the posterior which we define as their AE phenotype.

(2) *Hotspot Usage*

We wished to estimate the fraction of each individual's crossovers that occurred in hotspots identified previously from linkage disequilibrium (LD) data[4]. Hotspots were identified from genome-wide Phase II HapMap LD data and 32,991 hotspots were called that were active in at least two of the three constituent populations (CEU, YRI, JPT+CHB).

We used an approach that is similar in spirit to previous studies that have used this phenotype[5,6]. For each chromosome, we computed a hotspot map by assigning an equal recombination probability to all its hotspots and zero probability of recombination everywhere else (say, $\bar{r}^{\text{hotspots}}$). We also computed an alternative smoothed map (say, $\bar{r}^{\text{smooth}}$) using the HapMap2 population-averaged COMBINED LD-based map. We did this because long-range rates are remarkably similar in West African and European populations and the smoothed map offers a more plausible alternative for recombination outside LD hotspots than a uniform rate map. In detail, to create the smoothed map we replaced the rate in each SNP interval with the average rate in the 5 Mb neighborhood of that SNP interval.

For unrelated individuals, all filtered crossovers are used. For the pedigrees, we only use crossovers that are resolved within 100kb or better, as for the AE phenotype.

Similar to the analysis for the AE phenotype, we model recombination in every individual as a mixture of the hotspot map and the smoothed map with mixture proportion λ to obtain a mixture map $\lambda\bar{r}^{\text{hotspots}} + (1 - \lambda)\bar{r}^{\text{smooth}}$. We next calculate the likelihood $P(E|\lambda)$ up to a normalizing constant over a dense grid of $\lambda \in [0,1]$. We pick the value that maximizes the likelihood for each individual, and define this as their Hotspot Usage.

(3) Genome-wide rate

We count the total number of crossovers (without applying any filters) observed for each parent in the pedigrees. We do not test for genome-wide rate in unrelated individuals due to its strong dependence on the number of generations in every lineage in the genealogy of the individual, which is unknown.

(ii) Association testing is sensible in unrelated individuals even though they usually do not carry the same allele as the individual in whom a crossover event actually occurred

We tested SNPs for association to recombination phenotypes, both in the parents from the Pedigree Map, and in all the unrelated individuals from the AA Map.

In pedigrees, association between alleles and the phenotype can be tested directly because we have genotypes in the parents in whom the crossovers have occurred.

In the unrelated individuals, however, a subtlety is that the crossovers that we identify have not occurred in the individuals themselves but in their ancestors. While it is not possible to say which crossovers occurred in ancestors with the same allelic variants as the individual today, a simple calculation nevertheless illuminates the probability of such an occurrence.

Suppose that the number of generations of admixture in individual *i* is *n*. Each generation of ancestors, taken as a whole, passes on the same amount of genetic material to *i*, i.e., 46 chromosomes. The number of crossovers that are passed on is also, therefore, expected to be the same. In other words, the probability that a crossover detected in an individual today occurred in

an ancestor *k* generations ago has probability 1/*n*. Since there are 2^k ancestors *k* generations ago, the probability that a given crossover occurred in a specific ancestor *k* generations ago is $\frac{1}{n \cdot 2^k}$. Now, consider the allele individual *i* carries at any SNP. Each allele at a SNP is inherited from a unique ancestor in every generation. Therefore, the probability that a crossover detected in *i* occurred in the specific ancestor *k* generations ago whose allele was inherited by *i* is, as before, $\frac{1}{n \cdot 2^k}$. Summing over *n* generations we obtain the probability that a given crossover occurred in the same ancestor as one whose allele is inherited by *i* as:

$$\sum_{k=1}^{n} \frac{1}{n \cdot 2^k} = \frac{1}{n} \cdot \left(1 - \frac{1}{2^n}\right)$$

We estimate the number of generations of admixture in the populations used in our study to be approximately 6. Therefore, the expected number of crossovers that occurred in individuals with an allele that is ancestral to the one an individual carries today is 0.164, or approximately, 1/6.

These considerations suggest that association testing in the unrelated individuals is expected to be noisy, with 5/6 of the detected crossovers occurring on a genetic background that is unrelated to that of the individual in whom they are found by HAPMIX. Nevertheless, given the large sample size of almost 30,000 individuals, we proceeded with association testing.

(iii) Association testing

We performed association testing over 3,058,149 genotyped and imputed SNPs in the union of SNPs from the Affymetrix 6.0, Illumina 1M, Illumina Human610-Quad, HumanHap550 and HapMap2 YRI and CEU panels that met QC criteria in at least one of the seven constituent datasets (unrelated individuals from the CARe consortium, pedigrees from the CARe consortium, unrelated individuals from CHOP, pedigrees from CHOP, and unrelated individuals from the African American Prostate, Breast and Lung Cancer Consortia). We performed association testing separately for each of the seven datasets and then performed a meta-analysis across datasets to increase statistical power.

For the AE and hotspot usage phenotypes, we restricted the testing of unrelated individuals to those who had at least 35 events (approximately half of the mean number of events per unrelated individual). The goal of this requirement was to reduce statistical noise; it ensured that for all individuals included in the analysis, we had a sufficient number of crossovers to reliably estimate the phenotype. In pedigrees we re-scaled the phenotypes for male and female individuals to have the same mean and variance. We tested for association by performing linear regression of individuals‘ phenotypes against their genotypes, conditional on their genome-wide ancestry.

For the genome-wide rate phenotype, we tested mothers and fathers separately as well as together. For the joint test we rescaled the phenotypes for mothers and fathers to have the same mean and variance. We performed linear regression of their crossover counts with respect to their genotypes conditional on their genome-wide ancestry.

To perform a meta-analysis, we computed Z-scores from the P-values from all seven datasets, and then computed the combined P-value. The results for the AE and hotspot usage phenotypes are shown in Table S2. The top hit for both AE and hotspot usage phenotypes is SNP rs6889665 which is 4 kb upstream of *PRDM9* in chromosome 5. The combined P-value for all samples

including pedigree parents and unrelated individuals is 1.5×10^{-246}. In pedigrees alone, the p-value is 3.3×10^{-54}. This shows that there is considerable power to detect associations in this phenotype in unrelated individuals despite a noisy phenotype (Note S4 (ii)).We found no genome-wide significant associations with genome-wide rate.

To further explore variants in PRDM9 that influence the hotspot landscape, we tested association of SNPs with 20Mb of PRDM9 conditional on rs6889665. To do this, we performed linear regression as before, but with the genotype at rs6889665 as an additional predictor variable. Significant SNPs are reported in Figure 3C.

The P-values reported in Table S3 are corrected for genomic control. We calculated the genomic control inflation factor after removing SNPs within 10Mb of rs6889665. We estimated the inflation factor to be 1.046 for AE and 1.038 for hotspot usage. Finally, we used a threshold of 3.3×10^{-9} to determine genome-wide significance. This corresponds to a significance threshold of 0.01 after performing a Bonferroni correction for multiple hypothesis testing.

For the AE phenotype, we wished to estimate how much of the variation can be explained by variation in rs6889665. Due to large statistical noise in the phenotype of unrelated individuals as discussed above, we restrict the study to 158 pedigree parents where we could estimate their AE phenotype and could also genotype them at rs6889665. Allowing for different mean AE values in individuals carrying 0, 1 or 2 copies of rs6889665 explained 66% of the variance in the AE phenotype relative to a model without rs6889665 genotype. What is of greater biological interest, however, is the fraction of genetic or gamete specific variation explained (subtracting the noise in estimating the AE phenotype). Therefore, we estimated the noise in measuring the AE phenotype using jackknife on the set of crossovers for each individual. We use the mean of the variances in the AE phenotype across individuals as the noise estimate. When we subtract variance due to noise from both the total AE variance, and from the residual variance after the effect of rs6889665 is taken into account, we estimate that rs6889665 accounts for 82% of the underlying phenotypic variability.

For the genome-wide rate phenotype, we found that the C allele of SNP rs6889665 is associated with an increase in the total genetic map length. We estimate that the effect size is approximately 1.7 crossovers per copy of the C allele in males and 0.9 crossovers in females.

In 2008, Kong et al[7] demonstrated that haplotypes in the gene RNF212 are associated with genome-wide rate in both males and females in an Icelandic population. In both Europeans and West Africans, three haplotypes defined by SNPs rs3796619 and rs1670533 are found: [C,T], [T,C] and [T,T], however they segregate at very different frequencies. In females, Kong et al found that the haplotype [T,C] is associated with higher genetic map length relative to the haplotypes [C,T] and [T,T]. This haplotype is present at 19.2% frequency in CEU but at only 1.7% frequency in YRI. Due to the very low allele frequency of this variant, we do not expect to see a significant association in females in our study. In males, on the other hand, Kong et al observed that the haplotype [C,T] is associated with a strong increase in recombination rate relative to the haplotype [T,C] (effect size ~ 0.78M, $p=7\text{x}10^{-23}$ in 3135 males) and a somewhat weaker effect relative to the haplotype [T,T] (effect size ~ 0.55 M, $p = 3\text{x}10^{-7}$ in 3135 males). As mentioned above, the haplotype [T,C] is nearly absent in West Africans, and therefore we are not likely to see an association between these haplotypes. We believe that our inability to find a

signal between the [C,T] and [T,T] haplotypes is due both to the modest size of the effect and the small sample size in our testing (22 samples genotyped at rs3796619 and 80 imputed).

Baudat et al.[5] showed that *PRDM9* alleles influence hotspot usage patterns, and also found evidence that this gene influences genome-wide rate. They demonstrated that a *PRDM9* allele, called the I allele, significantly reduces hotspot usage in Hutterites who are a population of European descent. The I allele has a ~~–5~~/8" match to the 13-mer motif like the C allele, and other alleles, found in West Africans (Note S5) and is structurally similar to the C allele although with two additional zinc fingers. The I allele is rare and has not been found in West Africans or other Europeans. We are not able to assess, therefore, if it shares the haplotype background at rs6889665 with C-like alleles found in West Africans. The I allele does not show a significant association with genome-wide rate. Baudat et al.[5] found, however, that a different *PRDM9* allele, the B allele, significantly increases total genetic map length in both males and females.

References for Note S4

[1]The International Haplotype Map Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851-861 (2007).

[2] 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).

[3] Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-1103 (2010).

[4] Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.* **40**, 1124–1129 (2008).

[5] Baudat, F. *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**, 836–840 (2010).

[6] 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).

[7] Kong, A. *et al.* Sequence variants in the *RNF212* gene associate with genome-wide recombination rate. *Science* **319**, 1398-1401 (2008).

Note S5

Determination of *PRDM9* zinc finger array allele using 1000 Genomes Project data

(i) Computing a likelihood for combinations of previously identified *PRDM9* alleles

We took all Phase 1 pilot data[1] from the 1000 Genomes Project generated by Illumina sequencing, and whose best mapping in the genome was within the 4kb region (2kb upstream and 2kb downstream) surrounding the *PRDM9* zinc finger array (data kindly provided by A. Auton). These data, for 146 individuals, comprise mainly short paired-end reads, as well as some single-end reads, with read length of up to 101bp. Each individual has approximately 4-fold coverage genome-wide[1], and so we expect these reads to contain information about the *PRDM9* zinc finger (ZF) arrays each individual carries, although assembly of such short reads is a difficult task. We use the collection of 29 sequenced alleles for *PRDM9* previously identified by Berg et al.[2], assume that each of the two *PRDM9* alleles every individual carries was drawn from this collection, and then use a model-based approach to infer the likelihood of each of the 435 possible allelic pairs based on their short read data. Throughout, our allele labeling is the same as that used by Berg et al.[2]. Our approach cannot give the correct answer in the minority of cases where an individual carries *PRDM9* types not seen previously, but we hope that it chooses an allele in some sense ―close" to the truth even in such cases (in the sense that the allele that is chosen will be structurally related). Our likelihood calculation uses only variation within the array itself, and no information on other loci outside the ZF array.

First, we remapped all paired-end reads to the human genome reference sequence surrounding *PRDM9* (2kb upstream and 2kb downstream of the ZF array), in order to identify reads mapping within the ZF array, and to identify the insert size distribution. We used a custom approach to map reads inside the array, bearing in mind variability between *PRDM9* alleles. Specifically, to avoid discarding reads from individuals with non-reference *PRDM9* types, we constructed a ―template reference" to avoid mapping biases. This reference is made by concatenating the 2kb *PRDM9* upstream region, a 3-zinc-finger *PRDM9* template, and the 2kb *PRDM9* downstream region. The 3-zinc finger template is motivated by the fact that for paired-end reads at most 101 bp long as in our data, each short read can overlap at most 3 *PRDM9* zinc fingers, since each *PRDM9* zinc finger is 84 bp long, Therefore a template with 3 successive identical, generic zinc fingers is sufficient to cover and match any short read from any individual‘s ZF array. To construct the generic zinc fingers, we used the 19 previously identified *PRDM9* individual zinc finger types[2], which differ commonly at only 10 of the 84 bases within the zinc finger. We masked these bases by ―N"‘s, and used the shared sequence at the other 74-bp. For the reads, we masked bases with quality score <13 (corresponding to a 5% error rate) by N‘s, and then removed any reads (and corresponding mate pairs) whose best match to the template reference contained 5 or more mismatches (ignoring the ―N" bases) as potentially incorrect mappings. Reads were then mapped according to their first best match within this template.

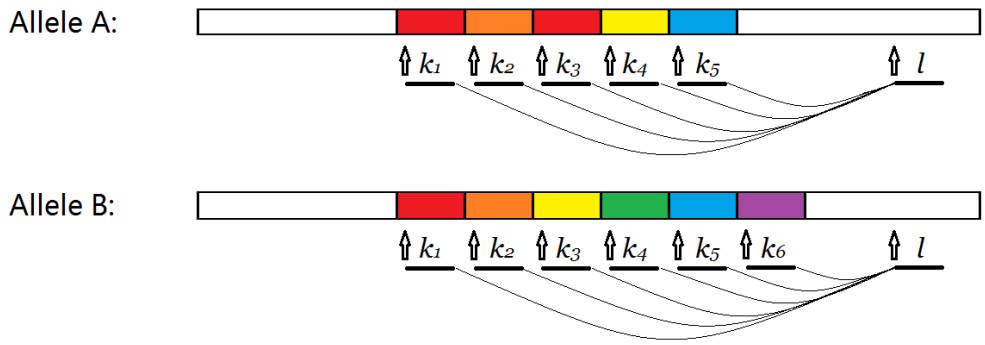
Each paired end read comes from one of 166 different read libraries. Separately for each library, we modeled its insert size distribution by using the local, empirical insert size distribution for each read library, based on the observed insert size distribution of all paired-end reads with both pair members mapped outside the ZF array of the reference sequence. Our modeled insert size distribution was a mixture of this empirical distribution (99%), and a uniform distribution (1%) to avoid overfitting. We used our mapped reads and insert size likelihoods to construct, for each

individual, the likelihood of their collection of reads, conditional on their having each of the 435 possible combinations of known alleles. We write the likelihood of a single read pair given a combination of known alleles A_1, A_2 (conditional on the read pair mapping within our region) as:

$$\mathrm{P}(\text{reads 1,2} \mid \text{alleles } A_1, A_2) = \sum_{k,l} \left[\frac{1}{2} \left(\frac{1}{\text{length(allele } A_1) + \text{length(allele } A_2)} \right) \cdot \mathrm{P}(\text{insert size: } k - l) \cdot \mathrm{P}(\text{read 1} \mid \text{starts at } k) \cdot \mathrm{P}(\text{read 2} \mid \text{starts at } l) \right]$$

where reads 1,2 are paired-end reads, and alleles A_1, A_2 are the genotype of the individual. Here, length(allele A_1) is the length of the ZF array and surrounding sequence for allele A_1. The likelihood for single-end reads is similar, but involves only one summation. In principle we must sum over all possible k and l in the sum above (viewing both alleles as having their own set of positions) but in practice most terms are negligible. Thus for reads mapping outside the array we assume their position is known and given by our mapping. For reads overlapping the ZF array, we perform the summation, but assume the position relative to the zinc finger start (for example, 23 bases into the zinc finger) is known from our mapping, while the particular zinc finger the read maps to is unknown and summed over.

As an example, suppose we want to know the likelihood of a read pair for an individual having allele A and allele B on the homologous chromosomes, with one read mapping inside the array and the other outside.



The upstream read is mapped inside the triplet template of zinc finger array, and the downstream read is mapped outside. Here the upstream read can take 5 different positions in the ZF array of allele A, and 6 different positions in the ZF array of allele B (shown), because of the number of tandem repeats in each allele. Each possible position of the upstream read gives a different insert size, and predicts a different sequence for the corresponding zinc finger, and we sum up the likelihood of each possibility for this read pair.

The final part of our algorithm is the likelihood calculation of each short read starting at a specific position inside the ZF array of a known read, i.e. $\mathrm{P}(\text{read 1} \mid \text{starts at } k)$. We calculate this likelihood for each base within a read, by performing the standard conversion of the read quality score at each base into a likelihood for the base ~~r~~ead" at that position (here we do not filter any particular quality score). We assume that given the base is read incorrectly, all 3 possible errors are equally likely so the likelihood of any mismatch is the same, and we also

assume no errors in the previously sequenced alleles. Finally, we multiply the likelihood across bases within the read, to give the likelihood of a read given a particular mapping position and sum across positions to give the likelihood for a given pair of alleles. To gain an overall likelihood $P(data\ |\text{alleles}\ A_1, A_2)$ of alleles A_1, A_2 for an individual, we multiply across read pairs and single-end reads, which it seems reasonable to model as independent. (To aid computation, we also pre-compute for each short read mapped inside the *PRDM9* ZF array the likelihood of coming from each possible triplet of zinc fingers. Many such triplets are found in multiple different *PRDM9* alleles, so this pre-computation greatly increases efficiency. If *N* is the number of short reads per individual, *n* is the number of known alleles and *L* is the number of tandem repeats for each allele, then the time complexity of our algorithm is $O(N \cdot n^2 \cdot L^2 + N \cdot n \cdot L \cdot l)$ where *l* is the maximum length of each short read.)

(ii) An EM algorithm to infer frequencies and posterior probabilities of *PRDM9* alleles

The procedure above produces a likelihood, for each of 146 individuals, of 435 possible *PRDM9* genotypes. However, since the frequency of alleles worldwide varies strongly, we sought to improve this analysis by inferring the worldwide frequency of the 29 different *PRDM9* alleles (we did not use previous frequency information to do this). Specifically, we modeled the probability an individual has an ordered pair of alleles numbered A_1, A_2 given their sequence data as proportional to the likelihood times the prior probability $P(A_1)P(A_2)$ of alleles A_1, A_2 under Hardy-Weinberg Equilibrium (HWE):

$$P(A_1, A_2|data) \propto P(data|A_1, A_2)P(A_1)P(A_2).$$

We then used the standard EM algorithm to obtain the allele frequencies maximizing the likelihood of the full dataset. We do not believe HWE will strictly hold here given variations in worldwide frequency, but also do not believe this will strongly affect our inferences and this approach avoids making assumptions regarding population frequency differences. Specifically, our EM algorithm is initialized with a uniform probability $P(A_1 = i) = P(A_2 = i) = \frac{1}{29} \forall i$ and then iterates two steps to convergence:

1. (E-step) Given the current estimated frequency $P_n(i)$ of each allele $i = 1,2,\cdots,29$ compute the posterior probability of each ordered allele pair A_1, A_2 for each individual in the dataset, using the previous equation.
2. (M-step) Given these probabilities $P_n(A_1, A_2|data)$, the estimator $P_{n+1}(i)$ of $P(i)$ of each allele i maximizing the conditioned log-likelihood is the expected proportion of all alleles in the data that are of type *i*. Set the new frequency estimate $P_{n+1}(i)$ to this value and return to step 1.

This approach allows us to maximize the likelihood of the observed data over all possible frequencies, and obtain a maximum likelihood estimator vector $\boldsymbol{P}$ of allele frequencies. Validating the general approach, the alleles identified as most common by this E-M approach were (in order) A (64%), B (13%), C (5%) which are the same alleles as those identified as most common by Berg et al., with inferred worldwide frequencies similar to those seen previously[2]. Other alleles were identified in the data but much less frequently.

We used the maximized allele frequencies to obtain the maximum likelihood estimate probabilities of the form $P(A_1, A_2|data_i)$ for each possible allele pair for each individual i. Summing over all allele pairs, it is straightforward to calculate the marginal expectations:

$$E_i(z) = E(\#\text{ copies of allele } z \text{ in individual } i\ |\boldsymbol{P})$$

The quantities $E_i(z)$ form a summary of the results, used to produce Figure S7 and indirectly in Figure 2.

(iii) Comparison of *PRDM9* inferred alleles and rs6889665 status

The analyses until this point are independent of the genotype an individual carries at rs6889665. We wished to compare the sequence of the *PRDM9* zinc finger array to this nearby SNP, and specifically to particular binding targets of *PRDM9*. To do this, we grouped *PRDM9* alleles by previous predictions[2] that inferred at how many bases their bioinformatically predicted binding targets matched the previously identified 13-bp motif CCNCCNTNNCCNC[3]. All alleles match 4-8 bases of this motif, with 8 bases representing a perfect match. The most frequent alleles A and B match 8/8 bases, while allele C, and an additional group of "C-like" alleles (below) match 5/8 bases of the motif and have similar binding targets. Other alleles have more varied predicted binding sequences and match 4/8, 6/8, 7/8 or 8/8 bases of the motif. From the HapMap Phase II data[4] we had genotype data for rs6889665 for 139 individuals, including 46 CEU and 44 YRI individuals. Each of these individuals carries two alleles of *PRDM9*: for each individual i we computed the expected number $Q_i(k)$ of these alleles that match k of 8 bases of the motif:

$$Q_i(k) = \sum_{z \text{ matches } k \text{ bases}} E_i(z)\,, k = 4,5,6,7,8$$

Finally, we ordered individuals by their rs6889665 genotype and used a bar plot to show $Q_i(k)$ for the CEU and YRI populations (Figure S7). The figure reveals that most alleles are 5/8 or 8/8 motif matches, as seen previously based on direct experimental measurement[2], and the number of copies of each type of allele is confidently inferred for most individuals. The number of copies of the 5/8 motif-matching alleles, referred to as "C-like" alleles below, correlates strongly with (and is almost identical to) the number of copies of allele "C" they carry at rs6889665 (correlation is 94%). Results for the combined JPT+CHB populations were very similar, with apparently complete correlation between heterozygosity of rs6889665 (5/49 individuals) and having a single *PRDM9* copy corresponding to the 5/8 motif match. Thus, we deduce that the SNP variant in the human genome that we identify as most strongly associated with increased usage of the African-enriched (AE) map over the Shared (S) map (or with low usage of Shared hotspots, which are active in Europeans) precisely marks *PRDM9* alleles predicted to bind a specific motif, and mismatching the previously identified 13-mer CCNCCNTNNCCNC. We discuss this motif, and comparisons with a motif we identify via examination of African-enriched hotspots, in more depth in Note S6.

(iv) A second EM algorithm used to generate Figure 2B

The above analysis revealed a strong relationship between rs6889665 and the alleles that an individual carries at *PRDM9*. To characterize this relationship more precisely (while allowing for uncertainty in the short read data) we repeated the EM algorithm analysis above, but this time

allowed *PRDM9* alleles to have differing underlying allele frequencies depending on the allelic rs6889665 background on which they occur. Specifically, we allowed for two probability vectors that give the frequency of each of the 29 alleles given type –C" or type –T" at this SNP on the same haplotype: $P_T(i)$, $P_C(i)$ respectively $i = 1,2,\cdots,29$). For an individual with types $S_1, S_2 \in \{T, C\}$ (for convenience, we order these alphabetically within each individual) at this SNP, their likelihood of possessing an ordered pair of alleles numbered A_1, A_2 is then:

$$P(A_1, A_2|data) \propto P(data|A_1, A_2)\left[P_{S_1}(A_1)P_{S_2}(A_2)\right]$$

We then modified our EM algorithm to obtain the allele frequencies maximizing the likelihood of the full dataset, again initializing both frequency vectors with a uniform probability and then iterating two steps to convergence:

1. (E-step) Given the current estimated frequencies $P_{T,n}(i)$ and $P_{C,n}(i)$ of each allele $i = 1,2,\cdots,29$ compute the posterior probability of each ordered allele pair A_1, A_2 (ordered with respect to the rs6889665 genotype S_1, S_2) for each individual in the dataset, using the previous equation

2. (M-step) Given these probabilities $P_n(A_1, A_2|data)$, the estimator $P_{T,n+1}(i)$ of $P(i)$ of each allele i maximizing the conditioned log-likelihood is the expected proportion of all type –T" alleles in the data that are of type i. The corresponding equation holds for type –C" alleles. Set the new frequency estimate $P_{T,n+1}(i)$ and $P_{C,n+1}(i)$ to these values and return to step 1.

We found after applying this algorithm (and thus allowing for uncertainty in the read data) an even stronger inferred association between rs6889665 and the corresponding *PRDM9* allele, with 100% of rs6889665 –C" haplotypes being predicted to carry *PRDM9* alleles corresponding to 5/8 motif matches. For –T" haplotypes, 71% of *PRDM9* alleles are inferred to be of the most common type –A", while 15% are inferred to be of the human reference sequence type –B". These alleles are inferred to be absent in rs6889665 –C" carriers, who have estimated 50% type –C" alleles, 20% type –L4" and 18% type –L6" *PRDM9* alleles, in reasonably close agreement to previously reported results[2]. We used motif predictions based on these inferred frequencies, as described in (iii) above, to produce the pie charts in Figure 2B.

References for Note S5

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

[2] Berg, I.L. *et al. PRDM9* variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genetics* **10,** 859-863 (2010)

[3] Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.* **40**, 1124–1129 (2008).

[4] The International Haplotype Map Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851-861 (2007).

Note S6

Discovery of a sequence motif associated with African American recombination

(i) Testing of words of length 5-9

We took the 2,454 candidate African-enriched hotspot regions (–AE regions") and 7,328 hotspot regions that are shared across populations (–S regions") (Methods) and extracted sequence for each region, separating out unique DNA from repeat sequence DNA using the RepeatMasker track in the UCSC genome browser. We then searched for particular non-degenerate words of length 5-9 that differed in frequency between the two sets of regions, reasoning that because we were comparing hotspots from one population to those from another (rather than previous approaches comparing hotspot to coldspot sequences[1]), identified motifs should confer different average hotspot activity in populations of West African and Northern European ancestry.

To identify significant motifs, we performed motif searching separately in repeat and unique sequence. For each motif of length 5-9, we counted its repeat/non-repeat occurrences (on either strand) in our AE regions and S regions, relative to the total number of repeat/non-repeat motifs of that length in each hotspot set, and tested for a frequency difference via a chi-squared association test (1 d.f.). We computed an overall P-value for each motif by converting (signed) repeat and non-repeat P-values to Z-scores with a $N(0,1)$ distribution under the null hypothesis of no association to African-enriched hotspots, and then summing these Z scores to obtain a statistic which we tested against a $N(0,2)$ null distribution to give a single overall P-value.

For a motif to be considered statistically significant, we required that it show a consistent enrichment signal ($P<0.01$) on both the repeat and non-repeat backgrounds, with effects in the same direction. Further, we required that the overall P-value should be statistically significant ($P<0.05$) after Bonferroni correction for the total number of tests performed (n=174,912). Finally, we separated out the resulting motifs on the basis of whether they were enriched in the African-enriched or Shared hotspots, and ordered motifs by the level of enrichment, measured as the overall odds ratio, in the appropriate set, to identify those motifs expected to have the strongest effect on recombination rate. After obtaining a set of statistically significant motifs, we removed all –redundant" short motifs entirely contained within a larger significant motif, but did not otherwise attempt to cluster different motifs.

(ii) Results of motif testing

We identified a single motif enriched in the African-enriched hotspot set and two motifs enriched in the Shared hotspot set ($P<0.05$ after Bonferroni correction) according to the above criteria. The two Shared motifs were similar; –CCTCCCT" (OR=1.14 for enrichment in the Shared set, $P=3.11\times10^{-9}$) (Bonferroni-corrected $P=5.4\times10^{-4}$) and –CCCCCC" (OR=1.18, $P=4.37\times10^{-13}$, Bonferroni corrected $P=7.6\times10^{-8}$). Both of these motifs have been previously identified as enriched in hotspots identified using LD patterns[1]. The first is a perfect match to the first seven bases of the previously determined 13-bp hotspot motif, –CCTCCCTNNCCAC", which is bound by the –A" and –B" alleles at *PRDM9*. The second mismatches only degenerate base 3 of this motif. This implies that comparison of our hotspot sets indeed has the power to identify genuine hotspot motifs. Further, it reveals that motifs bound by *PRDM9* alleles –A" and –B" are strongly

underrepresented in the African-enriched hotspot set. This is consistent with previous knowledge that the ―A" and ―B" alleles are common in non-African populations[2] and at lower frequency in West African populations, so we would expect such motifs to occur more rarely in African-enriched hotspots, as observed.

The motif identified as occurring more often in the African-enriched hotspot set is the 9-mer ―CCCCAGTGA" (OR=1.79, $P=2.24\times10^{-8}$, Bonferroni corrected P=0.004). The effect of this motif appears similar outside repeats (OR=1.63, $P=5.83\times10^{-5}$) and within repeats (OR=2.19, $P=1.01\times10^{-4}$). We verified that this effect did not relate to any potential G/C-content differences between the two hotspot sets by recalculating P-values in exactly the same way as before, except that we now compared the frequency of each motif in the two hotspot sets relative to all other motifs with identical G/C-content. After recalculating odds ratios and P-values in this way for each of the 20 candidate motifs, the signal for the motif ―CCCCAGTGA" was slightly strengthened (OR=1.82, $P=1.26\times10^{-8}$). Thus, we identified a single 9-bp motif with overwhelmingly statistically significant evidence of association with African-enriched hotspots.

(iii) Extension of 9-mer motif CCCCAGTGA

We wished to test whether sequence immediately flanking locations of the 9-mer motif might also influence local recombination rates at their positions, specifically in the populations of West African ancestry. To do this, we first identified all exactly matching locations of the motif in the genome (not required to be within either of the hotspot sets listed above). At each position within 50bp upstream or downstream of the motif, we marginally explored the effect of each possible nucleotide ―A", ―C", ―G" or ―T" on recombination rates. We looked at two independent measures of the component of the recombination rate that is enriched in populations of West African ancestry. First, we used the *difference* between the S Map and the AE Map rate, the two new maps obtained by the MCMC approach in Note S3, with the S Map representing an estimated component of recombination that is shared across worldwide populations, and the AE map the African-enriched component that we detect in the African American individuals. Both these maps are of course likely to contain errors, since they only estimate the genuine underlying S and AE maps. In particular since we do not identify unrelated African American individuals in our dataset that *only* use the true underlying AE map, our estimate of the AE map is expected to contain some Shared recombination, i.e. to represent a mixture between the true AE map and the true S map. Nevertheless, we aimed to use differences in the two maps to identify bases that increase the rate in the AE map relative to the S map, by promoting the occurrence of African-enriched hotspots. For a second comparison, we used the difference between the YRI estimated rate and the CEU estimated rate, again to determine bases that increase the rate only in the YRI map. The aim of this approach is to use as much information as possible from the genome-wide data, given that we expect only a subset of human hotspots to be African-enriched.

Using this idea, we first computed the difference in estimated rates for the 2-kb region centered on each occurrence of CCCCAGTGA in the human genome, by linear interpolation of the various genetic maps. We observe (Figure 1B-C, Figure S5) that in general there is a clear correspondence between rate peaks in our AA Map and the YRI map. However, the differences in rates are expected to be subject to much greater error due to the noise in the estimates being enriched after differencing, and in fact we found a Pearson correlation of only 24% between the

two measures of difference in rates suggesting that our phenotypic measurements may be noisy (or, this low correlation may possibly reflect more subtle evolution of recombination rate). Nevertheless, the excellent agreement that we obtained between consensus sequences (see below) suggests that we do have sufficient statistical power, from both map comparisons, to determine the optimal flanking sequences. More significant P-values were obtained (see below) using the YRI-CEU phenotype, suggesting that this phenotype may be subject to less noise, so we use this to define the consensus sequence and P-values reported in the main text. (However, in what follows we provide the consensus sequences determined using both comparisons, which are very similar.) We hypothesize that the YRI-CEU phenotype may be more informative because both the AA Map and the deCODE Map may well have higher errors than the LD Map rates at such fine scales as 2kb, leading to similar differences in resolution between the AE and S map.

Given either of our measures of rate difference, we define a consensus base at each position, for 50 bp upstream and downstream relative to the motif by choosing the base "A", "C", "G" or "T" that maximizes the mean difference in estimated rates when it occurs at the appropriate position relative to an exact match to "CCCCAGTGA". This results in a 109 bp "consensus" sequence. To check consistency across map comparisons, to verify we were not biased by the presence of genomic repeats, and to ensure that our consensus sequences were comparable between chromosomes, we constructed such consensus sequences in eight different ways. The first four motifs use the YRI-CEU map comparison, while the second four use the AE-S comparison.

1. We constructed a consensus, "Motif 1", using the YRI-CEU comparison and all occurrences of the 9-mer motif "CCCCAGTGA" that were further than 5 Mb from the telomere, and within the region covered by the LD Map, AA Map and the deCODE Map.
2. We constructed a consensus "Motif 2" in the same way as for Motif 1 and using the YRI-CEU comparison, but masked all occurrences of "CCCCAGTGA" occurring within 50bp of repeat DNA, and also removed occurrences of the 9-mer fewer than 500bp from the nearest such motif. This was designed to remove any non-independence issues caused by the presence of repeat DNA. The same map comparison and filtering criteria were later used for testing the significance of positions within and surrounding the motif.
3. We constructed a consensus "Motif 3" exactly as for Motif 1, but now using only motif occurrences from the *odd* chromosomes. This was designed to check that our motif was not providing an association signal specific to only part of the genome.
4. We constructed a consensus "Motif 4" exactly as for Motif 3, but using only motif occurrences from the *even* chromosomes.

5-8. We constructed consensus motifs "Motif 5-8" exactly as for Motif 1-4, but now using the AE-S map comparison.

We wished to test whether there was statistical support for different bases at a given position relative to "CCCCAGTGA" influencing the recombination rate. To do this, we used the data for constructing "Motif 2" and "Motif 6", i.e. we first masked repeat DNA and thinned our motif set to only allow occurrences of the 9-mer at least 500bp apart. In order to identify which bases genuinely play a role in influencing recombination rate (as opposed to exhibiting a higher rate by

chance), we performed significance testing separately for each position. To obtain a P-value at each position, we used the non-parametric Kruskal-Wallis test based on ranks, using our estimated difference in rates at each motif as the response, and the DNA base at a given position as a factor potentially influencing rates. This P-value measures the evidence that a given position relative to the motif is influential. (We used the rank-based test due to strong non-normality of the rate difference distribution.) The P-values obtained showed similar properties using either the YRI-CEU comparison or the AE-S comparison, but the former indicated a larger number of convincing ($P<0.01$) influential bases, so we believe there may be more information at the extremely fine scales provided by the YRI and CEU maps, and thus report consensus sequences and P-values defined this way in Figure 3A and Figures S8, S9. (In practice, results vary little if we instead make the choice of using the AE and S maps; see below.)

In constructing each motif, and in the testing, we also explored the effect of non-matching bases within the motif itself. To accomplish this, we identified all occurrence of the 9-mer mismatching at most one base within the motif, and used the same measures of rate difference as above. For bases outside the 9-bp motif, the analysis conditions on matching the 9 particular bases of the motif. To similarly condition on 9 matching bases in the analysis for positions *within* the 9-bp motif, we noted that the above analysis revealed a significant positive influence for possessing a "C" base 2-bp downstream of the motif, and so began by thinning our data to motifs matching this "C" at this position. Thus counts, and therefore statistical power, are expected to be similar when testing positions inside and outside the motif. Next, a consensus "within" the 9-mer was identified by choosing the base that maximized the mean difference in estimated rates, conditional on matching the other 8 bases within the motif, meaning that we conditioned on matching 9-bp of the consensus after taking into account the additional downstream "C" base. Note that the resulting motif could in principle mismatch any subset of the 9 positions within the motif initially used. We then obtained P-values for each base within the motif using the same data, exactly as described above. Specifically, we conditioned on non-repeat status and then performed a Kruskal-Wallis test for an influence of each position on rates.

To check the reliability of our consensus sequence inference, we were guided by three measures. First, we examined the positions of significant P-values. Second, we compared the consensus sequences defined by the two rate difference phenotypes. Third, we used the difference between YRI and CEU rates as the phenotype, and independently called a consensus based on the data for only even, or odd, chromosomes. We then compared the resulting consensus sequences.

(iv) Extension of 9-mer motif CCCCAGTGA: Results

The motifs as defined above gave eight aligned consensus sequences for enrichment of recombination in the African population.

For clarity, below we show a sub-region of the resulting sequences, of 25-bp. Outside of this region there is no obvious sequence similarity between the consensus motifs, and no convincing evidence for influential bases. (The signal is $P>0.01$ for all bases except one isolated position; we expect around 1 false positive in 109-bp using this cutoff and so the P-values outside the core 25-bp region suggest that there is little binding influence at these bases.) In the motifs listed below, the red bold font highlighted regions corresponding to the position of the identified 9-mer

(allowing mismatches), while, underneath, a consensus sequence is defined to show an upper case base if at least 7 of the 8 motifs agree at a given position, a lower-case base if at least 5 of the 8 motifs agree, and otherwise a dash. Finally, below the motifs, asterisks mark sites with significant support (P<0.01) for an influence of the particular base shown, based on an analysis of the occurrences used to produce Motif 2 (repeats and clustered motifs masked).

```
Motif 1: TCGACCCCAGTGAGCGTTGCCCCCG    (YRI-CEU all occurrences)
Motif 2: TAGTCCCCAGTGAGCGTTGCCAGAG    (YRI-CEU repeat masked)
Motif 3: TAGACCCCAGTGAGCGTTCCCCACA    (YRI-CEU odd chromosomes)
Motif 4: ACGACCCCAGTGAGCGTGGCCACCG    (YRI-CEU even chromosomes)
Motif 5: TGGACCCCAGTGAGCGTTGCTGACG    (AE-S all occurrences)
Motif 6: TTCACCCCGGTGAGCGTTACCTAGG    (AE-S repeat masked)
Motif 7: TTAACCCCTCTTAGCGTTGCTCGTG    (AE-S odd chromosomes)
Motif 8: ACCACCCCGGTGAGCGTGGCCGAGG    (AE-S even chromosomes)
         T-gACCCCaGTGAGCGTtgCc---G    (consensus)
           * **** ******** *  * *     (P<0.01)
```

As can be seen above, there is almost precise agreement in consensus sequence between the two maps, between all sequences and only non-repeat sequences, and between results for the completely independent data corresponding to odd and even chromosomes, at a series of bases. For example, using all motif occurrences (Motif 1, Motif 5), there is an 18-bp region "GACCCCAGTGAGCGTTGC" perfectly agreeing whether the YRI-CEU map comparison, or AE-S map comparison, is used, thus extending the original 9-mer motif by 9 additional matching bases. Further, positions at which there is strong agreement between the consensus sequences agree almost precisely with positions where there is statistical evidence (P<0.01) of influence of that base on the difference between rates in the YRI-CEU non-repeat comparison. Within the 18-bp region defined above and outside the 9-mer used for the conditioning, the AE-S odd vs. even comparison reveals matching of 7 of 9 bases and 13 of 18 bases overall (P=0.0013; $P=3.4\times10^{-5}$ respectively, assuming 25% match probability) while the YRI-CEU odd vs. even comparison also gives matching of 7 of 9 bases and 16 of 18 bases (P=0.0013; $P=2.1\times10^{-8}$). Thus, an extended motif marks a subset of African-enriched recombination hotspots.

In general, the region immediately downstream of "CCCCAGTGA" matches is most influential, while the sequence upstream of the 9-mer seems to be much less influential: the 16-bp sequence CCCCAGTGAGCGTTGC encompasses all bases both strongly matched and significant across comparisons, and so defines a natural consensus. However, bases immediately upstream and downstream of this sequence appear to have some relationship to recombination rates; for example a 17th base "C" is weakly signaled 1-bp downstream, as is the upstream sequence "GA". We note that our results are in fact consistent with any consensus motif length between 16 and 21 bases. We arbitrarily chose to report a 17-bp motif: CCCCAGTGAGCGTTGCC as our consensus in the main text, partly based on consideration of the *PRDM9* "C like" alleles that are enriched in West Africans, which suggests inclusion of the 17th base (see below). We emphasize, however, that slight modifications of our motif would not meaningfully impact results, and there is some uncertainty regarding the true motif end position based on our analysis.

(v) Exploration and quantification of degeneracy within the 17-bp motif

We noted considerable degeneracy within the 17-bp motif, despite the reasonable clarity of the consensus sequence at most bases. In particular, no matches to all 17-bp of the motif occur anywhere in the genome, so it is necessary to score mismatching bases to the motif in order to explore the effect of flanking sequence directly in terms of rate plots. To explore the effect of degeneracy, we produced a logo plot for the minimal 17-bp motif region as described previously, using the *makelogo* software[3] and the YRI-CEU rate difference (which we believe to be most powerful in determining the underlying motif). We used the matrix produced for the logo plot to score motif occurrences in the genome according to their flanking sequence, assuming independence across sites at motif binding sites. We have no very strong theoretical justification for this approach, other than the fact it gives greater rewards to bases showing stronger evidence of West African recombination enrichment.

In more detail, we reasoned that the difference in rates could be considered proportional to the probability of motif ―hotspot" activity given each base. Assuming a rough 25% frequency for each base in the genome, by Bayes' theorem the difference in rates is also proportional to the probability of each base given motif ―hotspot" activity. Specifically, given an observed set of mean rate differences $R_{A,i}$, $R_{C,i}$, $R_{G,i}$, $R_{T,i}$ at a position i relative to the motif, we set negative numbers (which correspond to more CEU recombination, and were rare) equal to zero, and then rescale the numbers to sum to one, giving relative letter height for e.g. letter A at position i of $R_{A,i}$ /($R_{A,i}$+ $R_{C,i}$+$R_{G,i}$+ $R_{T,i}$). As previously[4], we set the stack height proportional to $-\log_{10}(P_i)$ where P_i is the previously calculated P-value for association between the letter at base i and recombination activity. This approach was used to produce the logo plot shown in Figure 3A.

(vi) Scoring of flanking sequence surrounding the 17-bp motif

To explore directly the effect of the 8-bp flanking sequence region downstream of the 9-bp motif (and totaling 17-bp) (Figure 3B) we developed a means of scoring such flanking sequence. Assuming independence of each base, we define a copy of the motif as ―active" if it is bound (almost certainly by *PRDM9* C-like alleles – see below). The probability of observing a given sequence of bases under this assumption is thus simply the product of the corresponding ―probabilities" used for the logo plot at each base. Reversing the above argument, this is loosely proportional to the probability of a 17-bp region of the genome being bound given that it has that 17-bp sequence, and thus is a natural score for each motif occurrence (the scoring matrix is available on request). Given the scores, we identified all occurrences of the 9-mer in the genome where we had flanking SNP information and plotted rates (smoothed on a 2 kb scale, slid 500 bp along the region) for the AA Map and deCODE Map, and for the YRI Map and the CEU Map, separately (Figure S8A and Figure S8F respectively). Next, we ranked all these 9-mer occurrences by the resulting score and selected the top 500 motifs. We repeated the plots; Figure 3B shows the AA/deCODE rate comparisons, while Figure S8G shows the almost identical YRI/CEU rate comparisons. The AE and S maps yield plots (not shown) similar to those for the AA and deCODE maps from which they are derived, though with somewhat stronger peaks for the AE map.

One possible concern with this approach is that we used the same data to both fit the degeneracy matrix and explore the effect of flanking sequence via rate plots. Given many thousands of motif

occurrences within the genome, and the fact that the flanking sequence effect estimation is performed marginally at each base but then used jointly in the ranking, we do not expect this to be a strong effect. Nevertheless, to address this concern we performed cross-validation (Figure S8 panels D and E, with results without using flanking sequence information for comparison shown in Figure S8 panels B and C). First, we masked all odd chromosomes (~50% of the genome) and fit a degeneracy matrix as described above using only the even chromosomes. Then, we unmasked the odd chromosomes and explored the effect of selecting the top 250 ranking motifs (roughly corresponding to the top 500 motifs in the genome as above) on rates in this independent validation set. The results closely matched those for Figure S8A and Figure 3B (apart from more noise, and perhaps a suggestion of a slight degradation of the quality of the flanking sequence ranking), with a strong enrichment for "hotter" motif occurrences in the top 250 motifs, but peaks always restricted to populations with West African ancestry. Similarly, fitting based on only odd chromosomes, and testing using even chromosomes, produced the same result. Once again, the YRI/CEU and AA/deCODE comparisons exhibited very similar African-enriched rate rises (Figure S8).

(vii) Analysis of motifs occurring within repeats

Our scoring utilized both repeat and non-repeat regions of the genome, and thus is likely to include information from hotspots centered within repeat regions of the genome. To determine whether the motifs we have identified are strongly associated with African-enriched hotspot activity within particular repeat families, we identified all occurrences of the 9-mer "CCCCAGTGA" that occur within repeats. For individual repeat types in the genome (e.g. AluY) we then tested (via a t-test) whether there was evidence for an increase in rates (2kb scale) in the AE Map relative to the S Map, and in the YRI Map relative to the CEU Map, in cases where the motif is present. In general, we found a weak consistent elevation in African recombination rates at motif sites across many repeats, but few very strong rises for particular repeats. Thus, unlike the motif "CCNCCNTNNCCNC", which corresponds to the predicted binding target of the A/B alleles at *PRDM9*, and appears to promote strong hotspot activity within, for example, THE1 and L2 repeat family members[5], the motif "CCCCAGTGA" is not strongly associated with particular repeats (with an exception below). The following tables give average rates and evidence for a rate rise in the respective AE-S and YRI-CEU comparisons. The rows include all the repeat types, among the 100 repeats most commonly containing the motif, that show any evidence for enrichment (uncorrected $P<0.1$) in either comparison, and where the estimated African recombination rate (AE or YRI) exceeds 1.1cM/Mb. Rows in the table are ranked by the difference in rates between populations.

Rates around occurrences of CCCCAGTGA within repeats, in the YRI and CEU Maps

Repeat name	YRI mean	CEU mean	Occurrences	P-value
L1PA10	3.98	0.93	59	0.023
MLT1F	2.28	0.92	169	0.00078
L1M1	1.53	0.30	31	0.064
MLT1G1	1.39	0.36	27	0.014
MLT1H	1.40	0.41	37	0.020
L1M2	1.40	0.56	80	0.074
AluJb	1.51	0.69	74	0.022
MLT1F1	1.08	0.42	25	0.042
MER20	1.12	0.48	51	0.096
MER52D	1.34	0.86	51	0.097

ERVL-E	1.17	0.74	72	0.09

Rates around occurrences of CCCCAGTGA within repeats, in the AE and S maps

Repeat name	AE mean	S mean	Occurrences	P-value
L1PA10	2.53	0.76	59	0.062
L1PA13	2.38	0.85	122	0.0074
L2	2.41	1.88	899	0.086

There is thus little evidence for repeats consistently corresponding to intense African hotspots (evidenced by the only weakly significant signals, and modest AE/YRI map mean rate increases), and no strong consistency between the two map comparisons. However, it is notable that in the AE vs. S comparison the top two repeats, showing weak evidence for YRI-specific hotspots, are highly similar members of the L1PA family of L1 Long Interspersed Nuclear Elements, L1PA10 and L1PA13. L1PA10 also shows the largest African-enriched increase in average recombination rate in the YRI-CEU comparison. We therefore chose these repeats for further study and plotted average rates around occurrences of CCCCAGTGA within L1PA10 and L1PA13 repeats, using both the YRI vs. CEU and AE vs. S comparisons (Figure S10). In both cases, there was considerable background noise, but a sharp narrow peak centered exactly at the location of the 9-mer motif, and specific to the YRI and AA maps. Thus, we hypothesize that that motif occurrences within L1PA10 and L1PA13 can generate African-enriched hotspots, a hypothesis that is supported by further evidence below. Interestingly, L1 repeats in general have previously been observed to be unusually recombinationally *inactive*[1], but this study demonstrates that they can indeed contain hotspots.

To further test for whether these repeats truly harbor African-enriched hotspots, we explored the 8-bp sequence downstream from CCCCAGTGA in these repeats, in order to identify further influential bases in these repeats, and to attempt to understand why such hotspot activity might occur. We identify the base at each position downstream of CCCCAGTGA, and within only L1PA10 or L1PA13 cases, that gives the *largest* difference between the YRI and CEU rates (in the same way as for the 8 motifs obtained above, where the L1PA repeats considered here made little, and in the cases of motifs 2 and 6 no, contribution). We obtain the following consensus sequences, with positions where variation has a significant (P<0.01) influence on the difference in rates within L1PA13 elements by the Kruskal-Wallis test, showing asterisks. Below these new consensus sequences we show the 17-bp African-enriched hotspot motif identified above:

```
L1PA10 consensus:   CCCCAGTGAGTGTTGCT
L1PA13 consensus:   CCCCAGTGAGTGTTACT
17-bp consensus:    CCCCaGTGAGCGTtgCc
                             *    ***
```

The two new consensus sequences completely agree except at base 15, and strikingly, they also closely match the 17-bp consensus (15/17 and 14/17 bases respectively). For an example of the effect of the downstream sequence, on both backgrounds, it appears a ~~–G~~" at position 10 within this extended motif is critical for hotspot activity, with a mean YRI-CEU rate difference of 5.9cM/Mb for occurrences with a ~~–G~~" at this position, and <1cM/Mb otherwise. Restricting to cases with a ~~–G~~" in this position strongly enriches the signal for recombination activity at these

elements (Figure S10). The reason for the high average activity of the motif on these backgrounds therefore seems to be that L1PA10 and L1PA13 repeats frequently contain close matches to the 17-mer. Thus, hotspot activity occurs within L1PA10 and L1PA13 elements at motif occurrences, in a manner highly consistent with the genome as a whole.

(viii) Comparison with *PRDM9* allele predictions

Using our analysis based on the 1000 Genomes Project data, we found that in human populations, the SNP rs6889665 most associated with the AE phenotype is strongly associated with a group of *PRDM9* alleles. These alleles, which have an overall frequency of 36% in West Africa and just 1.3% in Europe, and include allele "C", the most common *PRDM9* allele specific to West African populations with frequency 13%[2], all share a similar group of C-terminal zinc fingers, and all are predicted to bind a common motif using these fingers, of length 17-bp: "CCgCNgtNNNCgtNNCC"[2]. Other than at these zinc fingers, the motif prediction is highly degenerate for allele C itself. Further, different C-like alleles vary in their upstream zinc fingers, and thus there is no longer consensus sequence for these *PRDM9* types.

We verified the 17-bp binding prediction by using a different bioinformatic approach to that used by Berg et al.[2] Following the approach of refs. 6 and 7, we predicted a consensus sequence for each of the alleles published by Parvanov et al. (2010)[8], matching these alleles to those identified by Berg et al. (2010)[2] where relevant. The "C" allele of Berg et al. (2010)[2] is identical in terms of DNA-contacting amino acids at positions -1,2,3,6 to the "CH3" allele of Parvanov et al. (2010). We then used the same approach as described in refs. 6 and 7, taking prediction parameters as for the reference B allele of *PRDM9*; i.e. A=-2.20, B=0.24, and predicted a consensus sequence for this allele as well as degeneracy levels within the motif. We reproduced an identical consensus sequence for the corresponding zinc fingers to that previously reported, with the exception that we predict stronger influence of bases 7 (T) and 13 (T) in the motif, but the identical optimal base. Putting in upper case more confidently inferred bases (entropy >0.48) in our prediction for the entire binding domain, we obtain:

```
Our allele C prediction:        aaCaagaCgaCgaagaagaaCCgCagTaagCgTaaCCgat
Berg et al. (2010) prediction:                  CCgC-gt---Cgt--CC
```

Similarly, our predictions for other C-like alleles were almost identical to those previously reported[2] and identically matched the same 17-bp motif. We chose, for consistency with previous work, to use the previously reported 17-bp motif in Figure 4, but here also compare the motif we predicted (independently) from the rate comparison, and described in the previous section, with our C-like allele prediction.

Comparing the 17-bp motif we identify in this work with the two predictions based only on the C-like allele predictions gives the following:

```
African-hotspot based:          CCCCaGTGAGCGTtgCc
Our C-like allele prediction:   CCgCagTaagCgTaaCC
Berg et al. (2010) prediction[2]: CCgC-gt---Cgt--CC
Match:                          . . . ..   ...  ..
```

The predictions are all almost identical, particularly at confidently inferred positions. Specifically, our *PRDM9* C-like allele motif prediction matches the African-enriched hotspot motif at all 8 strongly predicted positions within the 17-bp region of alignment ($P=1.53\times10^{-5}$) and 10 of 11 positions predicted by Berg et al.[2] ($P=8.1\times10^{-6}$).

To conclude, we have presented evidence for an African-enriched hotspot motif, which almost exactly matches a predicted binding motif for a group of "C-like" *PRDM9* alleles with frequency up to 36% in West Africans, and only 1.3% in Europeans, and which exhibit an extremely strong association with the SNP rs6889665. Adding the frequency of the A/B allele in these groups, the A/B allele and the C-like alleles together account for over 90% of *PRDM9* alleles in both Europeans and West Africans. In light of these findings, we believe that it may be difficult to identify further motifs for recombination in populations similar to those we are analyzing here. The remaining 8% of *PRDM9* alleles have diverse predicted binding targets that in the majority of cases are similar to, but distinct from, the A/B alleles, further complicating such analysis[2].

References for Note S6

[1] 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).

[2] Berg, I.L. *et al.* *PRDM9* variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genetics* **10,** 859-863 (2010)

[3] Schneider T.D. & Stephens R.M. Sequence Logos: A New Way to Display Consensus Sequences. *Nucl. Acids Res.* **18,** 6097-6100 (1990).

[4] Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.* **40**, 1124–1129 (2008).

[6] Myers, S. *et al.* Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. *Science* **327**, 876–879 (2010).

[7] Persikov, A.V., Osada, R., Singh, M. Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics* **2,** 22-29 (2009).

[8] Parvanov, E.D., Petkov, P.M. & Paigen, K. Prdm9 controls activation of mammalian recombination hotspots. *Science* **327**, 835 (2010).

Table S1

Comparison of the AA Map and deCODE Map using Pearson and Spearman correlations

Scale (interval size)	Spearman rank correlation of the AA Map (deCODE Map) to the specified LD Map		Pearson correlation of the AA Map (deCODE Map) to the specified LD Map, replicating Table 1 for easy comparison*	
	CEU	YRI	CEU	YRI
3 kb	0.64 (0.31)	0.69 (0.32)	0.66 (0.58)	0.71 (0.53)
10 kb	0.72 (0.47)	0.78 (0.48)	0.73 (0.70)	0.78 (0.65)
30 kb	0.83 (0.70)	0.87 (0.69)	0.78 (0.78)	0.83 (0.74)
100 kb	0.90 (0.88)	0.92 (0.86)	0.84 (0.85)	0.87 (0.81)
300 kb	0.92 (0.92)	0.94 (0.91)	0.89 (0.90)	0.92 (0.88)
1 Mb	0.95 (0.95)	0.96 (0.94)	0.94 (0.94)	0.95 (0.95)
3 Mb	0.97 (0.97)	0.97 (0.97)	0.97 (0.97)	0.98 (0.97)

Table S2

Results of genome-wide association testing for the AE and hotspot usage phenotypes

SNP	P-value	YRI freq.*	CEU freq. *	No. SNPs at P<3.3×10^{-7} within 10 Mb	Chromosome : Position in hg18	Genes within 50kb/notes
African-enrichment						
rs6889665	1.5×10^{-246}	0.29	0.02	506	5 :23,568,400	*PRDM9*
rs11888485	2.9×10^{-8}	0.92	1.00	4	2 : 118,403,702	*CCDC93*
Hotspot Usage						
rs6889665	1.8×10^{-52}	0.71	0.98	36	5 : 23,568,400	*PRDM9*
rs9987353	7.8×10^{-9}	1.00	0.70	4	8 : 9,153,759	Confirmed cis-effect; occurs within the copy-number rearranged *β-Defensin* cluster at 8p23.1§
rs10015037	1.5×10^{-7}	0.83	0.85	1	4: 98,101,343	

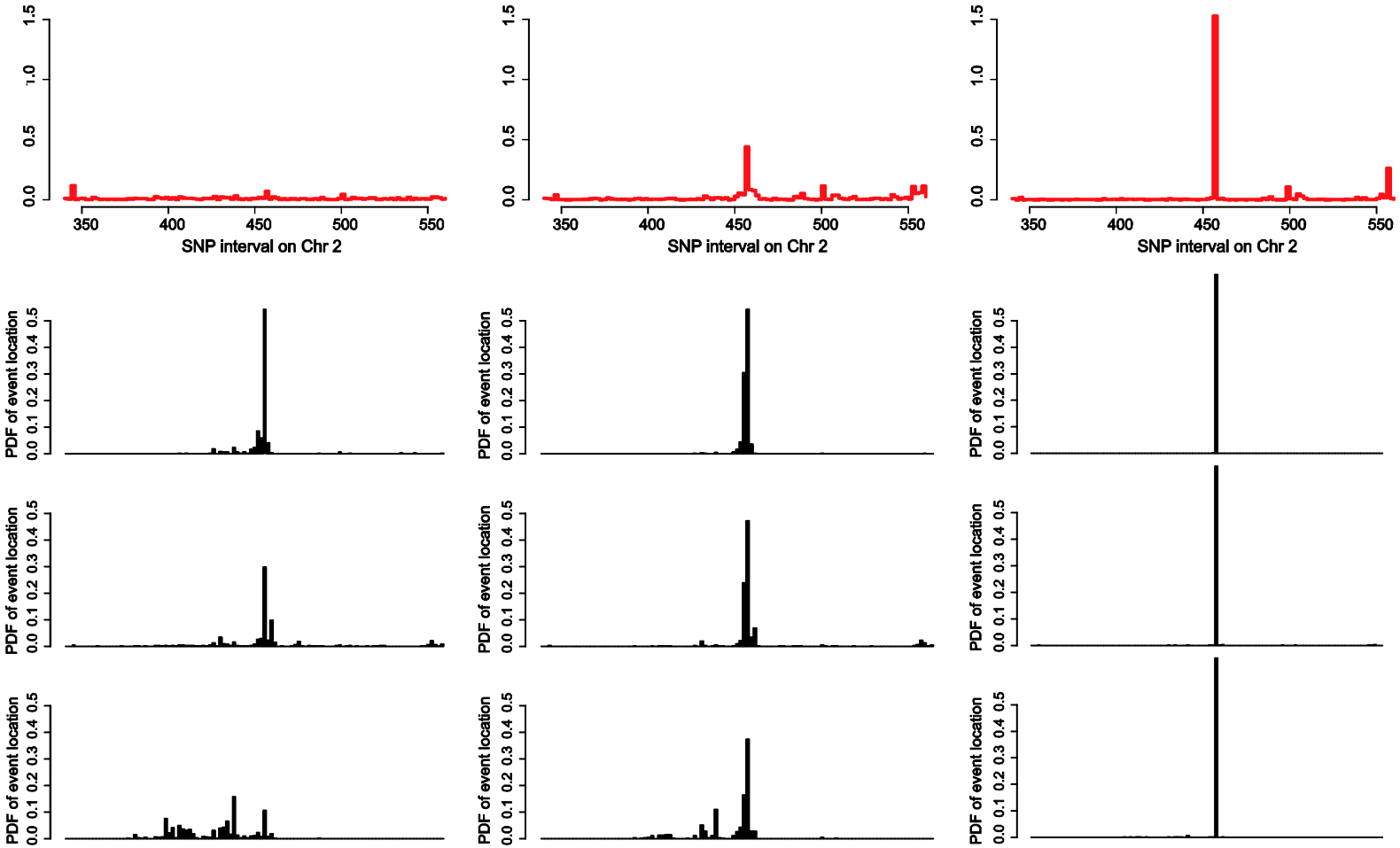
Note: We tested three recombination-related phenotypes: African-enrichment (AE), Hotspot Usage, and genome-wide recombination rate (genome-wide rate was tested pedigrees only). The table reports SNPs with a suggestive P <3.3×10^{-7} (less than one expected in scan in the absence of a true signal). Due to the clustering of significant SNPs, we pick the SNP with the smallest P-value to represent each locus (defined as a 10 Mb window in either direction of the most associated SNP).

* We quote the frequency of the allele that confers a higher value for the phenotype.

§ We retested SNPs in Chr 8 for association to a new AE phenotype defined only using crossovers and maps on other chromosomes 1-7 and 9-22. The association signal around rs9987353, which occurs within a region of Chromosome 8 known to harbor complex rearrangements in humans, disappeared confirming this as a cis-effect driven by unusual recombination patterns local to the SNP itself. This signal may be artefactual or due to an interaction between recombination and these rearrangements.

Figure S1

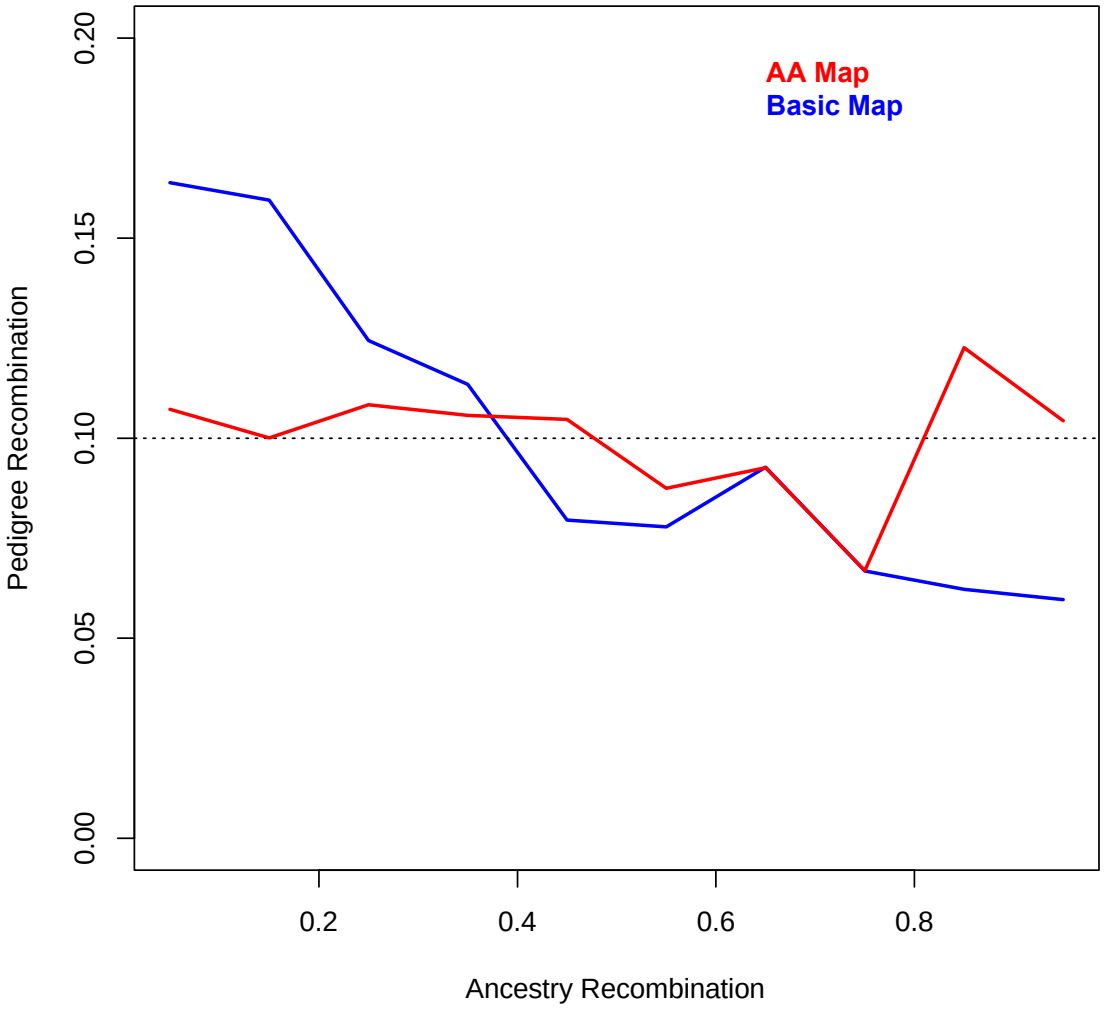
Localization of hotspots by the MCMC



Legend: The top panels show the recombination rate (cM) estimated by the MCMC at different stages in the chain for a small region in Chromosome 2. The bottom panels show three events occurring in different individuals in the same region of the genome. The x-axis shows SNP position and the y-axis shows the estimated probability distribution function (PDF) of localization of each event. The chain was started in the left panel using a uniform recombination rate per base of 1.1cM/Mb. The middle panel shows the state of the chain after 100 iterations and the final panel shows the state of the chain after 10,000 iterations.

Figure S2

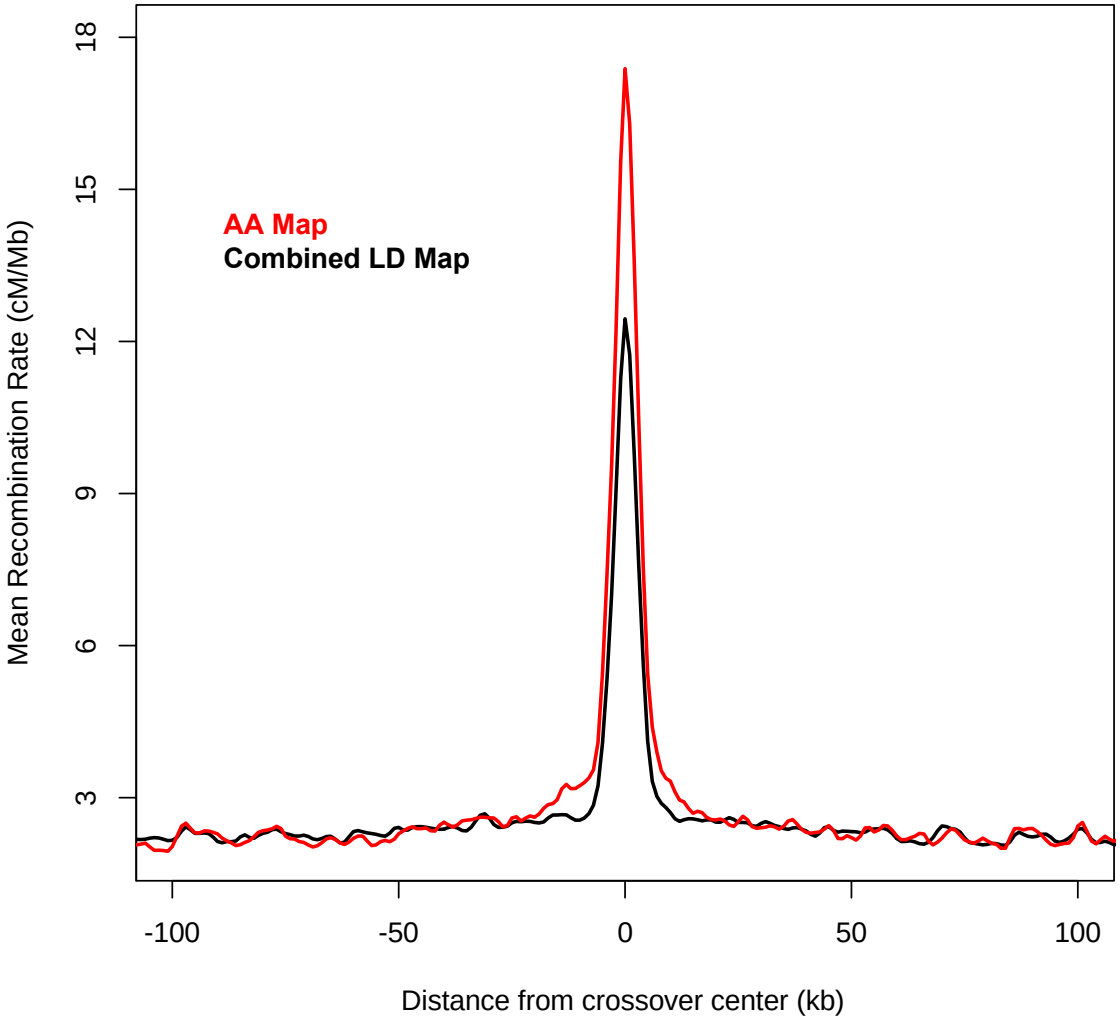
Comparing the calibration of the AA and Basic maps (based on the Pedigree Map)



Legend: Validation of the accuracy of recombination probabilities estimated by the AA Map using 370 crossover events observed in pedigrees as well as in unrelated individuals. The x-axis has 10% bins of recombination probability, with the left-most containing SNP intervals with the highest probability density per base of sequence and the right-most bin having the lowest. The y-axis shows the corresponding probability observed in the Pedigree Map crossovers. The slope of the curve of the Basic Map shows that it underestimates the recombination rates of hotspots, while the flat line for the AA Map suggests accurate estimation. (Details are included in Note S1).

Figure S3

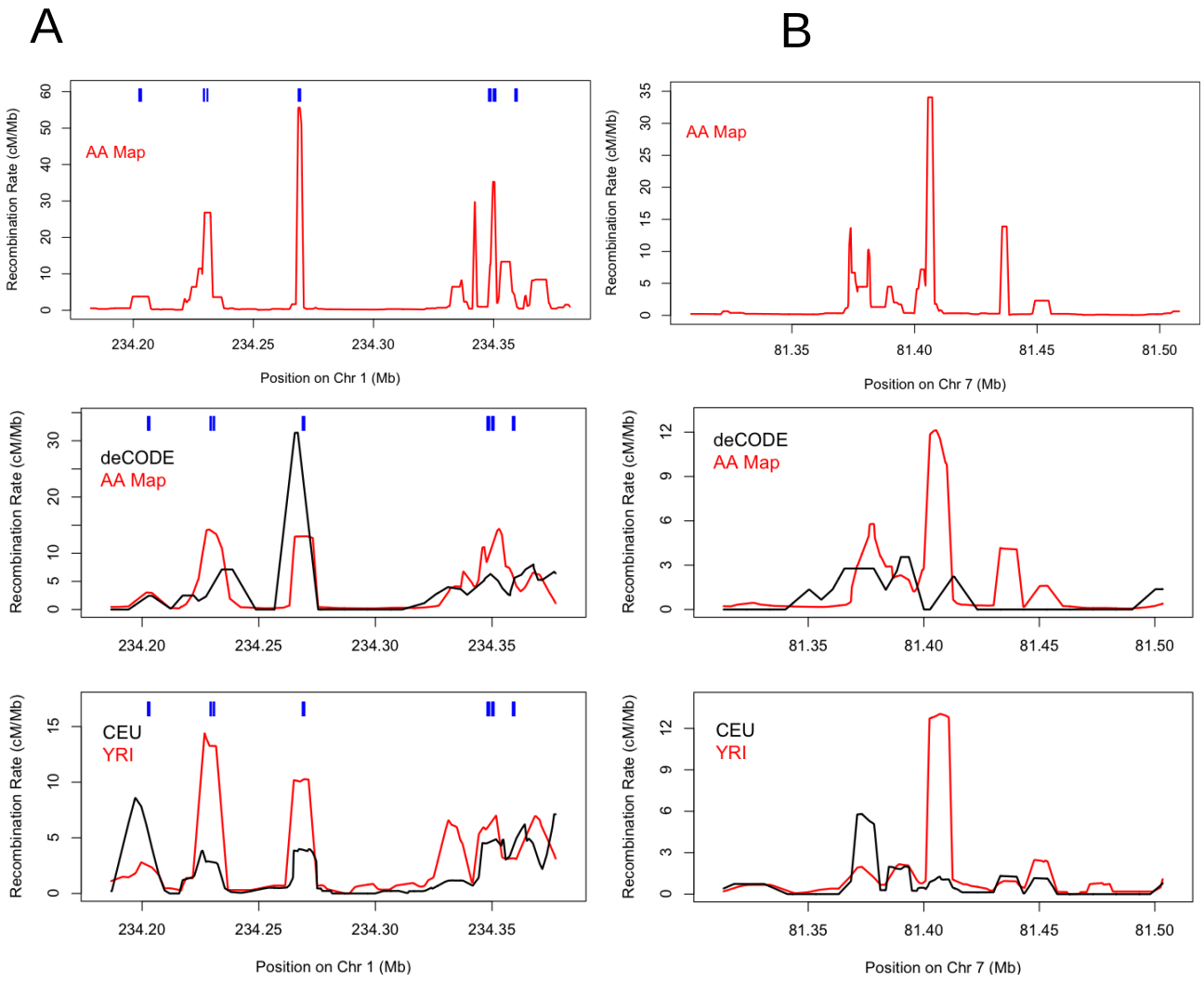
Recombination rates near crossovers detected in African American pedigrees



Legend: Mean recombination rate in the AA Map (red) and the population-averaged HapMap2 LD-based map (black) around 3,068 narrowly defined crossover events directly identified in African American families. The AA Map has a resolution comparable to the LD-based map, and also has a rate peak at least as high as the LD-based map.

Figure S4

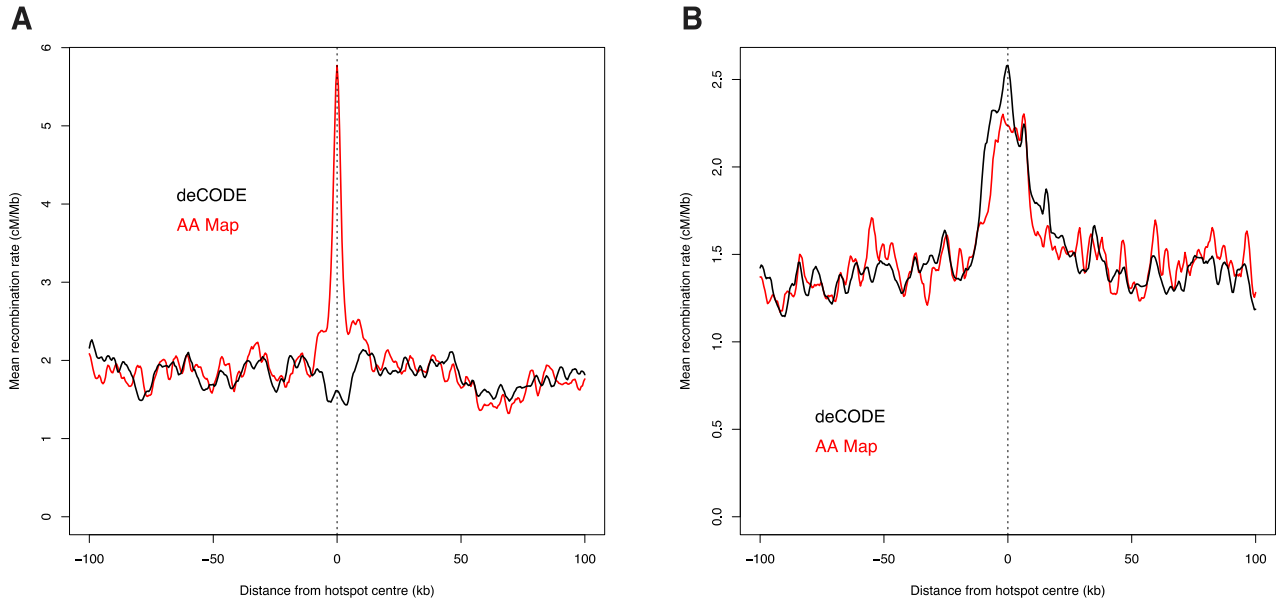
Detailed comparisons of genetic maps over selected 200kb regions



Legend: (A) A 200 kb region flanking the MS32 minisatellite where the positions of recombination hotspots in Europeans have previously been defined by sperm typing. The top panel shows our map at 1 kb resolution, whereas the bottom two panels show four maps smoothed to a 10 kb resolution to allow them to be compared (AA Map, deCODE Map, CEU LD Map, YRI LD Map). The higher resolution of the AA Map compared with the deCODE Map is evident. **(B)** A 200 kb region in chromosome 7 is presented as another illustration of an African-enriched hotspot that is inferred by both the AA and YRI Map. This site is cold with a low recombination rate in both the deCODE and CEU maps.

Figure S5

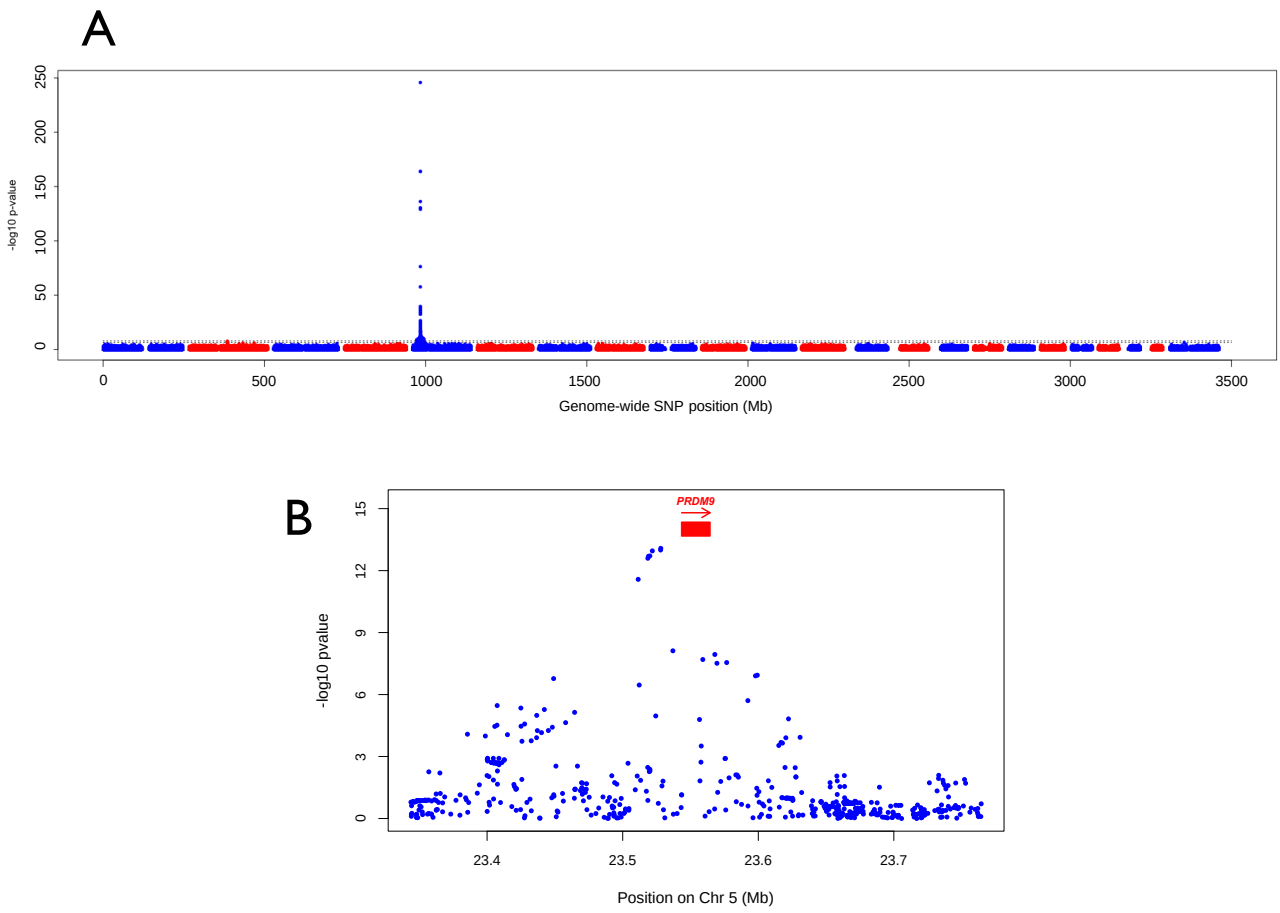
The AA and YRI maps show a shared signal for African-enriched hotspots not signaled in European maps



Legend: (A) Mean estimated rates in the AA Map (red) and the deCODE Map (black) (2 kb bins with a 500bp sliding window) surrounding 2,375 regions with high rates in the YRI Map (≥ 5 cM/Mb over 2kb) but not the CEU Map (<1 cM/Mb). The peak in the AA Map only (with no signal in the deCODE Map) validates the presence of genuine African-enriched hotspots in individuals with West African ancestry. **(B)** In the reciprocal experiment we searched for candidate European-specific hotspots by examining positions with high rate estimates in the CEU (≥ 5 cM/Mb over 2kb) but not in the YRI map (<1 cM/Mb). We plot the rates around the 1,263 such loci. The presence of similar and weak peaks in both deCODE and AA Maps suggests that these loci in fact represent weak hotspots active in both Europeans and West Africans, so the differing rate estimates in the LD maps at these loci are most likely due to statistical noise in the detection of weak hotspots, and there is no support for genuine European-specific hotspots.

Figure S6

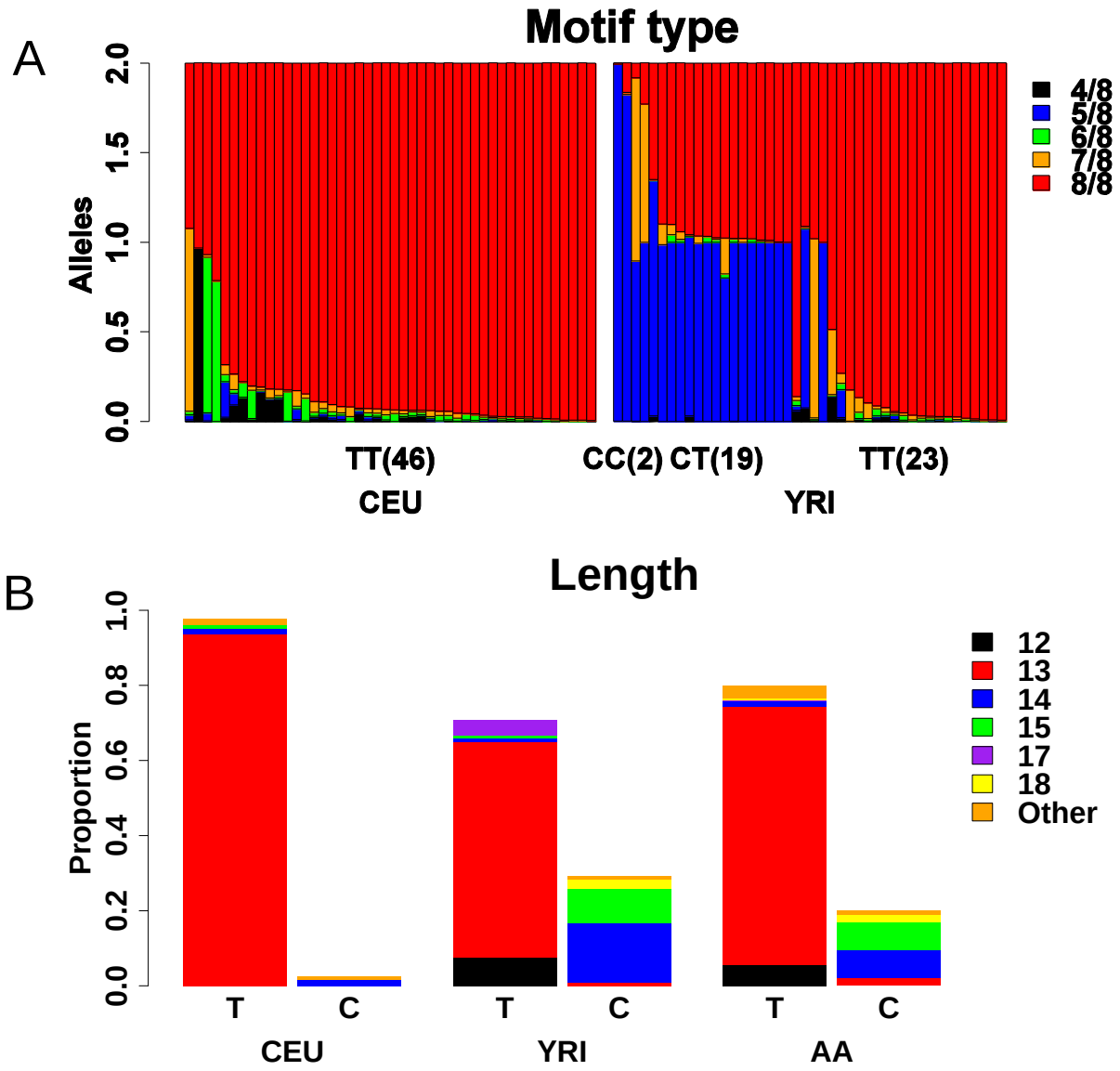
Variation at PRDM9 is associated with the African Enrichment phenotype



Legend: (A) Results of genome-wide association testing of more than 3 million SNPs for the African-enrichment (AE) phenotype in pedigrees and unrelated individuals. The y-axis corresponds to -$\log_{10}$ P-values for association. Changing colors indicate alternating chromosomes and dashed lines indicate thresholds for genome-wide suggestiveness (P<1 after Bonferroni correction) and significance (P<0.01). The most associated SNP is rs6889665 on chromosome 5 which is 4 kb downstream of *PRDM9* ($P=1.5\times10^{-246}$). **(B)** Blow-up of the *PRDM9* region with P-values after conditioning on the genotype of rs6889665. A significant residual signal is observed in several SNPs. The strongest residual association is with rs10043097 ($P= 8.3\times10^{-14}$), upstream of the *PRDM9* transcription start site.

Figure S7

Relationship between alleles at rs6889665 and *PRDM9 variants*

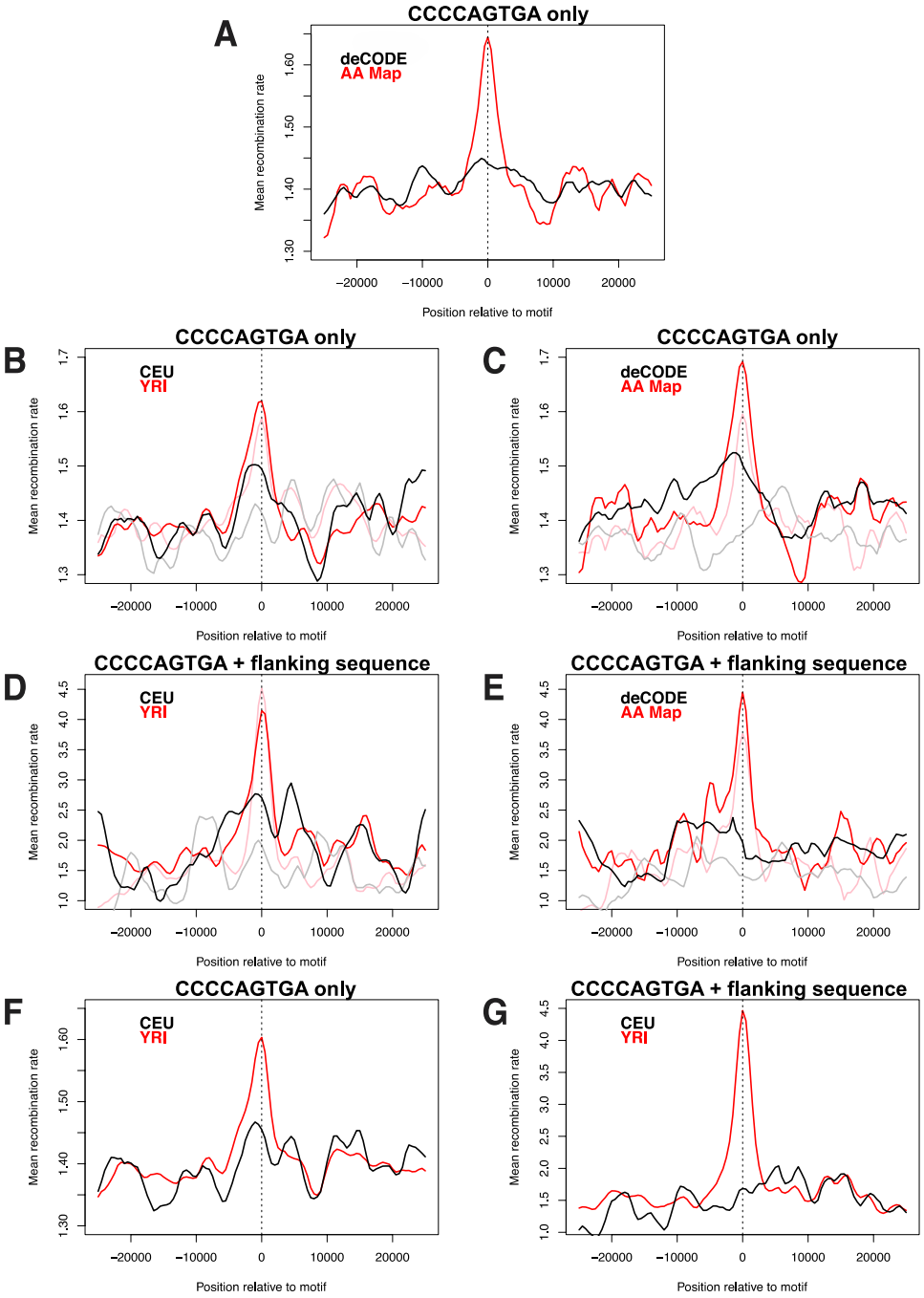


Legend: (A) Association of rs6889665 with predicted *PRDM9* binding targets. We used data from the 1000 Genomes Project (Note S5) to infer the *PRDM9* alleles for each of 46 CEU (left hand bar plot) and 44 YRI (right hand bar plot) individuals, among 29 *PRDM9* alleles identified previously via sequencing by Berg et al. (2010). Each individual carries 0-2 copies of each allele. We summed across *PRDM9* alleles according to how closely their previously predicted binding target defined in Berg et al. matches the recombination hotspot motif CCNCCNTNNCCNC, to give the expected number of copies (0-2) an individual carries of alleles matching 4,5,6,7 or all 8 non-degenerate bases of this motif. Each column corresponds to an individual and shows, of their two alleles, the expected number predicted from their 1000 Genomes Project data to match to each degree the motif. For example, the first CEU individual is predicted to have one *PRDM9* allele binding a target containing a 7/8 motif match and one whose target perfectly (8/8) matches the hotspot motif. Finally, we ordered individuals according to their genotype at SNP rs6889665, which is strongly associated with fine-scale recombination rates and which was not used in the *PRDM9* allele prediction. rs6889665 genotype is shown below the bar plot, with numbers of columns for the corresponding genotype bracketed. Note that individuals carrying *k* copies of the "C" allele at this SNP are almost always predicted to have *k PRDM9* alleles targeting a 5/8 motif match, for *k*=0,1,2, though there appear to be two exceptional YRI individuals among 23 carrying only T alleles at this SNP, but nevertheless with one predicted *PRDM9* allele targeting a 5/8 match. Note also that in both populations, *PRDM9* alleles in "TT" individuals at this SNP are typically predicted to bind exact 8/8 matches to the motif, but there are a number of cases where individuals appear to carry rarer alleles, predicted to bind 4/8,6/8 or 7/8 matches. **(B)** Relationship between rs6889665 and the length of *PRDM9* ZF array. We experimentally measured the number of zinc fingers in PRDM9 in 354 individuals including 166 African Americans (Methods). We find that the ancestral allele 'T' is almost always associated with the PRDM9 ZF array length < 14 (96%) while the derived allele 'C' is associated with an array of length ≥ 14 (93%).

doi:10.1038/nature10336

Figure S8

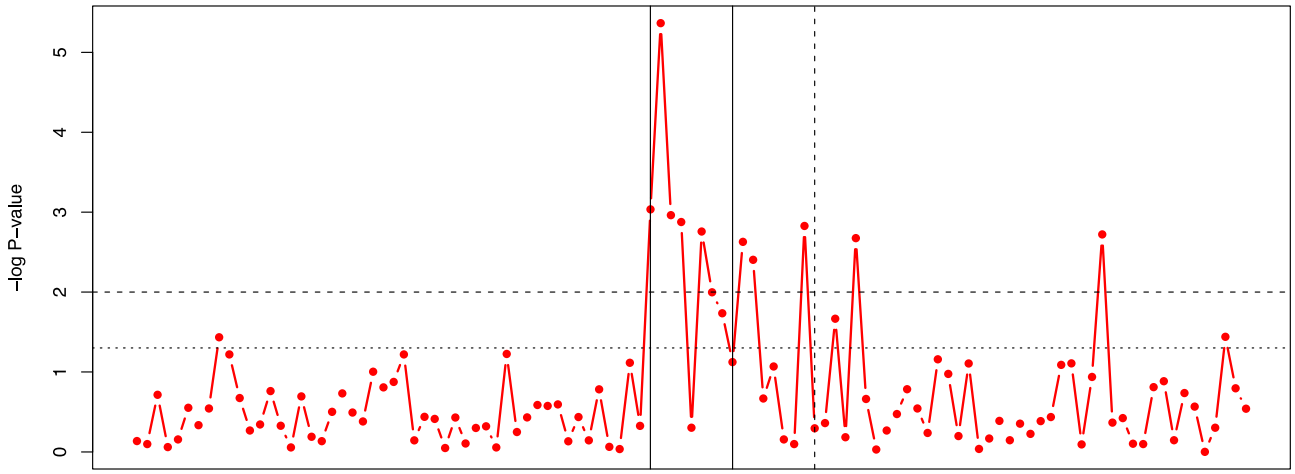
Hotspot activity around motif CCCCAGTGA and flanking sequence



Legend: Verification of the role of CCCCAGTGA and flanking sequence in determining hotspot activity. **(A)** Average recombination rates in the AA Map (red) and deCODE Map (black) surrounding exact genomic sequence matches to the motif CCCCAGTGA. We excluded motifs within 5 Mb of chromosome ends. **(B)** Average recombination rate around all occurrences of the motif CCCCAGTGA occurring within odd chromosomes (dark lines) and even chromosomes (lighter lines) estimated using the YRI (red) and CEU (black) maps. **(C)** As B, except the plot shows the AA Map mean rates (red lines) and deCODE mean rates (black lines). Note that the rates are similar to those for B. **(D)** As B, but the plot now shows average recombination rate around only the top 250 occurrences of the motif, ranked in terms of flanking sequence. The ranking used only a degeneracy matrix (Note S6) inferred from the even chromosomes, and applied to the odd chromosomes (dark lines), or inferred from the odd chromosomes and applied to the even chromosomes (lighter lines), so that for both pairs of lines the data used to fit the ranking procedure is independent of the plotted data. **(E)** As D, with rates plotted around the same motif occurrences, except the plot shows the AA Map mean rates (red lines) and deCODE mean rates (black lines). Note that the similarity to D again holds. **(F)** As A, rates plotted around occurrences of CCCCAGTGA in the genome except using YRI mean rates (red) and CEU mean rates (black). Note the similarity with A. **(G)** Average recombination rates plotted around the top 500 ranked occurrences of CCCCAGTGA in the genome, ranked in terms of flanking sequence using the YRI map mean rates (red) and CEU mean rates (black).

Figure S9

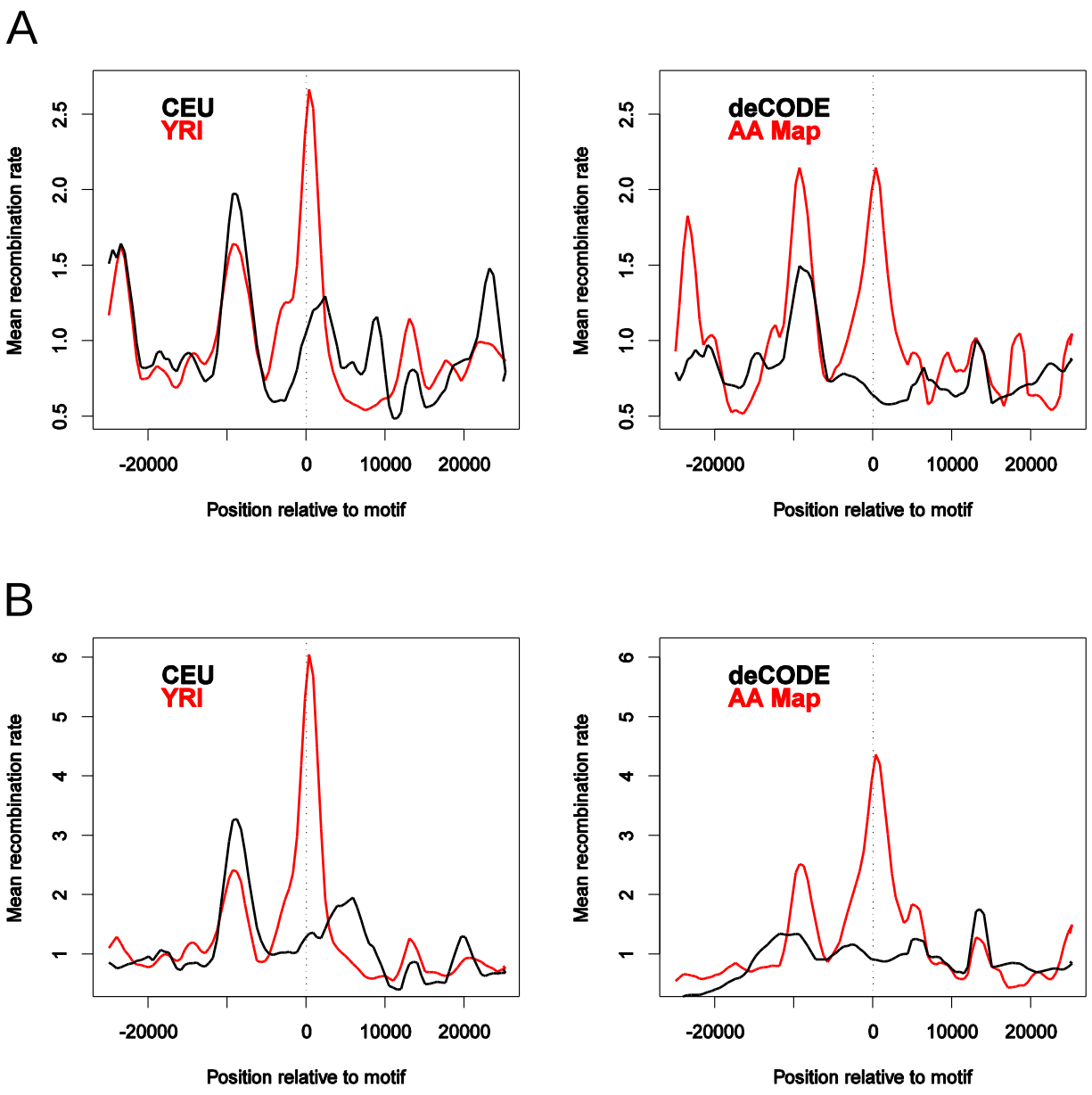
Influence of flanking bases around 9-mer motif CCCCAGTGA



Legend: -$\log_{10}$ P-values at positions surrounding the motif CCCCAGTGA (or within 1 bp mismatch of the motif) under a null hypothesis that recombination rate is independent of the base at that position. Vertical solid black lines mark the extent of the 9-mer. The vertical dashed black line marks the end of the 17 bp motif CCCCaGTGAGCGTtgCc which extends the core 9-mer towards the right. Horizontal lines mark P=0.01 (dashed) and P=0.05 (dotted).

Figure S10

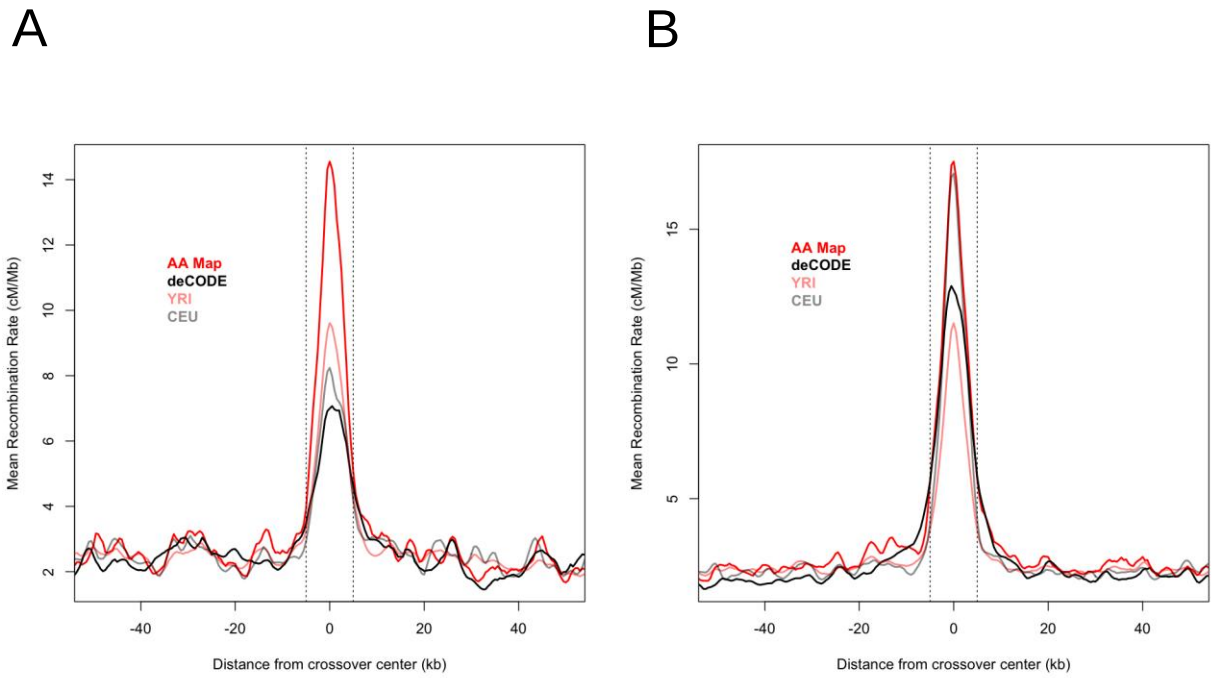
Rates around occurrences of CCCCAGTGA within L1PA10 and L1PA13 repeats



Legend: Exploration of recombination patterns around occurrences of CCCCAGTGA within L1PA10 and L1PA13 LINE repeats. **(A)** The two plots show average estimated recombination rate around all occurrences of the motif CCCCAGTGA within L1PA10 or L1PA13 repeats, in CEU and YRI (left plot) and the deCODE and AA maps (right plot). Only the peak located at position 0, and observed in the YRI and AA maps, is completely specific to these groups, and absent in the CEU and deCODE maps. **(B)** The two plots shown are identical to those in A, but show rates around occurrences of the 1 base extended motif CCCCAGTGAG within L1PA10 or L1PA13 repeats. This extension corresponds to a strong additional (~2-fold) increase in estimated African rate for both comparisons, and strengthens homology to the 17-bp consensus motif. (Additional homology strengthening bases further increase mean African rate within these repeats.)

Figure S11

Rates in different maps for pedigree individuals who are CT or TT at SNP rs6889665



Legend: We plot recombination rates in different maps around all crossover events resolved within 10kb in pedigree individuals. Vertical dotted lines show the maximum extent of included events (5kb on either side of crossover center). (**A)** 544 recombination events from 51 individuals heterozygous at rs6889665 were included. (**B)** 1,009 events from 102 individuals homozygous in the major allele 'T' were included. The rates in the European maps of deCODE and CEU show a significantly stronger hotspot in TT individuals than in CT individuals, consistent with the finding that African-enriched hotspots are activated by PRDM9 variants tagged by the C allele.