Recombination rates in admixed individuals revealed by ancestry–based inference

Supplemental Information

Contents

Part I
Supplementary Note

1 Samples and Genotyping

We outline here the sampling and genotyping procedures for the four projects this study drew data from. The data were shared among the participating institutions in this study via the data sharing protocol of the NIH STAMPEED project.

1.1 GeneSTAR

Participants (N= 1201; 415 after filtering relateds) were apparently healthy adult (21 to 80 year old) family members of African American probands hospitalized with premature coronary artery disease prior to 60 years of age. Eligibility criteria included being free of coronary disease and any life-threatening comorbidity. DNA was isolated from plasma and SNP genotyping was performed on the Illumina Human 1Mv1_C array platform at deCODE, Genetics, Inc. All samples have written informed consent and the analysis was approved by the Johns Hopkins Medical Institutions IRB. After applying our global SNP filters (see below), 875,418 SNPs remained. The average call rate per SNP was 99.8%, and minimum call rate was 87.2%. The sample is more fully described in a prior publication[1].

1.2 GENOA

The Genetic Epidemiology Network of Arteriopathy (GENOA) is one of four networks in the Family Blood Pressure Program (FBPP) which recruited hypertensive black and non-Hispanic white sibships for linkage and family-based association studies to investigate genetic contributions to blood pressure and, the cardiac and renal complications of, hypertension in multiple racial groups[2, 3]. Participant recruitment for GENOA (Exam 1, 1995–2000 and Exam 2, 2000–2005) was population-based in two geographic locations: Jackson, Mississippi and Rochester, Minnesota. African Americans in the study were located solely at the Jackson field center. Hypertensive probands were ascertained from the Jackson cohort of the Atherosclerosis Risk in Communities (ARIC) study if they were in a sibship with $\geq$ 2 individuals with essential hypertension (systolic BP $\geq$140 mm Hg or diastolic BP $\geq$90 mm Hg on the second and third clinic visit), diagnosed prior to age 60, and consented to participate. Index sib-pairs with possible secondary hypertension, including sib-pairs with previously diagnosed kidney disease (defined by serum creatinine level $>$ 2 mg/dL), were excluded. All subjects provided written informed consent and Institutional Review Board (IRB) approvals were obtained from each participating institution. The initial sample contained 1,263 African Americans, but with filtering for known related individuals based on a pedigree file, 583 individuals remained for analysis. DNA samples were genotyped by Affymetrix 6.0 platform at Mayo Clinic. After excluding SNPs with call rate less than 95% or with MAF less than 1%, there were 762,766 SNPs available for 583 African Americans. After applying our global SNP filters (see below), 659,781 SNPs remained. The average call rate per SNP was 99.4%, and minimum call rate was 92.7%

1.3 GRAAD

This study contains samples from the Baltimore-Washington, D.C. metropolitan (GRAADi) and Barbados (GRAADii). All subjects gave verbal and written consent as approved by the Johns Hopkins Institutional Review Board (IRB) and the Barbados Ministry of Health. Genotypes were generated by the Johns Hopkins University SNP Center at the Center for Inherited Disease Research (CIDR) for 665,352 polymorphic tagging SNPs using Illumina HumanHap650Y Versions 1 and 3 BeadChips and the Illumina Infinium II assay protocol[4]. While the overall study design is further described elsewhere[5], we summarize the two samples briefly.

1.3.1 GRAADi

This study includes 498 asthma cases and 500 non-asthmatic controls from the Baltimore-Washington, D.C. metropolitan area who self-reported as African American ethnicity. Of those, 935 unrelated individuals were included in the present study. These subjects comprised the consortium for "Genomic Researchon Asthma in the African Diaspora" and represent eight separate, NIH-funded studies of asthma in pediatric and adult African American populations, plus one study on healthy African Americans. Informed consent was obtained from each study participant, and the study protocol was approved by the institutional review board at either the Johns Hopkins University or Howard University. 935 individuals. After applying our global SNP filters (see below), 570,293 SNPs remained. The average call rate per SNP was 99.6%, and minimum call rate was 85.0%

1.3.2 GRAADii

Additionally, a population of 163 African–Caribbean families ascertained through asthmatic probands from Barbados and containing a total of 1,028 individuals. Of those, 299 unrelated individuals were included in our effort to infer ancestry switch points. In addition, 99 independant quartets (mother, father and two siblings) were used to infer a crude recombination map (see Section 4.1.2). Probands were recruited through referrals at local polyclinics or the Accident and Emergency Department at the Queen Elizabeth Hospital as previously described, and their nuclear and extended family members were recruited[6, 7]. After applying our global SNP filters (see below), 568,125 SNPs remained. The average call rate per SNP was 99.5%, and minimum call rate was 88.3%

1.4 SARP and CAG/CSGA

A total of 632 participants were recruited through the multicenter NHLBI Severe Asthma Research Program (SARP), University of Chicago as part of the Chicago Asthma Genetics (CAG) and the Collaborative Study on the Genetics of Asthma (CSGA) studies. Eligibility criteria for the CAG included adults and children with severe persistent asthma, and non-asthmatic control subjects over the age of 18. Eligibility criteria for the CSGA included subjects with mild to severe asthma, and non-asthmatic control subjects. All samples have written informed consent and the analysis was approved by the IRB at the University of Chicago. The samples are more fully described in[8]. DNA was isolated from whole blood. SNP genotyping on the Illumina 1M platform was conducted at Wake Forest University. After applying our global SNP filters (see below), 869,755 SNPs remained. The average call rate per SNP was 99.8%, and minimum call rate was 86.5%.

1.5 Merging the data: Global SNP filters

Since the inference of ancestry switch points is performed independently for each individual, we did not restrict our analysis to SNPs typed in all of the different studies. We, however, considered only SNPs which we included in the phased haplotypes used as reference panels (see below). A major concern was to have all SNPs matching the strand of the reference haplotypes. We therefore removed all A/T and G/C SNPs due to difficulties to match the strand. This solely affected the GENOA samples types on the Affymetrix platform. For these samples we removed a total of 98,037 SNPs (14.8% of all SNPs).

In order to identify conflicting SNPs, we compared the allele frequencies of all SNPs to the expected frequency from a a mixture of HapMap CEU and YRI samples with an 20% CEU contribution, as well as to the HapMap ASW sample of African–Americans[9]. This comparison reviled 41 obvious outliers which were subsequently removed from all samples. The final data set is given in Supplemental Table 1.

1.6 Choosing reference panels

In order to choose appropriate reference populations we performed a principal component analysis (Supplementary Figure 6) on all our samples together with several HapMap samples: the two samples of European origin CEU (Utah) and TSI (Tuscany); and the three African samples YRI (Yoruba from Nigeria), LWK (Luhya from Kenya) and MKK (Maasai from Kenya). The analysis was performed using smartPCA[10] and restricted to SNPs common to all samples. We found our African-descendant admixed individuals fall on a straight line connecting the YRI with the European samples, suggesting that the

YRI and any of the European samples are good candidates for reference populations. This is in line with a recent comparison of African segments of African American genome with contemporary West African populations, which revealed that the African segments are most similar to the genomes of non-Bantu Niger-Kordofanian-speaking populations such as the Igbo, Brong, and Yoruba[11]. For the European panel we chose the CEU over the TSI sample due to the larger number of markers with available genotype data for the CEU.

1.7 Genome–wide ancestry proportions

While calling switch–points, the HMM output of an individual can also be used to determine genome–wide proportions of European and African ancestry. We found that the sample had a mean African ancestry coefficient of ~ 0.81 with a 95% interquartile range of 0.54-0.96 (Supplementary Fig.5), a broad range that is generally consistent with the PC analysis reported above (Supplementary Figure 6) and previous studies of African diaspora descendant samples[11, 12, 13, 14]. We find a smaller average proportion of European ancestry among African–Caribbean individuals than among African-American individuals, in agreement with previous findings[15]. The mean European ancestry proportion among African American individuals that participated in this study is 19.4%, compared to 11.7% among African–Caribbeans.

Given the observed differentiation from the African-American samples, we investigated the Afro-Carribean sample in more detail. Of particular interest was potential admixture with Amerindian ancestors, which, if present, would impact our inference of ancestry switch–points. To infer the genome–wide proportion of Amerindian ancestry, we used `ADMIXTURE`, a program which implements a model-based maximum-likelihood approach to estimate the global ancestry proportions from unrelated individuals[16]. We applied this method to a merged sample consisting of the African–Caribbean sample (GRAADii) and the haplotypes of the CEU, YRI and Mexican (MEX) HapMap3 populations[9] (after pruning for linked variants, as suggested in the manual of `ADMIXTURE`). We found the Mexican individuals to be admixed, with a proportion of their ancestry shared with the European (CEU) samples, and a second component which we take to represent Amerindian ancestry (Supplementary Figure 12). The average Amerindian ancestry among the individuals of the GRAADii sample was less than 1%, and the Amerindian ancestry proportion was found to be larger than 2% for only 17 individuals ($\sim$ 5% of all African–Caribbean individuals in this study). These results are in good agreement with previous reports that suggest the populations of the outer Caribbean islands, such as Barbados, show generally only very limited Amerindian ancestry[15]. To check if those low levels of Amerindian ancestry (or any other features unique to the Afro-Carribean sample) would influence the recombination map inferred across all 2864 admixed individuals, we regenerated recombination maps only using the 2565 non–African–Caribbean individuals. We then computed both Pearson and Spearman correlations between the two inferred maps at all scales reported in the manuscript. All correlations were well above 99.9%, suggesting that the heterogeneity introduced by including African–Caribbean individuals does not have large effects on our inference.

2 Simulation Framework

In order to evaluate the performance of the proposed approach and to check the robustness of our inference to parameter choices we inferred relative recombination rates from a simulated data set. This data sets was generated to mimic our African American data set in many aspects. We simulated a total of 120 Mb, consisting of a 6 Mb segments randomly chosen from each of the chromosomes 1-20. We first simulated an ancestral African and an European population for each of these segments. African American haplotypes were then constructed from these reference panels according to a classic admixture model (see below).

2.1 Simulation of ancestral populations

To simulate the ancestral populations prior to admixture, we use the coalescent with recombination approach implemented in MACS[17]. We generated 10,000 haplotypic samples from both an African and an European population for each of these segments assuming the demographic model proposed by[18] and the HapMapCOMBINED recombination map[19]. The resulting SNPs were sub-sampled to match SNP density and the frequency spectra of the CEU and YRI HapMap samples. Of those samples we put 230

African and 234 European haplotypes aside and used them as reference panels in the later analysis. The remaining haplotypes were used to simulate admixed individuals.

2.2 Forward simulations of admixture

The African American population was assumed to have a constant size of 20,000 individuals and to have resulted from a single admixture event with 20% European contribution, followed by seven generations of random mating. We simulated such a diploid population based on our simulated African and European samples forward in time and kept track of all recombination events that were assumed to occur according to the HapMapCOMBINED recombination map[19], which is essentially a 50%/50% average of the HapMapCEU and the HapMapYRI maps.

To simulate admixture, we wrote a forward–in–time computer simulation of admixture between populations and recombination over several generations following a model widely in the population genetic literature for African-Americans[20, 21, 22]. The simulation program, simadmix, was written in C++.

simadmix allows the user to specify initial sub-population sizes, and genetic maps to use for recombination frequencies. In addition, population size and a mating distribution may be specified for each sub-population of each subsequent generation.

Each initial sub-population is a list of diploid individuals, where each individual has a list of chromosomes. Each chromosome has a unique identifier that encodes the sub-population id, the individual id, the pair id (e.g. 1-23 in humans), and which of the two chromosomes in the pair. The chromosome identifier allows precise tracking of recombination events over the course of the simulation.

Generations in the simulation are non-overlapping. Each sub-population of a new generation is created from the previous generation by drawing two parents at random to create a new individual. The parents are picked according to the user-specified mating distribution, which is a joint probability distribution that is used to choose the sub-populations of the parents.

Each parent donates a gamete, where each chromosome in the gamete is generated by random recombination of the corresponding pair of chromosomes in the parents. The random recombination is modeled as a heterogeneous Poisson process with density given by the user-specified genetic map. Specifically, the number of recombination events (possibly zero) is chosen first, according to the total rate of recombination over the chromosome. Then the windows of the recombination events are picked according to the genetic map, which specifies the rate of recombination between any two positions. The specific position in the window is picked uniformly at random. The resulting gametic chromosome is then either identical with one of the parental chromosomes, or a sequence of chromosomal blocks, alternating between the two parental chromosomes. Each chromosomal block retains the unique identifier of the chromosome in the first generation from which it originated.

At the end of the simulation, the haplotype patterns are substituted into the 2864 sampled individuals using haplotypes generated in the previous step (see above). Prior to do so we sub–sampled the SNPs to match the SNP density found among the African–American and African–Caribbean samples used in this study.

2.3 Testing the inference of recombination maps using simulations

We generated two sets of recombination maps based on this data: First, we used our ancestry switch–point method to estimate maps from admixed individuals generated using forward simulations. Second, we used LDhat[23] to estimate LD–based recombination maps in the two reference panels – thereby mimicking how the HapMap LD–based maps were obtained.

The switch–point based recombination maps were inferred as described in the main text. We performed the analysis on each of the 2864 individuals simulated using the forward approach described above. We then inferred relative recombination rates in an interval between markers at physical coordinates j and k using the naive estimator $\bar{c}_{jk}^{(i)}$ as well as the Empirical Bayes estimator r_{jk} for comparison purposes. Unless stated differently, all reported results are based on r_{jk}.

Since we report difficulties to correctly infer switch points close to the ends of the simulated segments (see Supplementary Figure 1), we discarded the first 500Kb from both ends of our simulated 6Mb segments for further comparison. We thus end up with a total of 100Mb, which corresponds roughly to the size of chromosome 15. For plotting we scaled the maps to the total map length used to simulate the segments.

We next ran LDhat to estimate recombination maps based on the African and European reference panel separately. For each segment, the rjMCMC chain implemented in LDhat was run for a total of $2 \cdot 10^8$ iterations, of which only every 2000^{th} was kept. The first 10^4 of those 10^5 samples were discarded and the remaining used to estimate the recombination maps. The rjMCMC was run with block penalty of 1, which is lower than the penalty of 5 used to generate the HapMap maps[24], but resulted in a higher correlation with the underlying map used to simulate the samples. Note that we only used 192 haplotypes from each reference panel, which allowed us to use the lookup table provided with LDhat. To match the switch–point based maps we also discarded the first 500Kb from both ends of each segment.

2.4 Monte Carlo estimation of $p(c^i_{jk}|d)$

The same set of admixed individuals was also used to obtain empirical distributions of the cumulative switch point probability stratified by the true number of switch points for various interval sizes. These distributions were discretized using $l = 600$ frequency bins within [0,3] and represented as a three dimensional matrix $L_{u,v,w}$, where u refers to the interval size, v to the number of switch points observed and w to a given frequency bin. To fill the matrix L we computed the cumulative switch point probability in overlapping intervals with start positions shifted by 1000 bp beginning at the first marker for each individual. For a given interval size b, these empirical distributions were used to obtain and estimate of $p(c^i_{jk}|d)$ as

$$p(c^i_{jk}|d) = \frac{1 + L_{b,d,f}}{l + \sum_{w=0}^{l} L_{b,d,w}}, \tag{1}$$

where f is the index of the frequency bin in which c^i_{jk} falls.

2.5 Calculating the expected number of ancestry switch points per individual

We estimated the total number of switch points per individual $E(S^{(i)}$ as

$$E(S^{(i)} = \sum_{j=1}^{L-1} E(s^{(i)}_{jk}|\bar{c}^{(i)}_{jk}), \tag{2}$$

where the sum runs over all intervals of our non-overlapping map on the 1 Mb scale. This leads to an estimate of 91.98 switch points per individual with standard deviation 33.68. Across the whole sample, this equates to a total of 263,431 switch points. As expected, we found individuals with a genome–wide African ancestry proportion close to 0.5 to harbor the largest number of switch points (results not shown).

We also used our simulation framework to estimate the expected total number of switch points per individual. Based on the simulated admixed population for the 20 independent 6Mb regions we find an average of 0.06659 switch points per Mb per individual. When extrapolating this to the total number of individuals (2864) and the total length of considered sequence (total sequenced spanned by markers after removing unsequenced regions and centromeres, roughly $2.5 \cdot 10^9$ base pairs) we expect to see a total of about 477,000 switch points. While of a similar order of magnitude, this prediction is higher than the number of switch points we estimated based on the data (∼263,000). This discrepancy might reflect a bias downwards in our estimation of the total number of switch points. For instance, our simulations show that the HMM misses about 7% of all isolated switch–points at the 1Mb scale (Supplementary Figure 2). In addition, the implemented empirical Bayes framework might not compensate for all switch–points that were masked by the multiple hits problem (see Main text). However, this is not a major concern as long as we infer relative rates well, which our simulations do suggest (Supplementary Figure 3).

A perhaps larger contributor to the discrepancy in expected versus observed switch points is the assumptions of the simulation model, which we based on a model used widely in the population genetic literature for African-Americans[20, 21, 22]. In particular, this model is fairly simplistic in that it has admixture as a one time event followed by random mating within the admixed population. In contrast, on-going admixture will lead to longer ancestry tracts and less switch points, perhaps explaining the observed discrepancy. Also, the simulations we used are based on African-American parameters. While only a small fraction of our sample, we might observe less switch points in African–Caribbeans because their average African ancestry proportion is closer to 1 than for African-Americans. Resolving the source of this discrepancy will require more accurate models for African-American/Afro-Carribean admixture to be developed.

3 Inference of Ancestry Switch Points

3.1 Source for phased reference panel data

The phased reference panel was downloaded from[9]:
http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2009–02_phaseIII/. We removed a total of 18,379 SNPs that were removed from later releases of unphased genotypes (HapMap3 Release #3), for which different genomic locations have been reported or were monomorphic across the CEU and YRI individuals[9].

3.2 Choices of HMM Parameters

Here we relied on the parameter values proposed by Price *et al.*[25] for African American samples with two exceptions: firstly, motivated by the huge variation in global European ancestry seen in our PC analysis we estimated this proportion for each studied chromosome using `ADMIXTURE`[16] (see above for details). Secondly, we did not use any existing recombination map to guide the inference of ancestry switch points between neighboring SNPs. Instead, we assumed the default recombination rate to depend solely on the physical distance between neighboring SNPs with a default rate of 1 centimorgan per megabase. The values used for the fixed parameters are given in Supplemental Table 4:

We evaluated the robustness of our inference to the choices of the fixed HMM parameters by comparing based on simulations. For our simulated segments we inferred computed the correlation between the underlying true map and maps inferred with different sets of parameters. These results suggests that our results are fairly robust to changes in those parameters and that the parameter values suggested by Price et al.[25] tend to be comparable or better than alternative values (Supplemental Figure 14). We also note that in principle one could infer the maximum likelihood estimates for each parameter in the HMM; however, given the complexity of our model and our sample size it is currently computationally very challenging to do so, as computing the likelihood for a single set of parameters is challenging.

3.3 Switch Points that are Identical by Descent

A concern when compiling inferred switch point probabilities across individuals is to give too much weight to switch points that are shared between individuals due to common ancestry. Individuals with a common ancestor several generations in the past are expected to share some segments of this ancestor's genome, including the switch points that were already present on these segments. While the probability of a switch point to be present on a shared segment is proportional to the local recombination rate, their presence adds statistical noise to estimates of local recombination rates. We thus used our simulation framework (see Section 2) to compute the fraction of switch points that are observed more than once in a sample from an admixed population African-American population (Supplementary Figure 4). While we find a considerable fraction of switch points to be shared between individuals of a random sample of 2864 individuals if the population size is assumed to be small, we found this fraction to be very small for more realistic size of the African-American population. In a sample taken from a population of size 200,000, for instance, this fraction is well below 0.2%. Note that we likely overestimate the fraction of switch points that are in IBD in our sample due to assumptions of the admixture model assumed. In particular, the model assumed is fairly simplistic in that it has admixture as a one time event followed by random mating within the admixed population. In contrast, on-going admixture will lead to less sharing of switch points between indiduals sampled.

4 Inference of Recombination Maps

4.1 Excluded Regions

After map compilation was completed across the whole genome, we excluded various regions from further analysis. Following the construction of the deCODE map[26] we discarded five mega bases on each side of each chromosome. This was additionally motivated by the observation that we generally underestimate recombination rates close to edges (see Supplemental Figure 1). We further excluded all intervals that overlapped with regions which are still unsequenced in the Human reference genome to date, again

following the deCODE map[26]. We finally exclude centromeric regions and the MHC region (as described in the following). All exclude regions are marked with gray bars in Supplemental Figures 10 through 10.

4.1.1 Centromeres

All large unsequenced regions are in centromeres. We found that those regions pose a challenge to infer switch points correctly. Generally, we infer inflated rates just outside such regions, suggesting that the assumed HMM does not work properly in such regions. We thus excluded an additional 5Mb on either side of all unsequenced regions spanning more than 1Mb. Two example is given in Supplemental Figure 9.

4.1.2 MHC

As shown in (Supplemental Figure 9C), the map we inferred shows a much higher recombination rate in the MHC region than previously published maps. We compared the AfAdm map to the crude map obtained from mapping recombination events in 122 quartets from families sampled in this study (see below). In order to understand the observed differences between the methods, it is important to recall a general feature of the assumed HMM model: under this model, each admixed chromosome is viewed as a mosaic of haplotypes from two reference panels. These panels contain 234 and 230 phased haplotypes from the CEU and YRI samples, respectively. The enormous divergence between haplotypes between as well as within populations is hard to capture with so few haplotypes. In such regions, the HMM appears not to be capable of accommodating mismatches by mutations alone, and seems to switch between populations at a much inflated rate. We thus greatly overestimate the recombination rate in the MHC region. We thus excluded the whole MHC region (25 – 35 Mb on chromosome 6) from further comparisons. We note also that the observed divergence pattern at the MHC locus is unmatched in the rest of the genome, so we do not expect similar problems elsewhere in the map.

Inference of Recombination Events from Family Quartets In addition we inferred a crude recombination map of chromosome 6 from quartets (consisting of a father, mother and two full siblings) identified in individuals from the GRAADIii (99 quartets) and GENESTAR (23 quartets) datasets. Recombination events attributed to each parents can be identified by examining the inheritance states of informative SNPs within the two siblings as relates to each individial parent. We apply a method that is similar in spirit to a recent application to full sequencing data[27], though we do not implement a HMM model for inheritance state inference and thus rely on a lower density of SNPs and thus lower resolution of positions of recombination events.

Informative SNPs are those that allow the definite determination of the inheritance state of the alleles passed on from one parent (for clarity of the explanation we refer to the father but the concept can equally be applied to the mother) to both siblings. Siblings can either inherit the same allele from their father (designated here an inheritance state of 1) or opposite alleles (designated an inheritance state of 0). Only if the father is heterozygote at a particular SNP is it possible to definitely identify this inheritance state and SNPs may still be uninformative depending on the genotype of the mother and the two siblings (for example 4 heterozygotes in uninformative as the alleles in the siblings cannot unambiguously be assigned to either a particular parents or a specific chromosome). Further, some genotype configurations violate the rules of Mendelian inheritance and can either be due to genotyping error (in this case we include the presence of potential Copy Number Variation as genotyping error) or more rarely, a mutation in one of the siblings. SNPs showing such configuration along with uninformative SNPs are ignored when subsequently inferring recombination events.

The inheritance state between the siblings at informative SNPs along a chromosome should remain the same for long stretches as either the siblings inherited the same or different chromosome regions as their father (or mother). However, a switch in the inheritance state (from 0 to 1 or 1 to 0) between two siblings suggests a recombination event has occurred in the father [or mother]. The resolution of this recombination event is limited to between the SNPs where the inheritance state appears to switch.

When the inheritance state switches twice within a region spanning a small number of informative SNPs (e.g. 0–1–0) the implication is that two independent closely space recombination events have occurred, which is expected to be rare. Alternative explanations are homologeous gene conversion in the region or genotyping error. Because we find such double events to be rare in most quartets but

very common (up to several thousands) in a few, they appear more likely to be due to be arising in a few samples with increased rates of genotyping error. We thus apply a tolerance factor where after an inheritance state switch, the new state must persist for at least the tolerance factor (e.g. a tolerance factor of 2 mean that the new state must persist for at least 2 SNPs) to be considered a true event. While this method may miss closely spaced events, the detection of false recombination events via genotyping error is more of a concern.

Depending on the specific dataset and quartet, approximately 1/5 of all genotyped SNPs were informative. Therefore the resolution of most recombination events should be approximately 20-30Kb. For the majority of quartets a tolerance factor of 1 was sufficient to identify most likely real recombination events, though in about 10% of cases within the GRADDii set a higher tolerance (we compared results for tolerance of 1 versus 5) was necessary to reduce likely spurious events (for example one quartet was reduced from 1212 to 81 paternal meiosis events using the greater tolerance). Amongst these 10% presumably at least one individual within the quartet had higher than average genotyping error. We thus used a tolerance factor of five for all individuals. As expected[28, 27], we detected more maternal than paternal-specific meiosis events (tol = 5, median maternal = 83, median paternal = 54) at a similar magnitude as observed previously.

We next compiled a recombination map for chromosome 6 with 1Mb intervals matching those used for all other maps. A relative recombination rate is obtained as the fraction of all 588 recombination events detected on this chromosome that fall within a given interval. Since the recombination events were inferred as an interval bounded by two informative SNPs, the relative recombination rate for an interval between positions i and j was computed as

$$q_{ef} = \frac{\sum_{l=1}^{L-1} \alpha_{l,i,j}}{L}, \tag{3}$$

where the sum runs over all inferred recombination events and $\alpha_{e,f,j}$ is the proportional overlap between the interval $[i, j]$ and the interval within which the recombination event l has been confined.

In most parts of chromosome 6 we find a good correspondence between the AfAdm and our quartet map. However, the quartet map does not show any indication of an elevated recombination rate in the MCH region but rather strikingly deviates from the AfAdm map (Supplemental Figure 9D). This corroborates evidence for spurious ancestry switch point inference in the MHC region. We thus excluded the MHC region from all further analysis.

5 Measuring similarity between published maps

When measuring similarity between maps, there is a concern that the similarity between the Hapmap-based LD maps and our AdAfm map may be inflated since both maps are built partly upon the same data (HapMap YRI and CEU haplotype panels) or that there may be bias in similarity towards one panel over another induced by the inference procedure. In sections 5.1-5.3 we describe our investigations of these issues.

As a summary, we do find a correlation of the estimation errors between the LD and admixture based maps, and it appears to be driven by small rate estimates. To protect against this problem in our main analysis, when we compare two maps we chose to use Pearson correlation on a natural scale, which gives less weight to small values and, for scales smaller than 1Mb, we additionally chose to trim the smallest rates in each map. We found this approach does not lead to artificial correlation between LD-based and admixture maps (section 5.1), nor any apparent reference panel bias (section 5.2), nor is it sensitive to outliers in the data (section 5.3).

These inferences are based on the 20 simulated 5Mb segments introduced earlier (see Section 2). While a through evaluation would ideally generate replicates of whole genomes, this is currently not feasible due to computational complexities. However, by choosing a random segment from 20 different chromosomes that total to 100Mb (about the size of chromosome 15), we expect the simulations will still be instructive about the whole genome analysis.

5.1 Accounting for correlated error between LD–based and switch–point based maps at small scales

The LD–based HapMap maps were inferred from the same set of haplotypes we use as reference panels in the current study. Sampling variation in this panel may thus affect both inferred maps, and may in turn inflate correlation statistics when comparing these maps. We thus investigated, based on the simulated segments, if the errors in the rate estimates are correlated between the LD–based and switch–point based maps. To be specific, we computed the estimation error of the interval between markers j and $k = j + 1$ of an estimated map, e_{jk}, as

$$e^{(jk)} = \log_{10}\left(\frac{o_{jk}}{t_{jk}}\right), \tag{4}$$

where o_{jk} is the estimated and t_{jk} the true recombination rate between markers j and k. We first computed the estimation errors of the switch–point based map, $e_{SP}^{(jk)}$, based on our estimator r_{jk}. Next we computed $e_{LD}^{(jk)}$, the estimation error of an 80%/20% average of the LD–based maps from the African and European reference panel. There were no intervals where the true or an inferred rate was zero.

At smaller scales we found strong correlations between $e_{SP}^{(jk)}$ and $e_{LD}^{(jk)}$. At 50Kb, for instance, the correlation was 0.208 and highly significant ($p < 10^{-15}$). Visual inspection (Supplementary Figure 7A) shows that most estimation errors are centered around the origin, with the $e_{LD}^{(jk)}$ being more dispersed than the $e_{SP}^{(jk)}$. The second mode at approximately (-1.5, 0) is likely due to the clustering of bins to equal rates that takes place in the model implemented in LDhat. A clear contributor to to the positive correlation in estimation errors is a fraction of intervals for which both maps underestimate the true rate substantially (lower left quadrant of the plot). Closer inspection revealed that those intervals are among those with the overall smallest estimated rates, a pattern that was confirmed at other scales.

Given that the correlation between the estimation errors $e_{SP}^{(jk)}$ and $e_{LD}^{(jk)}$ was mostly driven by the intervals with smallest rates, we investigated the correlation after trimming all intervals that are among the 20% smallest intervals in one of the maps. The correlation between the trimmed estimation errors is indeed much weakened and only significant at scales below 50Kb (green line in Supplementary Figure 7B). This is not due to a reduction in power due to smaller number of intervals, as random pruning does not lead to the removal of the correlation (dashed lines in Supplementary Figure 7B). In order to have a fair comparison between different maps in our main analysis, we thus use Pearson correlation on a natural scale, which gives little weight to intervals with small rates. In addition, we abstain from making comparisons below 50Kb and trim the lowest 20% intervals of each map for scales below 1Mb, prior to computing the Pearson correlation. For reference, we report the trimmed and untrimmed correlations at several scales in Supplementary Table 2.

5.2 No evidence for bias towards one of the reference panels

We also investigated if the switch–point based map might artificially correlate better with the map obtained from the reference panel representing the majority of the ancestry of the admixed individuals. To investigate the potential presence of such a reference panel bias we made use of our simulations where both reference panels were generated under the same recombination map. We asked if the switch–point based map correlates better with one of the two LD–based maps obtained from the reference panels. A bias of the switch–point based map towards the African reference panel, for instance, would lead to a artificially higher correlation between the switch–point based map and the African LD–map (r_{AFR}) than between the switch–point based map and the European LD–map (r_{EUR}).

Among the 20 simulated segments 12 (at 50Kb and 100Kb) or 9 (at 500Kb) correlated stronger with the African LD–map, which does not indicate a consistent bias ($p > 0.5$ in all cases using a sign test). To increase power we also concatenated the segments and estimated properties on the whole simulated 100Mb. We found r_{EUR} to be stronger than r_{AFR} for all scales except the smallest three (10Kb, 15Kb and 20Kb; though note that the scales are not independent and we thus except similar scales to show similar directions.)

We next assessed the significance of a difference in r_{AFR} and r_{EUR} by bootstrapping map intervals. To be specific, we constructed bootstrapped replicates by choosing map intervals at random with replacement until the number of chosen intervals matched the map length. We then computed r_{AFR} and r_{EUR} based on

these bootstrapped maps. This was repeated 10^5 times for each interval size and the resulting correlations were used to compute a p–value as the fraction of replicates where $r_{AFR} < r_{EUR}$. This procedure was repeated using both the complete maps and those obtained after trimming the intervals with the lowest 20% estimated rates from both maps (Supplementary Figure 7C). We did not find any significant bias towards one of the reference panel, independent of scale.

5.3 No effect of outlier intervals on Pearson correlation

To see if the Pearson correlations computed in this study are affected by outliers, we recomputed the correlations after trimming a fraction of intervals with largest Mahalanobis distance. As shown in Supplementary Figure 7D, when trimming up to 10% of all intervals the Pearson correlations computed are not affected. The results presented in Supplementary Figure 7D are based on the full data. Repeating the same analysis with the the maps obtained after trimming the intervals with the lowest 20% estimated rates from all maps (as mentioned above) revealed very similar results (data not shown).

6 Admixture Mapping of Hotspot Usage

We aimed at finding positions in the genome, at which the ancestry of an individual is associated with the preferential usage of HapMapCEU or HapMapYRI hotspots. We first identified hotspots in the HapMapCEU and HapMapYRI maps as the top 1% 50Kb intervals with largest recombination rates. Next we used our empirical Bayes approach to specifically infer the number of ancestry switch points in those intervals for each individual. Finally, we used the HMM estimated the local ancestry for each individual at every marker and used those to compute the average European ancestry in all 227,490 10Kb intervals throughout the genome. Since we observed a strong correlation between the genome–wide European ancestry proportion and the fraction of ancestry switch points inferred in HapMapCEU hotspots (correlation=0.12, $p < 10^{-7}$) we computed phenotypes as the residuals when regressing (i) the fraction of ancestry switch points found in HapMapCEU hotspots or (ii) the ratio of ancestry switch points found in HapMapCEu versus HapMapYRI hotspots against the genome–wide European ancestry proportion. None of the phenotypes showed a significant hit at a significance threshold of $p < 0.05/227490 \approx 2.2 \cdot 10^{-7}$. All 9 hits with $7.6 \cdot 10^{-7} < p < 1.0 \cdot 10^{-6}$) are found on chromosome 18 at between 6Mb and 9Mb.

7 Computational Details

The software `HAPMIX` implements an HMM model for ancestry switches and outputs the state probabilities for each marker[25]. Here we are, however, interested in getting the posterior probability that the ancestries at two neighboring SNPs are different and thus an ancestry switch occurred between these two SNPs ("switching probabilities"). Thus, we needed to compute novel quantities, and we chose to do so within the same model framework as `HAPMIX`. To accomplish this, we reimplemented the `HAPMIX` model and added the ability to compute the switching probabilities. Before describing our computations (Section 7.2), we review the `HAPMIX` model to introduce their notation and the requisite variables we needed to compute the switching probabilities.

7.1 The `HAPMIX` Model

7.1.1 Haploid Case

Given a haplotypic sample of an admixed individual and n_j haplotypes of the ancestral populations $j = \{1, \ldots, P\}$, the following model is assumed: an admixture event occurred at a single time T generations ago with a fraction μ_j of an admixed haplotype's ancestry drawn from population j, where $\sum_{j=1}^{P} \mu_j = 1$. Recombination events within the admixed population are modeled as a Poisson process along the chromosome at rate T per unit of genetic distance. The segments on both sides of such a recombination event are drawn randomly from population j with probability μ_j. Thus not all of these recombination events will lead to an ancestry switch. It is further assumed that the part of the genome with true ancestry from population j is a mosaic of haplotypes from population j with probability $1 - p_j$ and from any other population with probability p_j. The model thus allows for miscopying at rate p_j. Switches

between haplotypes from the ancestral pool (ancestral recombination events) are assumed to occur as a Poisson process with rate ρ_j. Finally, if a SNP is copied from an ancestral haplotype pool matching the true ancestry, pool j, a mutation may occur with probability θ_j. If a SNP is copied from the non-ancestral pool, a mutation may occur with probability $\theta_{miscopy}$.

Suppose we have S SNPs along a chromosome and the genetic distances $r_2, r_3, \ldots, r_S$ between adjacent pairs of sites. The hidden states in the haploid HMM are denoted by a triplet (ijk) where $i = \{1, \ldots, P\}$ represents the local ancestry, $j = \{1, \ldots, P\}$ the population the chromosome is copied from and $k = \{1, \ldots, n_j\}$ represents the individual from which the chromosomal segment is copied from. j and i may be different due to miscopying. Let $a^s(ijk; IJK)$ be the probability of transitioning from state (ijk) to state (IJK) between adjacent sites $s-1$ and s, then we have the following:

$$a^s(ijk;IJK) = \begin{cases} (1-e^{-r_s T})\mu_I \cdot \dfrac{1-p_I}{n_J} & \text{if } i \neq I \text{ and } J = I \\ (1-e^{-r_s T})\mu_I \cdot \dfrac{p_I}{n_m} & \text{if } i \neq I \text{ and } J \neq I \\ e^{-r_s T} \cdot (1-e^{-r_s \rho_I}) \cdot \dfrac{1-p_I}{n_J} & \text{if } i = I \text{ and } J = I \text{ and} \\ +(1-e^{-r_s T})\mu_I \cdot \dfrac{1-p_I}{n_J} & (j \neq J \text{ or } k \neq K) \\ e^{-r_s T} \cdot e^{-r_s \rho_I} & \text{if } i = I \text{ and } J = I \text{ and} \\ +e^{-r_s T} \cdot (1-e^{-r_s \rho_I}) \cdot \dfrac{1-p_I}{n_J} & j = J \text{ and } k = K \\ +(1-e^{-r_s T})\mu_I \cdot \dfrac{1-p_I}{n_J} & \\ e^{-r_s T} \cdot (1-e^{-r_s \rho_I}) \cdot \dfrac{p_I}{n_J} & \text{if } i = I \text{ and } J \neq I \text{ and} \\ +(1-e^{-r_s T})\mu_I \cdot \dfrac{p_I}{n_J} & (j \neq J \text{ or } k \neq K) \\ e^{-r_s T} \cdot e^{-r_s \rho_I} & \text{if } i = I \text{ and } J \neq I \text{ and} \\ +e^{-r_s T} \cdot (1-e^{r_s \rho_I}) \cdot \dfrac{p_I}{n_J} & j = J \text{ and } k = K \\ +(1-e^{-r_s T})\mu_I \cdot \dfrac{p_I}{n_J} & \end{cases} \tag{5}$$

Let t^s_{jk} denote the allele $p = 0, 1$ of haplotype k in the ancestral pool j at site s, and $b^s_{ijk}(p)$ the probability that the admixed haplotype has allele p at site s given the underlying hidden state. Then

$$b^s_{ijk}(p) = \begin{cases} (1-\theta_i) & \text{if } i = j \text{ and } t^s_{jk} = p \\ \theta_i & \text{if } i = j \text{ and } t^s_{jk} \neq p \\ (1-\theta_{miscopy}) & \text{if } i \neq j \text{ and } t^s_{jk} = p \\ \theta_{miscopy} & \text{if } i \neq j \text{ and } t^s_{jk} \neq p \end{cases} \tag{6}$$

7.1.2 Diploid Case

In the diploid case we use the same notation as above. However, the hidden state is a composite state for both underlying haplotypes, denoted by $(ijklmn)$. All parameters are assumed to be the same for both chromosomes of a diploid individual, including the time since admixture T and the proportion of ancestry from population j μ_j.

Let $a^s(ijklmn; IJKLMN)$ be the probability of transitioning between adjacent sites s and $(s+1)$ from state $(ijklmn)$ to state $(IJKLMN)$. Since the transitions of both chromosomes are independent, the transition probabilities are products of those given in (5):

$$a^s(ijklmn;IJKLMN) = a^s(ijk;IJK) \cdot a^s(lmn;LMN) \tag{7}$$

The emission probabilities, however, are different since the phase of the admixed haplotypes is unknown. Let $b^s_{ijklmn}(g)$ denote the probability that the admixed genotype is of type $g = 0, 1, 2$ at site s

given the underlying hidden state. Using the emission probabilities defined above we getting

$$b^s_{ijklmn}(0) = b^s_{ijk}(0) \cdot b^s_{lmn}(0) \tag{8}$$
$$b^s_{ijklmn}(1) = b^s_{ijk}(0) \cdot b^s_{lmn}(1) + b^s_{ijk}(1) \cdot b^s_{lmn}(0) \tag{9}$$
$$b^s_{ijklmn}(2) = b^s_{ijk}(1) \cdot b^s_{lmn}(1) \tag{10}$$

7.1.3 Speed-up through collapsing

Despite considerable tweaking of the code, the implementation of the diploid case resulted in very large computation times. `HAPMIX` use a collapsing method to speed up the computations, as outlined in the supplementary material of their paper[25]. The main idea is to collapse similar haplotypes within a window of size d and thus to reduce the state space considerably. We adopt this collapsing scheme and outline it in the following.

The parameter d is defined in units of genetic distance. At each site s, all haplotypes similar within a windows of size d centered at s are collapsed into a single type. At site s in population j, we define $T^k_j(s)$ to be the number of haplotypes collapsing to type k and n^s_j the total number of types. Since the types are defined independently for each site, a switch in type between site $s-1$ and s may happen with non-zero probability, even in the absence of recombination. Given $T^{kK}_j(s)$, the number of switches from type k at site $s-1$ to type K at site s in population j, the forward transition probability for such a switch is set to $T^{kK}_j(s)/T^k_j(s-1)$.

For the haploid case, the forward transition probability $a^s(ijk; IJK)$ from state (ijk) to state (IJK) between adjacent sites $s-1$ and s, where k and n denote collapsing types, are as follows (these solutions are slightly different from the ones published by Price *et al.*[25] due to obvious typos in the latter):

$$a^s(ijk;IJK) = \begin{cases} (1-e^{-r_s T})\mu_I \cdot (1-p_I) \cdot \dfrac{T^K_J(s)}{n_J} & \text{if } i \neq I \text{ and } J = I \quad \text{(a)} \\ (1-e^{-r_s T})\mu_I \cdot p_I \cdot \dfrac{T^K_J(s)}{n_J} & \text{if } i \neq I \text{ and } J \neq I \quad \text{(b)} \\ e^{-r_s T} \cdot (1-e^{-r_s \rho_I}) \cdot (1-p_I) \cdot \dfrac{T^K_J(s)}{n_J} & \text{if } i = I \text{ and } J = I \text{ and } j \neq J \quad \text{(c)} \\ \quad +(1-e^{-r_s T})\mu_I \cdot (1-p_I) \cdot \dfrac{T^K_J(s)}{n_J} & \\ e^{-r_s T} \cdot e^{-r_s \rho_I} \cdot \dfrac{T^{kK}_J(s)}{T^k_J(s-1)} & \text{if } i = I \text{ and } J = I \text{ and } j = J \quad \text{(d)} \\ \quad +e^{-r_s T} \cdot (1-e^{-r_s \rho_I}) \cdot (1-p_I) \cdot \dfrac{T^K_J(s)}{n_J} & \\ \quad +(1-e^{-r_s T})\mu_I \cdot (1-p_I) \cdot \dfrac{T^K_J(s)}{n_J} & \\ e^{-r_s T} \cdot (1-e^{-r_s \rho_I}) \cdot p_I \cdot \dfrac{T^K_J(s)}{n_J} & \text{if } i = I \text{ and } J \neq I \text{ and } j \neq J \quad \text{(e)} \\ \quad +(1-e^{-r_s})\mu_I \cdot p_I \cdot \dfrac{T^K_J(s)}{n_J} & \\ e^{-r_s T} \cdot e^{-r_s \rho_I} \cdot \dfrac{T^{kK}_m(s)}{T^k_J(s-1)} & \text{if } i = I \text{ and } J \neq I \text{ and } j = J \quad \text{(f)} \\ \quad +e^{-r_s T} \cdot (1-e^{-r_s \rho_I}) \cdot p_I \cdot \dfrac{T^K_J(s)}{n_J} & \\ \quad +(1-e^{-r_s T})\mu_I \cdot p_I \cdot \dfrac{T^K_J(s)}{n_J} & \end{cases} \tag{11}$$

We also constrained the total number of types by reducing the window size d until the total number of types in that window does not exceed a predefined constant n_{max}. The choice of n_{max} has a huge impact on the required computational effort because some of the computations needed scale with the square of

the number of types. Due to limits in available computational resources and based on an exploration of different values (see Supplemental Figure 14) we concluded to use $n_{max} = 25$.

7.1.4 Forward and Backward Variables

Denoting the forward variable at site s for state $(ijklmn)$ by $\alpha_s(ijklmn)$ and the observed genotype at site s by g_s, the induction of the forward variable is given by

$$\alpha_s(IJKLMN) = \left[\sum_{i=0}^{P} \sum_{j=0}^{P} \sum_{k=0}^{n_j^{s-1}} \sum_{l=0}^{P} \sum_{m=0}^{P} \sum_{n=0}^{n_m^{s-1}} \alpha_{s-1}(ijklmn) \cdot a^s(ijklmn; IJKLMN) \right] \cdot b^s_{IJKLMN}(g_s). \tag{12}$$

The initialization of the forward variable is given by

$$\alpha_1(IJKLMN) = \begin{cases} \mu_I(1-p_I)\frac{T_J^K(1)}{n_J} \cdot \mu_L(1-p_L)\frac{T_M^N(1)}{n_M} \cdot b^s_{IJKLMN}(g_1) & \text{if } J = I \text{ and } M = L \\ \mu_I \cdot p_I\frac{T_J^K(1)}{n_J} \cdot \mu_L(1-p_L)\frac{T_M^N(1)}{n_M} \cdot b^s_{IJKLMN}(g_1) & \text{if } J \neq I \text{ and } M = L \\ \mu_I(1-p_I)\frac{T_J^K(1)}{n_J} \cdot \mu_L \cdot p_L\frac{T_M^N(1)}{n_M} \cdot b^s_{IJKLMN}(g_1) & \text{if } J = I \text{ and } M \neq L \\ \mu_I \cdot p_I\frac{T_J^K(1)}{n_J} \cdot \mu_L \cdot p_L\frac{T_M^N(1)}{n_M} \cdot b^s_{IJKLMN}(g_1) & \text{if } J \neq I \text{ and } M \neq L \end{cases} \tag{13}$$

The induction of the backward variable for the transition fro sites $s+1$ to s is given by

$$\beta_s(IJKLMN) \tag{14}$$

$$= \sum_{i=0}^{P} \sum_{j=0}^{P} \sum_{k=0}^{n_j^{s+1}} \sum_{l=0}^{P} \sum_{m=0}^{P} \sum_{n=0}^{n_m^{s+1}} \beta_{s+1}(ijklmn) \cdot a^{s+1}(IJK; ijk) \cdot a^{s+1}(LMN; lmn) \cdot b^{s+1}_{ijklmn}(g_{s+1}) \tag{15}$$

The initialization of the backward variable is given by

$$\beta_1(IJKLMN) = 1. \tag{16}$$

7.2 Computing Switching Probabilities

We begin by describing a general solution and then show the equations for the `HAPMIX` model.

Let $\mathbf{S} = \{S_1, S_2, \ldots, S_N\}$ denote the hidden states of a HMM and the state at time t as q_t, where $1 \leq t \leq T$. We further denote the state transition probability distribution as $\mathbf{A} = \{a_{ij}\}$ where

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i), \quad 1 \leq i, j \leq N \tag{17}$$

We denote the distinct possible observations as $\mathbf{V} = \{v_1, v_2, \ldots, v_M\}$ and the sequence of observations as $\mathbf{O} = O_1 O_2 \ldots O_T$, where $1 \leq t \leq T$. Further, let $\mathbf{B} = \{b_j(k)\}$ denote the observation probability distribution, where

$$b_j(k) = P(v_k \text{ at } t | q_t = S_j), \quad 1 \leq j \leq N, \quad 1 \leq k \leq M \tag{18}$$

Finally, let $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ indicate the complete parameter set of the model, where $\boldsymbol{\pi} = \{\pi_j\}$ describes the initial state distribution with

$$\pi_j = P(q_1 = S_j), \quad 1 \leq j \leq N \tag{19}$$

We are interested in the probability that the state at time t is different from the state at time $t-1$, which is given by

$$P(q_t \neq q_{t-1} | \mathbf{O}, \lambda) = 1 - P(q_t = q_{t-1} | \mathbf{O}, \lambda) \tag{20}$$

$$= 1 - \sum_{j=1}^{N} P(q_t = S_j | q_{t-1} = S_j, \mathbf{O}, \lambda) \cdot P(q_{t-1} = S_j | \mathbf{O}, \lambda) \tag{21}$$

Let the forward variable $\alpha_t(i)$ be given as

$$\alpha_t(i) = P(O_1 \dots O_t, q_t = S_i|\lambda), \tag{22}$$

and the backward variable $\beta_t(i)$ as

$$\beta_t(i) = P(O_{t+1} \dots O_T|q_t = S_i, \lambda). \tag{23}$$

It is not hard to show that

$$P(q_t = S_j|q_{t-1} = S_j, \mathbf{O}, \lambda) = \frac{\beta_t(j)}{\beta_{t-1}(j)} \cdot b_j(O_t) \cdot a_{jj} \tag{24}$$

and

$$P(q_{t-1} = S_j|\mathbf{O}, \lambda) = \frac{\alpha_{t-1}(j) \cdot \beta_{t-1}(j)}{\sum_{k=1}^{N} \alpha_{t-1}(k) \cdot \beta_{t-1}(k)}. \tag{25}$$

We can thus rewrite (21) as

$$P(q_t \neq q_{t-1}|\mathbf{O}, \lambda) = 1 - \frac{1}{\sum_{k=1}^{N} \alpha_{t-1}(k) \cdot \beta_{t-1}(k)} \cdot \sum_{j=1}^{N} \beta_t(j) \cdot b_j(O_t) \cdot a_{jj} \cdot \alpha_{t-1}(j). \tag{26}$$

In our case we are interested in the probability that the ancestry state of one haplotype is different between sites s and $s+1$. Using the notation introduced above, we can write

$$\begin{aligned} &P(I \neq i|\mathbf{O}, \lambda) \\ &= 1 - \frac{\varepsilon_s}{\sum_{i=0}^{P}\sum_{j=0}^{P}\sum_{k=0}^{n_j^s}\sum_{l=0}^{P}\sum_{m=0}^{P}\sum_{n=0}^{n_m^s} \hat{\alpha}_s(ijklmn) \cdot \hat{\beta}_s(ijklmn)} \\ &\quad \cdot \sum_{i=0}^{P}\sum_{J=0}^{P}\sum_{K=0}^{n_J^{s+1}}\sum_{L=0}^{P}\sum_{M=0}^{P}\sum_{N=0}^{n_M^{s+1}} \hat{\beta}_{s+1}(iJKLMN) \cdot b_{iJKLMN}^{s+1}(g_{s+1}) \qquad (27) \\ &\quad \cdot \sum_{j=0}^{P}\sum_{k=0}^{n_j^s} a^{s+1}(ijk; IJK) \cdot \sum_{l=0}^{P}\sum_{m=0}^{P}\sum_{n=0}^{n_m^s} \cdot a^{s+1}(lmn; LMN) \cdot \hat{\alpha}_s(ijklmn) \end{aligned}$$

(28)

We then computed the expected number of changes in ancestry $c_{jk}^{(i)}$ between sites s and $s+1$ as

$$c_{jk}^{(i)} = P(I \neq i|\mathbf{O}, \lambda) + P(L \neq l|\mathbf{O}, \lambda). \tag{29}$$

Part II
Supplementary Tables

Sample	# of SNPs after QC	n	Sample locations	Genotyping rate
GeneSTAR	875,418	415	Baltimore, MD	0.999
GENOA	659,781	583	Jackson, MS	0.995
GRAADi	570,293	935	Baltimore, MD; Washington, DC	0.997
GRAADii	535,285	299	Cave Hill, Barbados	0.995
SARP/CSGA	869,755	632	Winston-Salem, NC; misc. others	0.998
Total		2864		

Supplementary Table 1. Summary of number of SNPs and number of individuals per sample within the project. The samples sizes (n) refer to the number of unrelated individuals used to infer recombination events.

Scale		HapMapCEU	HapMapYRI	HapMap8020	deCODE	AfAdm
3Mb	HapMap CEU	1	0.953	0.97	0.958	0.908
	HapMap YRI	0.935	1	0.998	0.957	0.916
	HapMap 80/20	0.959	0.997	1	0.965	0.921
	deCODE	0.941	0.941	0.951	1	0.921
	AfAdm	0.874	0.881	0.890	0.894	1
1Mb	HapMap CEU	1	0.922	0.951	0.939	0.900
	HapMap YRI	0.892	1	0.997	0.934	0.922
	HapMap 80/20	0.931	0.995	1	0.948	0.929
	deCODE	0.918	0.912	0.931	1	0.924
	AfAdm	0.865	0.893	0.904	0.897	1
500Kb	HapMap CEU	1	0.89	0.932	0.917	0.869
	HapMap YRI	0.85	1	0.995	0.907	0.903
	HapMap 80/20	0.904	0.993	1	0.927	0.913
	deCODE	0.891	0.879	0.904	1	0.901
	AfAdm	0.819	0.869	0.879	0.869	1
100Kb	HapMap CEU	1	0.807	0.882	0.849	0.727
	HapMap YRI	0.759	1	0.99	0.816	0.795
	HapMap 80/20	0.854	0.987	1	0.853	0.808
	deCODE	0.816	0.772	0.820	1	0.780
	AfAdm	0.669	0.752	0.767	0.732	1
50Kb	HapMap CEU	1	0.774	0.865	0.816	0.662
	HapMap YRI	0.738	1	0.987	0.772	0.740
	HapMap 80/20	0.844	0.985	1	0.817	0.753
	deCODE	0.789	0.734	0.788	1	0.711
	AfAdm	0.611	0.697	0.712	0.666	1
10Kb	HapMap CEU	1	0.718	0.838	0.701	0.505
	HapMap YRI	0.702	1	0.981	0.645	0.586
	HapMap 80/20	0.829	0.980	1	0.699	0.599
	deCODE	0.695	0.636	0.691	1	0.553
	AfAdm	0.476	0.552	0.566	0.537	1

Supplementary Table 2. Pearson's correlations between published recombination maps. Above the diagonal Pearson's correlation coefficients are reported for the full data. Below the diagonal Pearson's correlation coefficients after trimming the intervals with lowest 20% estimated rates from both maps. Due to correlated estimation errors at small scales between the LD–based and switch–point based maps, we chose to ignore correlations below 50Kb and to use the trimmed correlations below 1Mb.

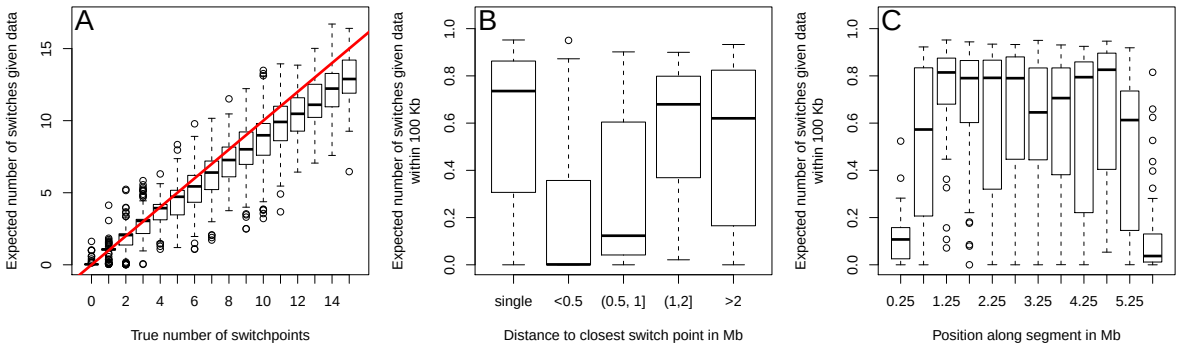
Chr	Position [Mb]	Difference [cM][a]	Structural variations
1	15.9 - 17.7	-1.25	
16	53.0 - 55.3	-1.12	100Kb inversion[29, 30, 31]
5	13.1 - 14.7	1.08	
3	120.9 - 122.8	-1.02	
16	77.1 - 78.4	0.94	
10	117.5 - 118.7	0.87	
2	8.9 - 10.1	-0.87	

Supplementary Table 3. Seven regions where visual inspection reveals the difference between the AfAdm map and the 80%/20% map is not due to a feature of the AfAdm map. In these intervals the AfAdm rate is much like the HapMapCEU and deCODE values and the HapMapYRI rate is an outlier. We suspect these intervals arise due to sampling error in the HapMapYRI map or potential population-specific recombination features or selection events. We suspect similar outlier intervals exist in the HapMapCEU map, but because we are comparing the AfAdm map to a 80%/20% average of the HapMapYRI/HapMapCEU maps, one will preferentially detect the HapMapYRI outlier intervals. See Table in main text for all other regions. The reported structural variations are validated in surveys of structural variations in random samples of European or African individuals and not further than 1Mb away from the focus regions. In addition, CNVs had to be at least 500Kb to be included. The intervals prior to collapsing are marked in Supplementary Figures 10 through 10.
[a] Largest difference per region. Negative values imply lower rates in the HapMapYRI map.

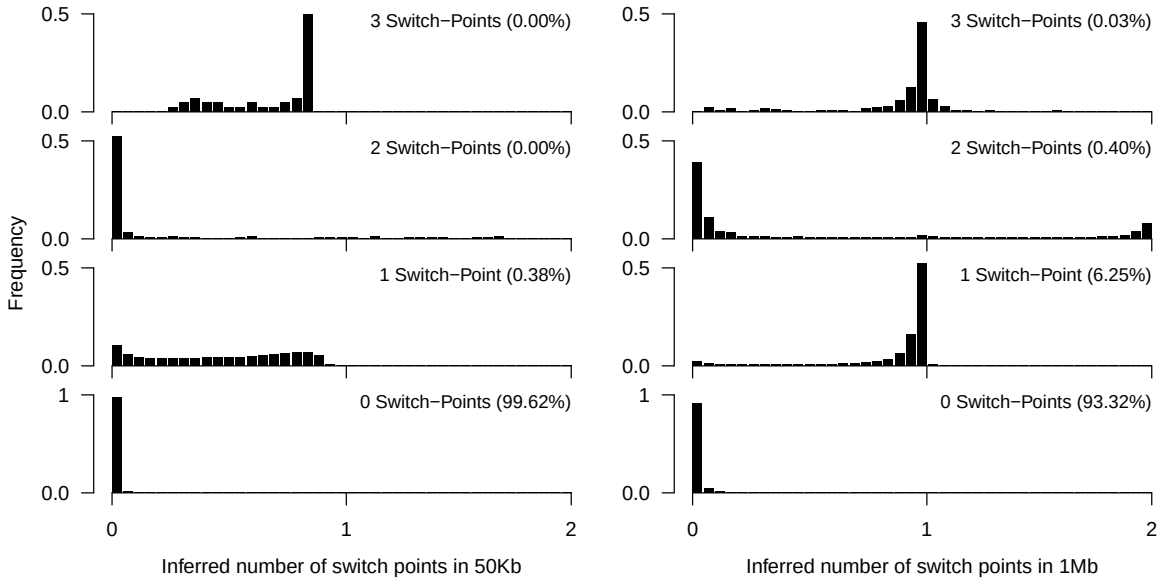
HMM parameter	used value
Default recombination rate	1 cm/Mb
Time since admixture	7 generations
Proportion of African Ancestry	individually estimated
European population recombination rate [a]	60000/117 = 512.08
African population recombination rate [a]	90000/115 = 782.06
European population mutation rate [a]	0.2/0.2+117) = 0.0017
African population mutation rate [a]	0.2/(0.2+115) = 0.0015
Miscopy rate	0.05
Miscopy mutation rate	0.01
Collapsing distance	0.2 cM
Maximum of allowed haplotypes when collapsing	25

Supplementary Table 4. Used values of fixed HMM parameters. See text and Price et al.[25] for details. [a] These parameters are scaled by the number of individuals in the reference panel.

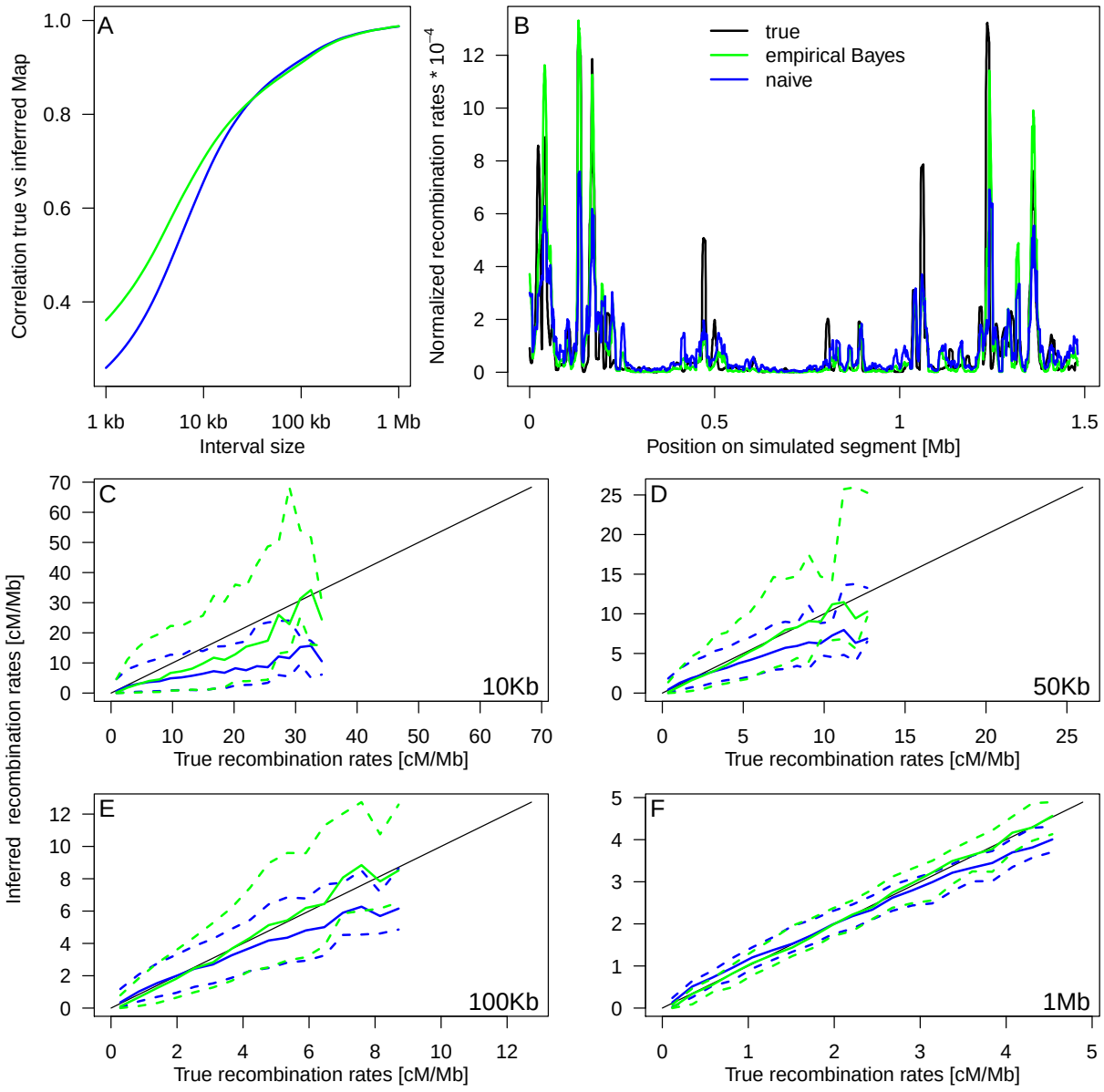
Part III
Supplementary Figures



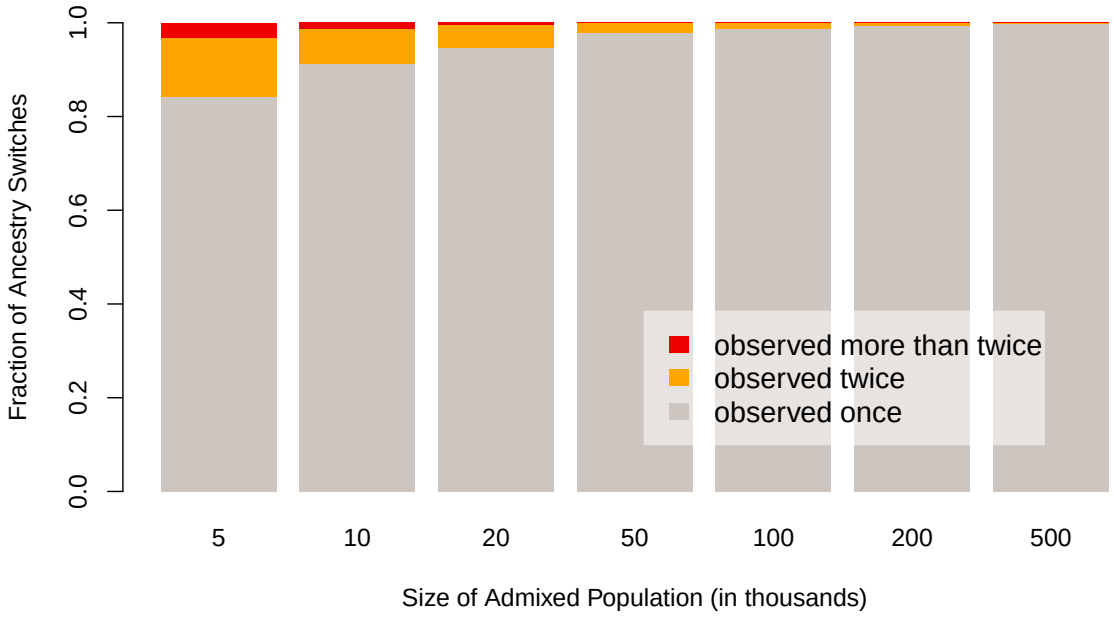
Supplementary Fig. 1. Accuracy of switch point inference. The 20 simulated regions of 6 Mb each allowed us to asses the accuracy of inferring ancestry switches. A) Comparison of the true number of ancestry switches per individual against the inferred number of ancestry switches computed as the expected number of switches given the data $\sum_j \bar{c}_{jk}^{(i)}$ across all segments. B) The estimator $\sum_j \bar{c}_{jk}^{(i)}$ is interfering with nearby switch points: in close proximity to another switch point the estimator is strongly biased towards too small values. We refer to this issue as the "multiple hits" problem and address it using an Empirical Bayes approach. C) Demonstration of the edge effect near the boundaries of the 6Mb segment. The inference of switch points in individual with a large number of true switch points is more likely to be affected by both effects, hence the underestimation in those individuals.



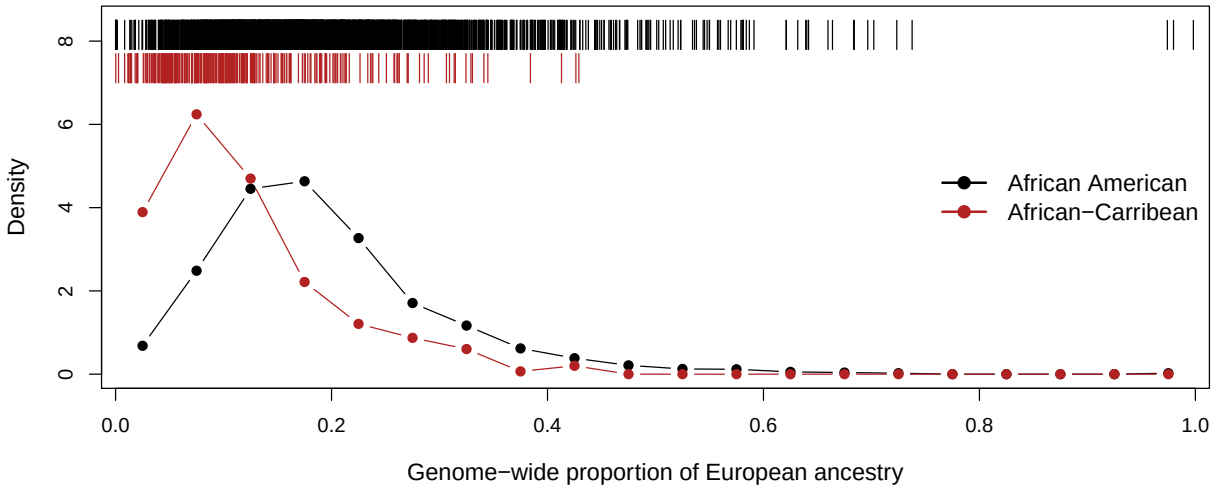
Supplementary Fig. 2. Simulation-based assessment of the performance of the recombination inference approach. The impact of multiple switch points within intervals. Each panel shows the distribution of the inferred number of switch points $c_{jk}^{(i)}$ as function of the simulated number of switch points in intervals. Left column is for 50Kb intervals; right column for 1Mb intervals. The fraction of randomly chosen intervals with zero or more switch points in our simulations is given as percentages, suggesting that intervals with multiple switch points are generally rare. For instance, among all 1Mb windows with at least one switch point, only 6.0% harbor two and 0.4% more than two switch points.



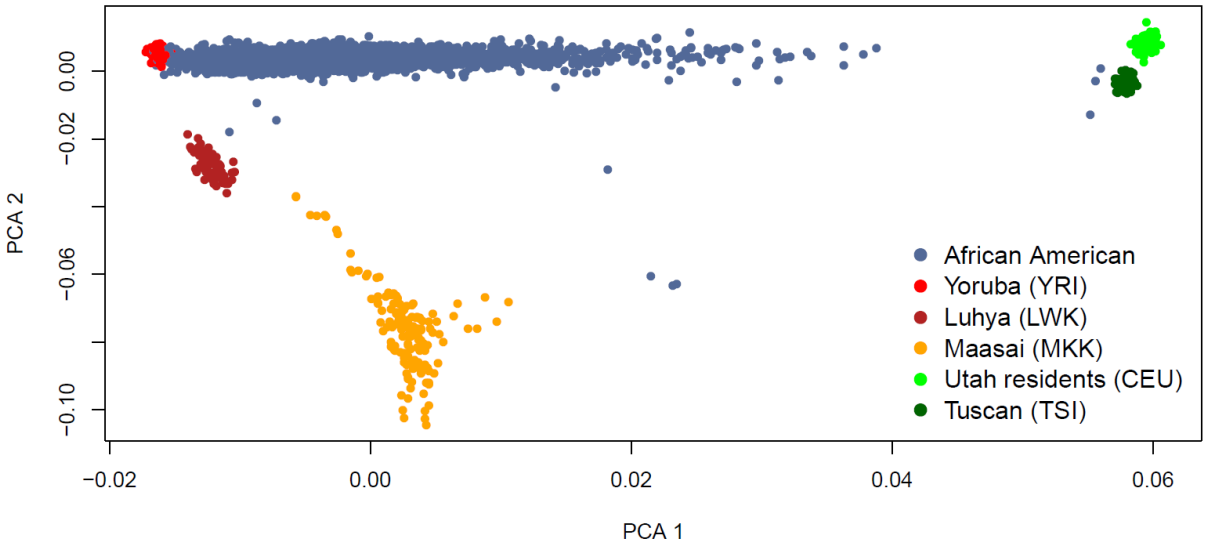
Supplementary Fig. 3. Comparison of naive estimator to empirical Bayes on simulated data. (A) Pearson correlations computed between the true underlying map used to generate the simulated data and the map inferred either with the naive (blue) or empirical Bayes (green) approach. Especially for small interval sizes we see a substantial improvement when using the empirical Bayes as compared to the naive approach. (B) True, naive and empirical Bayes 10Kb–map for an example region of 1.5Mb takes from a simulated segment. "Multiple hits" are most likely to occur in regions with very large recombination rates. The naive approach is therefore generally underestimating those regions. For some of those regions the empirical Bayes approach is able to effectively pool information across individuals, resulting in much better estimates. For this particular region the Spearman correlation increases from 0.70 to 0.80 when using the Empirical Bayes approach (the Pearson correlation is increasing from 0.71 to 0.80). (C) through (F) Comparison of normalized recombination rates between true and inferred maps across all simulated segments for different interval sizes. We find a good agreement between the empirical Bayes rates and the true rates for interval sizes of 5Kb and above.



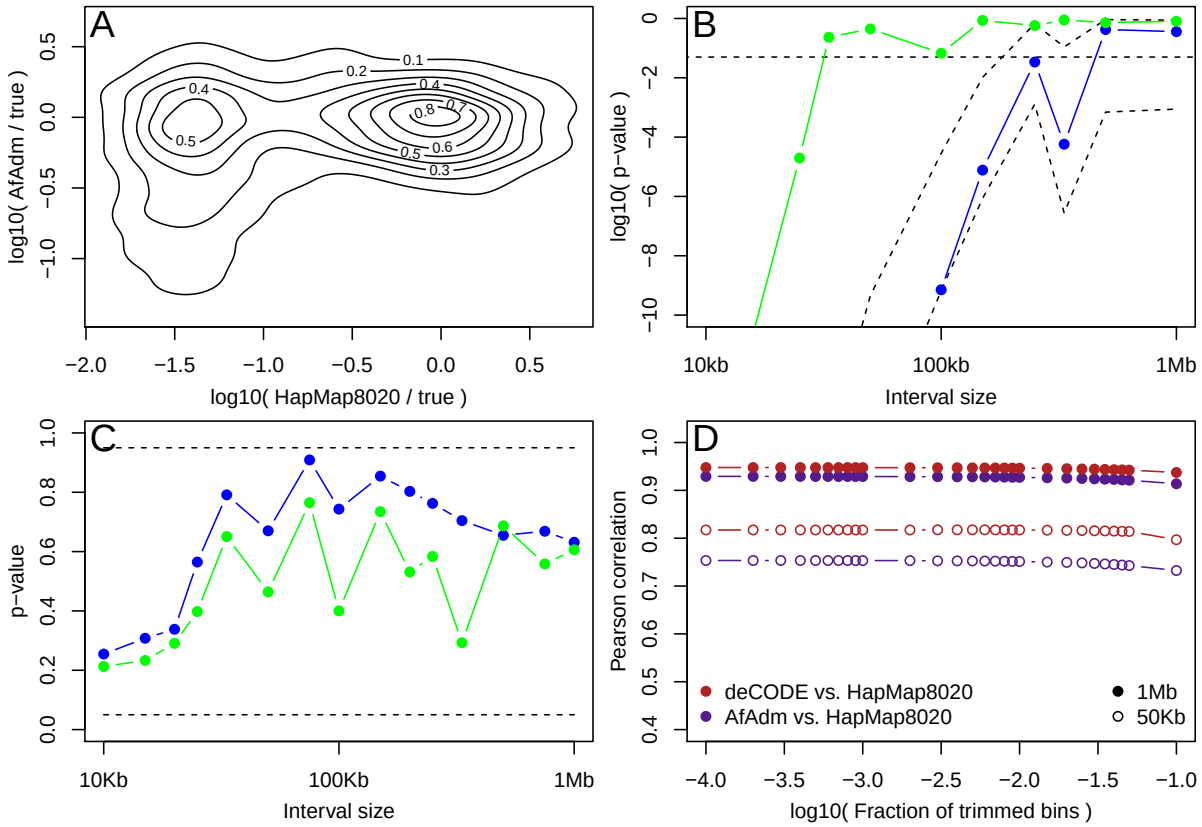
Supplementary Fig. 4. Number of times a switch point is expected to be observed in a sample of 2864 African Americans We simulated an admixture process of seven generations with an European contribution of 20% in the first generation for different sizes of the admixed population (see Section 2).The fraction of switch points observed multiple times is more than 11% if the sample is taken from an admixed population only 10,000 individuals. However, for more realistic size of the African-American population, this fraction is very small. In a sample taken from a population of size 200,000 this fraction is well below 0.2%.



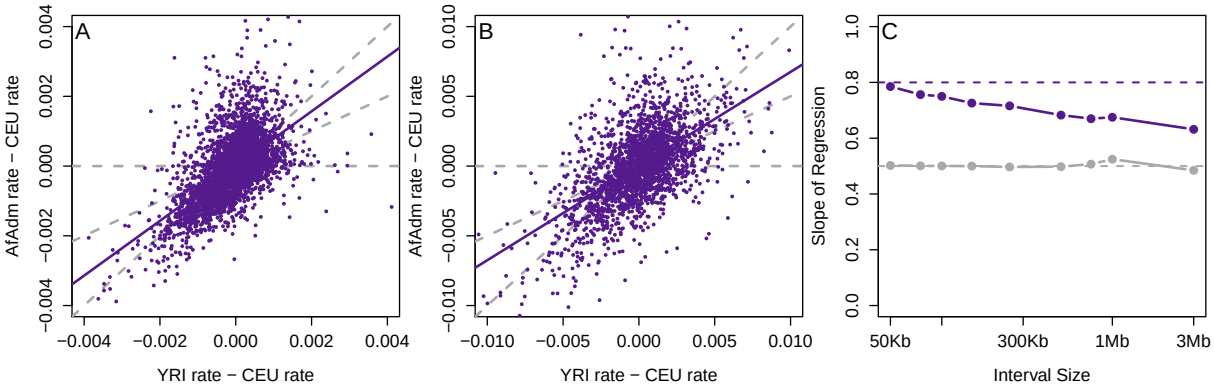
Supplementary Fig. 5. Proportion of genome–wide European ancestry. Based on the individual HMM results we can obtain an estimate of the proportion of the genome that is of European ancestry per individual. The mean Europeran ancestry proportion among African American individuals is 19.4%, compared to 11.7% among African-Carribeans. Above the histograms, each individual is represented by a tick mark.



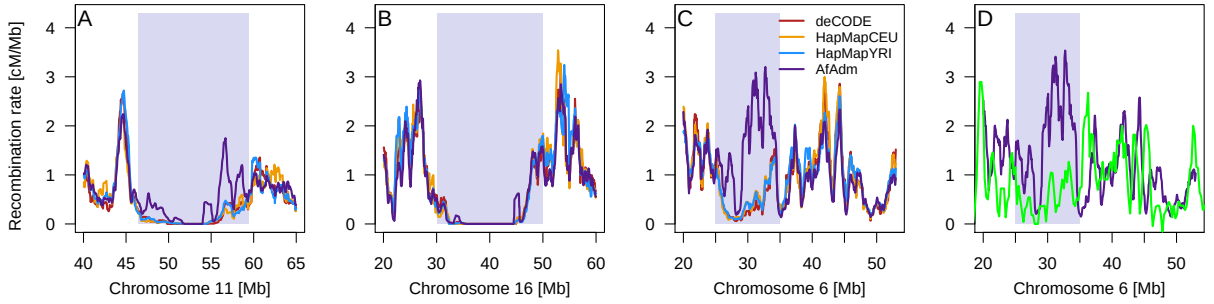
Supplementary Fig. 6. PCA of samples along with HapMap populations. We performed a principal component analysis on all our samples together with several HapMap samples: the two samples of European origin CEU (Utah) and TSI (Tuscany); and the three African samples YRI (Yoruba from Nigeria), LWK (Luhya from Kenya) and MKK (Maasai from Kenya). The analysis was restricted to SNPs in common to all studies.



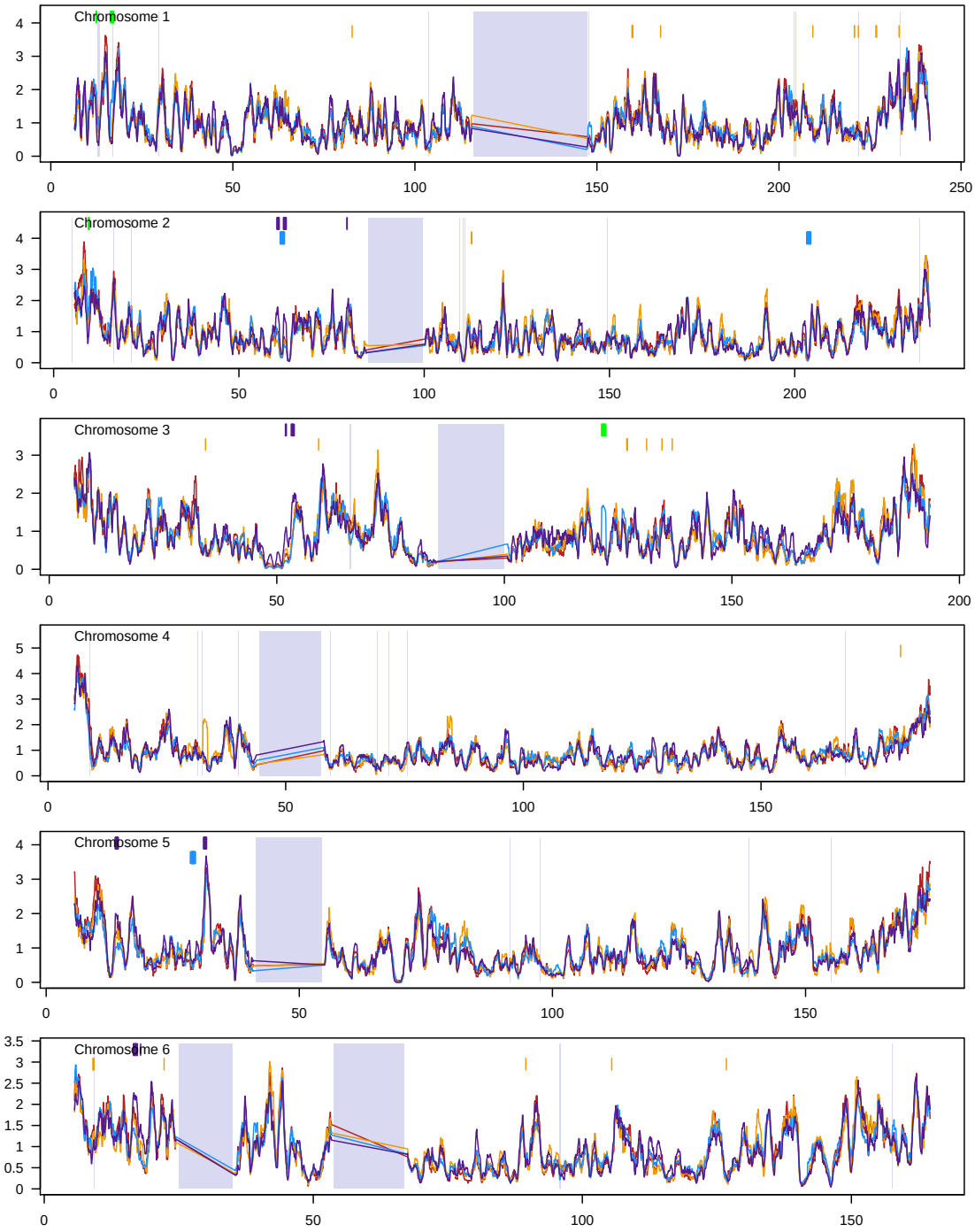
Supplementary Fig. 7. Robustness of correlation statistics. (A) Joint distribution of the estimation error of the switch–point based map $e_{SP}^{(jk)}$ and the LD–based 80/20 map $e_{LD}^{(jk)}$ at 50Kb. We find $e_{LD}^{(jk)}$ to be more dispersed than $e_{SP}^{(jk)}$, probably due to the rjMCMC framework implemented in LDhat. The distribution is symmetric around $e_{SP}^{(jk)} = 0$, except for a fraction intervals for which both maps underestimate the true rate substantially. (B) The significance of the correlation between e_{jk}^{SP} and $e_{LD}^{(jk)}$ for the whole maps (blue) and those obtained after trimming the intervals with the lowest 20% estimated rates from both maps (green). The trimming greatly reduces the correlation between the estimation errors. The 2.5% and 97.5% quantiles of the p–value obtained after trimming the same number of intervals at random are given as dashed lines. (C) Significance of differences between the correlation of the switch–point based map to the LD–based map obtained from either the African (r_{AFR}) or European (r_{EUR}) reference panel. The p–value was assessed as the fraction of bootstrapped replicates where $r_{AFR} < r_{EUR}$. The dashed lines mark the 5% and 95% significance levels. (D) The Pearson correlation between the deCODE and the HapMap8020 (red) and between the AfAdm and the HapMap8020 maps (purple) after trimming a fraction (10^{-4} to 10^{-1}) of the bins based on the Mahalanobis distance. Trimming up to 10% of the data did not influence the obtained correlations, independent of scale (50Kb or 1Mb).



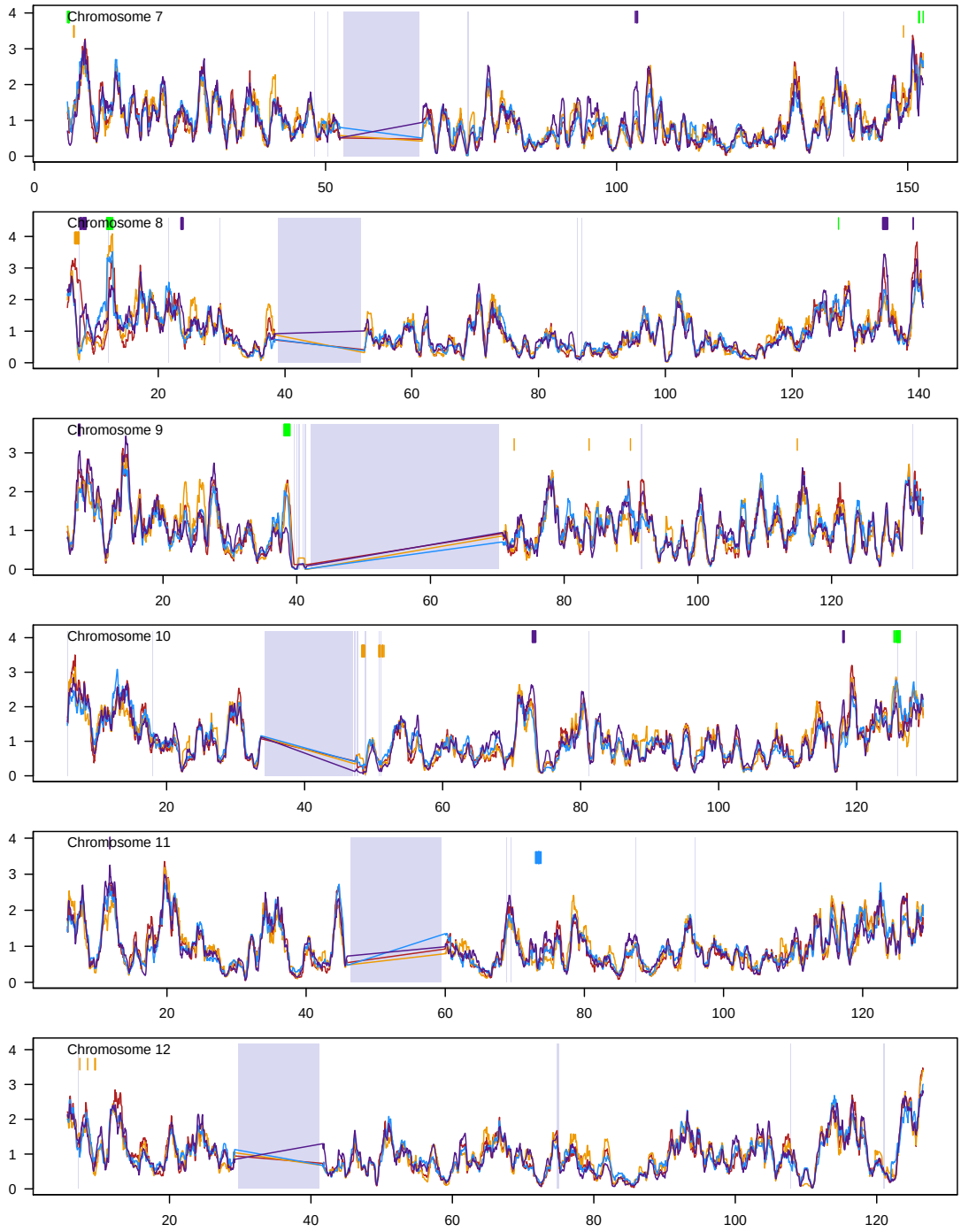
Supplementary Fig. 8. Estimates of the proportional weighting of the HapMapYRI and HapMapCEU that best predicts the AfAdm map. We fit a model in which the AfAdm map is a convex linear combination of the HapMapYRI and HapMapCEU maps: AfAdm= a∗HapMapYRI + $(1-a)$∗HapMapCEU. To estimate a, we use a linear regression of (AfAdm-HapMapCEU) on (HapMapYRI-HapMapCEU). We show visualization of such an analysis at 50Kb (A) and 1Mb (B). The regression line is shown as a solid, purple line and the dashed, gray lines indicate slopes of 0, 0.5 and 1. R^2 was 0.38 at 50Kb and 0.34 at 1Mb. (C) The estimated proportional weight is shown as a function of map interval size (on a $\log_{10}$ scale) for the AfAdm map (purple) and for simulations under a "null" scenario where all three maps are identical (gray). The dashed horizontal purple line is the expectation for an 80%/20% proportional weighting; the dashed horizontal gray line is expected for a 50%/50% proportional weighting.



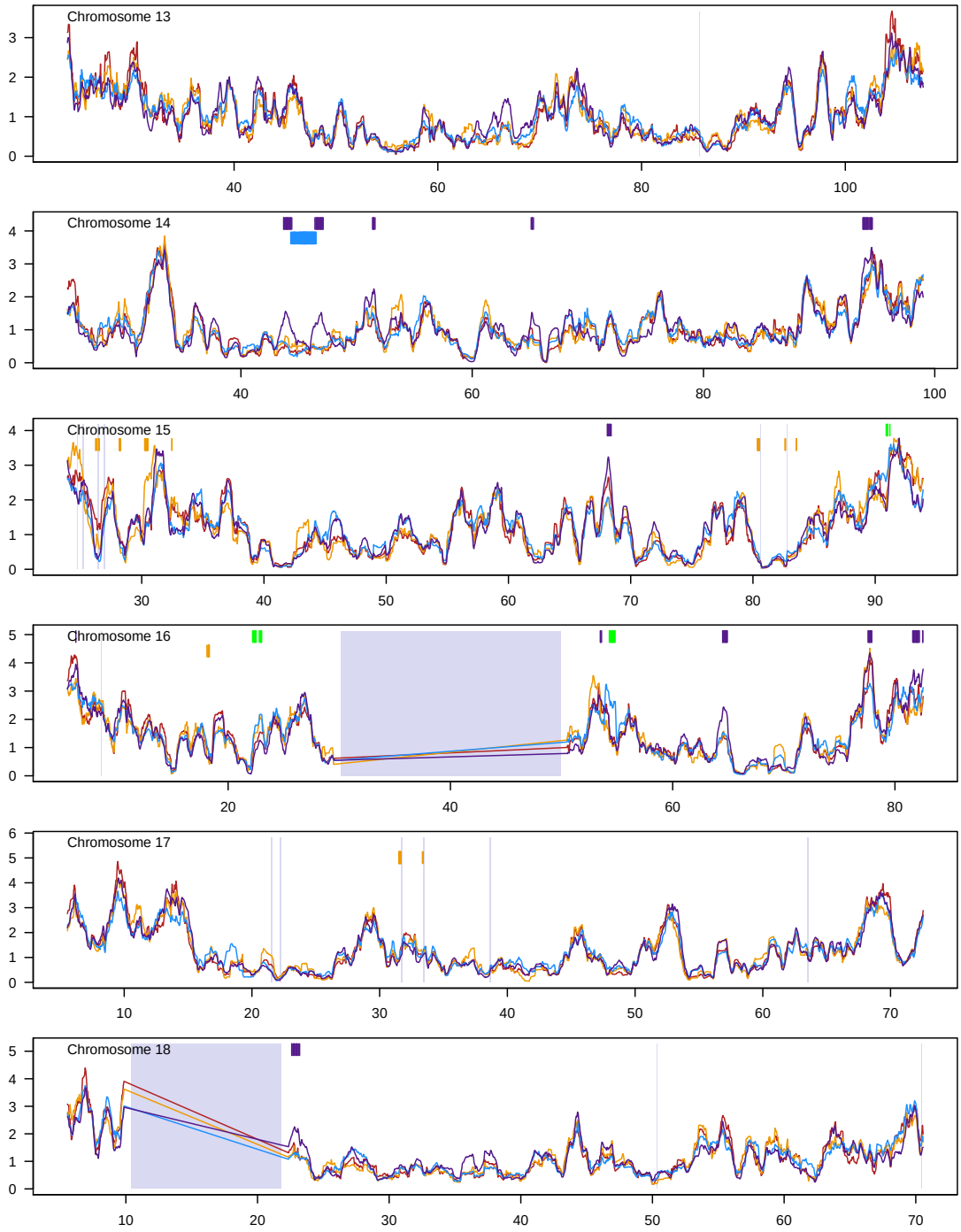
Supplementary Fig. 9. Regions for which switch point inference proved difficult. (A) and (B) Examples of two centromeres. The gray areas mark the regions we excluded for downstream analysis. (C) MHC region. The gray area marks the region we chose to exclude for downstream analysis. (D) Comparing the recombination map based on admixture (purple) with the crude map obtained from quartets (green, see Supplemental Note). The quartet map does not show elevated recombination rates in the MHC region.



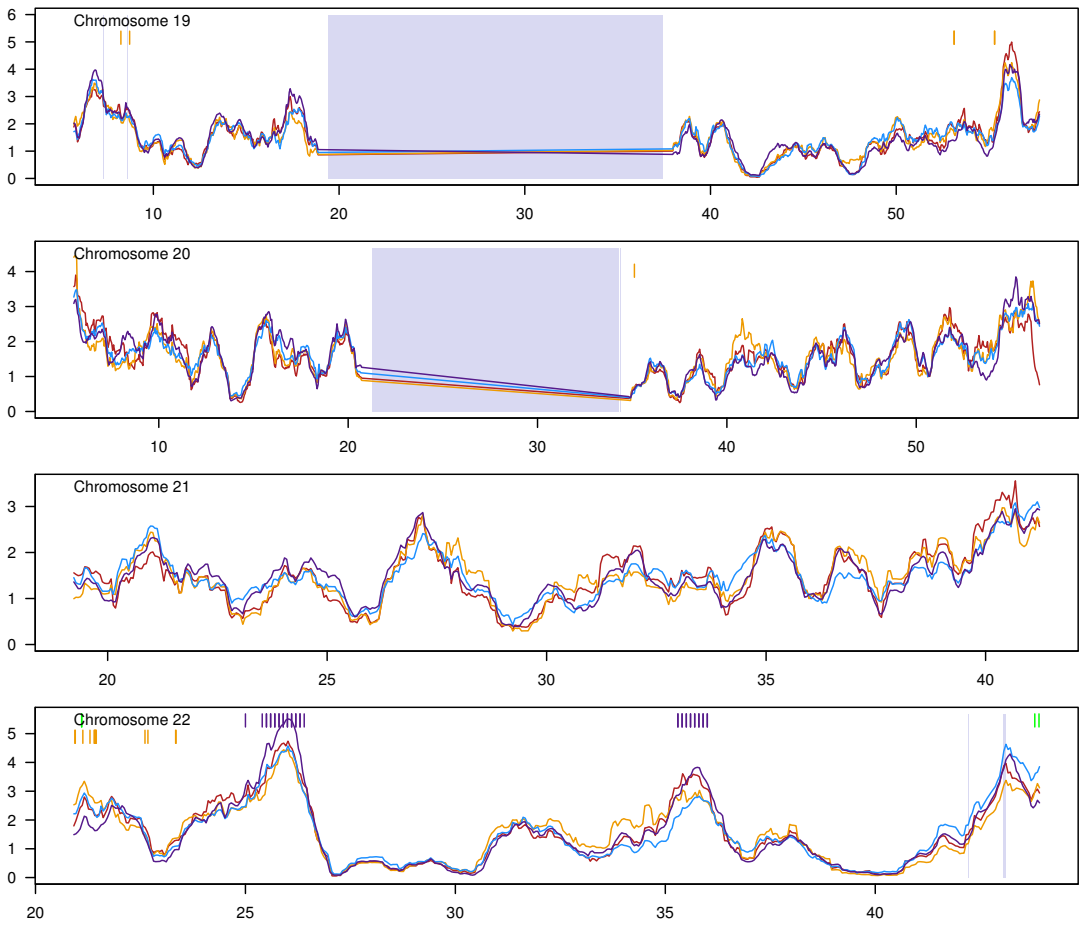
Supplementary Fig. 10. Comparison of recombination maps at a 1 Mb scale. Chromosomes 1 through 6. See last panel for details and legend.



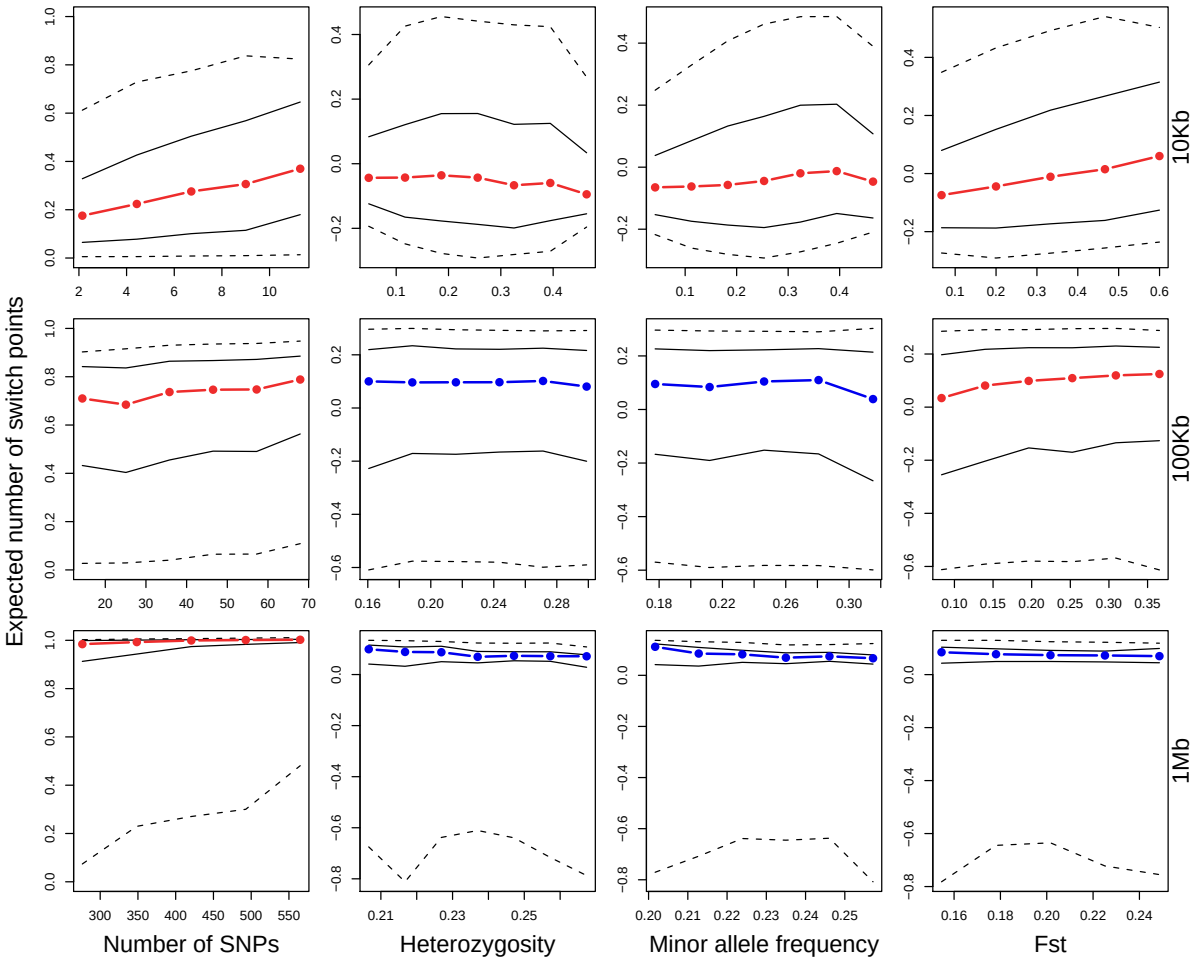
Supplementary Fig. 10. Comparison of recombination maps at a 1 Mb scale. Chromosomes 7 through 12. See last panel for details and legend.



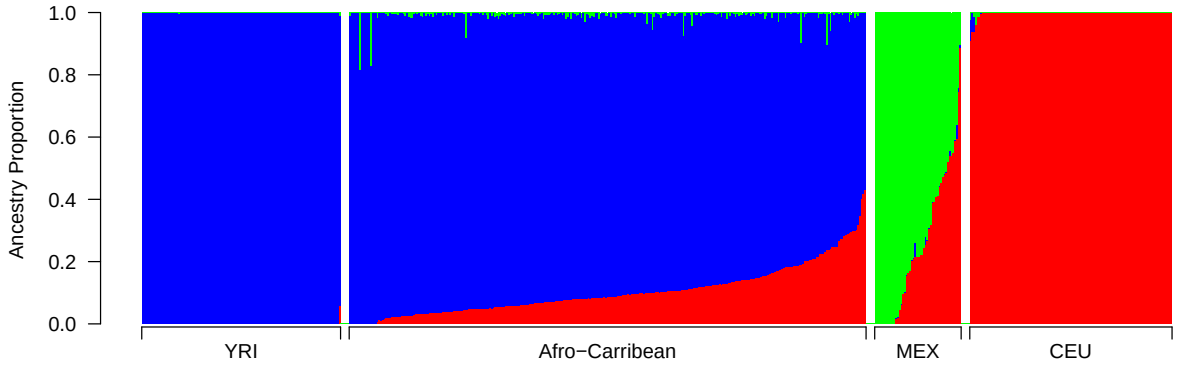
Supplementary Fig. 10. Comparison of recombination maps at a 1 Mb scale. Chromosomes 13 through 18. See last panel details and for legend.



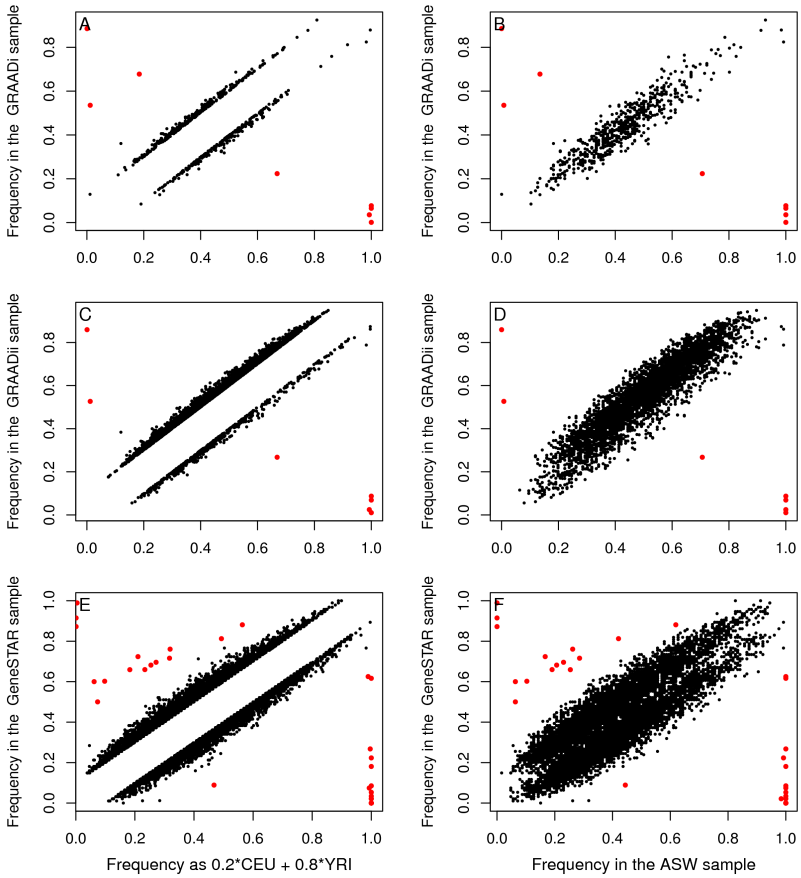
Supplementary Fig. 10. Comparison of recombination maps at a 1 Mb scale. Chromosomes 19 through 22. The plots also show all excluded areas (gray rectangles) and several outlier intervals: on the top row the 1% 1Mb intervals with largest absolute differences between the AfAdm map and an 80%/20% average of the HapMapCEU and HapMapYRI maps. Intervals where the AfAdm map shows a larger recombination rate are given in purple, the others in green. On the second row we show the 0.1% 10Kb intervals with highest (orange) and lowest (blue) average proportion of African ancestry across individuals.



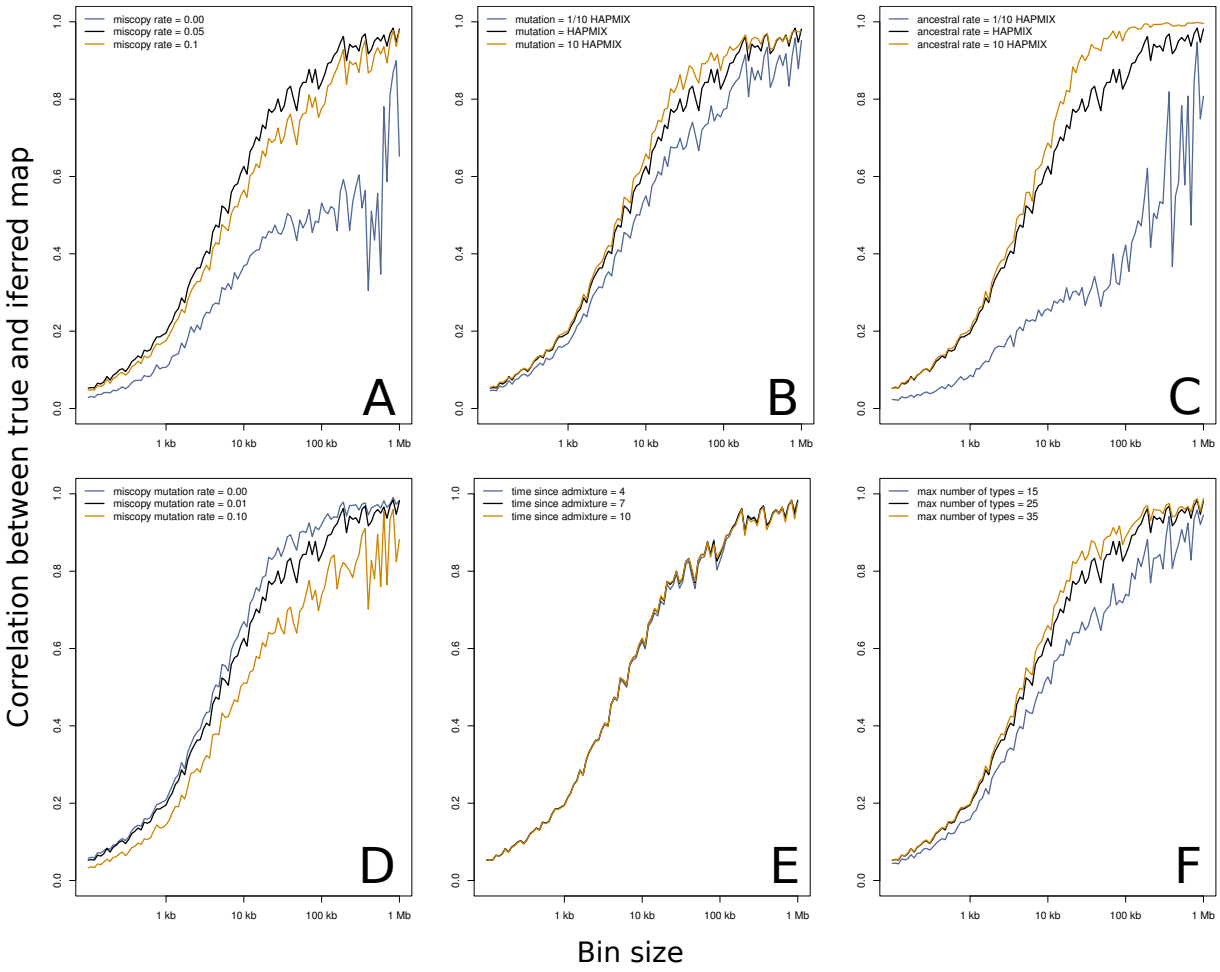
Supplementary Fig. 11. Sensitivity of inference to SNP density, heterozygosity, minor allele frequency and F_{ST}. We stratified the expected number of switch points $c_{jk}^{(i)}$ inferred in intervals of varying sizes symmetrically around true switch point locations by features of the SNPs in those intervals. Each row contains results obtained for a specific interval size (10Kb, 100Kb or 1Mb). We generally found $c_{jk}^{(i)}$ to be larger in intervals with more SNPs (first column), suggesting that a higher SNP density is helpful in placing ancestry switch points. This effect, however, is much smaller with larger interval sizes. We next evaluated the sensitivity of $c_{jk}^{(i)}$ to the local genetic diversity. Since we observed more variance in genetic diversity among intervals with low SNP densities, we based our evaluation of heterozygosity, minor allele frequencies and F_{ST} on the residuals of $c_{jk}^{(i)}$ after regressing against the number of SNPs per interval. In each subplot we show the 2.5%, 25%, 50%, 75% and 97.5% quantiles of $c_{jk}^{(i)}$ (or the residuals) for different bins covering the range of values found in our simulations. Significant correlations (after Bonferroni correction) are shown in red, non–significant in blue.



Supplementary Fig. 12. Proportion of genome–wide ancestry among African–Caribbean individuals. Estimates of individual ancestry proportions as inferred by `ADMIXTURE` based on a merged sample consisting on the African–Caribbean sample (GRAADii) and the haplotypes of the CEU, YRI and MEX HapMap3 populations and K=3.



Supplementary Fig. 13. Outlier SNPs with inconsistent allele frequencies. We compared the SNP frequencies to those of an artificial 20%/80% admixture of HapMap CEU and YRI individuals and of the HapMap ASW sample. For each SNP we computed the absolute difference between the observed allele frequency and the expected allele frequency in the artificially admixed sample. Visual inspection suggested to drop all SNPs with an absolute frequency difference >0.3 (red dots in left column). We then confirmed the outlier status of these SNPs by comparing their allele frequency against the frequency from the HapMap ASW sample (right column). Outlier SNPs were dropped from all samples. Note that we only plotted SNPs with an absolute difference between observed and expected frequencies of at least 0.1 (black dots). Only three of our samples showed outlier SNPs: (A) and (B) GRAADi (C) and (D) GRAADii (E) and (F) GeneSTAR



Supplementary Fig. 14. Effect HMM parameters on inference. We used our simulated data sets to compare the power to infer the true map when using different HMM parameters. Starting with the values suggested by Price et al.[25] (black line), we also used values that were substantially lower (blue) and substantially higher (yellow) values for one parameter at the time while keeping all others constant. We tried hard to mimic many aspects of our data set when generating the simulations, we have to stress that the simulated data is not suited to infer the proper HMM parameters to be used, they differ in several aspects (e.g. no genotyping error). We rather use these data sets to get a general understanding of how robust the inference is to the parameter choices and to check if the suggested parameters behave unexpectedly. See the description of the HMM for details on the different parameters. A) miscopy rate, B) mutation rate, C) ancestral population sizes, D) miscopy mutation rate and E) number of maximally allowed haplotypes when collapsing.

References

1. Johnson, A. D. *et al.* Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists. *Nat Genet* **42**, 608–13 (2010).

2. FBPP Investigators. Multi-center genetic study of hypertension: The family blood pressure program (fbpp). *Hypertension* **39**, 3–9 (2002).

3. Daniels, P. R. *et al.* Familial aggregation of hypertension treatment and control in the genetic epidemiology network of arteriopathy (genoa) study. *Am J Med* **116**, 676–81 (2004).

4. Gunderson, K. L. *et al.* Whole-genome genotyping. *Methods Enzymol* **410**, 359–76 (2006).

5. Mathias, R. A. *et al.* A genome-wide association study on African-ancestry populations for asthma. *J Allergy Clin Immunol* **125**, 336–346.e4 (2010).

6. Barnes, K. C. *et al.* Linkage of asthma and total serum IgE concentration to markers on chromosome 12q: evidence from Afro-Caribbean and Caucasian populations. *Genomics* **37**, 41–50 (1996).

7. Zambelli-Weiner, A. *et al.* Evaluation of the CD14/-260 polymorphism and house dust endotoxin exposure in the Barbados Asthma Genetics Study. *J Allergy Clin Immunol* **115**, 1203–9 (2005).

8. Moore, W. C. *et al.* Characterization of the severe asthma phenotype by the National Heart, Lung, and Blood Institute's Severe Asthma Research Program. *J Allergy Clin Immunol* **119**, 405–13 (2007).

9. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–8 (2010).

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

11. Bryc, K. *et al.* Genome-wide patterns of population structure and admixture in west africans and african americans. *Proc Natl Acad Sci U S A* **107**, 786–91 (2010).

12. Murray, T. *et al.* African and non-african admixture components in african americans and an african caribbean population. *Genet Epidemiol* **34**, 561–8 (2010).

13. Parra, E. J. *et al.* Estimating african american admixture proportions by use of population-specific alleles. *Am J Hum Genet* **63**, 1839–51 (1998).

14. Tang, H. *et al.* Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am J Hum Genet* **81**, 626–33 (2007).

15. Benn-Torres, J. *et al.* Admixture and population stratification in African Caribbean populations. *Ann Hum Genet* **72**, 90–8 (2008).

16. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664 (2009).

17. Chen, G. K., Marjoram, P. & Wall, J. D. Fast and flexible simulation of DNA sequence data. *Genome Research* **19**, 136–142 (2009).

18. Schaffner, S. F. *et al.* Calibrating a coalescent simulation of human genome sequence variation. *Genome Research* **15**, 1576–1583 (2005).

19. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).

20. Long, J. C. The genetic structure of admixed populations. *Genetics* **127**, 417–28 (1991).

21. Pfaff, C. L. *et al.* Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am J Hum Genet* **68**, 198–207 (2001).

22. Pool, J. E. & Nielsen, R. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**, 711–9 (2009).

23. McVean, G. A. T. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004). URL `http://www.sciencemag.org/content/304/5670/581.abstract`. `http://www.sciencemag.org/content/304/5670/581.full.pdf`.

24. 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).

25. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* **5**, e1000519 (2009).

26. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–103 (2010).

27. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–9 (2010).

28. Petkov, P. M., Broman, K. W., Szatkiewicz, J. P. & Paigen, K. Crossover interference underlies sex differences in recombination rates. *Trends Genet* **23**, 539–42 (2007).

29. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727–32 (2005).

30. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).

31. McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19**, 1527–41 (2009).