SUPPLEMENTARY DISCUSSION

RHH visualization compared to SABER and extended to the Illumina 550K array

Although RHH and SABER produced nearly identical patterns of genome mosaicism (Figs.1b, 2b, 2d and Supplementary Figs. S27-S29), there were minor differences that may impact visualization of true admixture that spans very short chromosomal lengths (<3 Mb). Specifically, the RHH-Illm550K visualizations have somewhat sparser rare-het density than their RHH-Affy500K counterparts suggesting that Affy500K may be more effective in visualizing genuine but very short non-Caucasian segments (e.g. see Fig. 1b whose tiny rare-het segment at bottom of chromosome 20 is absent from Supplementary Fig. S27 for RHH-Illm550K). SABER output also contains multiple, very short regions of putative non-Caucasian DNA whose validity is mostly uncertain, though a few of these peaks appear to be false-positives because they recur in areas of sparse genotyping (e.g. on chromosome 8 at ~45 Mb in Supplementary Figs. S27 and S29) or occur with one genotyping array but not the other (e.g. in Supplementary Fig. S27, the Affy500K peak on chromosome 1 at ~104 Mb is absent in the Illm550K SABER output). However, other small SABER peaks are likely true-positives since they occur in output from both arrays and also correspond to small RHH rare-het segments (e.g. compare Fig. 1b tiny rare-het segment near bottom of chromosome 20 with corresponding SABER peaks in Supplementary Fig. S27). Although most genuine admixture appears to reside in continuous chromosomal segments spanning at least 5 Mb, genuine admixture in much shorter segments (<3 Mb long) is likely to evoke future work to more conclusively visualize and characterize the properties of such segments.

RHH performs effectively for "small" datasets of 50-200 subjects

Although C_{het}=0.5% is near-optimal for RHH mosaic visualization and ethnic outlier detection, Supplementary Figures S30-S33 also indicate that C_{het} values of 1% and higher still enable statistical detection of admixture. This is important for datasets with fewer than 200 subjects in which it would be impossible to achieve a rare-het frequency below C_{het}=0.5%=1/200. Therefore to explore RHH performance in smaller datasets, we combined the 25 "set B" outliers shown in Table 1 with 25, 50, or 100 other 58BC subjects randomly selected from set B. For these datasets with 125, 75, or 50 subjects, RHH detected statistically excess rare-het counts for ~95%, ~70%, or ~35% of the 25 outliers, respectively, at both the lowest possible C_{het} value and C_{het}=5% with either Affy500K or Illm550K genotypes. Furthermore, admixed chromosome mosaicism remained clearly visualizable for all mosaic subjects at the lowest possible C_{het} values as well as being recognizable at C_{het}=5% for all three datasets and both genotyping arrays (see Supplementary Figs. S36-S39). Though not comprehensive, these results indicate that RHH can produce valuable results from smaller-sized datasets.

Mathematical Model Results for Unthinned Rare-Het Counts

In applying the mathematical model to *unthinned* rare-het counts, we found that 1Mb "thinning" of rare-hets does not substantially reduce RHH ability to detect ethnic outliers. Graphs analogous to "thinned" results but evaluating unthinned rare-het counts are given in Supplementary Figures S40-S43 and imply that 1Mb thinning elevates the lowest detectable admixture by only ~1% of the genome for most parameter combinations. For example, minimum detectable Asian admixture is ~4.3% of the genome for 1Mb-thinned counts at C_{het}=0.5% and Y=0.1%

(Supplementary Figs. S31, S33), but if unthinned counts are evaluated as if they are independent (Supplementary Figs. S41, S43) detectable Asian admixture would be ~3.3% of the genome.

RHH applied to a single "test" subject added to a large panel of unadmixed individuals

Another inference from the mathematical model relates to evaluating a single "test" subject by running RHH analysis with the test subject added to a large panel of reference subjects who lack admixture. The Y value for this dataset would be very low (<<0.1%) since only the test subject contributes admixture and the p-value for excess rare counts could be set higher (at 0.001 or 0.05) than the Bonferroni-corrected p-value (0.001/1500) in Supplementary Figures S30-S33 since only the test subject is being evaluated. For datasets with this substantial decrease in Y and increase in p-value, calculations like those for Supplementary Figures S30-S33 indicate that a statistical excess of rare-het counts would be generated by African admixture covering only ~0.25% of the genome. Since RHH only evaluates DNA between the two most widely separated SNPs genotyped on each autosome, the total genome length evaluated by RHH is ~2780 Mb implying that ~0.25% of the genome would be ~7Mb, a length of rare-het segment easily visualized by RHH (Figs. 1-2). Thus a statistical excess of rare-hets could be used to flag subjects possibly carrying a small but easily visualized segment of admixed DNA. Alternatively, such subjects could be initially detected simply by compiling locations of their rare hets to flag subjects with possible rare-het clustering and to direct subsequent inspection of RHH visual output.

Effect of Modest Admixture on Association P-values

Our comparisons with conventional PC-MDS and PLINK Z-score analyses found that RHH can detect many additional outliers with modest amounts of admixture (see Table 1). For example, more than 1% of the 1500 controls in the 58BC dataset are outliers with modest African admixture (mean genome coverage: 2.5%, range: 0.5% to 7.1%) which are undetected by PC-MDS or PLINK Z-score but are readily identified by RHH. This raises the important question of whether GWA p-values could be substantially inflated by not excluding outliers carrying modest admixture (i.e. covering under 10% of an outlier's genome) who are detected by RHH but not by less sensitive analyses such as PC-MDS or PLINK Z-score.

This question is too complex to comprehensively address here, but we have investigated the increase in false-positive p-values below 10^{-3} for several models of modest case-control admixture to provide the broad outlines of an answer (Supplementary Table S10). Association results (p-value≤0.001) for common SNPs (MAF≥0.05) appear largely unaffected by multiple outliers with modest admixture since insufficient outliers are admixed at the same genomic position to markedly perturb a common SNP away from its true MAF. But modestly admixed outliers can increase the dataset frequency of a rare minor allele (0.00025≤MAF≤0.01) by 10% to 200% if the minor allele is relatively common (MAF>0.1) in the outlier population; and this can, in turn, markedly inflate GWA p-values when the modest admixture differs in cases and controls.

For example, in Supplementary Table S10, the three model datasets assume 5% of the genome is covered by African admixture in 1000 case outliers versus only 1% being covered in 1000 control outliers. Consequently, SNPs which are rare in Caucasians (0.00025≤MAF≤0.05) but more common in Africans (0.2≤MAF) are

shown to add 16% to 79% more false-positive associations (p-value≤0.001) to the ~500 GWA false-positives expected by chance among the ~500,000 SNPs of the Affy500K or Illm550K array. Many more such false-positive associations would be likely if future GWA studies target large numbers of rare SNPs now being discovered by large-scale sequencing (Supplementary Table S10). Although the relatively high percentages of modestly admixed outliers in these model datasets (5% to 20%) might be argued to rarely occur, GWA studies are expanding along the following three dimensions that may make significant amounts of modest admixture both more likely to be discovered and more likely to occur in future datasets: (1) expanding focus on genome-wide testing of rare variants (MAF<<0.05) for disease association, (2) expanding of sample sizes to detect weaker disease loci, either by increasing genotyping or by cobbling together smaller and sometimes disparate datasets to perform meta-analyses, and (3) expansion to ethnic groups worldwide, some of which may contain more widespread admixture and less ethnic homogeneity than the European-descent datasets primarily used in the first wave of GWA studies. Given the potential of modest admixture to inflate disease association statistics as illustrated in Supplementary Table S10, more work is needed both to explore additional models of modest admixture and to determine the magnitudes of modest admixture in real datasets.

SUPPLEMENTARY METHODS

Genotyping and Sample QC

Genotyping of 500,000 SNP markers densely covering the genome was conducted for the WTCCC samples with the Affymetrix GeneChip 500K array (Affy500K) as previously described (12). We used genotype calls available through the WTCCC and its website which were provided by Affymetrix and are based on their application of the BRLMM calling algorithm. This algorithm provides a confidence score between 0 and 0.5 for each called genotype such that the genotype call is considered less certain as the confidence score increases and genotypes with confidences above 0.5 were dropped (i.e. not called) by the Affymetrix BRLMM protocol applied to WTCCC samples (see http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf). Following the same criteria applied in the WTCCC study (12), we omitted a small number of DNA samples with genome-wide genotype call rates below 97% since low call rate indicates poor DNA quality and increased likelihood of genotyping error. To model RHH performance and demonstrate visualization of admixed chromosome mosaicism with the Illumina HumanHap550 array (Illm550K), we also used Illm550K genotypes assayed by the Sanger Institute on ~1500 subjects of the 58BC cohort (sample set B).

Tabulation of rare-het and rare-hom counts

Our method identifies ethnic outliers as subjects with a high total number of rare heterozygote and/or rare homozygote genotypes compared to other subjects in the same dataset. Rare-het counts included all heterozygous genotypes for any SNP that: (a) had *zero* counts for the rarer homozygote and (b) a heterozygote genotype

frequency in the data set of 0.005 or below. The cutpoint or "C_{het}" of 0.5% specifying the highest allowable frequency of rare-hets was settled upon through empirical observation and is supported by: (a) Supplementary Figures S34-S35 which show clear visualization of mosaicism in admixed chromosomes for C_{het}=0.5% compared to other C_{het} values and (b) Supplementary Figures S30-S33 and the "Mathematical Model of RHH Performance" section in DISCUSSION showing that C_{het}=0.5% provides near-optimal statistical detection of admixture across a wide range of parameters influencing RHH performance. Although our experience and evidence presented here shows C_{het}=0.5% to be a highly effective cutpoint frequency for defining rare hets, RHH software is flexible in allowing the user to specify the C_{het} value used for RHH analysis and mosaic visualizations.

In initially conceiving and applying the method, the rarer homozygote genotypes of a SNP were included among the "rare hom" counts only if *no* subjects in the dataset were heterozygous at that SNP (i.e., the SNP had zero heterozygote counts and non-zero but typically very few counts for the rarer homozygote). However, for analyses described here, we relaxed the initial criterion so that rare-hom counts were included for any SNP if: (a) the genotype frequency of its rarer homozygote exceeded the genotype frequency of its heterozygote, and (b) the heterozygote genotype frequency in the dataset was 0.002 or lower. The initial criterion of allowing no heterozygotes at rare-hom SNPs was relaxed to the very low heterozygote frequency of 0.002 so that SNPs with a rare homozygote in an ethnic outlier were not excluded due to a second ethnic outlier in the dataset being heterozygous at the same SNP. We have found that the initial and relaxed criteria both identify the same subset of ethnic outliers exhibiting a statistically significant excess of rare-hom counts, but that the relaxed criterion increases the total number of rare-hom counts observed in these

outliers ~1.5 to 3 fold. RHH software is flexible in allowing the user to change the allowable heterozygote frequency for rare-hom SNPs.

If two SNPs are in linkage disequilibrium (LD), rare-hom counts (or het counts) contributed to a subject by the two SNPs may not be independent of each other. Therefore, to better assess the statistical significance of observed clustering of rare-homs or rare-hets within individual samples, we totalled rare-hom and rare-het counts at two levels of inter-SNP distance to account for possible interdependence due to LD. For level-1, counts were totalled without requiring that the contributing SNPs be a minimum distance apart ("All SNPs" in Tables 1 and 3). For level-2, a rare count was added to the total counts for a subject (i.e. hom *or* het counts) only if other SNPs contributing a count to that subject were at least 1 Mb away ("1 Mb apart" in Tables 1 and 3). This 1 Mb spacing was chosen since the half-length of LD as measured by D' is 50 kb or less in human populations and hence pairwise LD should be zero or negligible between SNPs at least 1 Mb apart (5, 18) implying that the "thinned" counts from such SNPs can be considered independent. As explained in the SUPPLEMENTARY DISCUSSION, our Mathematical Model results imply that 1 Mb thinning of rare-het counts does not substantially reduce RHH ability to detect ethnic outliers; however, RHH software allows the user to specify a different thinning distance for comparison with results from unthinned counts as illustrated in Tables 1 and 3.

Ethnic SNPs and "non-Caucasian" alleles in Chromosomal Maps and Tables

We used two panels of "ethnic" SNPs defined as SNPs which are monomorphic in HapMap Caucasian (CEU) subjects but have minor allele frequency (MAF) of at least 0.4 in HapMap Yoruban (YRI) subjects ("YRI SNPs") or in

HapMap Chinese (CHB) subjects ("CHB SNPs") (5,6). These "YRI" or "CHB SNPs" would rarely be heterozygous in a non-admixed Caucasian, but we chose their near-maximal MAF (≥0.4) in YRI and/or CHB since the SNPs would be among those most likely to be heterozygous or homozygous for the "non-Caucasian" allele in DNA of non-Caucasian ancestry. There were a total of 2,397 YRI SNPs and 343 CHB SNPs genotyped on the Affy500K chip; and these SNPs are denoted by tiny purple and green triangles, respectively, in the chromosomal maps of individual subjects if the SNP is heterozygous or homozygous for the allele *not* observed in HapMap CEU subjects (hence implying the presence of a "non-Caucasian" allele). To calculate p-values based on het counts from SNPs unlikely to be in LD (see above), we also identified a subset of each panel (838 YRI SNPs, 139 CHB SNPs) that contained only SNPs at least 1 Mb away from the nearest neighbor in the subset. Counts of heterozygotes in these two pre-selected subsets of YRI and CHB SNPs are shown under " 'Ethnic' Het Counts" in Table 1.

Statistical significance of observed counts in individual subjects

In RESULTS, DISCUSSION, and in describing Tables 1 and 3, we refer to permutation-derived thresholds and p-values for different types of genotype count observed in individual subjects. These 6 different count types are each shown in Table 1: rare-het counts or rare-hom counts from "All SNPs" or from SNPs "1 Mb apart" (defining 4 count types) and counts of heterozygotes at YRI SNPs or CHB SNPs at least 1 Mb apart (defining the final 2 count types). In this section of SUPPLEMENTARY METHODS, we therefore generically refer to "counts" (of thinned or unthinned rare hets, thinned or unthinned rare homs, or hets at YRI or CHB SNPs 1 Mb apart) since the permutation procedure and software for calculating p-

value thresholds were the same for each type of count. The goal of these permutations was to calculate an integer count threshold whose probability (p-value) of being reached or exceeded by *one or more* subjects of a dataset was less than 0.001 or, alternatively, less than 0.05 under the null hypothesis. We assumed for our null hypothesis that all N subjects in a WTCCC dataset derived from a homogeneous population in which each subject had an equal (1/N) probability of inheriting an observed count. In Tables 1 and 3, observed counts that equalled or exceeded the count threshold for $p<0.001$ are in **bold** and underlined whereas observed counts that reached the threshold for $p<0.05$ but not for $p<0.001$ are underlined but not in bold. Reaching these thresholds is taken as evidence that the subject differs from the majority of other subjects who do not exhibit excess counts.

To calculate the specific thresholds to apply to a particular WTCCC dataset and type of count, we first determined the total number of counts (T) observed in all N subjects in the dataset as well as the number of SNPs that contributed each number of integer counts (1, 2, 3, etc.) to the total T. We then ran 1000 permutations in which each permutation randomly distributed the T counts among the N subjects as follows: For each SNP contributing n=1, 2, 3, etc. counts, n of the N subjects were randomly selected and one count was added to the number of counts already obtained by the selected subject(s) from SNPs processed earlier in the permutation. After the T counts from all SNPs had been randomly distributed to the N subjects, we determined the distribution for the permutation defined by the number of subjects observed to have received each possible number of integer counts (0 to T). Each observed combination of c counts in s subjects was tallied across all 1000 permutations to determine the number of instances per 1000 trials that exactly c counts were observed in *one or more* subjects.

This distribution of instances for c counts observed in one or more subjects typically had the highest number of instances near T/N, the mean counts expected in an individual subject, while *zero* instances per 1000 were typically observed at integer count thresholds above and below T/N (i.e. *all* subjects in *all* permutations exhibited counts between these thresholds). Based on this distribution, the threshold for $p<0.001$ was set at the lowest integer count above T/N that was never observed in any subject in any of the 1000 permutations. The threshold for $p<0.05$ was the lowest integer count observed in one or more subjects for fewer than 50 of the 1000 permutations. In this way, we were able to designate $p<0.001$ and $p<0.05$ thresholds that were specific for each type of count and each WTCCC dataset.

Simulation of HapMap-Derived Ethnically Admixed Subjects in Table 3

Phased haplotypes for release 21 of the HapMap data set were obtained from the HapMap website (www.hapmap.org), along with estimates of genetic distances between markers. Five separate lineages of five generations were generated by initially crossing one CEU individual with either a CHB or YRI individual, then sequentially backcrossing this simulated offspring with further CEU individuals. One individual representing each generation was selected from independent lineages to be included in the RHH and PLINK analyses shown in Table 3. A second set of individuals was created by intercrossing the "F1" offspring of two separate CEU×CHB outcrosses or two CEU×YRI outcrosses.

Offspring were simulated by selecting one chromosome to be inherited at random from each parent. Recombinant chromosomes were simulated in each parent and for each generated offspring as follows: For each pair of adjacent markers on each chromosome, a random number between 0 and 100 was generated; if this number was

less than the genetic distance in centimorgans between the two markers, the two parental haplotypes 3' of the 5' marker in the pair were switched.

Mathematical Model of RHH performance

RHH performance for the Affymetrix500K and Illumina550K SNP arrays

Given specific values of Y, p and q, equation (1) in the main text can determine if a SNP's F_{het} value falls at or below the rare-het frequency cutpoint (C_{het}) which qualifies the SNP to contribute rare-het counts to individual subjects and the entire dataset. By thus identifying rare-het SNPs among the larger pool on a commercial SNP array, RHH performance can be modeled under various combinations of Y, C_{het}, outlier ethnicity and Affymetrix or Illumina array since each of these parameters determines the SNP subset which qualifies as rare-het SNPs. To generate our modeling results, we took SNPs on chromosome 20 to be representative of the whole genome and used all chromosome 20 SNPs on the Affy500K array (12400 SNPs) and Illm550K array (14107 SNPs) that were also found in HapMap (which contained ~98% of the chromosome 20 SNPs on either array). To accurately model F_{het} values for rare-het SNPs which, by definition, must have very low Caucasian MAFs ($\leq C_{het}$=0.1% to 5%), we estimated p (Caucasian MAF) from ~1400 Caucasians of the 58BC cohort which had no admixture on chromosome 20 and which had been genotyped on both the Affy500K and Illm550K arrays. Each SNP's corresponding value for q (allele frequency in Africans or Asians) was estimated from ~60 HapMap Africans (YRI) or ~90 HapMap Asians (CHB+JPT).

As a preliminary step in modeling RHH performance under the different parameter combinations in Supplementary Figures S30-S33, we created Perl software that calculated the expected (mean) number of counts per dataset subject from rare-het

SNPs on chromosome 20. This expected number (E_{het}^{20}) was calculated by summing each F_{het} value from equation (1) for all chromosome 20 SNPs that qualified as being rare-het SNPs for the particular parameter combination being evaluated. The same software also processed only rare-het SNPs to calculate expected rare-het counts in a subject whose chromosome 20 was covered by (a) 100% non-Caucasian admixture or (b) 0% admixture. The expected counts for (a) and (b) were calculated from the component parts of equation (1), i.e., by summing each rare-het SNP's value for p(1-q)+q(1-p) to obtain (a) and by summing each SNP's value for 2p(1-p) to obtain (b). To estimate *whole genome* counts corresponding to (a), (b) and E_{het}^{20}, our software multiplied the (a), (b) and E_{het}^{20} values by 44.6≈100/2.25 since the length of chromosome 20 is ~2.25% of the genotyped autosomes (as determined by base pair distance between the two genotyped SNPs at the ends of each autosome). We denote the whole genome estimates corresponding to (a), (b) and E_{het}^{20} as E_{100}, E_0 and E_{null}, respectively. "E_{null}" is used for the whole genome value because E_{het}^{20} equals the expected number of chromosome 20 counts in a randomly selected subject under the null hypothesis (see below). A related way of understanding E_{null} is that equation (1) implies that E_{null} would be the number of rare-het counts expected in a subject whose percent of genome admixture equals the mean subject admixture in the dataset (i.e. Y×100).

Supplementary Figures S30-S33 and Tables S2-S5 model "1Mb apart" RHH results like those shown in Tables 1 and 3 in which rare-het counts are "thinned" for each subject so that the SNPs contributing rare-het counts to that subject are at least 1Mb apart. To estimate thinned rare-het counts corresponding to E_{100}, E_0 and E_{null}, we calculated the percent of rare-het counts that survived 1Mb thinning in individual

simulated and real subjects for runs of RHH software across the range of rare-het frequency cutpoints (C_{het}=0.1% to 5%) shown in Supplementary Figures S30 to S33. To estimate thinning of E_{100}, our index subjects were the simulated outliers in Table 3 having 100% genome coverage by admixture from one African parent (YRI×CEU) or one Asian parent (CHB×CEU) and other simulated and real subjects in Tables 1 and 3 with high het counts and admixed ethnicity similar to the YRI×CEU or CHB×CEU subject. To estimate thinning of E_0, we also used the RHH runs to calculate percent het survival in index subjects with no evidence of non-Caucasian admixture. We found the survival percent for each RHH run to be similar among index subjects of the same type enabling us to estimate thinned counts by multiplying E_{100} and E_0 for each parameter combination by corresponding het survival percent for E_{100} or E_0. The thinned E_{100} and E_0 values (denoted E^*_{100} and E^*_0) were then used to calculate the thinned E_{null} value for the same model combination and its corresponding Y value with the equation $E^*_{null} = (1-Y)E^*_0 + Y(E^*_{100})$ in accordance with the relationship implied by equation (1).

In the SUPPLEMENTARY METHODS section entitled "**Statistical significance of observed counts in individual subjects**", we explained that our null hypothesis assumes a homogenous dataset consisting of N subjects with equal (1/N) probability of inheriting rare-het counts. This section also describes a permutation procedure in RHH software that determines an integer count threshold with 0.001 probability of "being reached or exceeded by *one or more* subjects" in a null-hypothesis dataset. These permutation thresholds are very precise and 0.001 is the tail probability that the *entire dataset* would exhibit one or more outliers by chance; but as we now explain, the same 0.001 thresholds can be accurately estimated analytically and RHH performance thereby modeled by calculating the probability that a *single*

random subject has total rare-het counts whose null-hypothesis probability is 0.001/1500 or lower. To see this, note that each rare-het SNP contributes i counts to the dataset (where i=1, 2, etc. such that $i/1500 \leq C_{het}$) and thus each rare-het SNP can be considered to be a binomial trial with null hypothesis probability i/1500 of contributing a rare-het count to a random subject. When rare-het SNPs are thinned so that SNPs contributing counts to the *same* subject are at least 1Mb apart and hence unlikely to be in LD (see above), then each SNP's rare-het contribution can be considered independent implying that the binomial trials are independent. The probability of "success" is not equal for each binomial trial since i/1500 varies, but i/1500 is always *very low* ($\leq C_{het}$) and it is well known that a series of unequal but low binomial probabilities are very accurately estimated by a "generalized binomial distribution" (GBD) with a single binomial probability of "success" equal to the mean of the probabilities of each trial (see main text reference 29). Therefore the GBD can used to determine any threshold T_x and corresponding tail probability P_x that a random null-hypothesis subject would exhibit rare-het counts that equal or exceed T_x. Furthermore, the T_x and P_x values for a single random subject that correspond to a p-value of 0.001 that the *entire dataset* would exhibit at least one subject with rare-hets at or above T_x can be found by solving the equation $1-(1-P_x)^{1500}=0.001$ which yields $P_x=6.67\times10^{-7}$ (which is virtually identical to the Bonferroni corrected p-value 0.001/1500).

Therefore using the GBD corresponding to each parameter combination (Y, C_{het}, ethnicity, SNP array), we determined each combination's threshold (T_x) specifying the number of rare-het counts a subject must exhibit to generate a p-value of 0.001/1500. We then calculated the minimum fraction (F_{min}) of genome coverage by admixture that would generate exactly T_x rare-het counts in an outlier by solving

the equation $T_x = (1 - F_{\min})E_0^* + F_{\min}E_{100}^*$ (where E_0^* and E_{100}^* are expected thinned rare-het counts for a genome with 0% and 100% coverage by admixture as discussed above). The F_{min} values were converted to percentages and displayed in Supplementary Figures S30-S33 which therefore show the smallest percentage of genome admixture detectable in an outlier at a significance level of p<0.001/1500. For comparison, analogous results evaluated in the same manner but based on the *unthinned* rare-het counts are shown in Supplementary Figures S40-S43 and Tables S6-S9.

A final detail concerning the creation of Supplementary Figures S30-S33 and Tables S2-S5 is that T_x for each GBD was calculated using Poisson approximation of binomial tail probabilities since, for each GBD, the generalized binomial probability (P_{GB}) is very low (<C_{het}) and total number (N_{GB}) of trials (i.e. rare-het SNPs) is very large (>1000). The GBD is therefore very accurately estimated by the Poisson distribution with Poisson intensity $\lambda=(N_{GB})(P_{GB})$ where λ is the mean of the GBD (see main text reference 29). λ equals E_{null}^* (expected rare-het counts in a random subject under the null hypothesis) and λ for each GDB can be calculated from the equation $\lambda = E_{null}^* = (1-Y)E_0^* + Y(E_{100}^*)$ where Y, E_0^* and E_{100}^* are specific values for the parameter combination being evaluated. The corresponding T_x value was calculated by summing the Poisson probabilities of observing exactly s binomial "successes" (s=0, 1, etc.) using the R software function "dpois(s, λ, log=FALSE)" (http://cran.r-project.org/). T_x equals (s+1) where s is the first integer value for which $[1 - \sum_{S} \text{dpois}(s, \lambda)] \leq 0.001/1500$.

Model datasets and calculation of values in Supplementary Table S10

Table S10 shows the increase in false-positive disease associations (p-value≤0.001) caused by modest ethnic admixture in three model datasets with an equal number (N) of disease cases and controls (N=5000, 10000 or 20000). All subjects are unadmixed except for 1000 case outliers with admixed DNA covering 5% of the subject's genome and 1000 control outliers with admixed DNA covering 1% of their genomes, implying that modest ethnic outliers represent 20%, 10% or 5% of subjects in the three model datasets. (Details about admixture "covering" a subject's genome are the same as defined in **"Mathematical Model of RHH performance"** in the main text). To create Table S10, the probability of p-values≤0.001 was calculated for a "neutral" (i.e. non-disease causing) SNP (denoted "S") with minor allele frequency (MAF) in the non-outlier population equaling one of the 7 values shown in Table S10 (MAF=0.00025, 0.0005, 0.001, 0.005, 0.01, 0.03, or 0.05) and corresponding frequency of the same allele equaling 0.2 in the outlier population from which admixed DNA was derived.

To calculate the probability of p-values≤0.001, we first determined "Prob(C)", the probability that, in a sample of N cases or N controls, the genomic position of neutral SNP S is covered by outlier DNA on exactly C chromosomes (and hence not covered on 2N-C chromosomes). Since the number of admixed cases or controls is 1000 and since outlier DNA is assumed to be carried on at most one of the admixed subject's two homologous chromosomes, Prob(C) is given by the following binomial distribution:

$$\text{Prob(C)} = \left(\frac{1000!}{C!(1000 - C)!} \right) X^{C} (1 - X)^{1000-C} \quad \text{equation (2)}$$

where X is the fraction of the genome covered by admixture (recall that X=0.05 in admixed cases and X=0.01 in admixed controls).

As in equation (1) of the main text, let "p" be the MAF of S in the non-outlier population and "q" be the frequency of the same allele in the outlier population. Furthermore, for N cases or N controls, let "L" be total counts of the minor allele that derive from the C chromosomes covered by outlier DNA and let "M" be total counts of the minor allele from the (2N-C) chromosomes not covered by outlier DNA at S. For each possible value of C ($0 \leq C \leq 1000$), the conditional probabilities of L and M are given by the binomial distributions:

$$\text{Prob}(L|C) = \left(\frac{C!}{L!(C - L)!} \right) q^{L} (1 - q)^{C-L} \qquad \text{equation (3)}$$

$$\text{Prob}(M|C) = \left(\frac{(2N - C)!}{M!(2N - C - M)!} \right) p^{M} (1 - p)^{2N-C-M} \qquad \text{equation (4)}$$

Based on equations (2), (3), and (4), we determined a probability distribution for the number of minor allele counts (L+M) observed in N cases and a second probability distribution for (L+M) in the corresponding N controls under each combination of model dataset values of N and p as shown in Table S10 (with q fixed at 0.2 as stated). From each pair of distributions for the probability of minor allele counts (i.e. L+M) in cases and in controls, we then summed the joint probabilities for (L+M) combinations in cases and controls that were sufficiently different in magnitude to give a p-value of 0.001 or lower to obtain the increased rate of false-positive associations (shown in column 2 of Table S10). For the extremely low MAFs (p=0.00025 to 0.001), Fisher's exact test was used to determine L+M combinations in cases and controls giving p-values of 0.001 or lower. For higher MAFs ($p \geq 0.005$) at which there was negligible probability of minor allele counts

below 20 in either cases or controls, the chi-square test for association and corresponding normal approximations were used to determine the increased rate of false-positive p-values≤0.001. The Fisher's exact test and chi-square approaches gave similar results when compared for the three model datasets at p=0.005, thereby confirming the accuracy of the calculations.

The final three columns in Table S10 show a lower bound for the actual number of *additional* SNPs giving false-positive associations (p-value≤0.001) if a GWA scan was performed and the non-outlier population was European Caucasian and outlier DNA was African. The number of additional false-positive SNPs are shown for GWA scans with the Affymetrix 500K or Illumina 550K array, or if all HapMap SNPs were genotyped. As explained in footnote d of Table S10, genotyping all HapMap SNPs was used to roughly estimate the much larger number of rare false-positive SNPs likely in future GWA scans aimed at testing many more rare SNPs (now being discovered by large-scale sequencing projects such as '1000 Genomes').

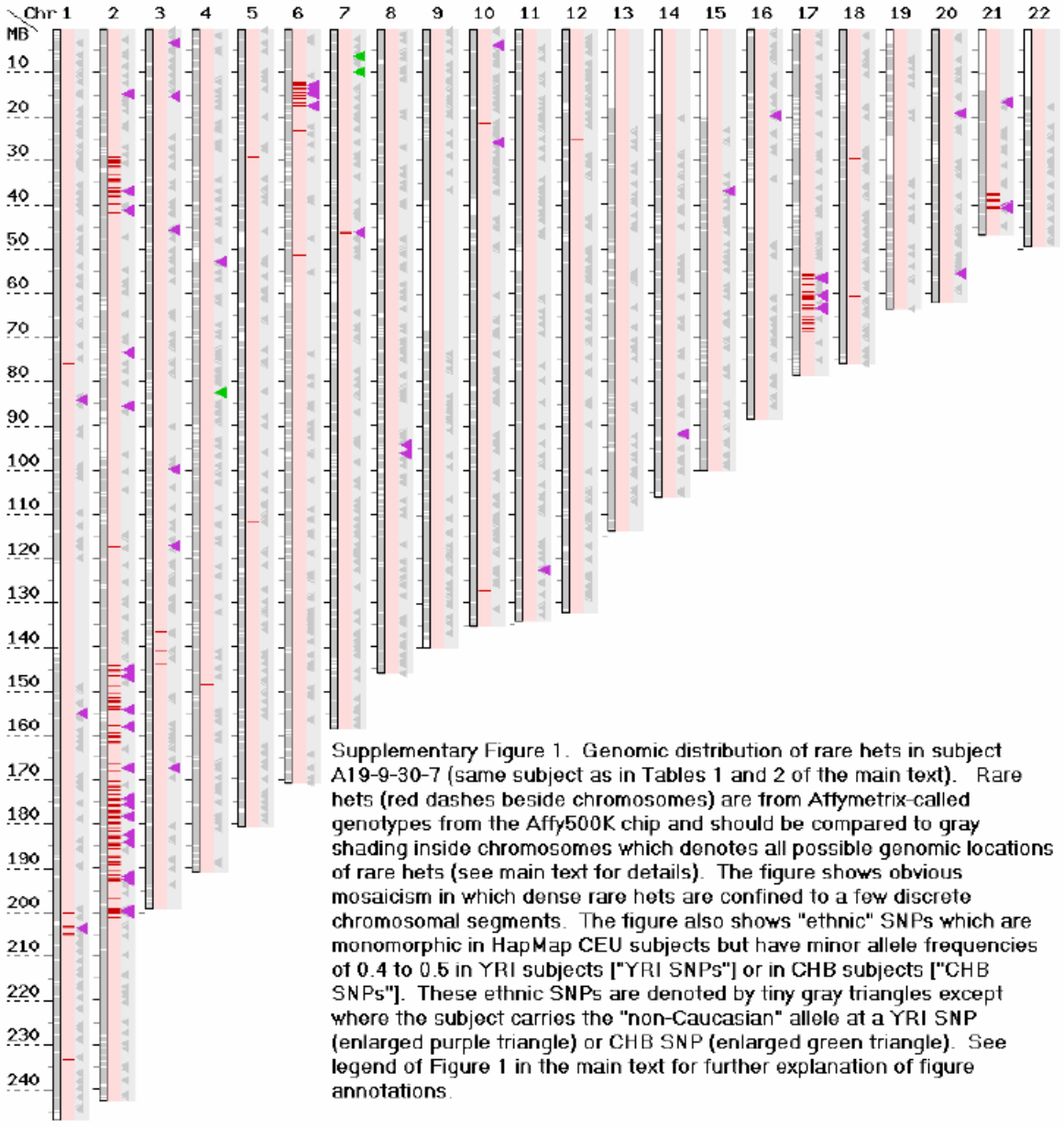
To estimate the number of additional false-positive SNPs, we took SNPs on chromosome 20 to be representative of the genome and estimated whole-genome SNP numbers by multiplying the chromosome 20 results by 44.6≈100/2.25 since the length of chromosome 20 is ~2.25% of the genotyped autosomes (for more details see **Mathematical Model of RHH performance** in SUPPLEMENTARY METHODS above). As explained under "**Mathematical Model ..**", the Caucasian (i.e. non-outlier) MAFs of chromosome 20 SNPs on the Affymetrix 500K and Illumina 550K arrays were determined from ~1400 58BC Caucasian controls who had been genotyped on both arrays, and the African (outlier) MAFs for the same SNPs were determined from HapMap Yorubans (YRI). Chromosome 20 SNPs whose Caucasian MAF fell into each frequency "bin" in Table S10 were counted if their MAF in

HapMap YRI was 0.2 or higher. For example, SNPs with YRI MAF≥0.2 were counted as belonging to the 0.005 MAF bin shown in Table S10 if their Caucasian (58BC) MAF was 0.005 or lower, but greater than 0.001 (upper limit for the next MAF bin in Table S10). Note here that SNPs were counted if YRI MAF was *above* 0.2 because such SNPs would yield an even higher rate of false-positive p-values than "neutral" SNP S (used to calculate the false-positive rate in Table S10 assuming q=0.2). Similarly, SNPs whose 58BC MAF was *below* 0.005 were counted in the 0.005 bin since they also would yield a higher false-positive rate than calculated for SNP S (which assumed a non-outlier MAF of p=0.005).

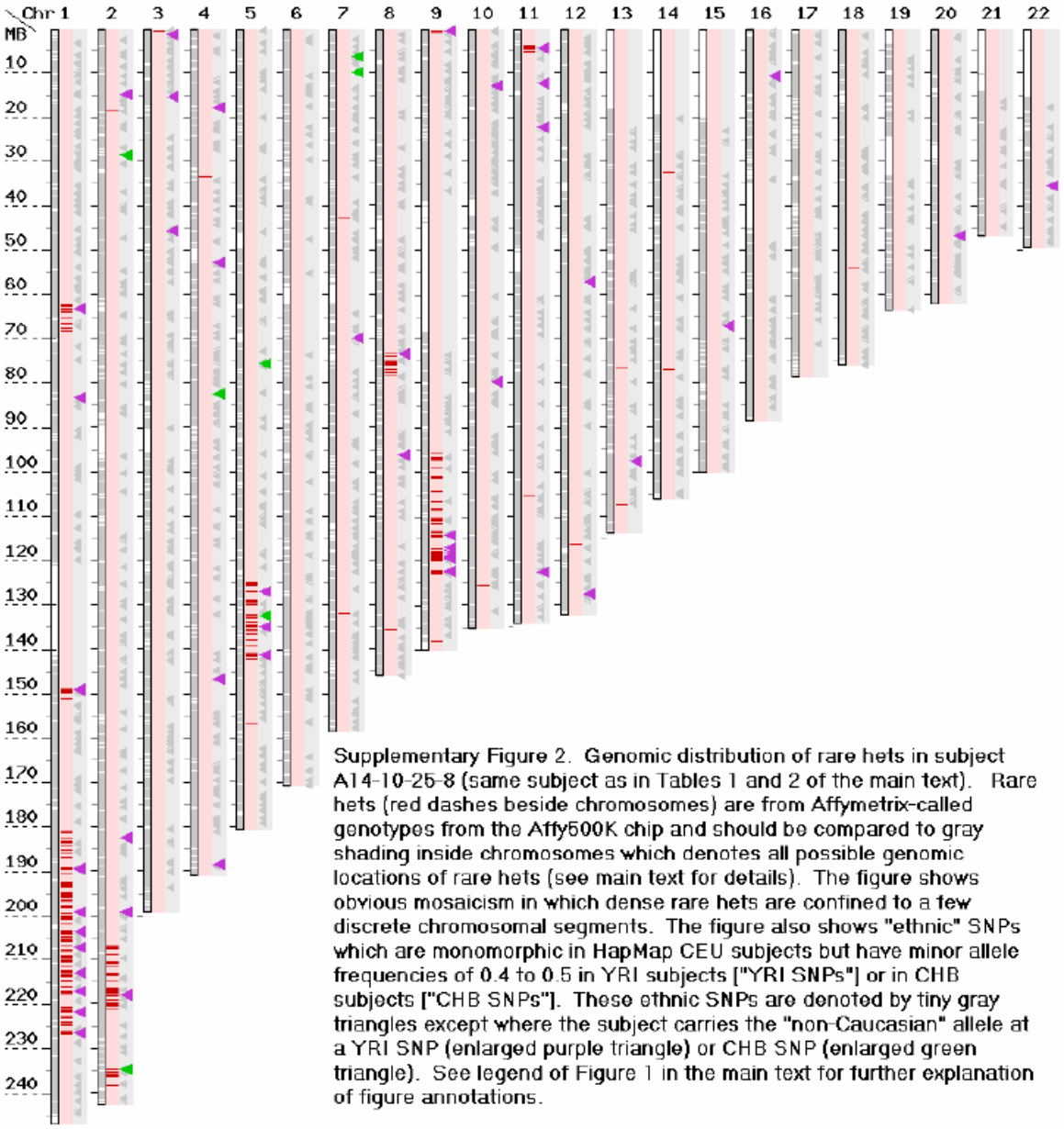
Thus each MAF bin's increase in false-positive rate (in column 2 of Table S10) is a *lower bound* for the SNPs being counted in the bin. As explained in footnote b of Table S10, a "1-fold" increase shown in column 2 means that, for every 1000 SNPs counted in the bin, there would be at least 1 additional SNP giving a p-value≤0.001 (implying an overall false positive rate of 2 SNPs per 1000 rather than correct type 1 error of approximately 1 SNP per 1000). Therefore, to determine a corresponding lower bound giving the number of additional GWA SNPs with p-values≤0.001 on the Affy500K or Illm550K chip for each MAF bin in Table S10, the number of Affy500K or Illm550K SNPs estimated to be in the bin was divided by 1000 and then multiplied by the "-fold" increase shown in column 2.

To estimate the number of HapMap SNPs falling into each bin, we examined 2122 "rare" HapMap SNPs on chromosome 20 with MAF<0.01 in HapMap CEU parents but MAF≥0.2 in HapMap YRI parents. Approximately 18% (386) of the 2122 rare SNPs had been genotyped on the Affy500K and/or Illum550K chips in the ~1400 58BC Caucasian controls. Of these 386 SNPs, approximately 73% (282) were polymorphic in the ~1400 Caucasian 58BC sample, enabling us to assign these SNPs

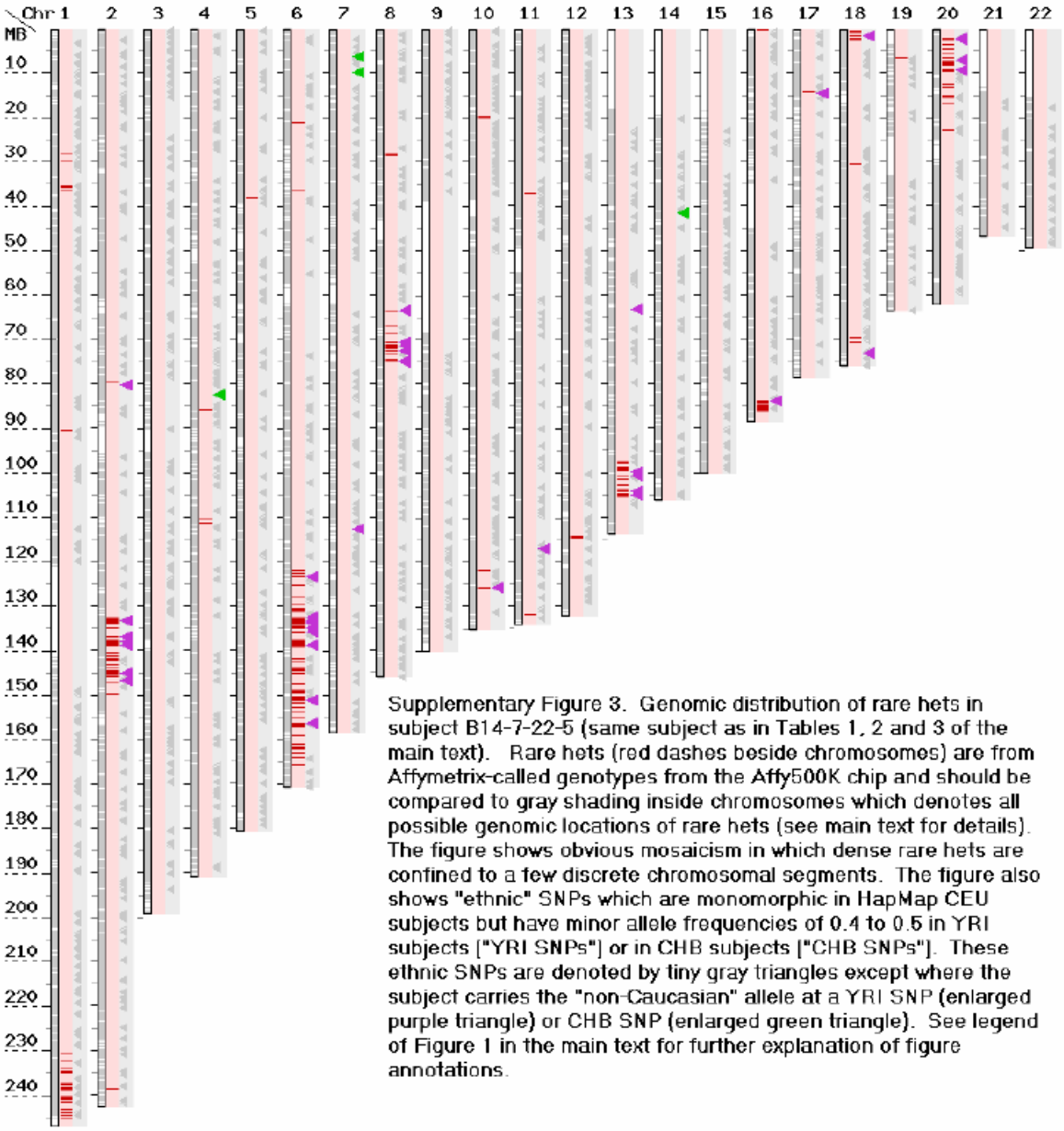
to a specific MAF bin. To estimate the number of *all* HapMap SNPs in each MAF bin of Table S10, we assumed that the same rare MAF spectrum (i.e. bin percentages) in 58BC Caucasians also applied to the ~82% (1732) of rare HapMap SNPs on chromosome 20 (MAF<0.01 in CEU, MAF≥0.2 in YRI) which were *not* genotyped on the Affy500K or Illm550K chips. In accordance with the rationale explained above, the chromosome 20 SNP numbers were multiplied by 44.6 to obtain the whole-genome estimates given in Table S10.



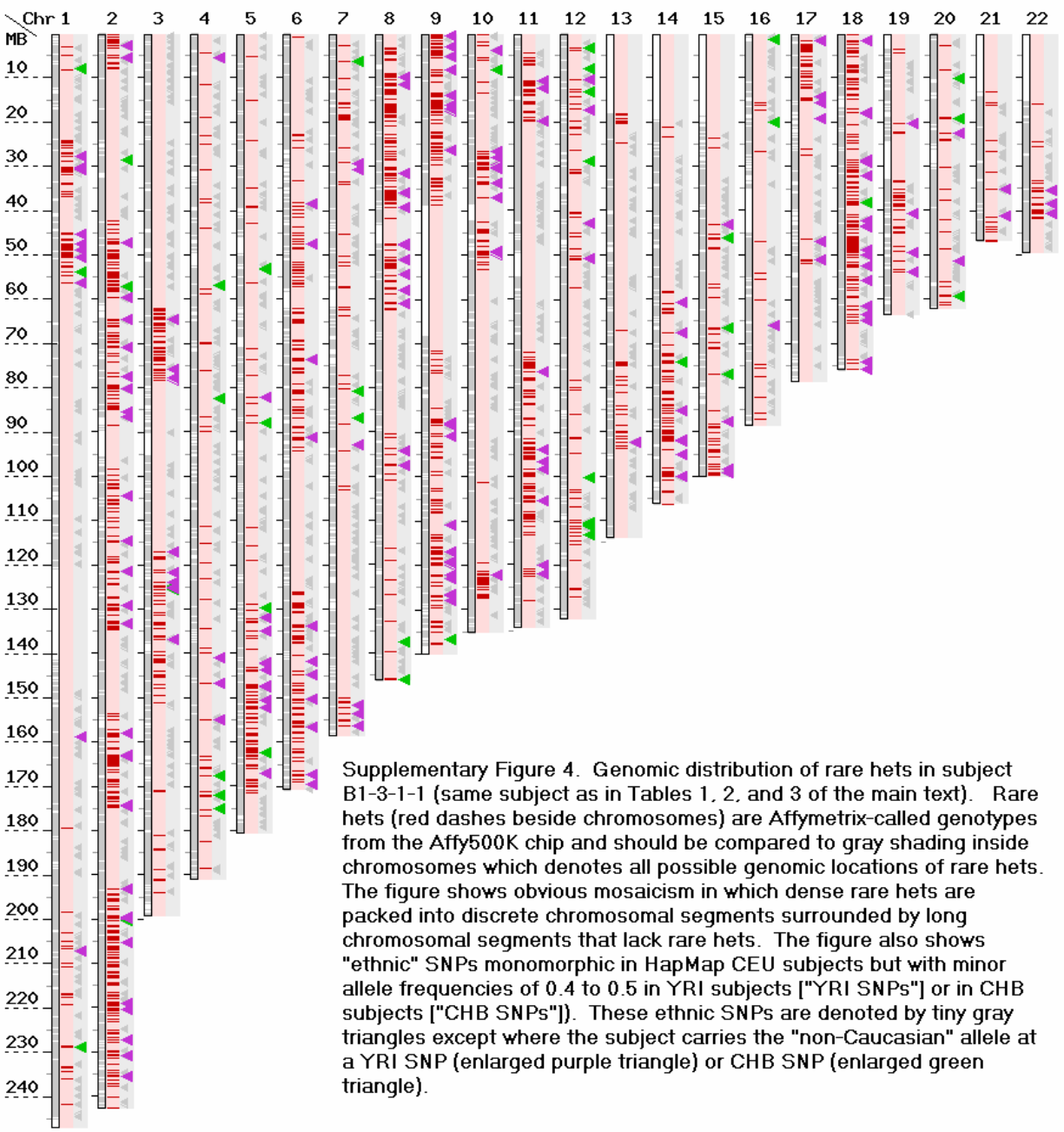
Supplementary Figure 1. Genomic distribution of rare hets in subject A19-9-30-7 (same subject as in Tables 1 and 2 of the main text). Rare hets (red dashes beside chromosomes) are from Affymetrix-called genotypes from the Affy500K chip and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (see main text for details). The figure shows obvious mosaicism in which dense rare hets are confined to a few discrete chromosomal segments. The figure also shows "ethnic" SNPs which are monomorphic in HapMap CEU subjects but have minor allele frequencies of 0.4 to 0.5 in YRI subjects ["YRI SNPs"] or in CHB subjects ["CHB SNPs"]. These ethnic SNPs are denoted by tiny gray triangles except where the subject carries the "non-Caucasian" allele at a YRI SNP (enlarged purple triangle) or CHB SNP (enlarged green triangle). See legend of Figure 1 in the main text for further explanation of figure annotations.



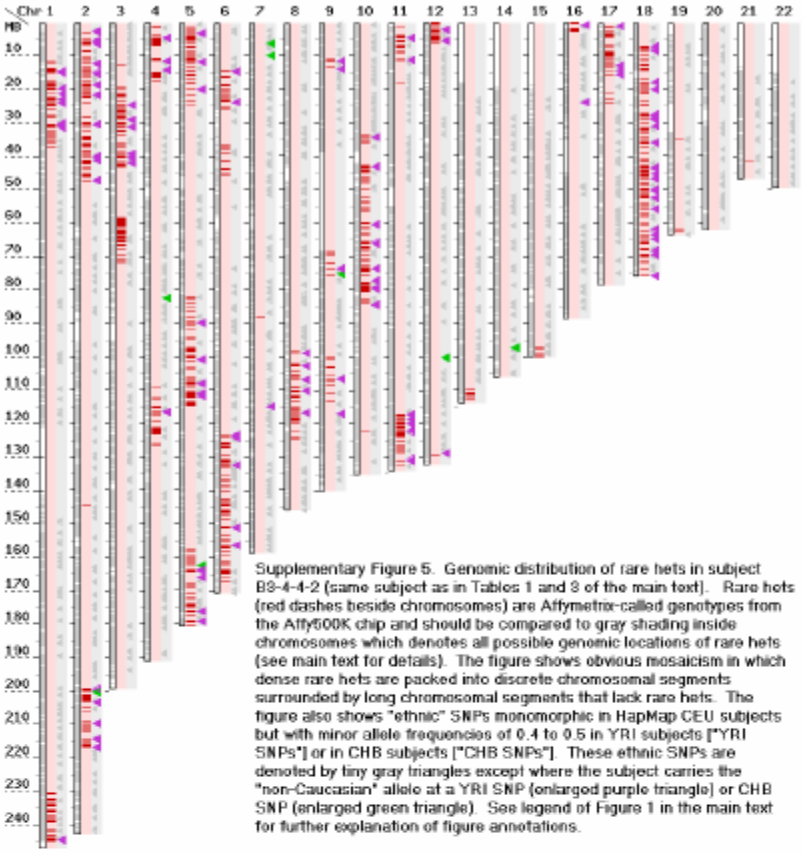
Supplementary Figure 2. Genomic distribution of rare hets in subject A14-10-25-8 (same subject as in Tables 1 and 2 of the main text). Rare hets (red dashes beside chromosomes) are from Affymetrix-called genotypes from the Affy500K chip and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (see main text for details). The figure shows obvious mosaicism in which dense rare hets are confined to a few discrete chromosomal segments. The figure also shows "ethnic" SNPs which are monomorphic in HapMap CEU subjects but have minor allele frequencies of 0.4 to 0.5 in YRI subjects ["YRI SNPs"] or in CHB subjects ["CHB SNPs"]. These ethnic SNPs are denoted by tiny gray triangles except where the subject carries the "non-Caucasian" allele at a YRI SNP (enlarged purple triangle) or CHB SNP (enlarged green triangle). See legend of Figure 1 in the main text for further explanation of figure annotations.



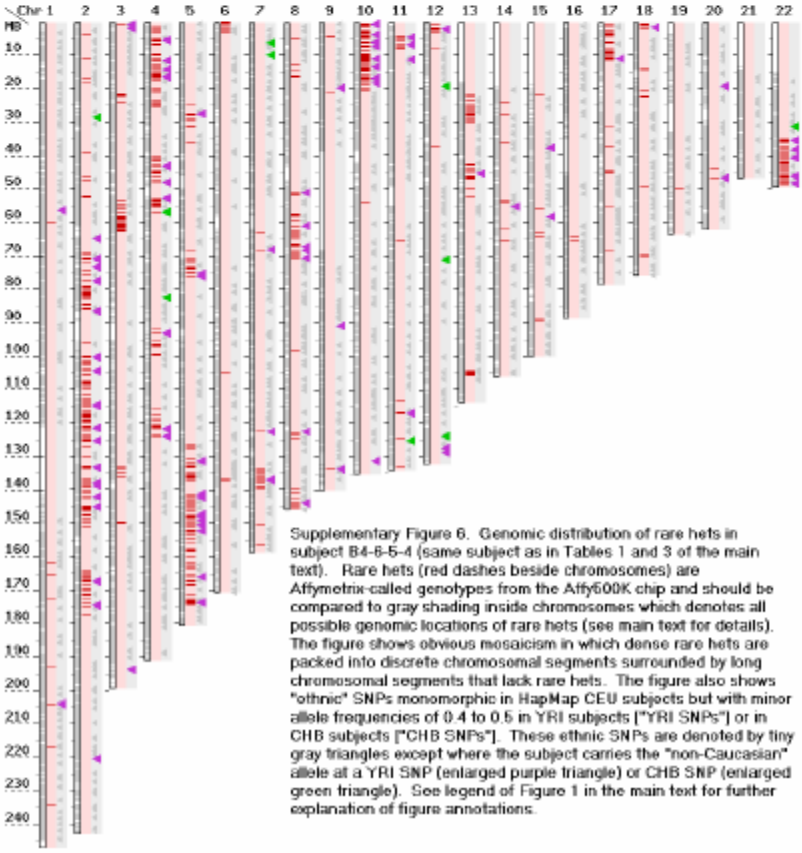
Supplementary Figure 3. Genomic distribution of rare hets in subject B14-7-22-5 (same subject as in Tables 1, 2 and 3 of the main text). Rare hets (red dashes beside chromosomes) are from Affymetrix-called genotypes from the Affy500K chip and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (see main text for details). The figure shows obvious mosaicism in which dense rare hets are confined to a few discrete chromosomal segments. The figure also shows "ethnic" SNPs which are monomorphic in HapMap CEU subjects but have minor allele frequencies of 0.4 to 0.5 in YRI subjects ["YRI SNPs"] or in CHB subjects ["CHB SNPs"]. These ethnic SNPs are denoted by tiny gray triangles except where the subject carries the "non-Caucasian" allele at a YRI SNP (enlarged purple triangle) or CHB SNP (enlarged green triangle). See legend of Figure 1 in the main text for further explanation of figure annotations.



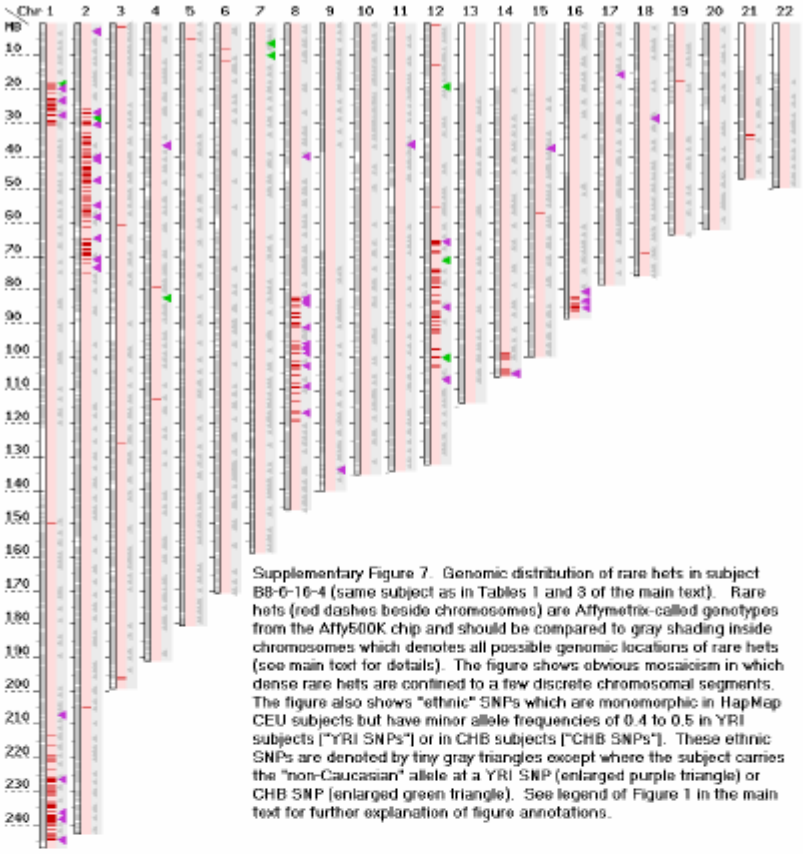
Supplementary Figure 4. Genomic distribution of rare hets in subject B1-3-1-1 (same subject as in Tables 1, 2, and 3 of the main text). Rare hets (red dashes beside chromosomes) are Affymetrix-called genotypes from the Affy500K chip and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets. The figure shows obvious mosaicism in which dense rare hets are packed into discrete chromosomal segments surrounded by long chromosomal segments that lack rare hets. The figure also shows "ethnic" SNPs monomorphic in HapMap CEU subjects but with minor allele frequencies of 0.4 to 0.5 in YRI subjects ["YRI SNPs"] or in CHB subjects ["CHB SNPs"]). These ethnic SNPs are denoted by tiny gray triangles except where the subject carries the "non-Caucasian" allele at a YRI SNP (enlarged purple triangle) or CHB SNP (enlarged green triangle).



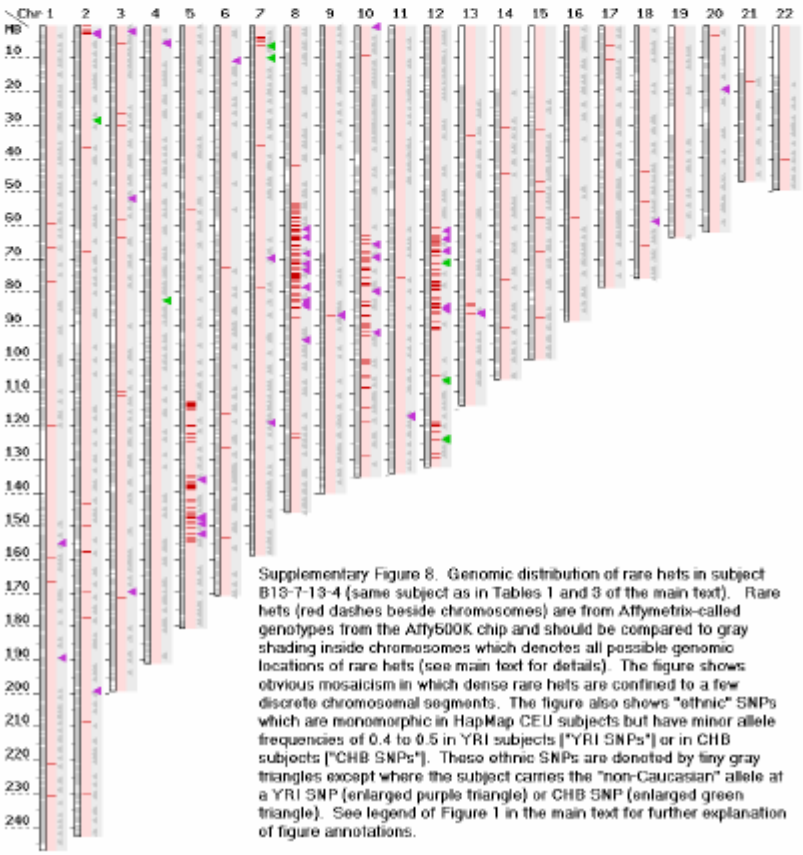
Supplementary Figure 5. Genomic distribution of rare hets in subject B3-4-4-2 (same subject as in Tables 1 and 3 of the main text). Rare hets (red dashes beside chromosomes) are Affymetrix-called genotypes from the Affy600K chip and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (see main text for details). The figure shows obvious mosaicism in which dense rare hets are packed into discrete chromosomal segments surrounded by long chromosomal segments that lack rare hets. The figure also shows "ethnic" SNPs monomorphic in HapMap CEU subjects but with minor allele frequencies of 0.4 to 0.5 in YRI subjects ["YRI SNPs"] or in CHB subjects ["CHB SNPs"]. These ethnic SNPs are denoted by tiny gray triangles except where the subject carries the "non-Caucasian" allele at a YRI SNP (enlarged purple triangle) or CHB SNP (enlarged green triangle). See legend of Figure 1 in the main text for further explanation of figure annotations.



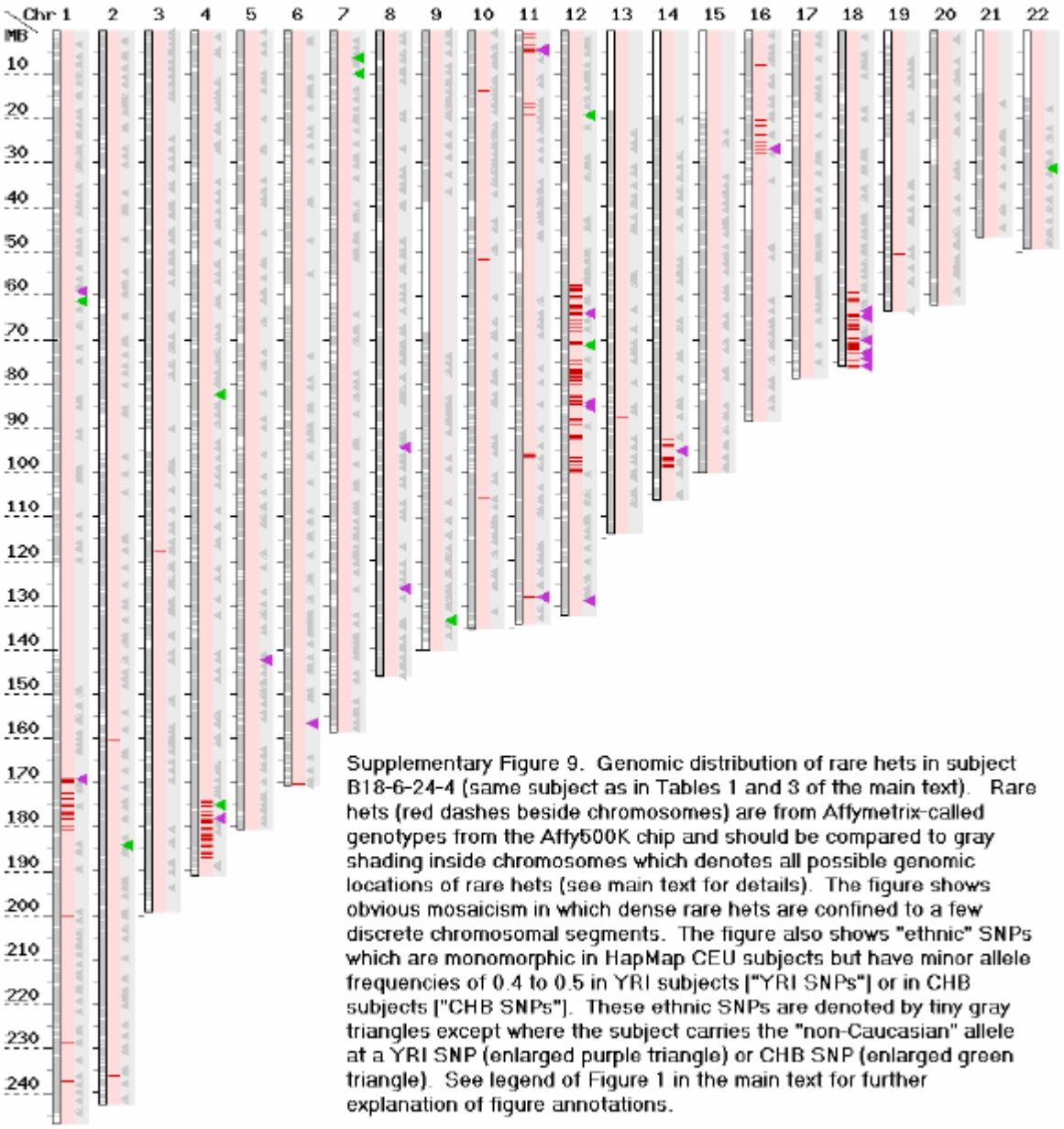
Supplementary Figure 6. Genomic distribution of rare hets in subject B4-6-5-4 (same subject as in Tables 1 and 3 of the main text). Rare hets (red dashes beside chromosomes) are Affymetrix-called genotypes from the Affy600K chip and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (see main text for details). The figure shows obvious mosaicism in which dense rare hets are packed into discrete chromosomal segments surrounded by long chromosomal segments that lack rare hets. The figure also shows "ethnic" SNPs monomorphic in HapMap CEU subjects but with minor allele frequencies of 0.4 to 0.5 in YRI subjects ["YRI SNPs"] or in CHB subjects ["CHB SNPs"]. These ethnic SNPs are denoted by tiny gray triangles except where the subject carries the "non-Caucasian" allele at a YRI SNP (enlarged purple triangle) or CHB SNP (enlarged green triangle). See legend of Figure 1 in the main text for further explanation of figure annotations.



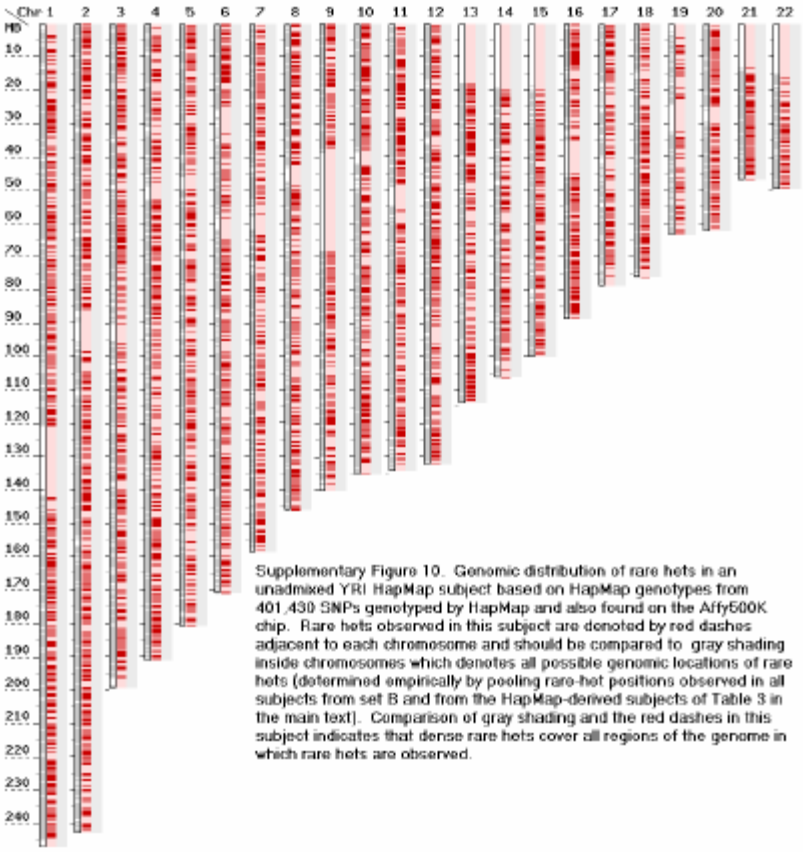
Supplementary Figure 7. Genomic distribution of rare hets in subject B8-6-16-4 (same subject as in Tables 1 and 3 of the main text). Rare hets (red dashes beside chromosomes) are Affymetrix-called genotypes from the Affy500K chip and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (see main text for details). The figure shows obvious mosaicism in which dense rare hets are confined to a few discrete chromosomal segments. The figure also shows "ethnic" SNPs which are monomorphic in HapMap CEU subjects but have minor allele frequencies of 0.4 to 0.5 in YRI subjects ["YRI SNPs"] or in CHB subjects ["CHB SNPs"]. These ethnic SNPs are denoted by tiny gray triangles except where the subject carries the "non-Caucasian" allele at a YRI SNP (enlarged purple triangle) or CHB SNP (enlarged green triangle). See legend of Figure 1 in the main text for further explanation of figure annotations.



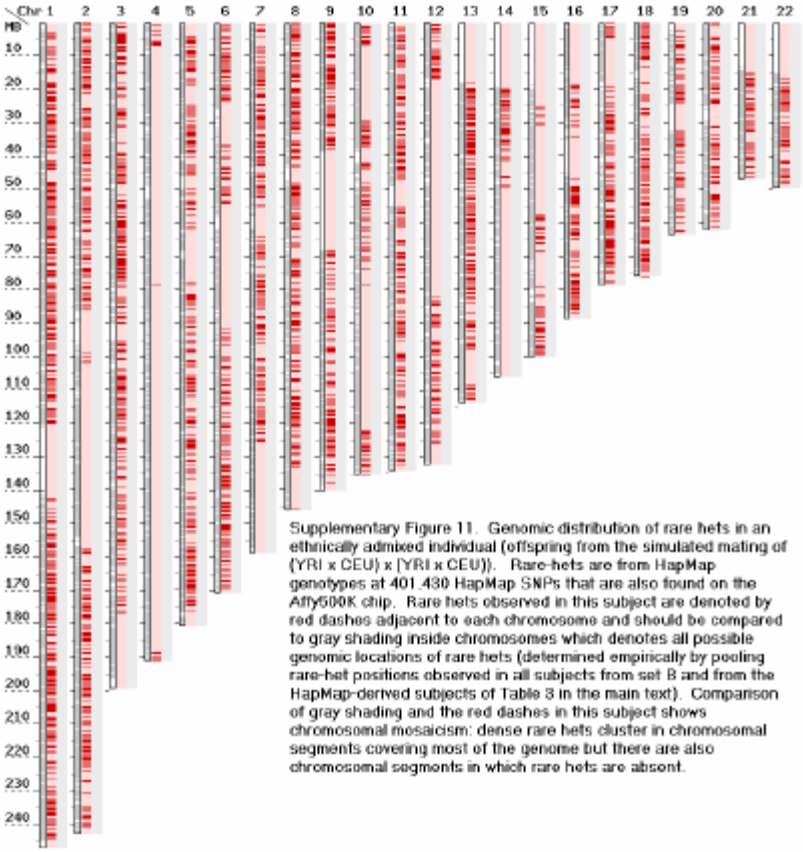
Supplementary Figure 8. Genomic distribution of rare hets in subject B13-7-13-4 (same subject as in Tables 1 and 3 of the main text). Rare hets (red dashes beside chromosomes) are from Affymetrix-called genotypes from the Affy500K chip and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (see main text for details). The figure shows obvious mosaicism in which dense rare hets are confined to a few discrete chromosomal segments. The figure also shows "ethnic" SNPs which are monomorphic in HapMap CEU subjects but have minor allele frequencies of 0.4 to 0.5 in YRI subjects ["YRI SNPs"] or in CHB subjects ["CHB SNPs"]. These ethnic SNPs are denoted by tiny gray triangles except where the subject carries the "non-Caucasian" allele at a YRI SNP (enlarged purple triangle) or CHB SNP (enlarged green triangle). See legend of Figure 1 in the main text for further explanation of figure annotations.



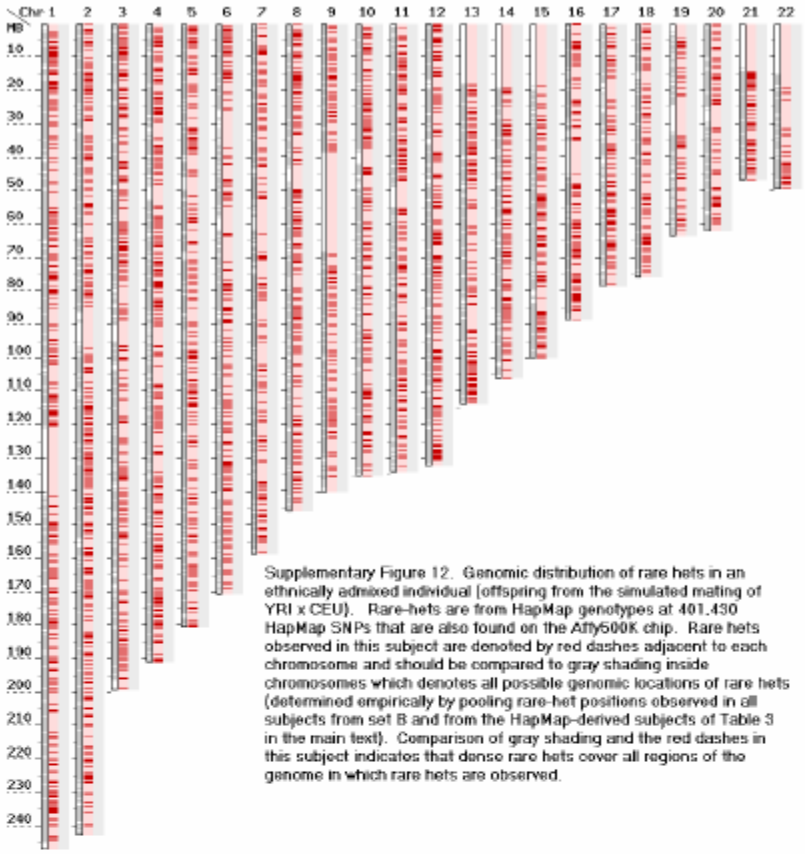
Supplementary Figure 9. Genomic distribution of rare hets in subject B18-6-24-4 (same subject as in Tables 1 and 3 of the main text). Rare hets (red dashes beside chromosomes) are from Affymetrix-called genotypes from the Affy500K chip and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (see main text for details). The figure shows obvious mosaicism in which dense rare hets are confined to a few discrete chromosomal segments. The figure also shows "ethnic" SNPs which are monomorphic in HapMap CEU subjects but have minor allele frequencies of 0.4 to 0.5 in YRI subjects ["YRI SNPs"] or in CHB subjects ["CHB SNPs"]. These ethnic SNPs are denoted by tiny gray triangles except where the subject carries the "non-Caucasian" allele at a YRI SNP (enlarged purple triangle) or CHB SNP (enlarged green triangle). See legend of Figure 1 in the main text for further explanation of figure annotations.



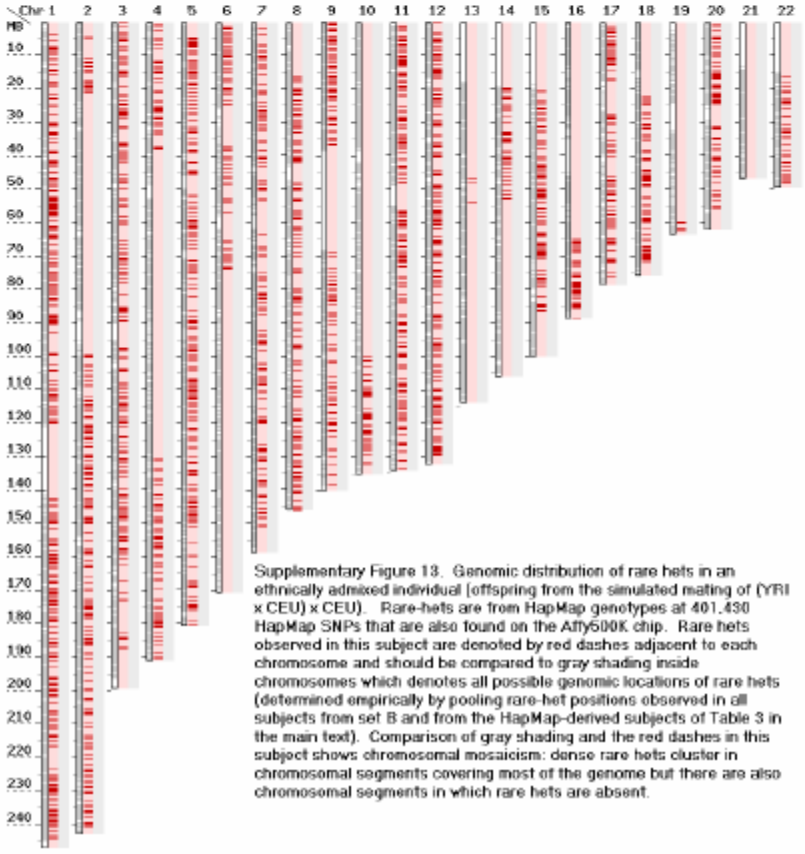
Supplementary Figure 10. Genomic distribution of rare hets in an unadmixed YRI HapMap subject based on HapMap genotypes from 401,430 SNPs genotyped by HapMap and also found on the Affy500K chip. Rare hets observed in this subject are denoted by red dashes adjacent to each chromosome and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (determined empirically by pooling rare-het positions observed in all subjects from set B and from the HapMap-derived subjects of Table 3 in the main text). Comparison of gray shading and the red dashes in this subject indicates that dense rare hets cover all regions of the genome in which rare hets are observed.



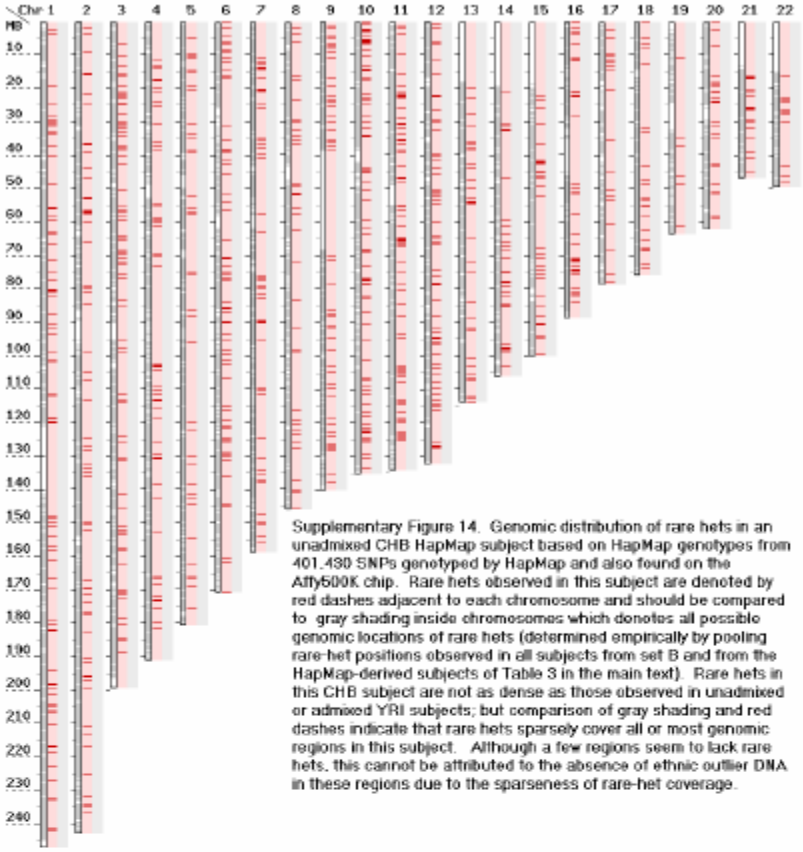
Supplementary Figure 11. Genomic distribution of rare hets in an ethnically admixed individual (offspring from the simulated mating of (YRI x CEU) x (YRI x CEU)). Rare-hets are from HapMap genotypes at 401,430 HapMap SNPs that are also found on the Affy500K chip. Rare hets observed in this subject are denoted by red dashes adjacent to each chromosome and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (determined empirically by pooling rare-het positions observed in all subjects from set B and from the HapMap-derived subjects of Table 3 in the main text). Comparison of gray shading and the red dashes in this subject shows chromosomal mosaicism: dense rare hets cluster in chromosomal segments covering most of the genome but there are also chromosomal segments in which rare hets are absent.



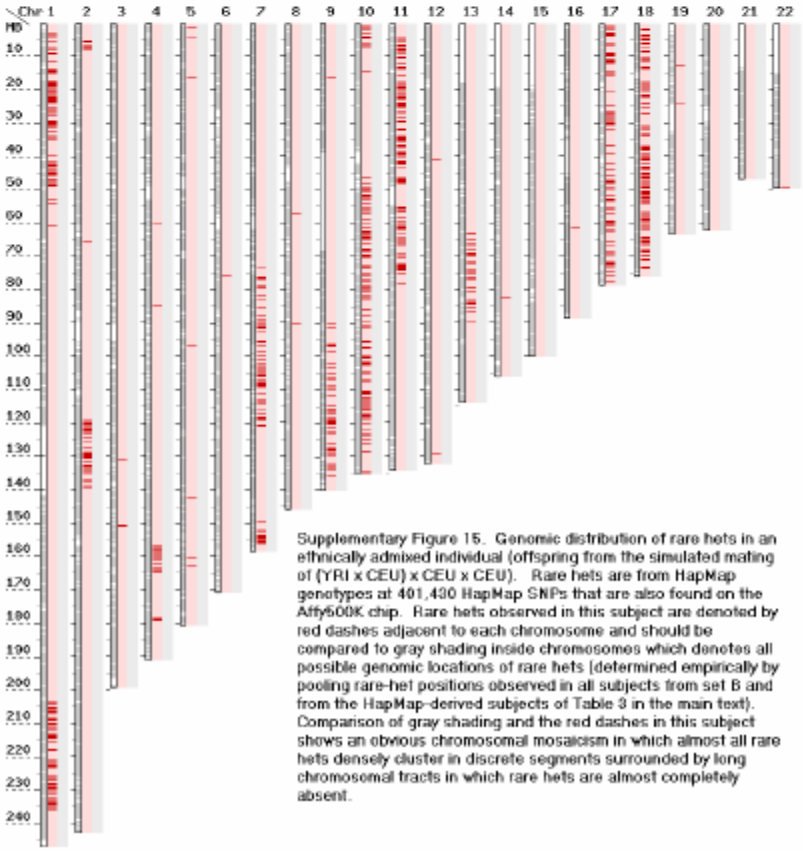
Supplementary Figure 12. Genomic distribution of rare hets in an ethnically admixed individual (offspring from the simulated mating of YRI x CEU). Rare-hets are from HapMap genotypes at 401,430 HapMap SNPs that are also found on the Affy500K chip. Rare hets observed in this subject are denoted by red dashes adjacent to each chromosome and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (determined empirically by pooling rare-het positions observed in all subjects from set B and from the HapMap-derived subjects of Table 3 in the main text). Comparison of gray shading and the red dashes in this subject indicates that dense rare hets cover all regions of the genome in which rare hets are observed.



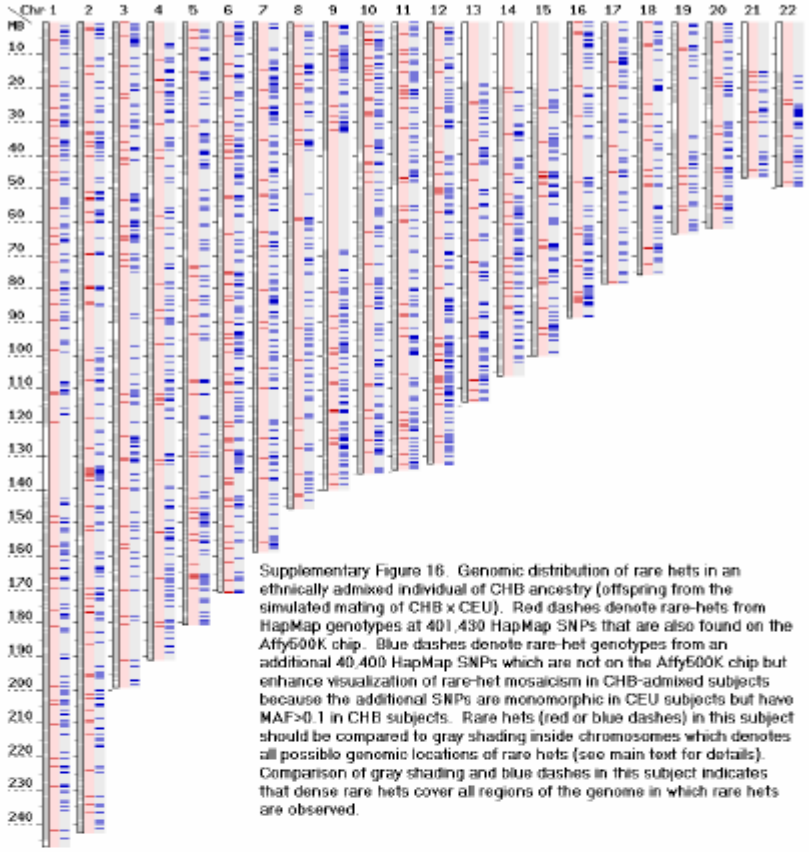
Supplementary Figure 13. Genomic distribution of rare hets in an ethnically admixed individual (offspring from the simulated mating of (YRI x CEU) x CEU). Rare-hets are from HapMap genotypes at 401,430 HapMap SNPs that are also found on the Affy500K chip. Rare hets observed in this subject are denoted by red dashes adjacent to each chromosome and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (determined empirically by pooling rare-het positions observed in all subjects from set B and from the HapMap-derived subjects of Table 3 in the main text). Comparison of gray shading and the red dashes in this subject shows chromosomal mosaicism: dense rare hets cluster in chromosomal segments covering most of the genome but there are also chromosomal segments in which rare hets are absent.



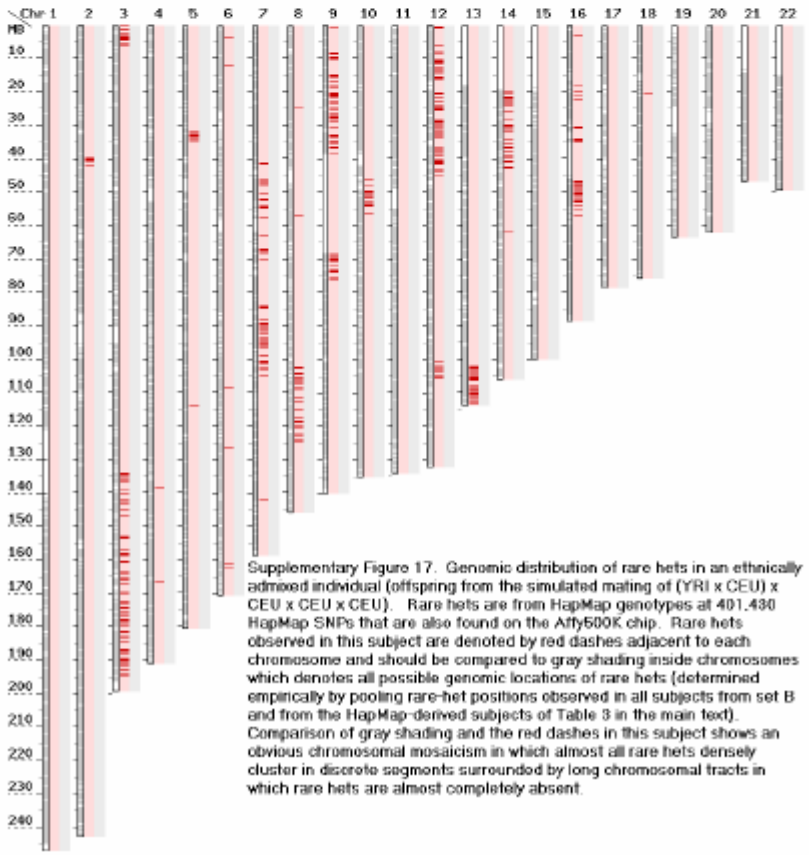
Supplementary Figure 14. Genomic distribution of rare hets in an unadmixed CHB HapMap subject based on HapMap genotypes from 401,430 SNPs genotyped by HapMap and also found on the Affy600K chip. Rare hets observed in this subject are denoted by red dashes adjacent to each chromosome and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (determined empirically by pooling rare-het positions observed in all subjects from set B and from the HapMap-derived subjects of Table 3 in the main text). Rare hets in this CHB subject are not as dense as those observed in unadmixed or admixed YRI subjects; but comparison of gray shading and red dashes indicate that rare hets sparsely cover all or most genomic regions in this subject. Although a few regions seem to lack rare hets, this cannot be attributed to the absence of ethnic outlier DNA in these regions due to the sparseness of rare-het coverage.



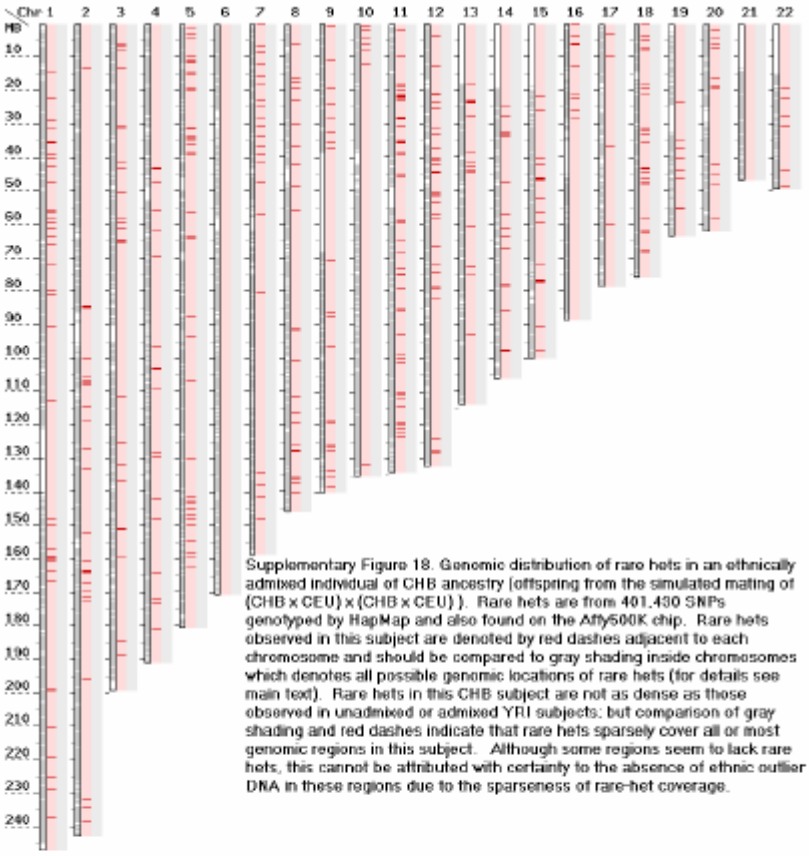
Supplementary Figure 15. Genomic distribution of rare hets in an ethnically admixed individual (offspring from the simulated mating of (YRI x CEU) x CEU x CEU). Rare hets are from HapMap genotypes at 401,430 HapMap SNPs that are also found on the Affy600K chip. Rare hets observed in this subject are denoted by red dashes adjacent to each chromosome and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (determined empirically by pooling rare-het positions observed in all subjects from set B and from the HapMap-derived subjects of Table 3 in the main text). Comparison of gray shading and the red dashes in this subject shows an obvious chromosomal mosaicism in which almost all rare hets densely cluster in discrete segments surrounded by long chromosomal tracts in which rare hets are almost completely absent.



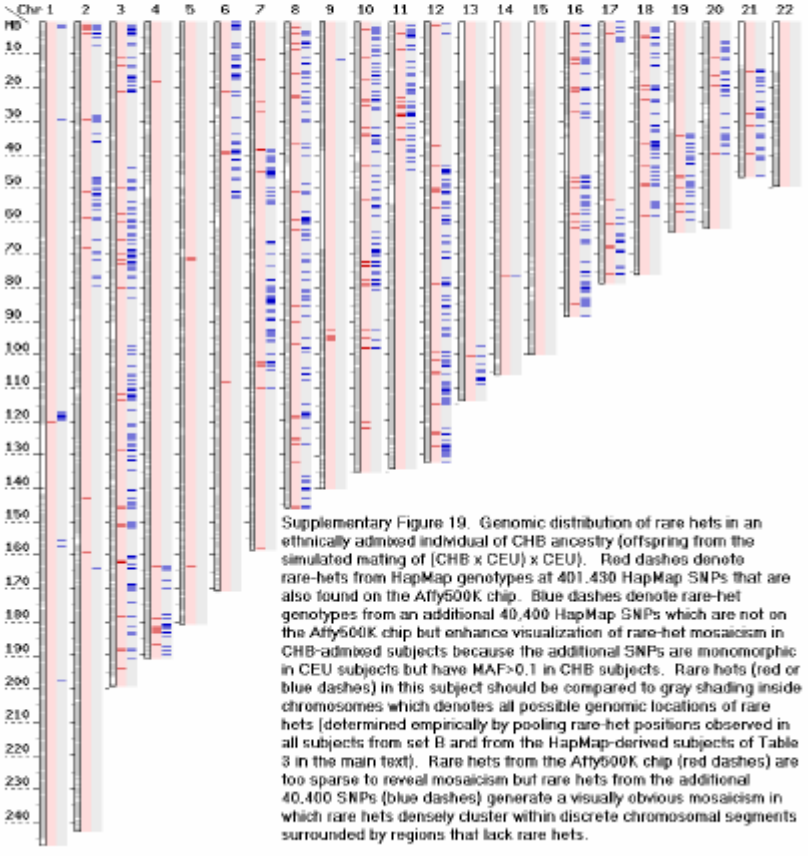
Supplementary Figure 16. Genomic distribution of rare hets in an ethnically admixed individual of CHB ancestry (offspring from the simulated mating of CHB x CEU). Red dashes denote rare-hets from HapMap genotypes at 401,430 HapMap SNPs that are also found on the Affy500K chip. Blue dashes denote rare-het genotypes from an additional 40,400 HapMap SNPs which are not on the Affy500K chip but enhance visualization of rare-het mosaicism in CHB-admixed subjects because the additional SNPs are monomorphic in CEU subjects but have MAF>0.1 in CHB subjects. Rare hets (red or blue dashes) in this subject should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (see main text for details). Comparison of gray shading and blue dashes in this subject indicates that dense rare hets cover all regions of the genome in which rare hets are observed.



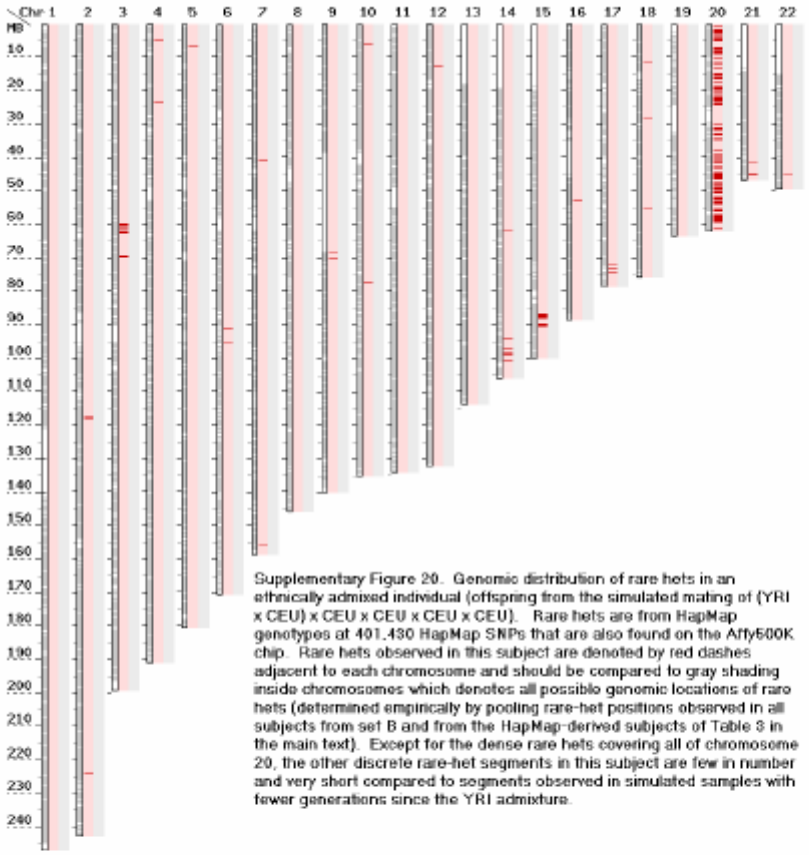
Supplementary Figure 17. Genomic distribution of rare hets in an ethnically admixed individual (offspring from the simulated mating of (YRI x CEU) x CEU x CEU x CEU). Rare hets are from HapMap genotypes at 401,430 HapMap SNPs that are also found on the Affy500K chip. Rare hets observed in this subject are denoted by red dashes adjacent to each chromosome and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (determined empirically by pooling rare-het positions observed in all subjects from set B and from the HapMap-derived subjects of Table 3 in the main text). Comparison of gray shading and the red dashes in this subject shows an obvious chromosomal mosaicism in which almost all rare hets densely cluster in discrete segments surrounded by long chromosomal tracts in which rare hets are almost completely absent.



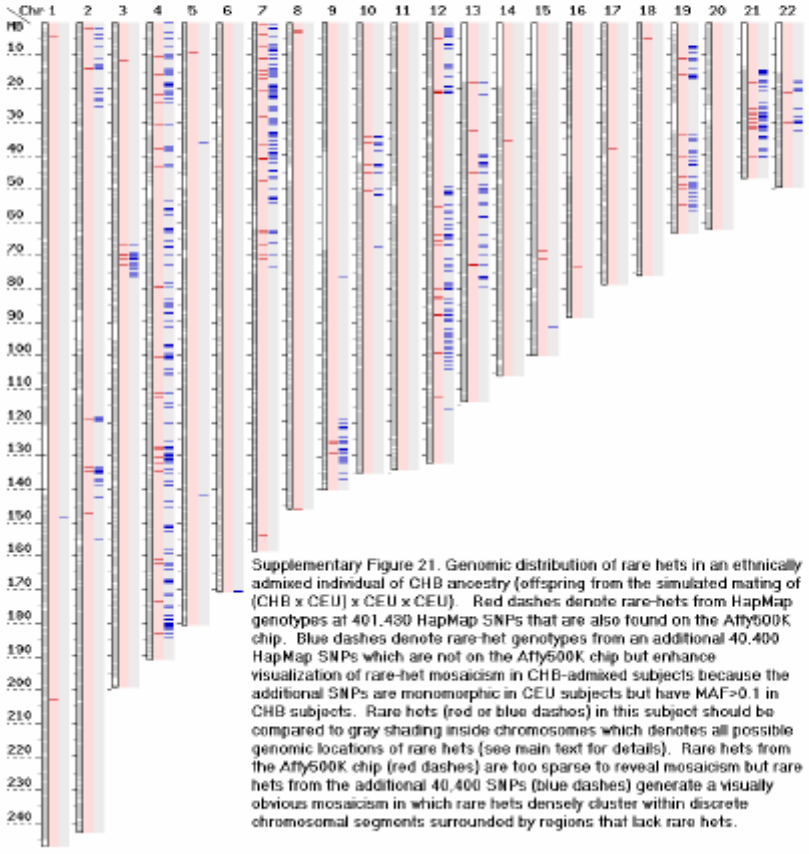
Supplementary Figure 18. Genomic distribution of rare hets in an ethnically admixed individual of CHB ancestry (offspring from the simulated mating of {CHB x CEU} x {CHB x CEU}). Rare hets are from 401,430 SNPs genotyped by HapMap and also found on the Affy500K chip. Rare hets observed in this subject are denoted by red dashes adjacent to each chromosome and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (for details see main text). Rare hets in this CHB subject are not as dense as those observed in unadmixed or admixed YRI subjects; but comparison of gray shading and red dashes indicate that rare hets sparsely cover all or most genomic regions in this subject. Although some regions seem to lack rare hets, this cannot be attributed with certainty to the absence of ethnic outlier DNA in these regions due to the sparseness of rare-het coverage.



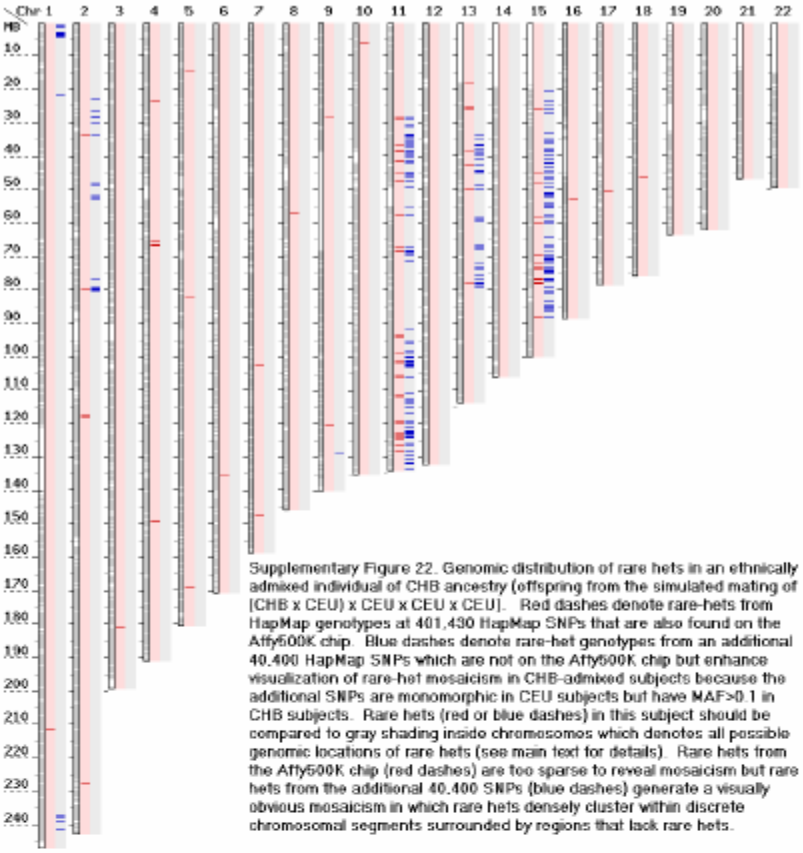
Supplementary Figure 19. Genomic distribution of rare hets in an ethnically admixed individual of CHB ancestry (offspring from the simulated mating of {CHB x CEU) x CEU). Red dashes denote rare-hets from HapMap genotypes at 401,430 HapMap SNPs that are also found on the Affy500K chip. Blue dashes denote rare-het genotypes from an additional 40,400 HapMap SNPs which are not on the Affy500K chip but enhance visualization of rare-het mosaicism in CHB-admixed subjects because the additional SNPs are monomorphic in CEU subjects but have MAF>0.1 in CHB subjects. Rare hets (red or blue dashes) in this subject should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (determined empirically by pooling rare-het positions observed in all subjects from set B and from the HapMap-derived subjects of Table 3 in the main text). Rare hets from the Affy500K chip (red dashes) are too sparse to reveal mosaicism but rare hets from the additional 40,400 SNPs (blue dashes) generate a visually obvious mosaicism in which rare hets densely cluster within discrete chromosomal segments surrounded by regions that lack rare hets.



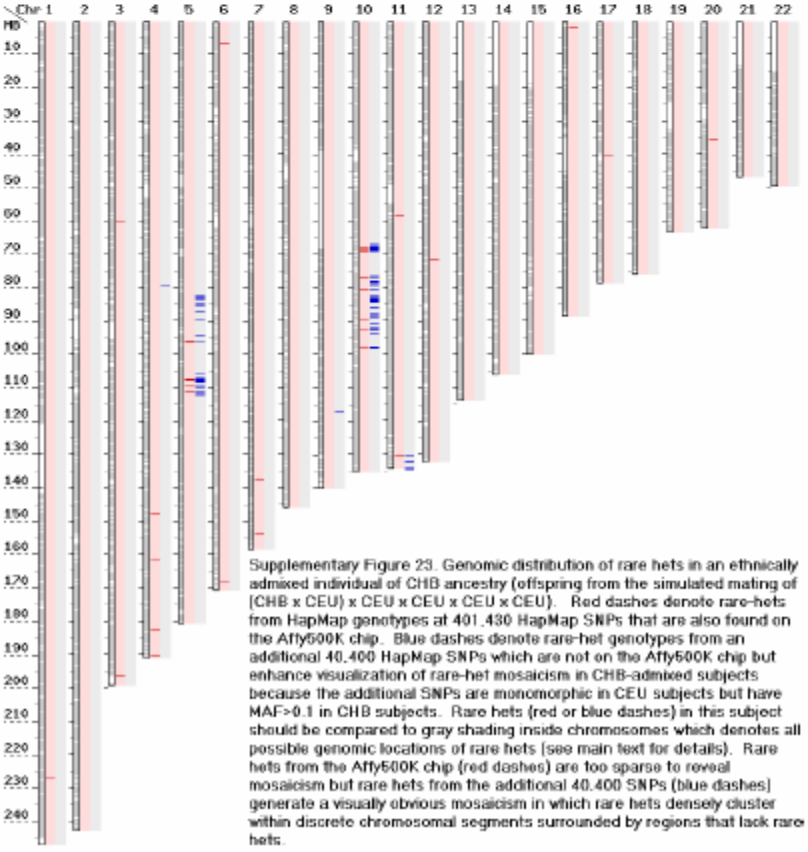
Supplementary Figure 20. Genomic distribution of rare hets in an ethnically admixed individual (offspring from the simulated mating of (YRI x CEU) x CEU x CEU x CEU x CEU). Rare hets are from HapMap genotypes at 401,430 HapMap SNPs that are also found on the Affy500K chip. Rare hets observed in this subject are denoted by red dashes adjacent to each chromosome and should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (determined empirically by pooling rare-het positions observed in all subjects from set B and from the HapMap-derived subjects of Table 3 in the main text). Except for the dense rare hets covering all of chromosome 20, the other discrete rare-het segments in this subject are few in number and very short compared to segments observed in simulated samples with fewer generations since the YRI admixture.



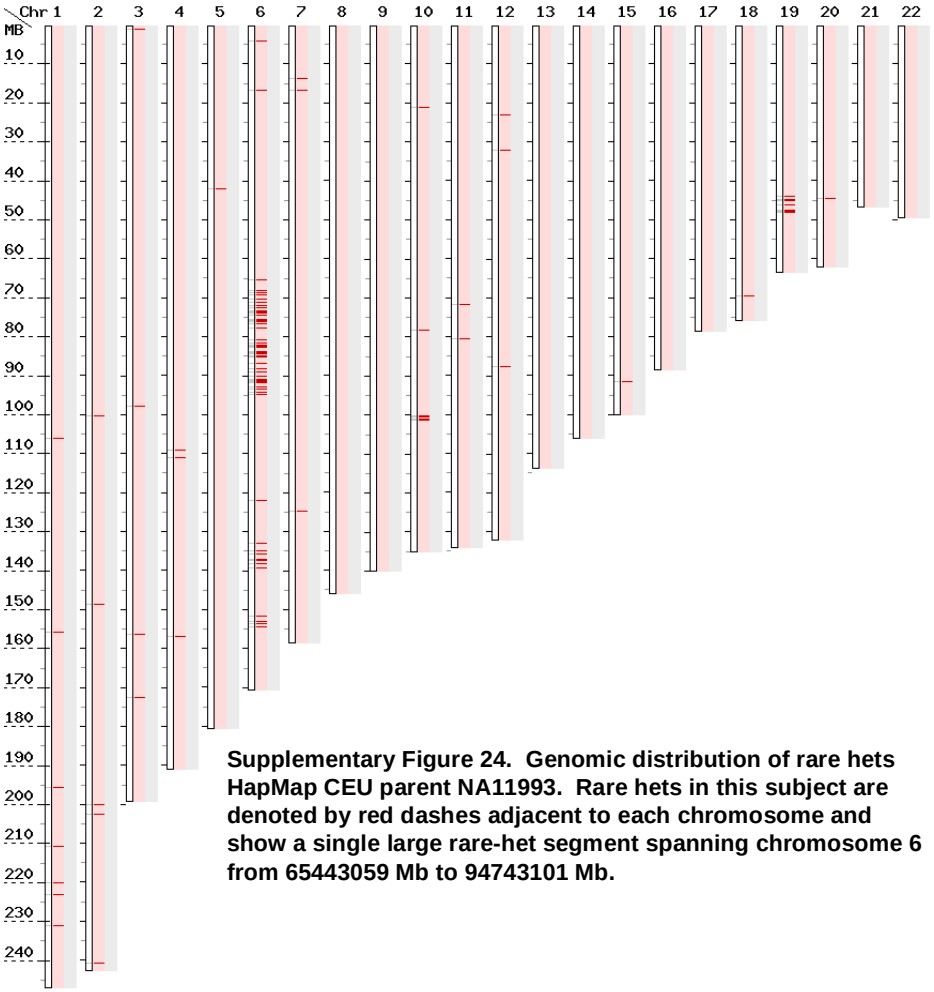
Supplementary Figure 21. Genomic distribution of rare hets in an ethnically admixed individual of CHB ancestry (offspring from the simulated mating of (CHB x CEU) x CEU x CEU). Red dashes denote rare-hets from HapMap genotypes at 401,430 HapMap SNPs that are also found on the Affy500K chip. Blue dashes denote rare-het genotypes from an additional 40,400 HapMap SNPs which are not on the Affy500K chip but enhance visualization of rare-het mosaicism in CHB-admixed subjects because the additional SNPs are monomorphic in CEU subjects but have MAF>0.1 in CHB subjects. Rare hets (red or blue dashes) in this subject should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (see main text for details). Rare hets from the Affy500K chip (red dashes) are too sparse to reveal mosaicism but rare hets from the additional 40,400 SNPs (blue dashes) generate a visually obvious mosaicism in which rare hets densely cluster within discrete chromosomal segments surrounded by regions that lack rare hets.



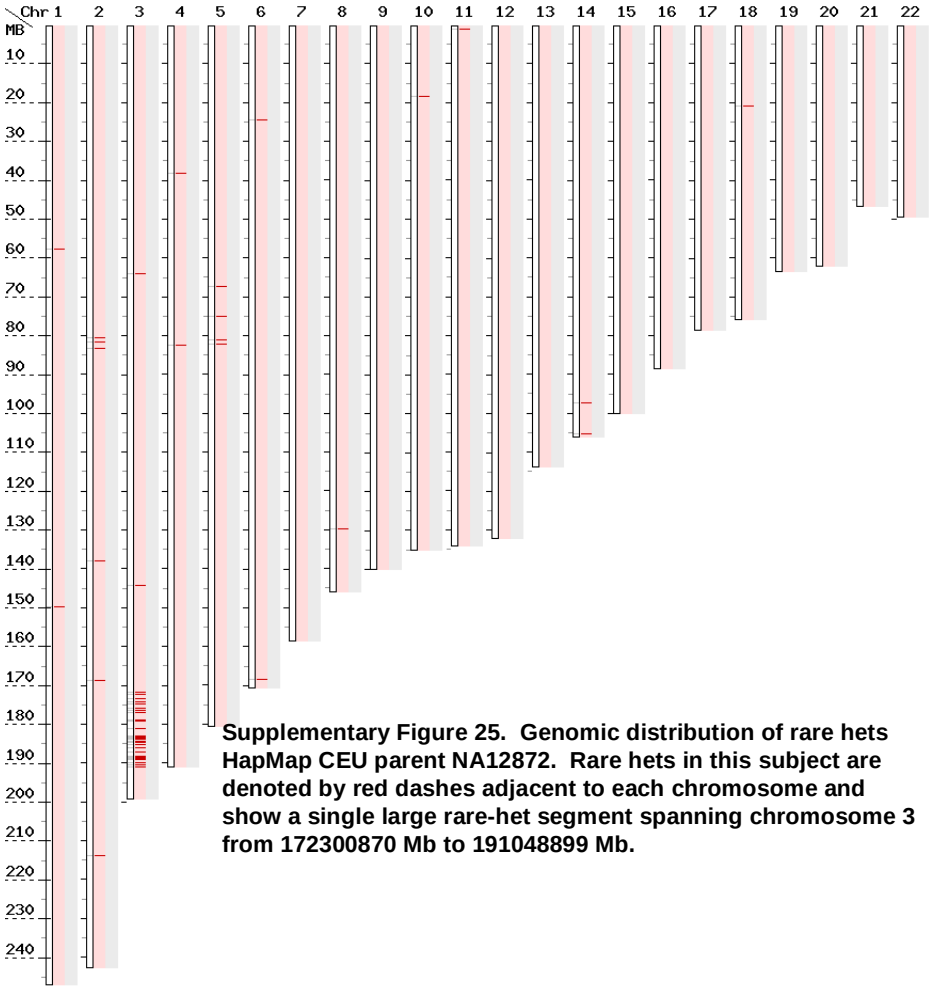
Supplementary Figure 22. Genomic distribution of rare hets in an ethnically admixed individual of CHB ancestry (offspring from the simulated mating of [CHB x CEU) x CEU x CEU x CEU]. Red dashes denote rare-hets from HapMap genotypes at 401,430 HapMap SNPs that are also found on the Affy500K chip. Blue dashes denote rare-het genotypes from an additional 40,400 HapMap SNPs which are not on the Affy500K chip but enhance visualization of rare-het mosaicism in CHB-admixed subjects because the additional SNPs are monomorphic in CEU subjects but have MAF>0.1 in CHB subjects. Rare hets (red or blue dashes) in this subject should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (see main text for details). Rare hets from the Affy500K chip (red dashes) are too sparse to reveal mosaicism but rare hets from the additional 40,400 SNPs (blue dashes) generate a visually obvious mosaicism in which rare hets densely cluster within discrete chromosomal segments surrounded by regions that lack rare hets.



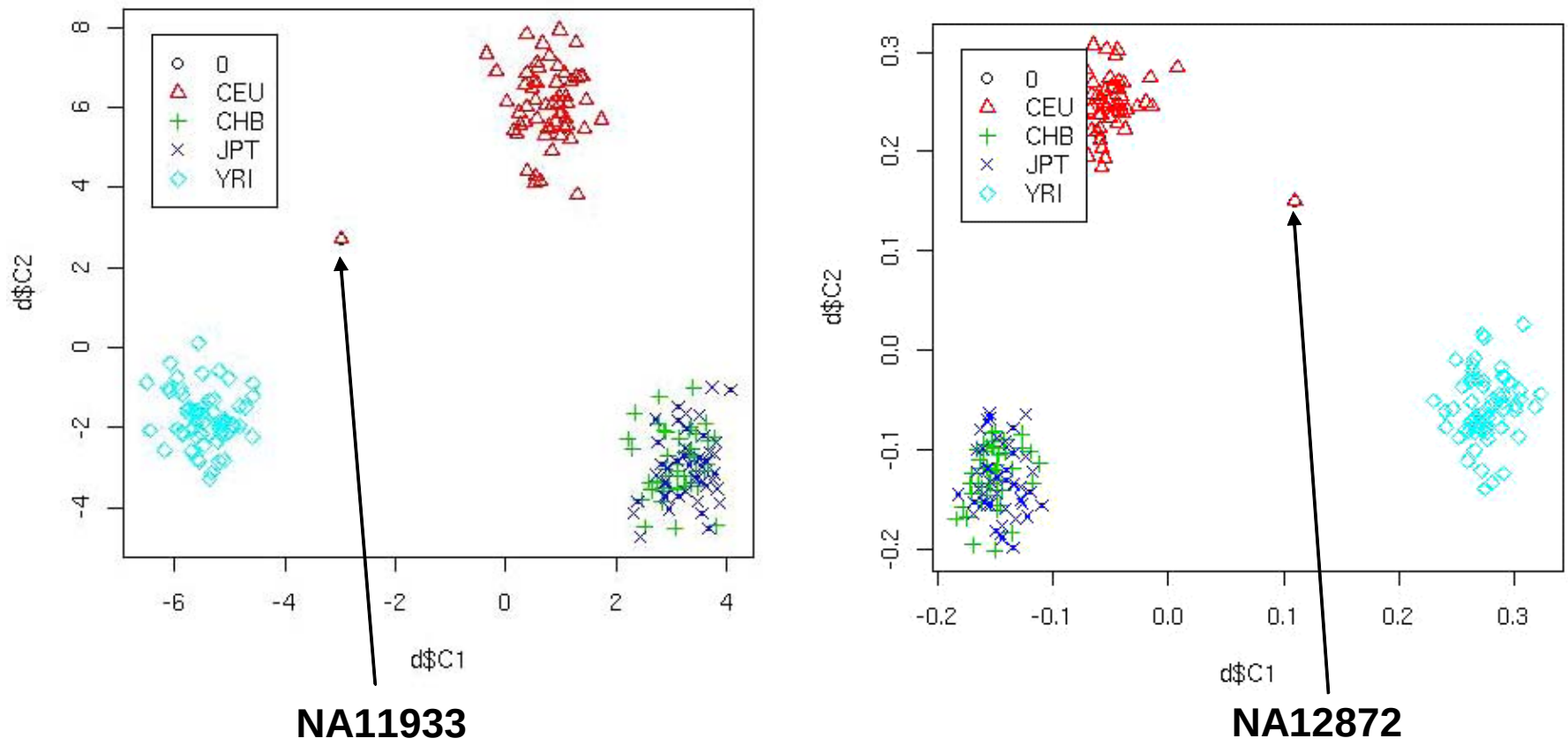
Supplementary Figure 23. Genomic distribution of rare hets in an ethnically admixed individual of CHB ancestry (offspring from the simulated mating of [CHB x CEU) x CEU x CEU x CEU x CEU]. Red dashes denote rare-hets from HapMap genotypes at 401,430 HapMap SNPs that are also found on the Affy500K chip. Blue dashes denote rare-het genotypes from an additional 40,400 HapMap SNPs which are not on the Affy500K chip but enhance visualization of rare-het mosaicism in CHB-admixed subjects because the additional SNPs are monomorphic in CEU subjects but have MAF>0.1 in CHB subjects. Rare hets (red or blue dashes) in this subject should be compared to gray shading inside chromosomes which denotes all possible genomic locations of rare hets (see main text for details). Rare hets from the Affy500K chip (red dashes) are too sparse to reveal mosaicism but rare hets from the additional 40,400 SNPs (blue dashes) generate a visually obvious mosaicism in which rare hets densely cluster within discrete chromosomal segments surrounded by regions that lack rare hets.



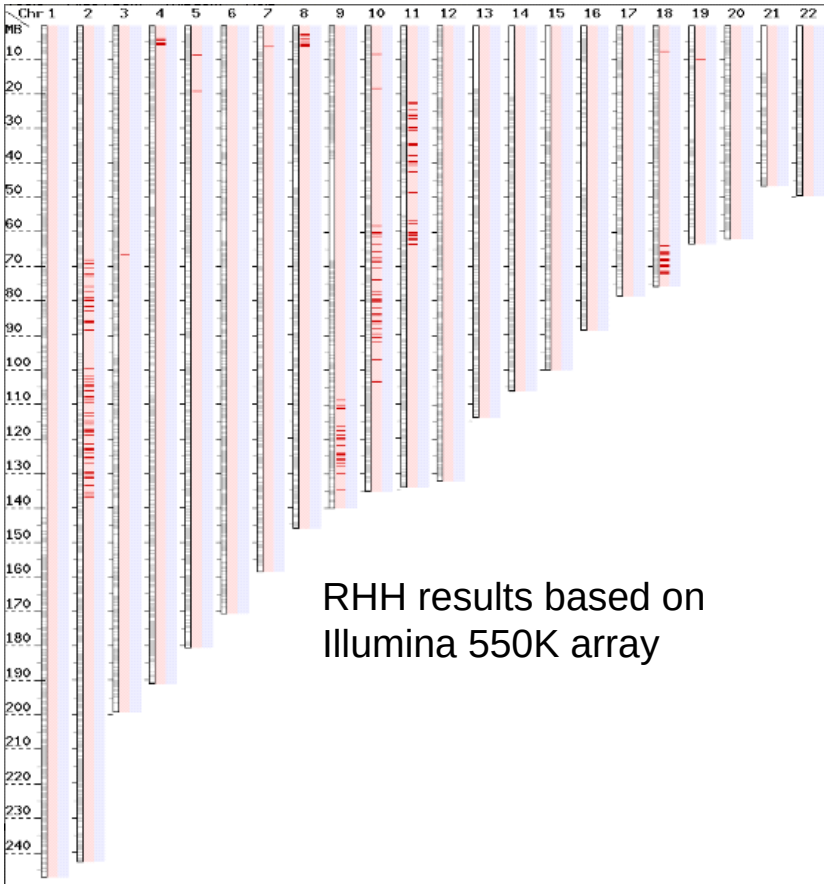
Supplementary Figure 24. Genomic distribution of rare hets HapMap CEU parent NA11993. Rare hets in this subject are denoted by red dashes adjacent to each chromosome and show a single large rare-het segment spanning chromosome 6 from 65443059 Mb to 94743101 Mb.



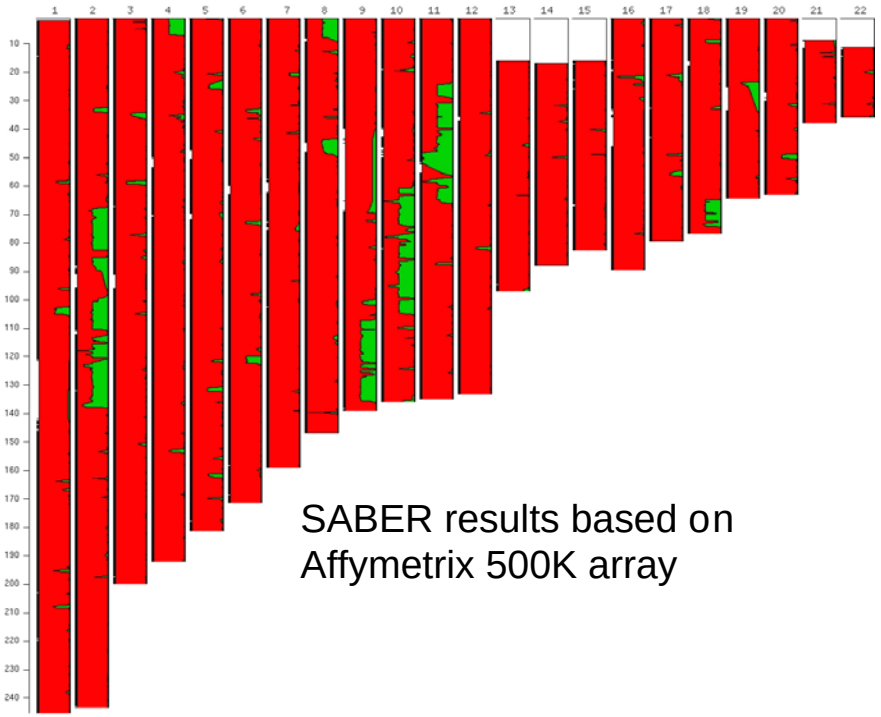
Supplementary Figure 25. Genomic distribution of rare hets HapMap CEU parent NA12872. Rare hets in this subject are denoted by red dashes adjacent to each chromosome and show a single large rare-het segment spanning chromosome 3 from 172300870 Mb to 191048899 Mb.



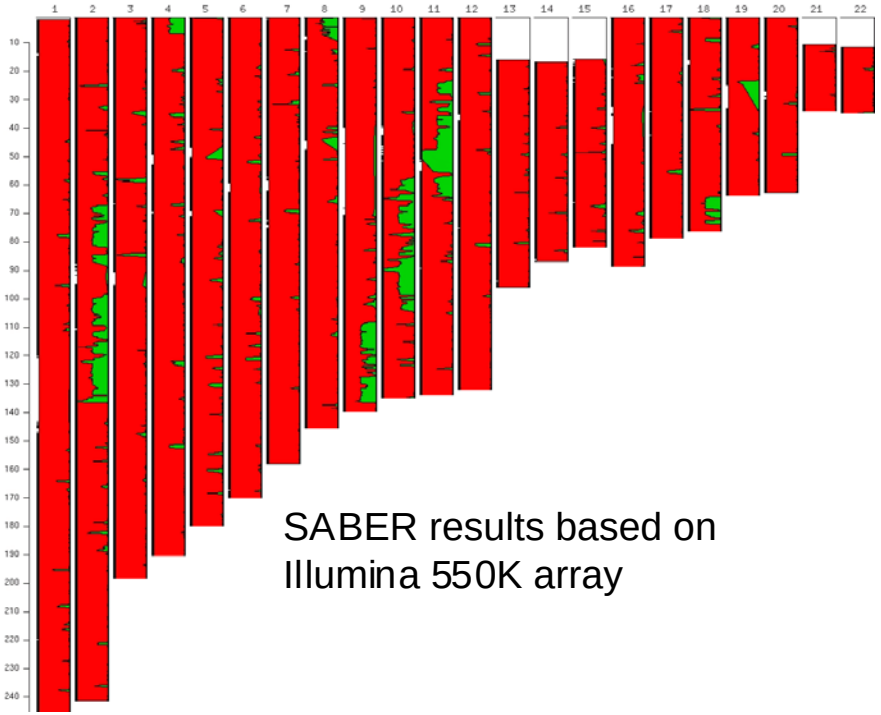
Supplementary Figure 26. HapMap samples plotted for the first two principle components of multidimensional scaling (MDS) of pairwiise identity-by-state (IBS) genotype distances between samples. In the left panel, genotypes were evaluated only from the region of a single rare-het segment in subject NA11993 (chrm. 6 from 65443059 to 94743101 Mb). In the right panel, genotypes were evaluated only from a single rare-het segment in subject NA12872 (chrm 3 from 172300870 to 191048899 Mb). In both panels, each HapMap ethnic group is tightly clustered with all others of the same ethnicity except for the HapMap CEU parent that carries the rare-het segment (NA11993, NA12872) who is located halfway between the Caucasians (CEU) and the Yorubans (YRI). This implies that each rare-het segment is likely to contain admixed non-Caucasian DNA of African origin.



RHH results based on
Illumina 550K array

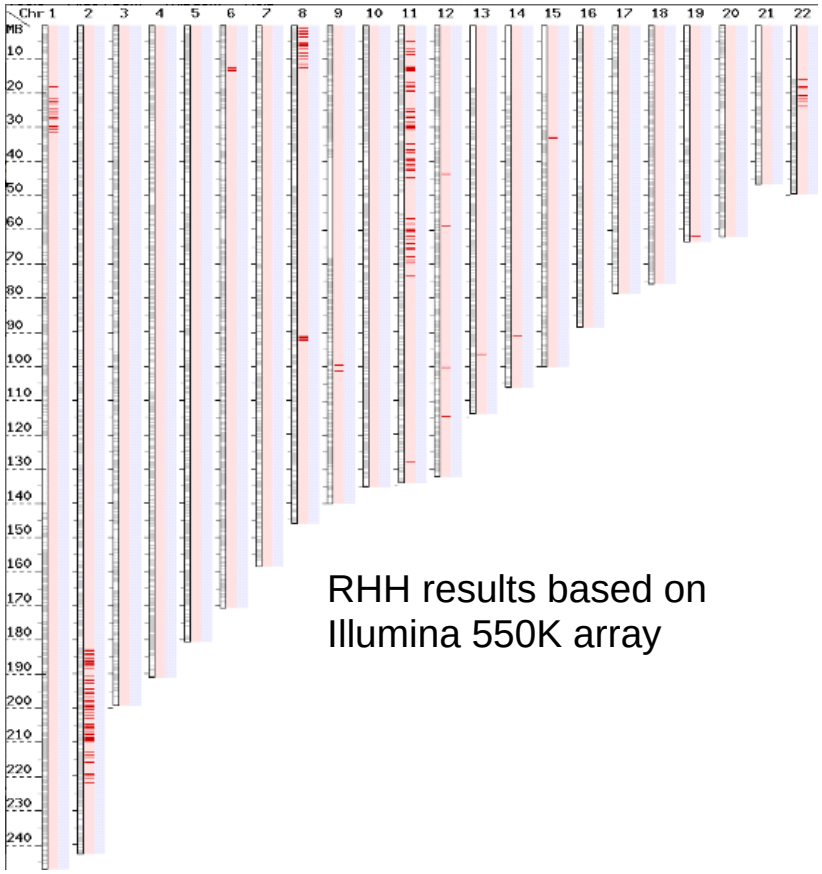


SABER results based on
Affymetrix 500K array

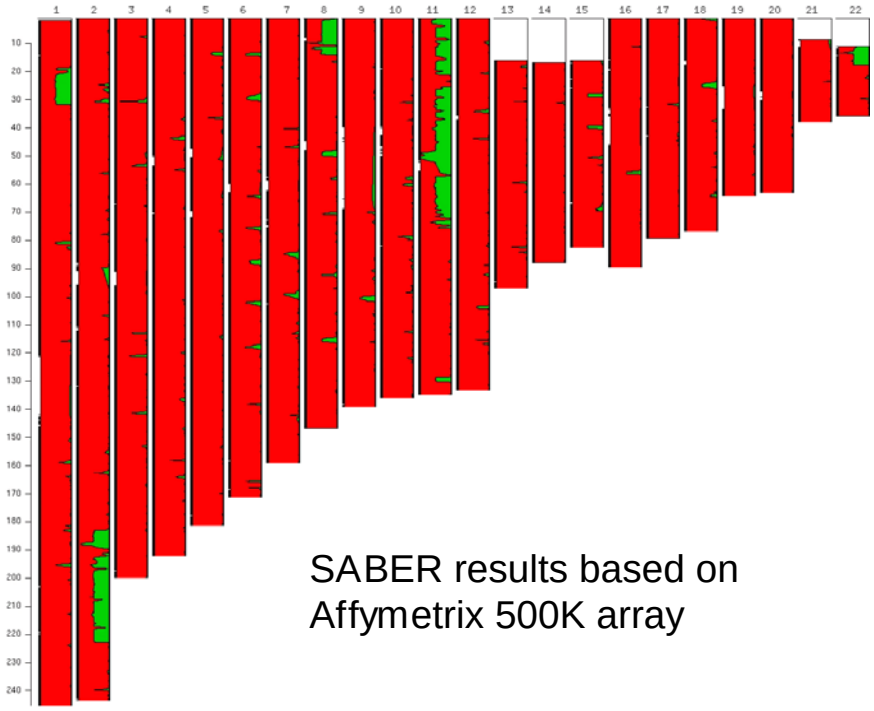


SABER results based on
Illumina 550K array

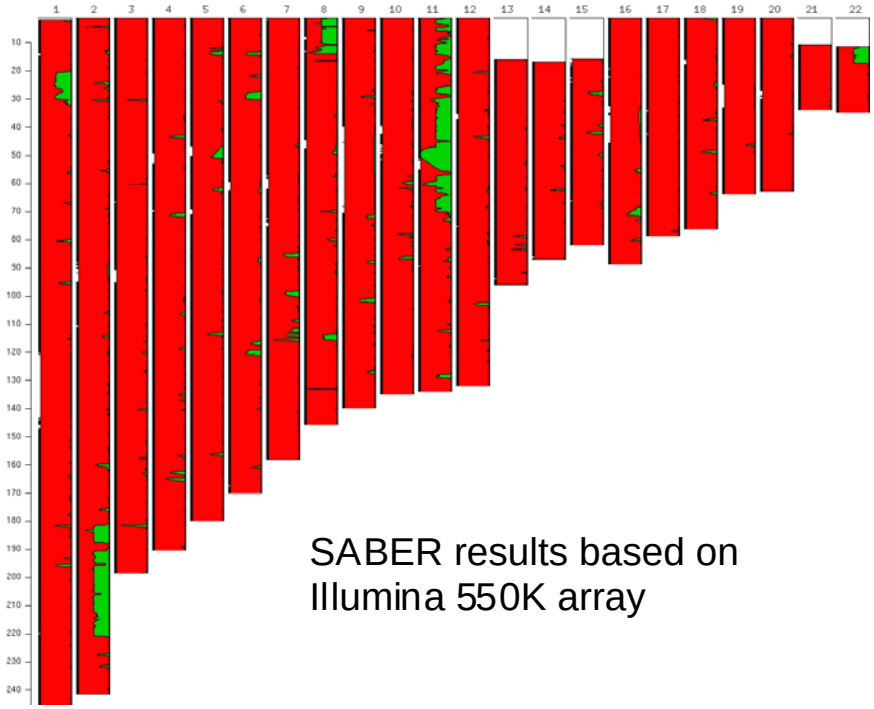
Supplementary Figure S27. Chromosome mosaicism in subject B5-4-12-2 (see Figure 1b and Table 1) as visualized by RHH using the Illm550K array or by SABER using the Affy500K or Illm550K arrays. In the RHH visualization, rare hets are denoted by red dashes next to a chromosome such that clusters of rare hets mark chromosomal segments of non-Caucasian DNA. In the corresponding visualizations produced by SABER software, green shading marks inferred genomic location(s) of African DNA and red marks inferred Caucasian DNA such that a roughly equal left-right split between red and green implies a genomic region in which DNA on one homologous chromosome is Caucasian while DNA on the other chromosome is African. In the SABER visualizations, centromeric regions without SNP genotyping are denoted by a "gap" in the lefthand boundary of the chromosome whereas the same regions in the RHH visualizations are denoted by extended absence of gray shading *inside* chromosomes. Comparison of this figure with Figure 1b shows that the overall pattern of mosaicism for subject B5-4-12-2 is almost identical for RHH and Saber using either the Affy500K or Illm550K chip. However, as described in the Discussion, there are minor differences in the results that may impact visualization of very short (<3Mb long) segments of admixture.



RHH results based on
Illumina 550K array

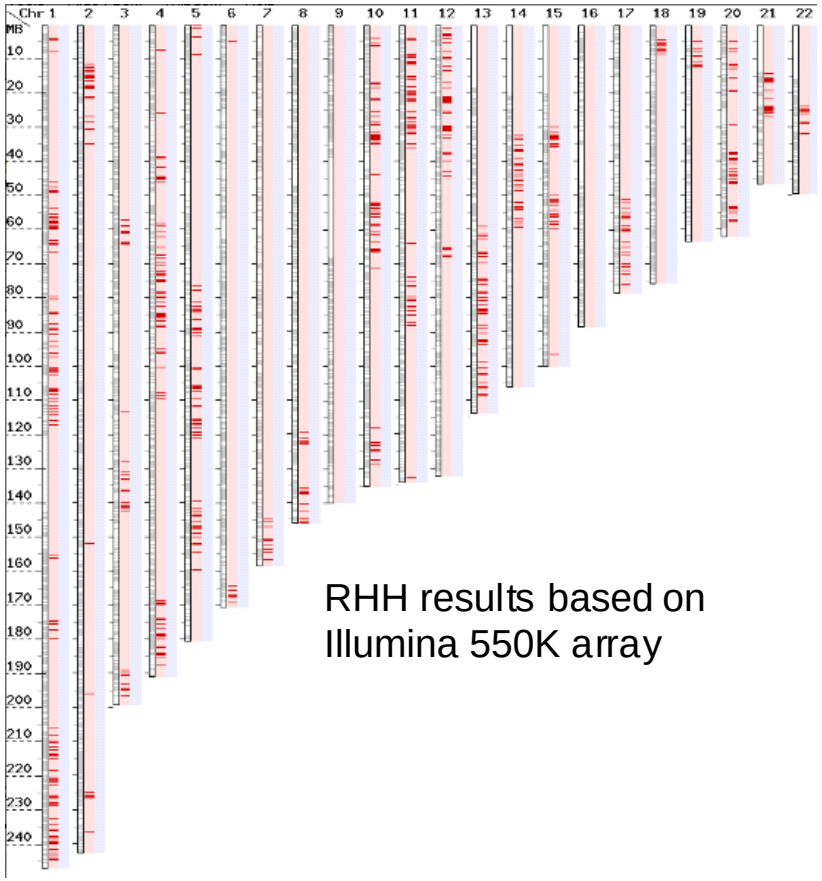


SABER results based on
Affymetrix 500K array

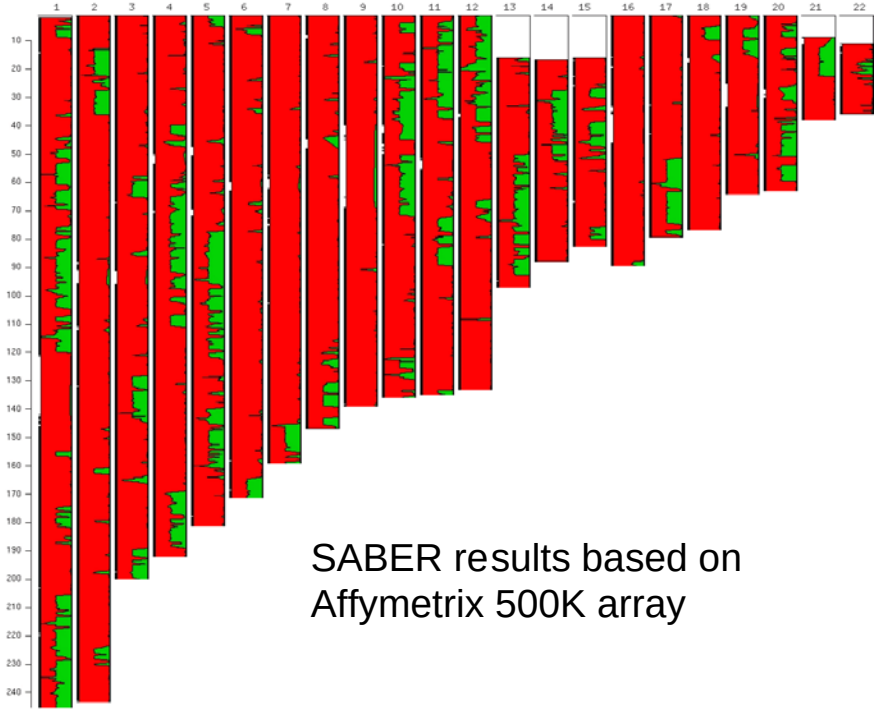


SABER results based on
Illumina 550K array

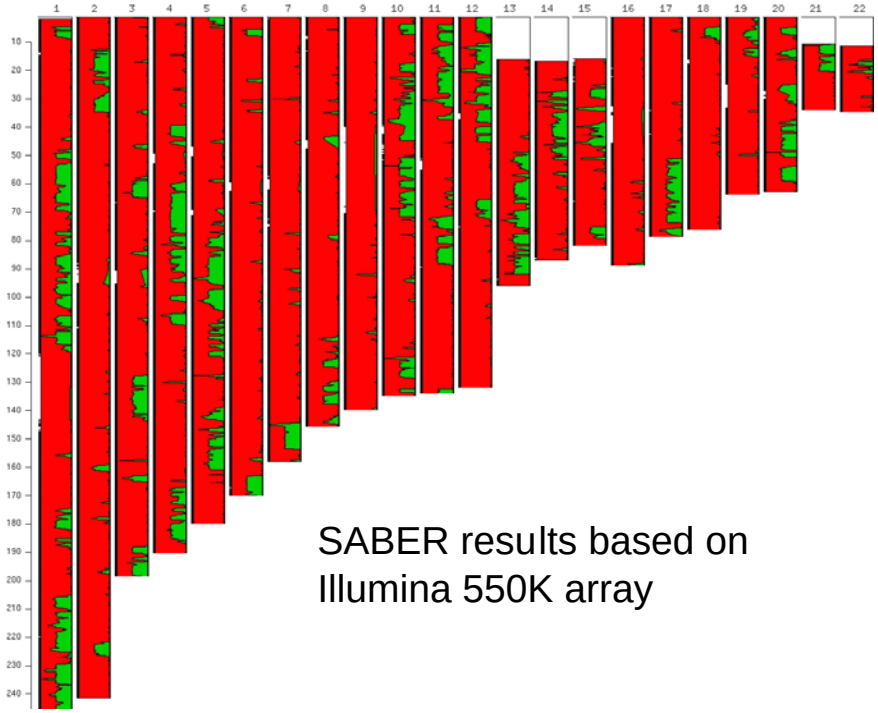
Supplementary Figure S28. Chromosome mosaicism in subject B7-7-18-5 (see Figure 2b and Table 1) as visualized by RHH using the Illm550K array or by SABER using the Affy500K or Illm550K arrays. In the RHH visualization, rare hets are denoted by red dashes next to a chromosome such that clusters of rare hets mark chromosomal segments of non-Caucasian DNA. In the corresponding visualizations produced by SABER software, green shading marks inferred genomic location(s) of African DNA and red marks inferred Caucasian DNA such that a roughly equal left-right split between red and green implies a genomic region in which DNA on one homologous chromosome is Caucasian while DNA on the other chromosome is African. In the SABER visualizations, centromeric regions without SNP genotyping are denoted by a “gap” in the lefthand boundary of the chromosome whereas the same regions in the RHH visualizations are denoted by extended absence of gray shading *inside* chromosomes. Comparison of this figure with Figure 2b shows that the overall pattern of mosaicism for subject B7-7-18-5 is almost identical for RHH and Saber using either the Affy500K or Illm550K chip. However, as described in the Discussion, there are minor differences in the results that may impact visualization of very short (<3Mb long) segments of admixture.



RHH results based on
Illumina 550K array

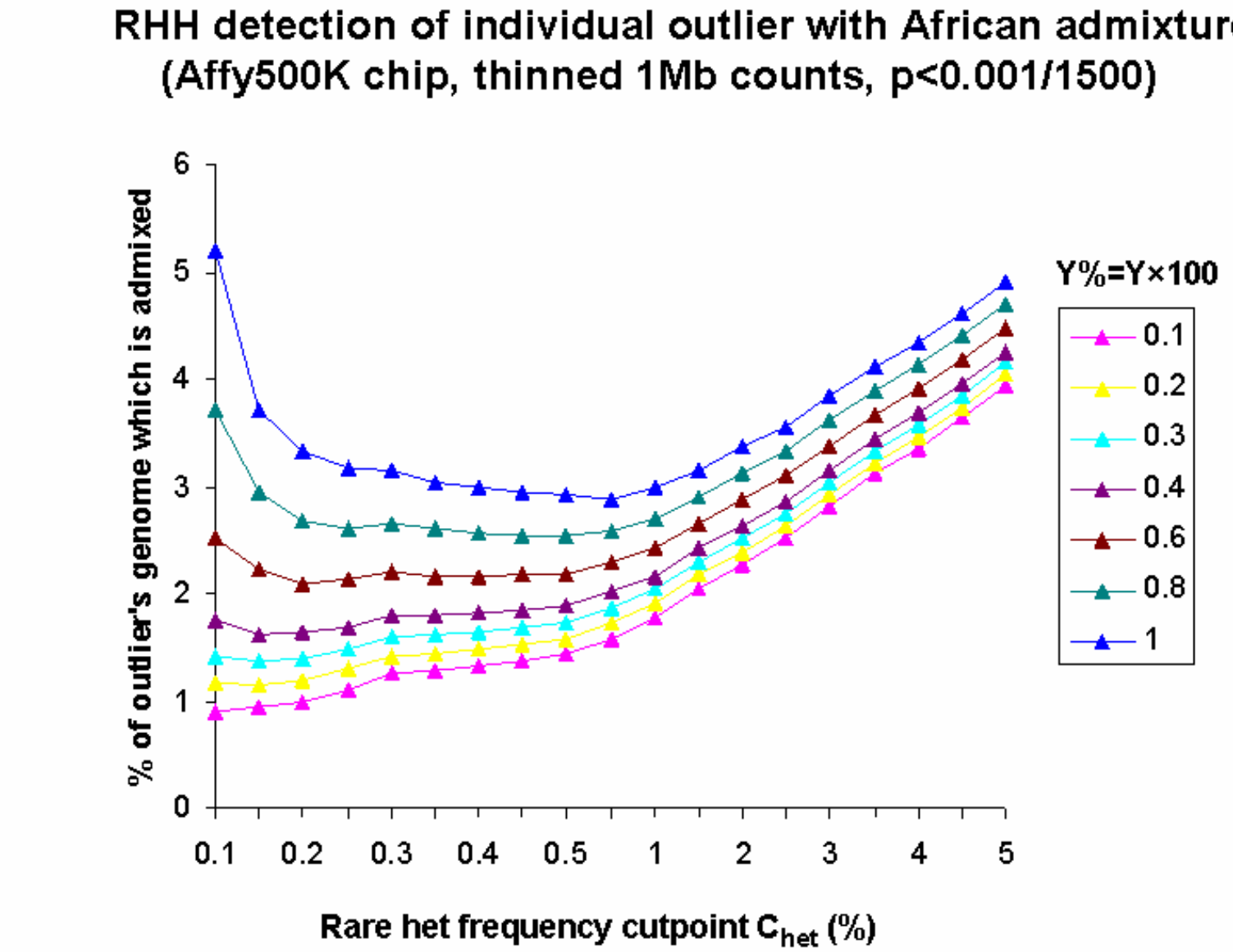


SABER results based on
Affymetrix 500K array

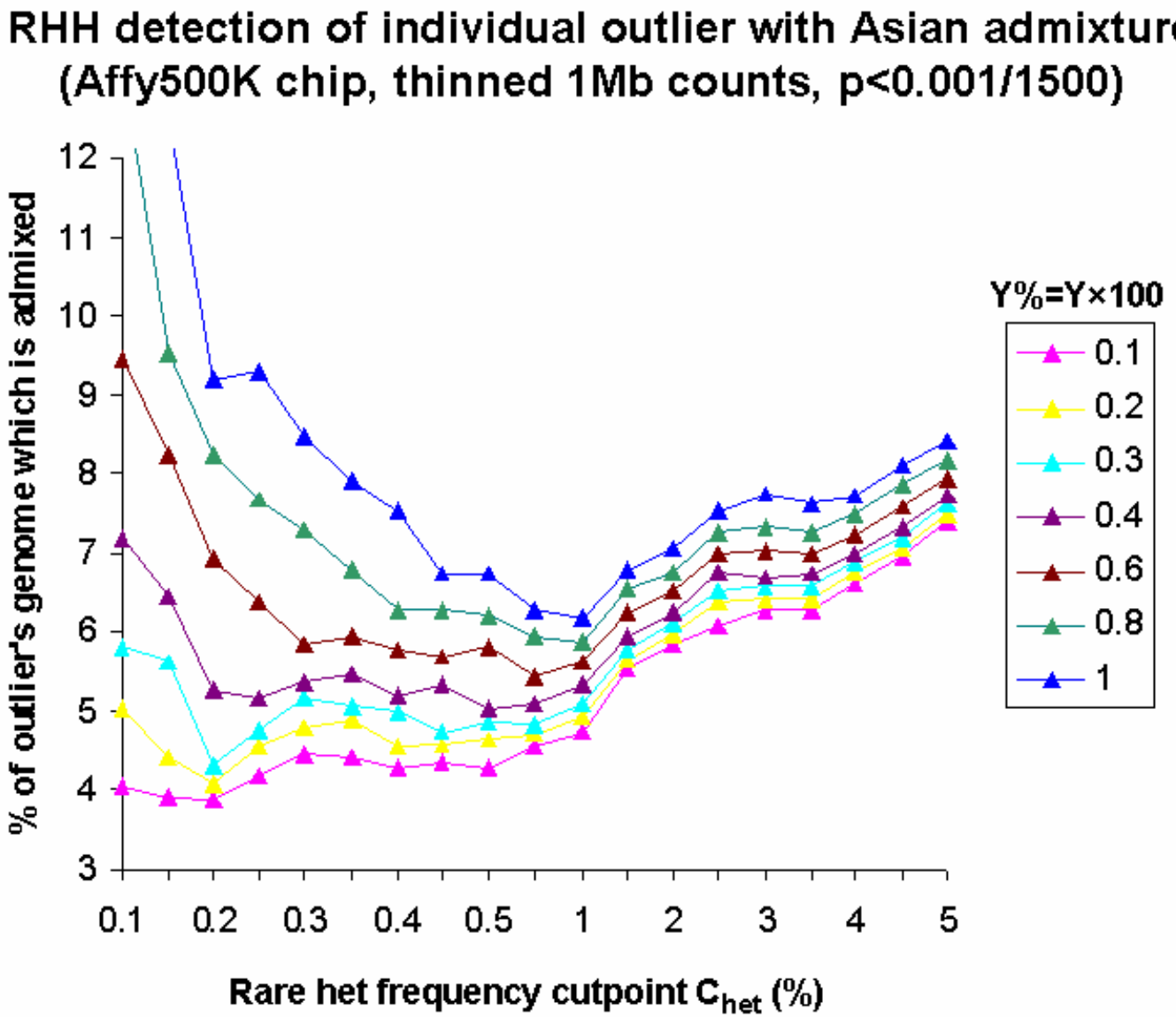


SABER results based on
Illumina 550K array

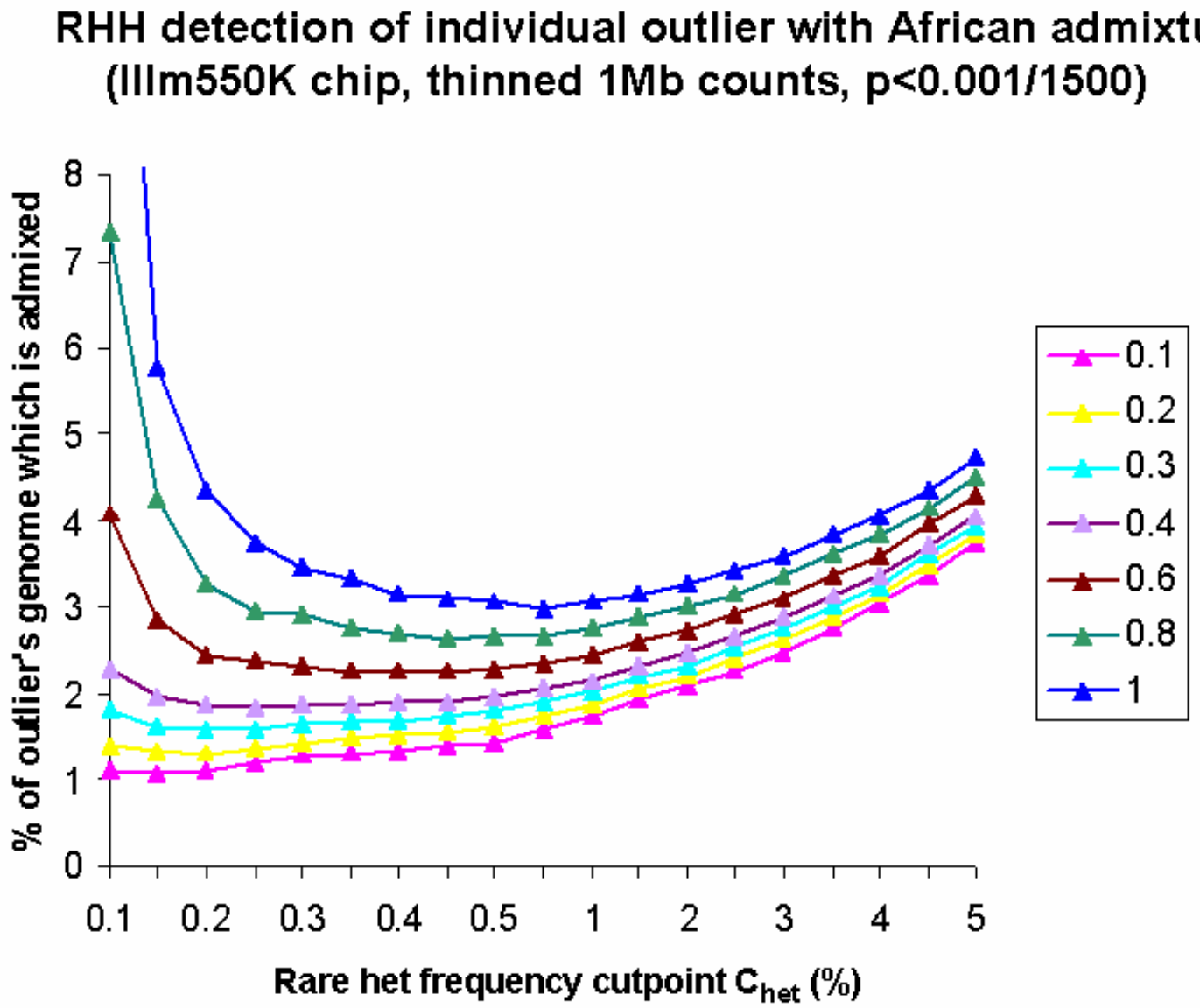
Supplementary Figure S29. Chromosome mosaicism in subject B2-5-2-3 (see Figure 2d and Table 1) as visualized by RHH using the Illm550K array or by SABER using the Affy500K or Illm550K arrays. In the RHH visualization, rare hets are denoted by red dashes next to a chromosome such that clusters of rare hets mark chromosomal segments of non-Caucasian DNA. In the corresponding visualizations produced by SABER software, green shading marks inferred genomic location(s) of African DNA and red marks inferred Caucasian DNA such that a roughly equal left-right split between red and green implies a genomic region in which DNA on one homologous chromosome is Caucasian while DNA on the other chromosome is African. In the SABER visualizations, centromeric regions without SNP genotyping are denoted by a "gap" in the lefthand boundary of the chromosome whereas the same regions in the RHH visualizations are denoted by extended absence of gray shading *inside* chromosomes. Comparison of this figure with Figure 2d shows that the overall pattern of mosaicism for subject B2-5-2-3 is almost identical for RHH and Saber using either the Affy500K or Illm550K chip. However, as described in the Discussion, there are minor differences in the results that may impact visualization of very short (<3Mb long) segments of admixture.



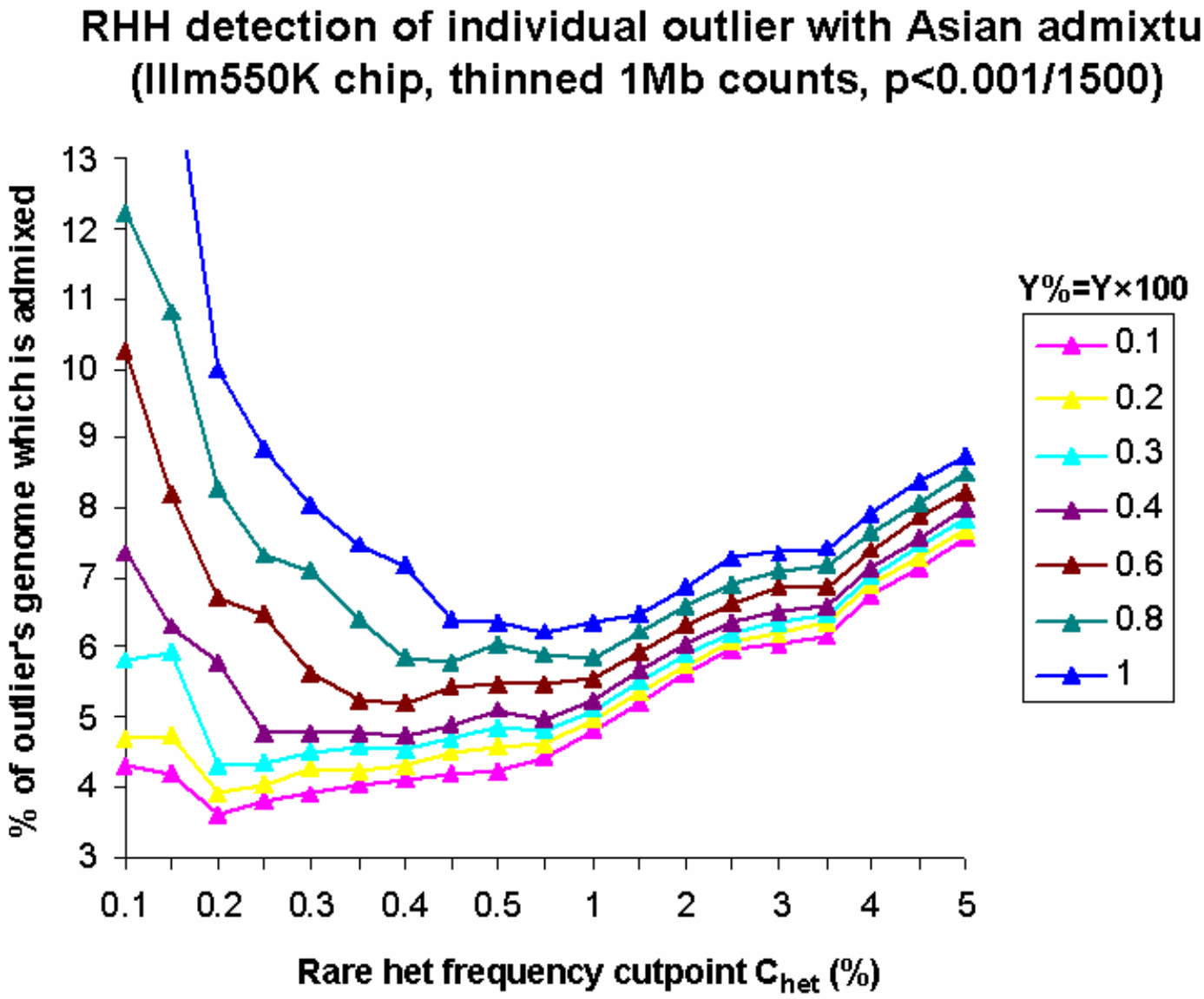
Supplementary Figure S30. Graph showing the lowest percentage of genome admixture in a subject (Y-axis) which is detectable by RHH at p<0.001/1500 for different values of C_{het}, the rare-het frequency cutpoint (X-axis). In this graph, non-Caucasian admixture is from Africa (HapMap Yoruban=YRI) and rare-het counts are from 1Mb-thinned SNPs of the Affy500K array. A hypothetical dataset of 1500 subjects is assumed but each curve in the graph represents a different dataset defined by a specific value of Y%=Y×100, the mean percentage of dataset subjects who are admixed at any genomic position (see Methods).



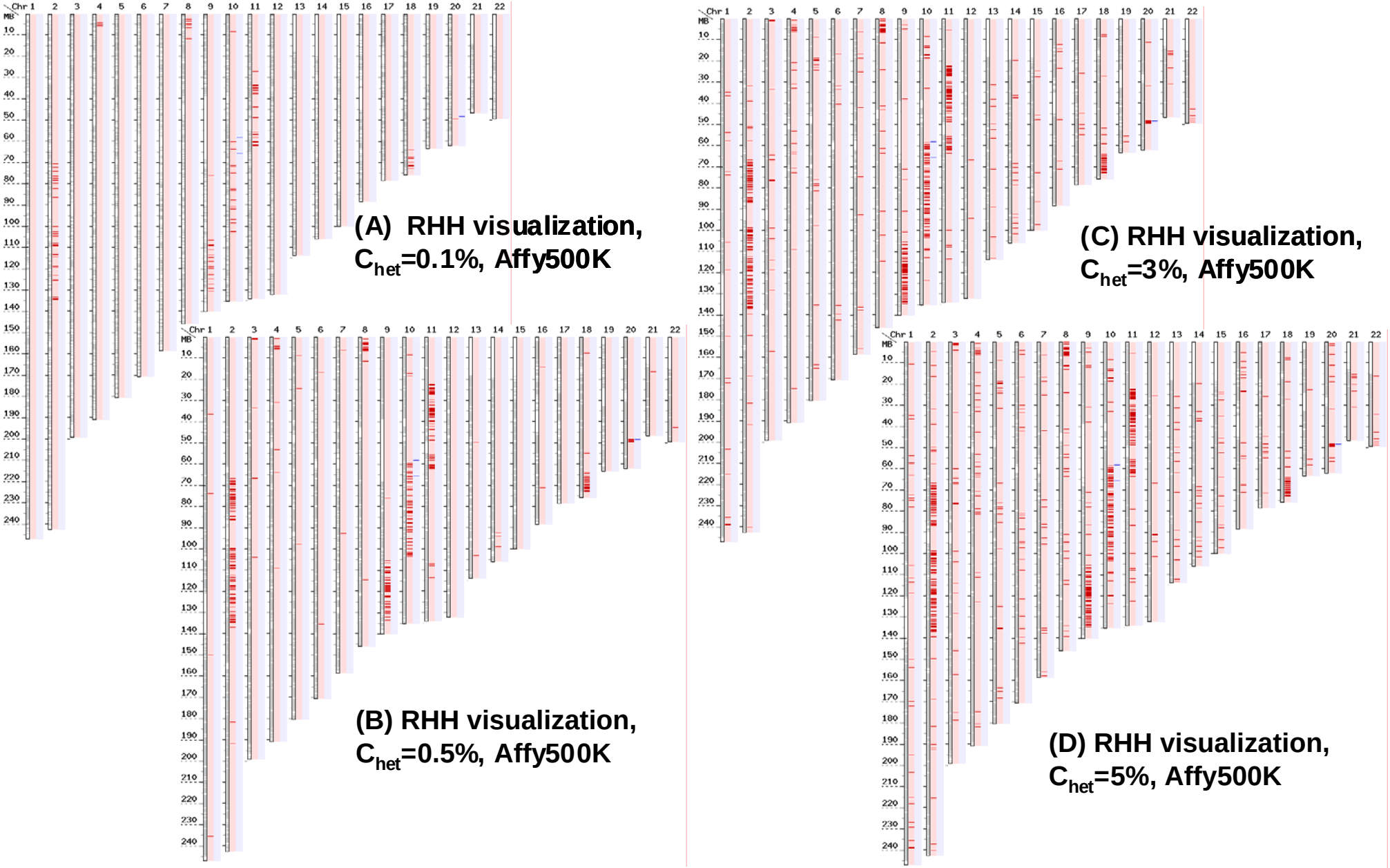
Supplementary Figure S31. Graph showing the lowest percentage of genome admixture in a subject (Y-axis) which is detectable by RHH at p<0.001/1500 for different values of C_{het}, the rare-het frequency cutpoint (X-axis). In this graph, non-Caucasian admixture is from Asia (HapMap CHB+JPT) and rare-het counts are from 1Mb-thinned SNPs of the Affy500K array. A hypothetical dataset of 1500 subjects is assumed but each curve in the graph represents a different dataset defined by a specific value of Y%=Y×100, the mean percentage of dataset subjects who are admixed at any genomic position (see Methods).



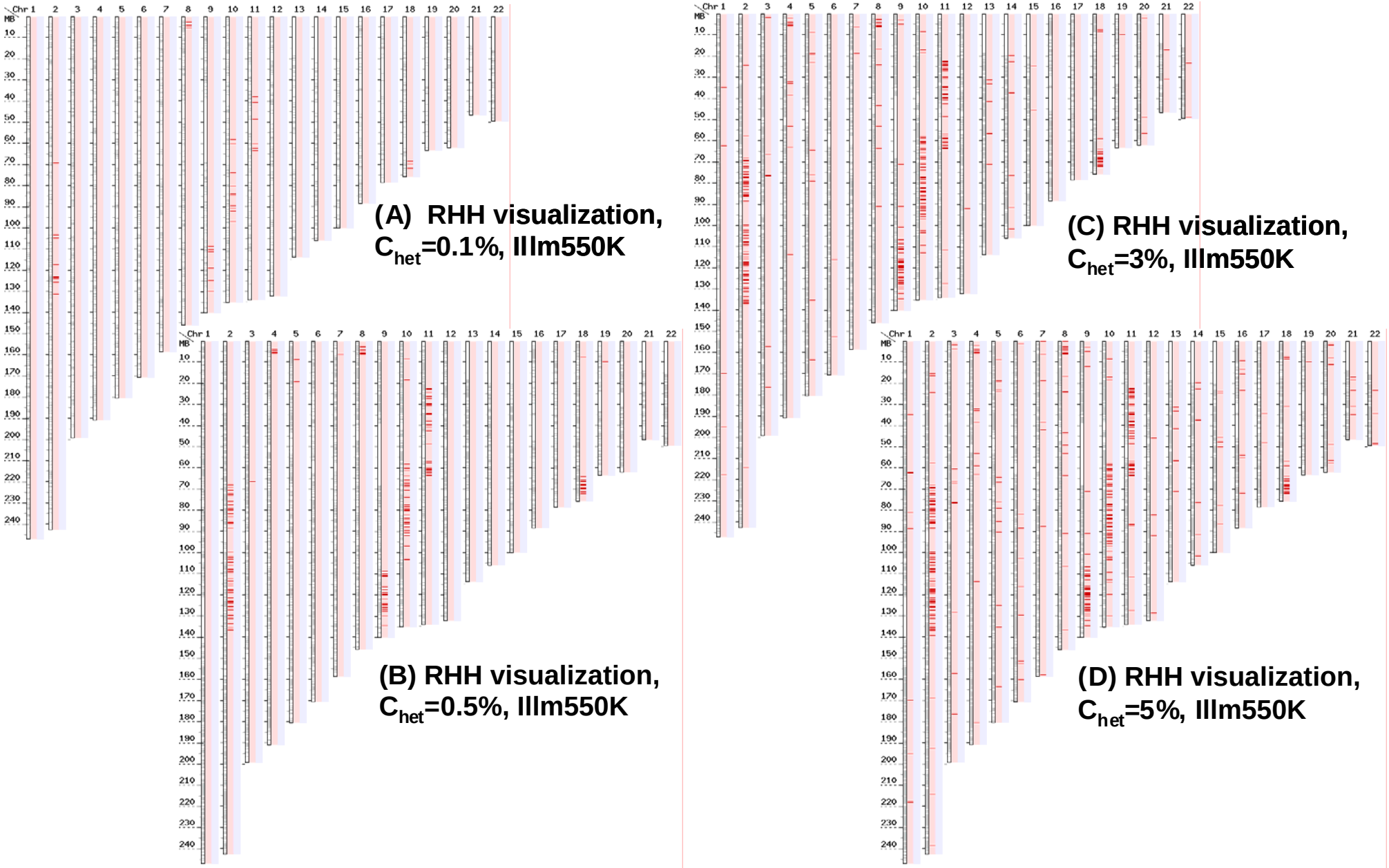
Supplementary Figure S32. Graph showing the lowest percentage of genome admixture in a subject (Y-axis) which is detectable by RHH at p<0.001/1500 for different values of C_{het}, the rare-het frequency cutpoint (X-axis). In this graph, non-Caucasian admixture is from Africa (HapMap Yoruban=YRI) and rare-het counts are from 1Mb-thinned SNPs of the Illm550K array. A hypothetical dataset of 1500 subjects is assumed but each curve in the graph represents a different dataset defined by a specific value of Y%=Y×100, the mean percentage of dataset subjects who are admixed at any genomic position (see Methods).



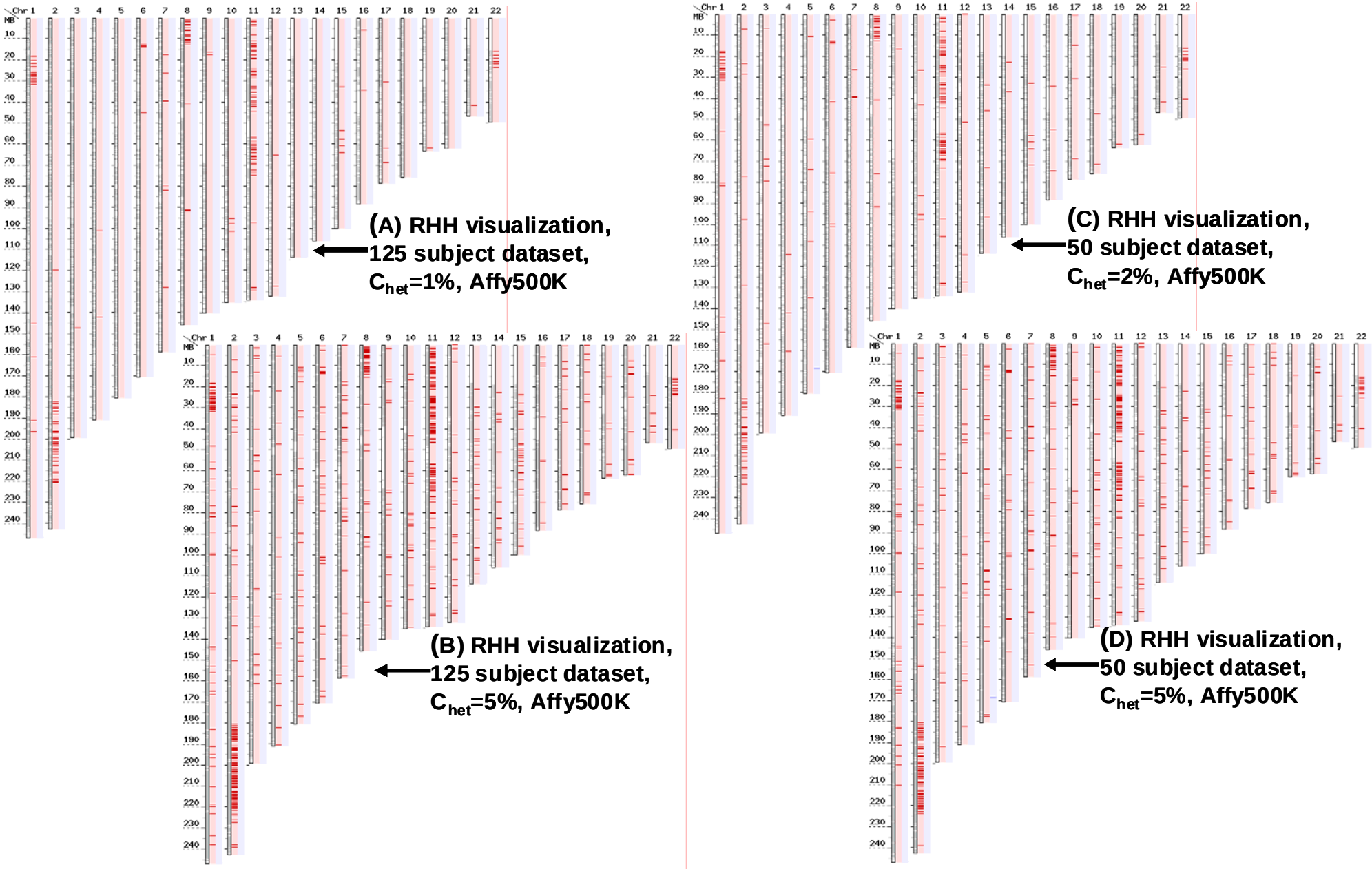
Supplementary Figure S33. Graph showing the lowest percentage of genome admixture in a subject (Y-axis) which is detectable by RHH at p<0.001/1500 for different values of C_{het}, the rare-het frequency cutpoint (X-axis). In this graph, non-Caucasian admixture is from Asia (HapMap CHB+JPT) and rare-het counts are from 1Mb-thinned SNPs of the Illm550K array. A hypothetical dataset of 1500 subjects is assumed but each curve in the graph represents a different dataset defined by a specific value of Y%=Y×100, the mean percentage of dataset subjects who are admixed at any genomic position (see Methods).



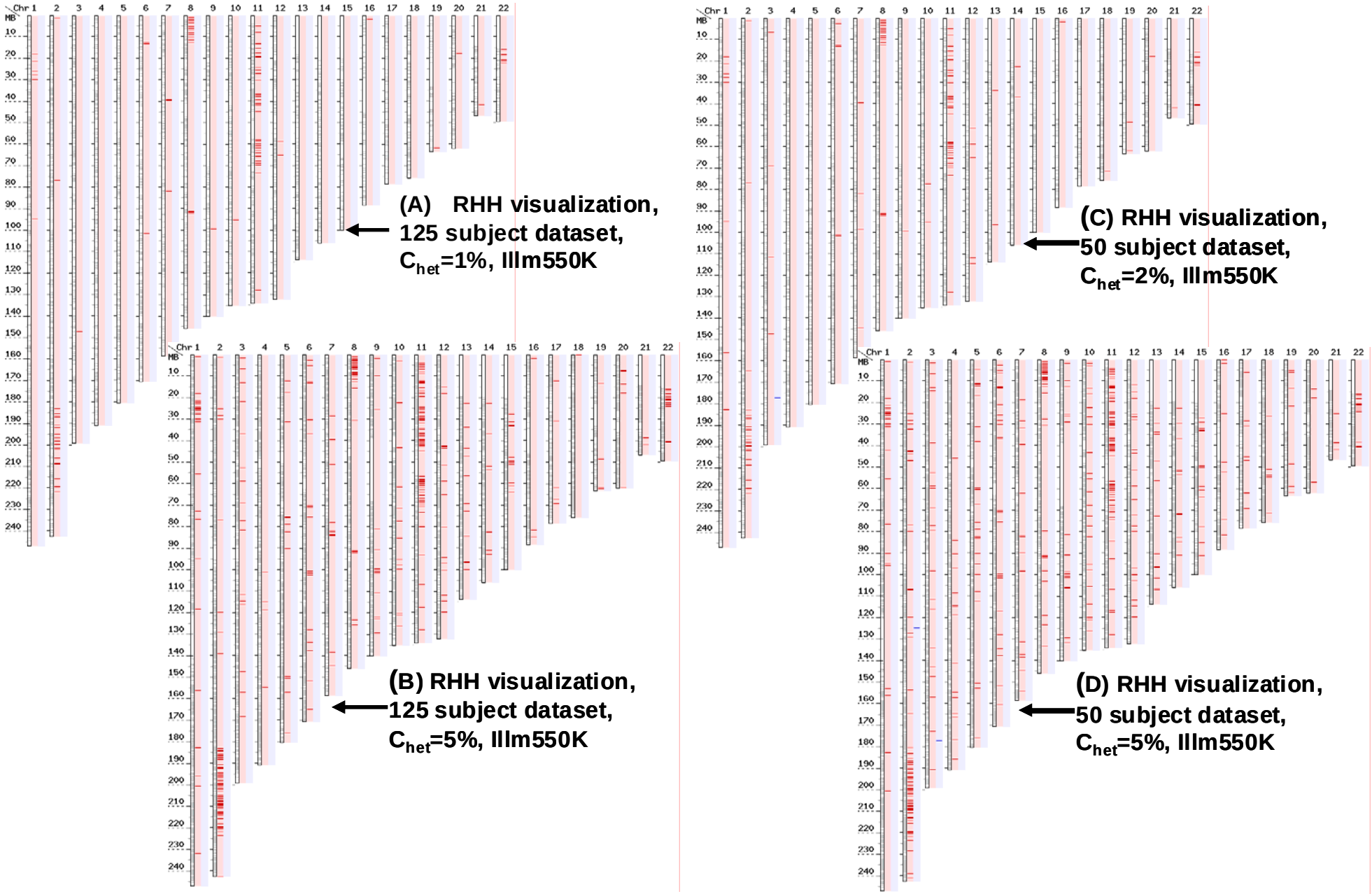
Supplementary Figure S34. Subject B5-4-12-2 (see Fig.1b and Table 1) is used to illustrate visualization of mosaicism at different rare-het frequency cutpoints (C_{het}) for the Affy500K array. For C_{het}=0.1% (panel A), mosaicism is evident but rare hets marking non-Caucasian DNA are somewhat sparse; for C_{het}=0.5% (panel B), mosaicism is clearer with much denser rare-het coverage of non-Caucasian DNA segments; for C_{het}=3% (panel C) and C_{het}=5% (panel D), there is a large increase in sporadic rare-hets falling outside the long segments of non-Caucasian DNA but rare-het density inside the segments is not dramatically increased and thus the mosaicism is visually less well-defined. Visualizations for C_{het} values from 0.3% to 1% inclusive look similar to panel B (C_{het}=0.5%) but outside this range of C_{het} values, the visual clarity of mosaicism decreases as illustrated here. This supports our observationally derived use of C_{het}=0.5% as a generally applicable rare-het frequency cutpoint; however RHH software allows user selection of any other C_{het} value desired.



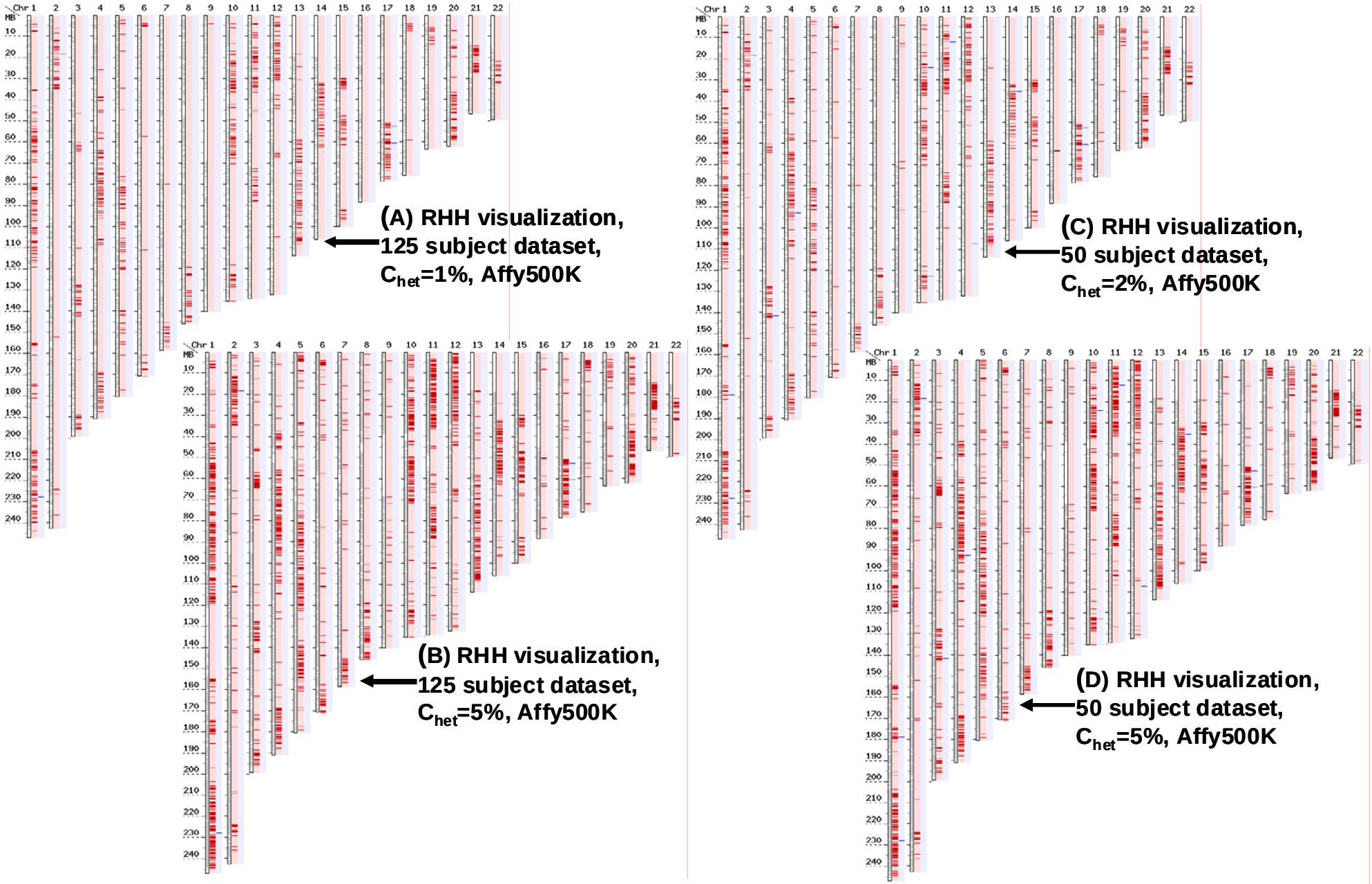
Supplementary Figure S35. Subject B5-4-12-2 (see Fig.1b and Table 1) is used to illustrate visualization of mosaicism at different rare-het frequency cutpoints (C_{het}) for the Illm550K array. For C_{het}=0.1% (panel A), mosaicism is evident but rare hets marking non-Caucasian DNA are too sparse; for C_{het}=0.5% (panel B), mosaicism is clearer with much denser rare-het coverage of non-Caucasian DNA segments; for C_{het}=3% (panel C) and C_{het}=5% (panel D), there is a large increase in sporadic rare-hets falling outside the long segments of non-Caucasian DNA but rare-het density inside the segments is not dramatically increased and thus the mosaicism is visually less well-defined. Visualizations for C_{het} values from 0.3% to 1% inclusive look similar to panel B (C_{het}=0.5%) but outside this range of C_{het} values, the visual clarity of mosaicism decreases as illustrated here. This supports our observationally derived use of C_{het}=0.5% as a generally applicable rare-het frequency cutpoint; however RHH software allows user selection of any other C_{het} value desired.



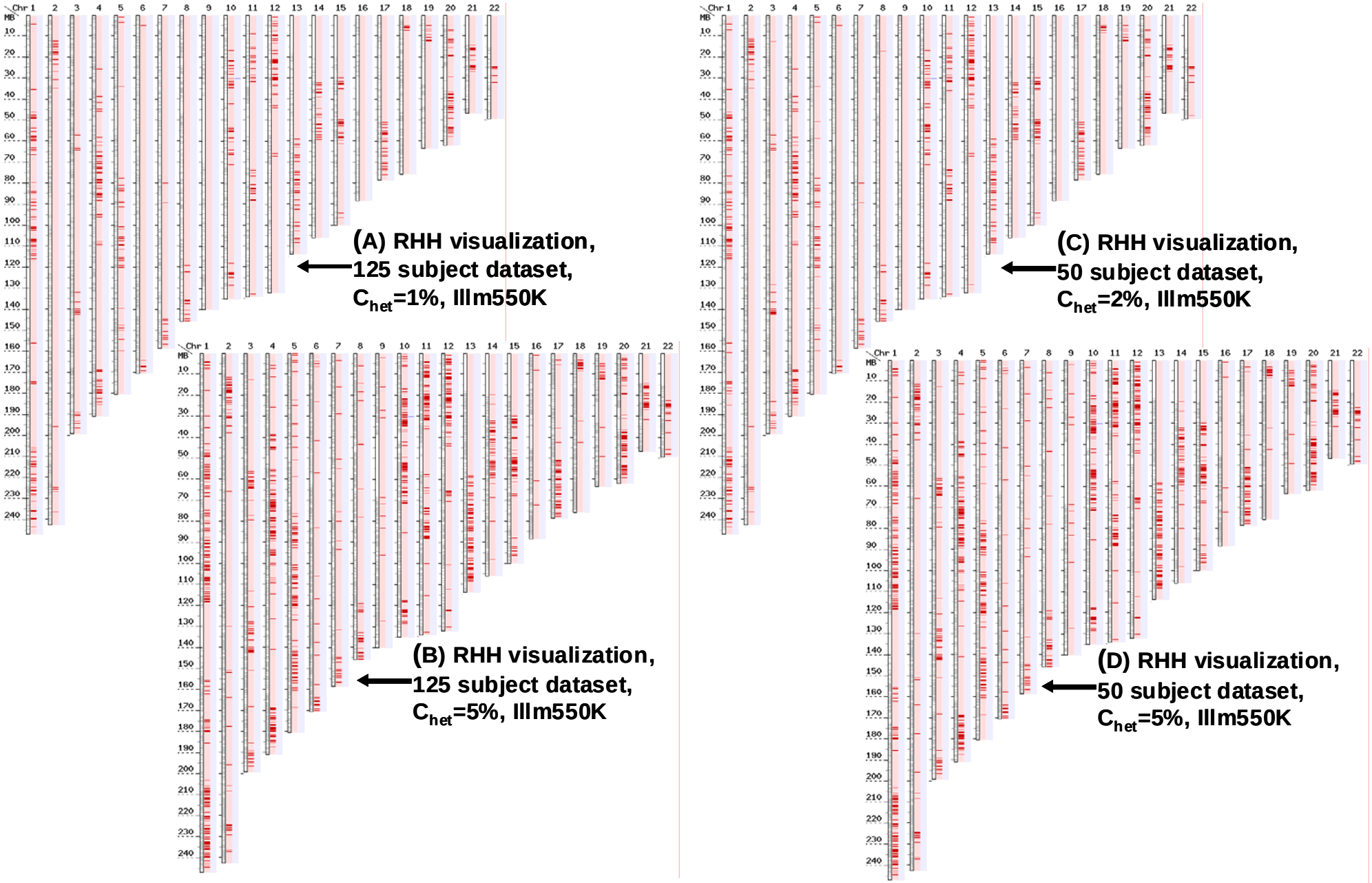
Supplementary Figure S36. Subject B7-7-18-5 (see Fig.2b and Table 1) is used to illustrate visualization of mosaicism in a dataset with 125 or 50 subjects using the Affy500K array and either the lowest possible C_{het} value or C_{het}=5%. For the lowest integer C_{het} values possible with 125 or 50 subjects (panels A and C), mosaicism is clearly visualizable despite the presence of sporadic rare-hets outside regions of admixture marked by dense rare-het segments. When C_{het}=5% (panels B and D), mosaicism and individual regions of admixture are still clearly identifiable using both the 125 or 50-subject dataset but the mosaicism is less visually distinct due to the increase of sporadic rare hets outside regions of admixture.



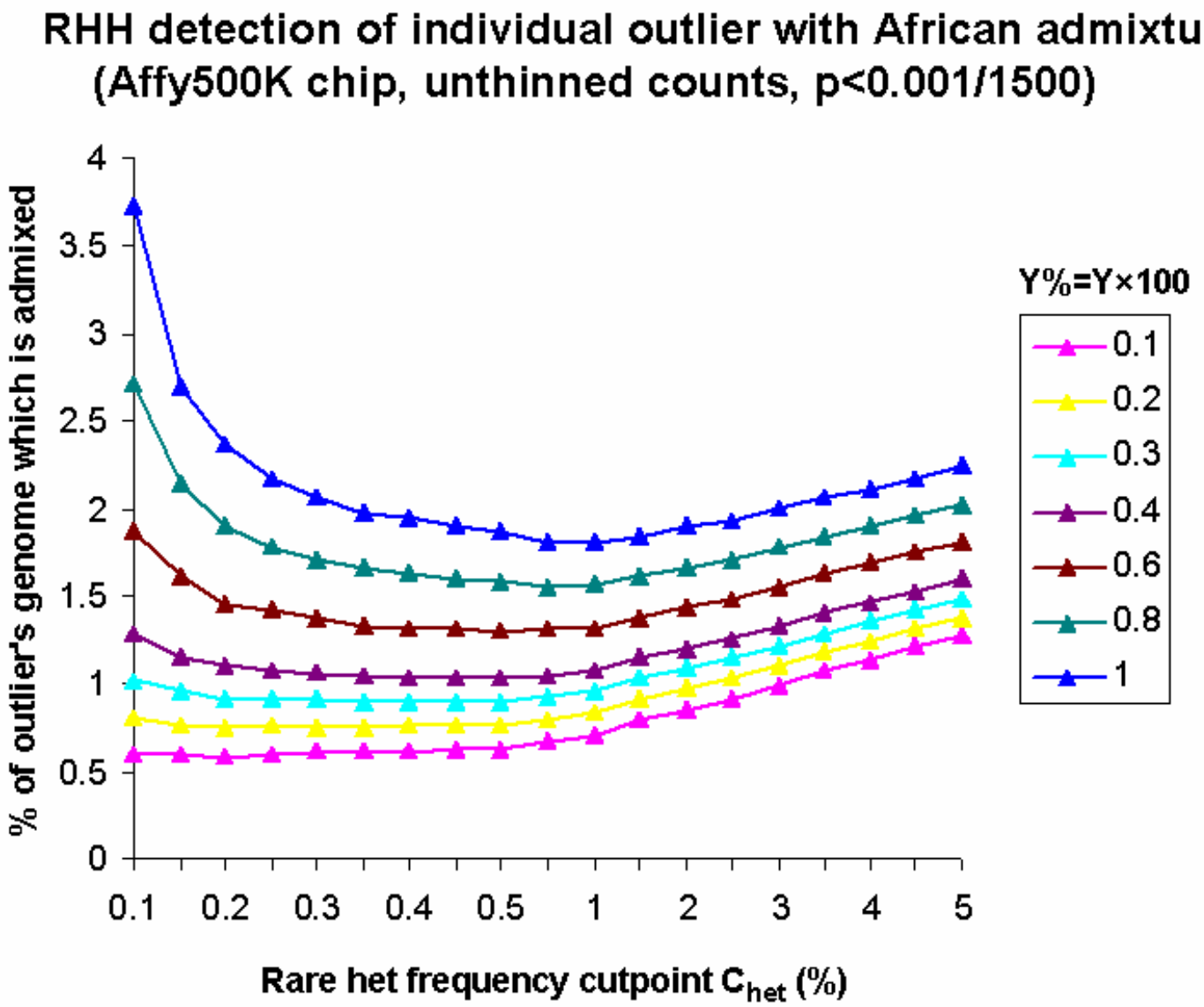
Supplementary Figure S37. Subject B7-7-18-5 (see Fig.2b and Table 1) is used to illustrate visualization of mosaicism in a dataset with 125 or 50 subjects using the Illm550K array and either the lowest possible C_{het} value or C_{het}=5%. For the lowest integer C_{het} values possible with 125 or 50 subjects (panels A and C), mosaicism is clearly visualizable despite occassional sporadic rare-hets outside regions of admixture marked by dense rare-het segments. When C_{het}=5% (panels B and D), mosaicism and individual regions of admixture are still clearly identifiable using both the 125 or 50-subject dataset but the mosaicism is less visually distinct due to the increase of sporadic rare hets outside regions of admixture.



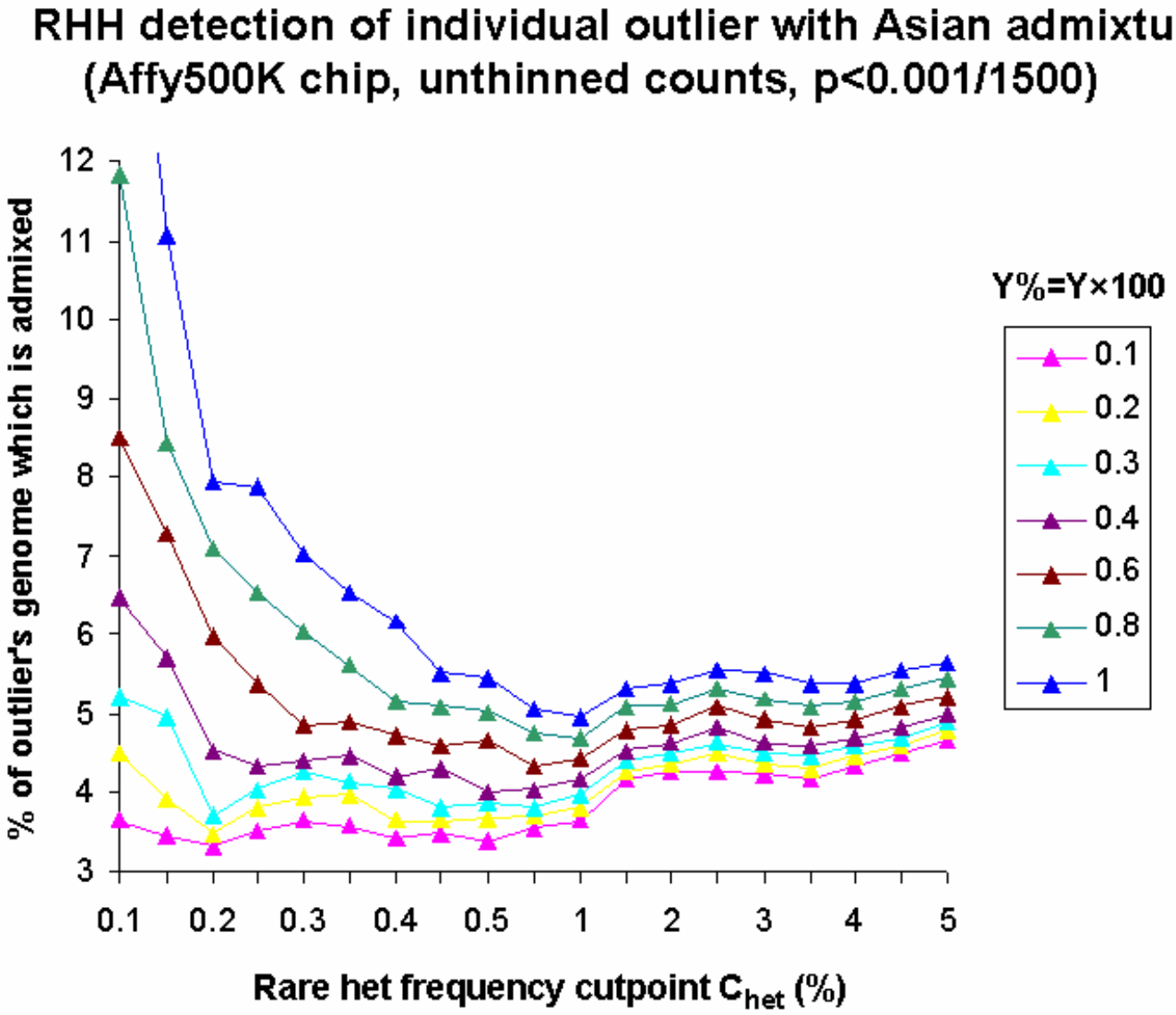
Supplementary Figure S38. Subject B2-5-2-3 (see Fig. 2d and Table 1) is used to illustrate visualization of mosaicism in a dataset with 125 or 50 subjects using the Affy500K array and either the lowest possible C_{het} value or C_{het}=5%. For the lowest integer C_{het} values possible with 125 or 50 subjects (panels A and C), mosaicism is clearly visualizable despite the presence of sporadic rare-hets outside regions of admixture marked by dense rare-het segments. When C_{het}=5% (panels B and D), mosaicism and individual regions of admixture are still clearly identifiable using both the 125 or 50-subject dataset but the mosaicism is less visually distinct due to the increase of sporadic rare hets outside regions of admixture.



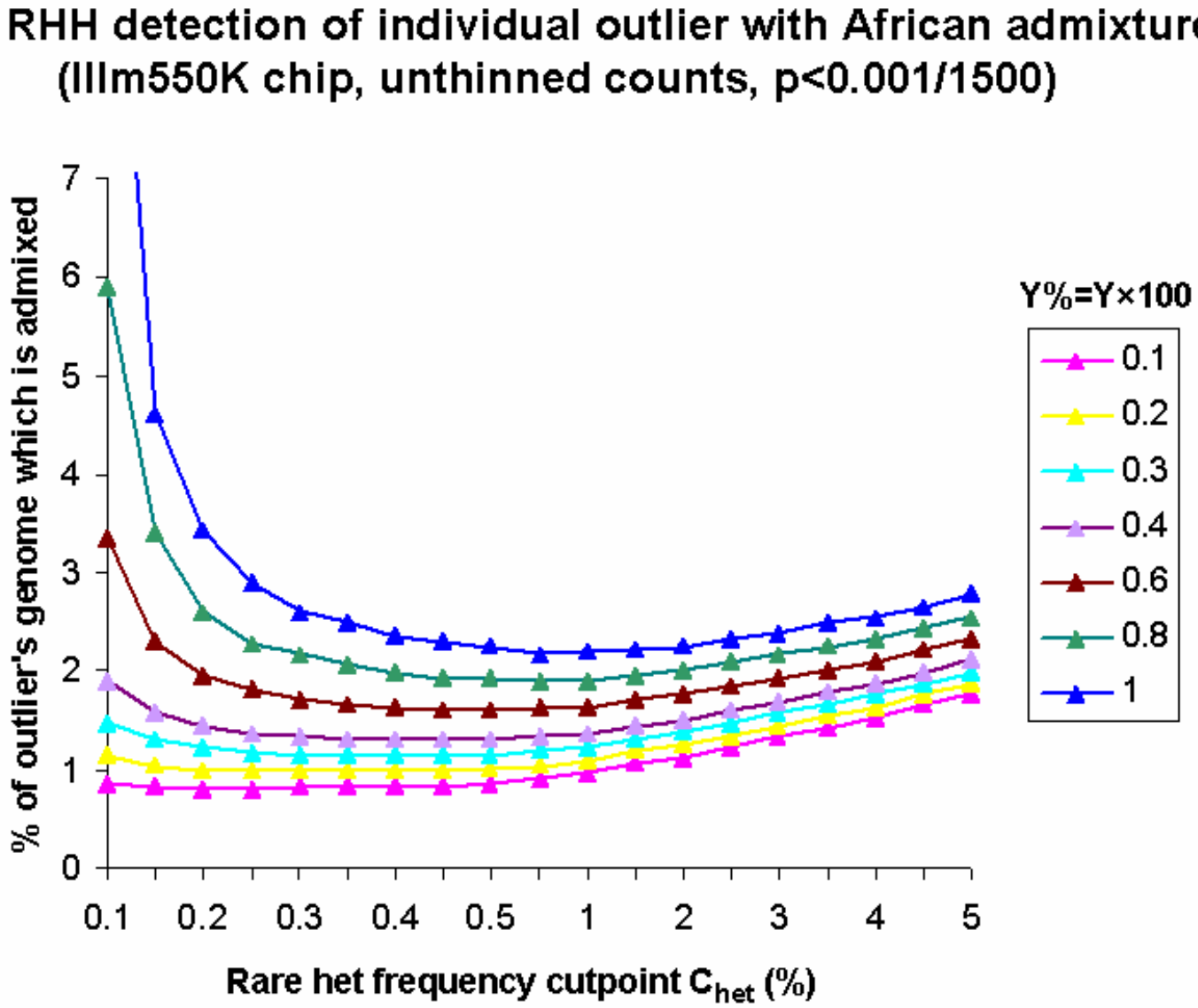
Supplementary Figure S39. Subject B2-5-2-3 (see Fig. 2d and Table 1) is used to illustrate visualization of mosaicism in a dataset with 125 or 50 subjects using the Illm550K array and either the lowest possible C_{het} value or C_{het}=5%. For the lowest integer C_{het} values possible with 125 or 50 subjects (panels A and C), mosaicism is clearly visualizable despite the presence of sporadic rare-hets outside regions of admixture marked by dense rare-het segments. When C_{het}=5% (panels B and D), mosaicism and individual regions of admixture are still clearly identifiable using both the 125 or 50-subject dataset but the mosaicism is less visually distinct due to the increase of sporadic rare hets outside regions of admixture.



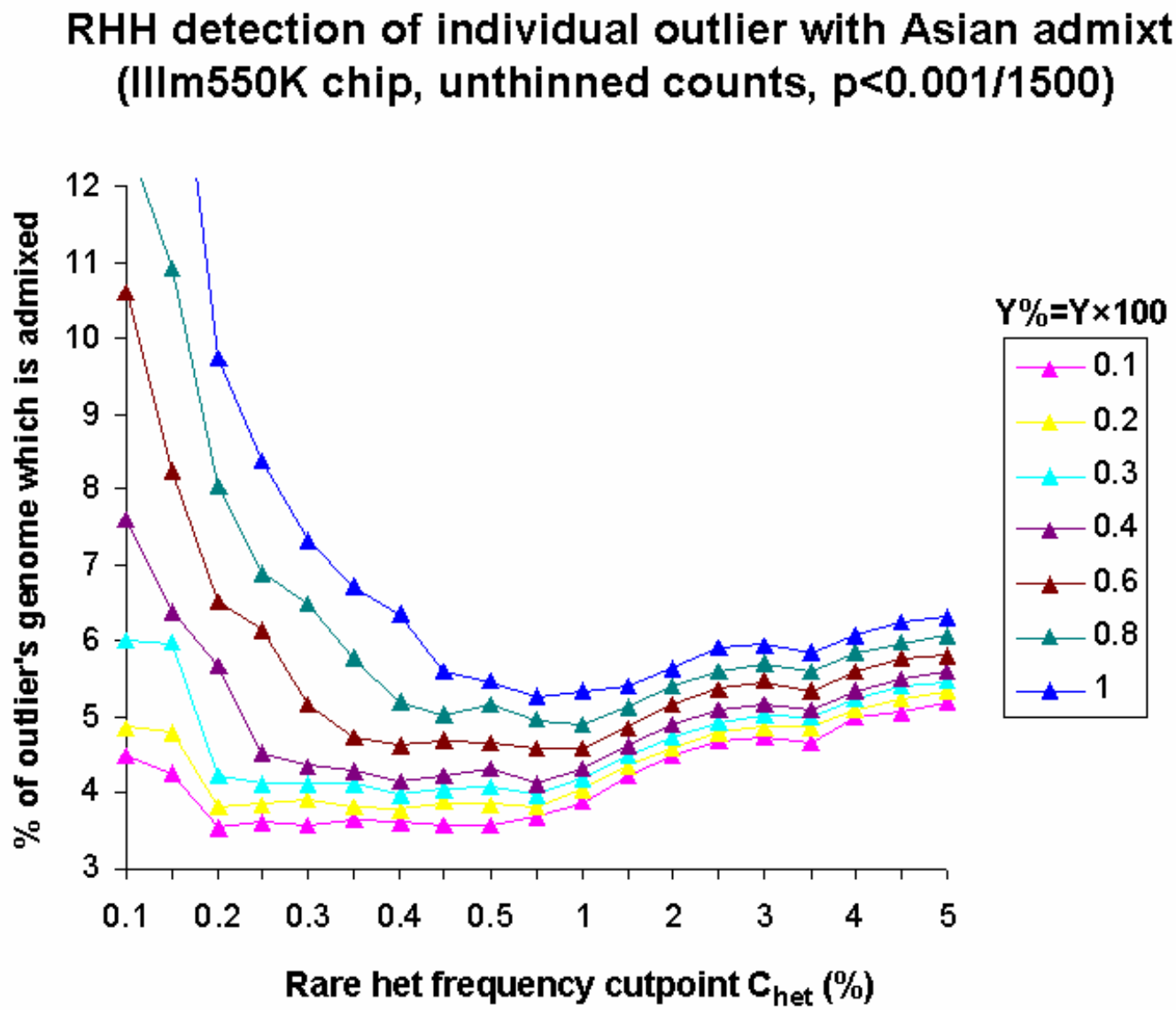
Supplementary Figure S40. Graph showing the lowest percentage of genome admixture in a subject (Y-axis) which is detectable by RHH at p<0.001/1500 for different values of C_{het}, the rare-het frequency cutpoint (X-axis). In this graph, non-Caucasian admixture is from Africa (HapMap Yoruban=YRI) and rare-het counts are from unthinned SNPs of the Affy500K array. A hypothetical dataset of 1500 subjects is assumed but each curve in the graph represents a different dataset defined by a specific value of Y%=Y×100, the mean percentage of dataset subjects who are admixed at any genomic position (see Methods).



Supplementary Figure S41. Graph showing the lowest percentage of genome admixture in a subject (Y-axis) which is detectable by RHH at p<0.001/1500 for different values of C_{het}, the rare-het frequency cutpoint (X-axis). In this graph, non-Caucasian admixture is from Asia (HapMap CHB+JPT) and rare-het counts are from unthinned SNPs of the Affy500K array. A hypothetical dataset of 1500 subjects is assumed but each curve in the graph represents a different dataset defined by a specific value of Y%=Y×100, the mean percentage of dataset subjects who are admixed at any genomic position (see Methods).

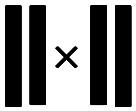
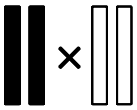
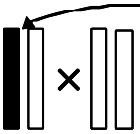
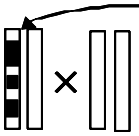
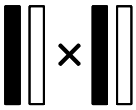
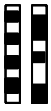


Supplementary Figure S42. Graph showing the lowest percentage of genome admixture in a subject (Y-axis) which is detectable by RHH at p<0.001/1500 for different values of C_{het}, the rare-het frequency cutpoint (X-axis). In this graph, non-Caucasian admixture is from Africa (HapMap Yoruban=YRI) and rare-het counts are from unthinned SNPs of the Illm550K array. A hypothetical dataset of 1500 subjects is assumed but each curve in the graph represents a different dataset defined by a specific value of Y%=Y×100, the mean percentage of dataset subjects who are admixed at any genomic position (see Methods).



Supplementary Figure S43. Graph showing the lowest percentage of genome admixture in a subject (Y-axis) which is detectable by RHH at p<0.001/1500 for different values of C_{het}, the rare-het frequency cutpoint (X-axis). In this graph, non-Caucasian admixture is from Asia (HapMap CHB+JPT) and rare-het counts are from unthinned SNPs of the Illm550K array. A hypothetical dataset of 1500 subjects is assumed but each curve in the graph represents a different dataset defined by a specific value of Y%=Y×100, the mean percentage of dataset subjects who are admixed at any genomic position (see Methods).

Supplementary Table S1. Different types of ethnic outlier from different matings involving two ethnic groups such as HapMap Yorubans (YRI) and Caucasians (CEU)

Type of Mating[a]	Parents[b]	Type of offspring (outlier)[c]	==Genome Distribution== Excess Rare Hets[d]	Excess Rare Homs[d]
Within ethnic group ("Incross") YRI × YRI		 Unadmixed YRI	Entire genome	Entire genome
"Outcross" YRI × CEU		 F1 YRI-CEU hybrid	Entire genome	none
"Backcross" F1 YRI-CEU with CEU (YRI × CEU) × CEU		 One chromosome is mosaic for YRI-CEU, the other is 100% CEU	Partial genome (mosaic)	none
2nd CEU Backcross (YRI × CEU) × CEU {twice}		 Mosaic chromosome has less YRI DNA and rare hets than parent	Partial genome (mosaic)	none
"Intercross" two F1 YRI-CEU hybrids (YRI × CEU) × (YRI × CEU)		 Each chromosome is a different mosaic of YRI and CEU DNA	Partial genome (mosaic)	Partial genome (mosaic)

[a] "X" symbolizes mating and HapMap populations (YRI, CEU) are here considered analogous to mouse strains. Names for different types of mouse crosses are used to specify the corresponding type of mating involving HapMap subjects and their simulated offspring ("Outcross", "Backcross", etc.)

[b] A homologous pair of chromosomes is shown for each parent in the mating (black=YRI DNA, white=CEU DNA)

[c] The expected composition of each chromosome in the outlier offspring (e.g. 100% YRI or YRI-CEU mosaic) is shown and is consistent with the genome distribution of excess rare homs and hets observed in simulated outliers of Table 3 in the main text.

[d] Summary of results for simulated outliers in Table 3. Note that excess rare homs were observed only in simulated outliers who inherited outlier DNA (YRI or CHB) from father and mother (see "Incross" and "Intercross"). Furthermore, excess rare hets were limited to part of the genome (implying chromosomal mosaicism) only in simulated outliers with at least one parent who was ethnically admixed (see "Backcross" and "Intercross"). By contrast, two unadmixed parents of the same ethnic group ("Incross") or different groups ("Outcross") produced outliers with excess rare hets across the entire genome.

Supplementary Table

Table S2. RHH detection of individual outlier with African admixture (1Mb-thinned rare-het counts, Affy500K chip, p<0.001/1500)

Mean % admixture (Y) in the entire dataset	Rare-het frequency cutpoints (C_{het})					
Y% (=Yx100)	C_{het}=0.1%	C_{het}=0.3%	C_{het}=0.5%	C_{het}=1%	C_{het}=3%	C_{het}=5%
1.0%	284; 1; 15	1043; 6; 38	1460; 12; 54	1867; 29; 84	2519; 153; 244	2979; 388; 515
0.8%	439; 1; 17	1218; 7; 39	1597; 13; 53	1925; 31; 82	2531; 155; 241	2983; 389; 510
0.6%	722; 1; 19	1404; 8; 38	1735; 15; 53	1953; 33; 79	2554; 158; 239	3000; 394; 511
0.4%	1053; 1; 19	1594; 9; 37	1801; 16; 50	1984; 34; 76	2563; 160; 235	3003; 397; 508
0.3%	1303; 2; 20	1664; 10; 36	1814; 17; 48	2001; 36; 76	2587; 164; 237	3003; 397; 505
0.2%	1501; 3; 20	1771; 11; 36	1835; 17; 46	2018; 37; 74	2601; 166; 237	3003; 397; 502
0.1%	1857; 4; 21	1816; 12; 34	1858; 18; 44	2032; 37; 72	2601; 166; 234	3016; 402; 505

The 3 values separated by semi colons in each cell of the table are:
(A) E^*_{100}, total genome-wide rare-het counts expected in a subject with 100% genome admixture from Africa (HapMap Yorubans)
(B) E^*_0, total genome-wide rare-het counts expected in an unadmixed Caucasian subject with 0% non-Caucasian admixture
(C) T_X, rare-het count threshold whose probabilty is p<0.001/1500 of being reached or exceeded under the null hypothesis

Note: As described in Methods, each minimum detectable percentage of genome admixture (F_{min}) shown in Supplementary Figures S30-S33 was found by solving the equation $T_X=(1-F_{min})E^*_0+F_{min}E^*_{100}$ for F_{min}, i.e., $F_{min}=(T_X-E^*_0)/(E^*_{100}-E^*_0)$ and then multiplying the result by 100 to convert F_{min} to a percentage. Counts in table are rounded to nearest integer which may produce a slight difference between F_{min} calculated from table counts and F_{min} in the corresponding figure.

Supplementary Table

Table S3. RHH detection of individual outlier with Asian admixture (1Mb-thinned rare-het counts, Affy500K chip, p<0.001/1500)

Mean % admixture (Y) in the entire dataset	Rare-het frequency cutpoints (C_{het})					
Y% (=Yx100)	C_{het}=0.1%	C_{het}=0.3%	C_{het}=0.5%	C_{het}=1%	C_{het}=3%	C_{het}=5%
1.0%	80; 4; 18	296; 11; 35	459; 17; 47	703; 36; 77	1139; 159; 235	1753; 394; 508
0.8%	111; 4; 18	338; 12; 35	481; 17; 46	710; 36; 76	1172; 161; 235	1758; 395; 507
0.6%	156; 4; 18	416; 12; 36	491; 17; 45	710; 36; 74	1188; 163; 235	1762; 396; 505
0.4%	203; 4; 18	433; 12; 34	548; 18; 44	710; 36; 72	1215; 165; 236	1766; 397; 503
0.3%	251; 4; 18	436; 12; 34	549; 18; 44	730; 37; 72	1215; 165; 234	1770; 400; 504
0.2%	286; 4; 18	455; 12; 33	563; 18; 43	736; 37; 72	1224; 166; 234	1777; 402; 505
0.1%	344; 4; 18	480; 12; 33	599; 19; 43	745; 37; 71	1231; 167; 233	1782; 403; 505

The 3 values separated by semi colons in each cell of the table are:
(A) E^*_{100}, total genome-wide rare-het counts expected in a subject with 100% genome admixture from Asia (HapMap Chinese and Japanese)
(B) E^*_0, total genome-wide rare-het counts expected in an unadmixed Caucasian subject with 0% non-Caucasian admixture
(C) T_X, rare-het count threshold whose probabilty is p<0.001/1500 of being reached or exceeded under the null hypothesis

Note: As described in Methods, each minimum detectable percentage of genome admixture (F_{min}) shown in Supplementary Figures S30-S33 was found by solving the equation $T_X=(1-F_{min})E^*_0+F_{min}E^*_{100}$ for F_{min}, i.e., $F_{min}=(T_X-E^*_0)/(E^*_{100}-E^*_0)$ and then multiplying the result by 100 to convert F_{min} to a percentage. Counts in table are rounded to nearest integer which may produce a slight difference between F_{min} calculated from table counts and F_{min} in the corresponding figure.

Supplementary Table

Table S4. RHH detection of individual outlier with African admixture (1Mb-thinned rare-het counts, Illm550K chip, p<0.001/1500)

Mean % admixture (Y) in the entire dataset	Rare-het frequency cutpoints (C_{het})					
Y% (=Yx100)	C_{het}=0.1%	C_{het}=0.3%	C_{het}=0.5%	C_{het}=1%	C_{het}=3%	C_{het}=5%
1.0%	58; 0; 8	760; 3; 29	1123; 6; 40	1444; 15; 59	1997; 75; 144	2227; 205; 300
0.8%	130; 0; 10	886; 3; 29	1220; 7; 39	1504; 17; 58	2005; 77; 141	2230; 206; 297
0.6%	276; 0; 11	1100; 4; 30	1291; 8; 37	1520; 17; 54	2005; 77; 137	2237; 207; 294
0.4%	572; 0; 13	1271; 5; 29	1339; 8; 35	1547; 18; 51	2032; 80; 136	2255; 212; 295
0.3%	739; 1; 14	1308; 5; 27	1339; 8; 32	1575; 19; 51	2032; 80; 133	2264; 213; 294
0.2%	919; 1; 14	1373; 6; 26	1365; 9; 31	1577; 20; 49	2032; 80; 131	2265; 215; 293
0.1%	1188; 2; 15	1394; 6; 24	1375; 9; 29	1594; 21; 49	2032; 80; 129	2267; 216; 292

The 3 values separated by semi colons in each cell of the table are:
(A) E^*_{100}, total genome-wide rare-het counts expected in a subject with 100% genome admixture from Africa (HapMap Yorubans)
(B) E^*_0, total genome-wide rare-het counts expected in an unadmixed Caucasian subject with 0% non-Caucasian admixture
(C) T_X, rare-het count threshold whose probabilty is p<0.001/1500 of being reached or exceeded under the null hypothesis

Note: As described in Methods, each minimum detectable percentage of genome admixture (F_{min}) shown in Supplementary Figures S30-S33 was found by solving the equation $T_X=(1-F_{min})E^*_0+F_{min}E^*_{100}$ for F_{min}, i.e., $F_{min}=(T_X-E^*_0)/(E^*_{100}-E^*_0)$ and then multiplying the result by 100 to convert F_{min} to a percentage. Counts in table are rounded to nearest integer which may produce a slight difference between F_{min} calculated from table counts and F_{min} in the corresponding figure.

Supplementary Table

Table S5. RHH detection of individual outlier with Asian admixture (1Mb-thinned rare-het counts, Illm550K chip, p<0.001/1500)

Mean % admixture (Y) in the entire dataset	Rare-het frequency cutpoints (C_{het})					
Y% (=Yx100)	C_{het}=0.1%	C_{het}=0.3%	C_{het}=0.5%	C_{het}=1%	C_{het}=3%	C_{het}=5%
1.0%	79; 2; 14	241; 6; 24	377; 8; 32	522; 20; 52	821; 77; 131	1180; 212; 296
0.8%	100; 2; 14	268; 6; 24	377; 8; 31	552; 20; 51	821; 77; 129	1180; 212; 294
0.6%	117; 2; 14	339; 6; 25	402; 9; 30	552; 20; 50	821; 77; 128	1195; 215; 295
0.4%	162; 2; 14	385; 6; 24	407; 9; 29	552; 20; 48	838; 78; 128	1195; 215; 293
0.3%	204; 2; 14	395; 6; 24	422; 9; 29	552; 20; 47	846; 79; 128	1203; 216; 294
0.2%	253; 2; 14	407; 7; 24	432; 10; 29	564; 21; 48	858; 80; 128	1212; 217; 294
0.1%	263; 2; 14	429; 7; 23	451; 10; 28	574; 22; 48	865; 81; 128	1212; 217; 293

The 3 values separated by semi colons in each cell of the table are:
(A) E^*_{100}, total genome-wide rare-het counts expected in a subject with 100% genome admixture from Asia (HapMap Chinese and Japanese)
(B) E^*_0, total genome-wide rare-het counts expected in an unadmixed Caucasian subject with 0% non-Caucasian admixture
(C) T_X, rare-het count threshold whose probabilty is p<0.001/1500 of being reached or exceeded under the null hypothesis

Note: As described in Methods, each minimum detectable percentage of genome admixture (F_{min}) shown in Supplementary Figures S30-S33 was found by solving the equation $T_X=(1-F_{min})E^*_0+F_{min}E^*_{100}$ for F_{min}, i.e., $F_{min}=(T_X-E^*_0)/(E^*_{100}-E^*_0)$ and then multiplying the result by 100 to convert F_{min} to a percentage. Counts in table are rounded to nearest integer which may produce a slight difference between F_{min} calculated from table counts and F_{min} in the corresponding figure.

Supplementary Table

Table S6. RHH detection of individual outlier with African admixture (Unthinned rare-het counts, Affy500K chip, p<0.001/1500)

Mean % admixture (Y) in the entire dataset	Rare-het frequency cutpoints (C_{het})					
Y% (=Yx100)	C_{het}=0.1%	C_{het}=0.3%	C_{het}=0.5%	C_{het}=1%	C_{het}=3%	C_{het}=5%
1.0%	557; 1; 21	3363; 12; 81	5215; 25; 123	7181; 60; 190	10076; 289; 485	12414; 746; 1007
0.8%	861; 1; 24	3930; 14; 81	5703; 28; 118	7402; 64; 179	10125; 293; 468	12428; 748; 984
0.6%	1416; 1; 28	4530; 17; 79	6195; 32; 112	7512; 67; 165	10216; 299; 453	12498; 758; 971
0.4%	2064; 2; 28	5141; 20; 74	6431; 35; 101	7630; 70; 151	10251; 301; 433	12513; 763; 951
0.3%	2555; 2; 28	5369; 21; 70	6478; 35; 94	7697; 73; 146	10346; 309; 431	12513; 763; 938
0.2%	2944; 4; 28	5713; 24; 67	6552; 37; 87	7763; 75; 139	10403; 313; 424	12513; 763; 925
0.1%	3642; 6; 27	5857; 26; 61	6637; 38; 80	7814; 76; 131	10403; 313; 413	12566; 774; 924

The 3 values separated by semi colons in each cell of the table are:
(A) E_{100}, total genome-wide rare-het counts expected in a subject with 100% genome admixture from Africa (HapMap Yorubans)
(B) E_0, total genome-wide rare-het counts expected in an unadmixed Caucasian subject with 0% non-Caucasian admixture
(C) T_X, rare-het count threshold whose probabilty is p<0.001/1500 of being reached or exceeded under the null hypothesis

Note: As described in Methods, each minimum detectable percentage of genome admixture (F_{min}) shown in Supplementary Figures S40-S43 was found by solving the equation $T_X=(1-F_{min})E_0+F_{min}E_{100}$ for F_{min}, i.e., $F_{min}=(T_X-E_0)/(E_{100}-E_0)$ and then multiplying the result by 100 to convert F_{min} to a percentage. Counts in table are rounded to nearest integer which may produce a slight difference between F_{min} calculated from table counts and F_{min} in the corresponding figure.

Supplementary Table

Table S7. RHH detection of individual outlier with Asian admixture (Unthinned rare-het counts, Affy500K chip, p<0.001/1500)

Mean % admixture (Y) in the entire dataset	Rare-het frequency cutpoints (C_{het})					
Y% (=Yx100)	C_{het}=0.1%	C_{het}=0.3%	C_{het}=0.5%	C_{het}=1%	C_{het}=3%	C_{het}=5%
1.0%	104; 5; 22	510; 25; 59	819; 36; 79	1279; 74; 133	2277; 301; 410	3730; 758; 926
0.8%	144; 5; 22	582; 25; 59	859; 37; 78	1291; 74; 131	2343; 304; 410	3740; 760; 922
0.6%	203; 6; 22	717; 26; 59	877; 37; 76	1291; 74; 128	2376; 307; 409	3748; 762; 918
0.4%	264; 6; 23	747; 26; 58	978; 38; 76	1291; 74; 125	2429; 312; 410	3757; 764; 914
0.3%	326; 6; 23	751; 26; 57	981; 38; 75	1328; 75; 125	2429; 312; 408	3767; 769; 916
0.2%	371; 6; 23	784; 26; 56	1006; 39; 74	1338; 76; 124	2448; 313; 407	3781; 773; 917
0.1%	447; 6; 22	828; 27; 56	1069; 40; 75	1354; 76; 123	2462; 315; 406	3791; 776; 916

The 3 values separated by semi colons in each cell of the table are:
(A) E_{100}, total genome-wide rare-het counts expected in a subject with 100% genome admixture from Asia (HapMap Chinese+Japanese)
(B) E_0, total genome-wide rare-het counts expected in an unadmixed Caucasian subject with 0% non-Caucasian admixture
(C) T_X, rare-het count threshold whose probabilty is p<0.001/1500 of being reached or exceeded under the null hypothesis

Note: As described in Methods, each minimum detectable percentage of genome admixture (F_{min}) shown in Supplementary Figures S40-S43 was found by solving the equation $T_X=(1-F_{min})E_0+F_{min}E_{100}$ for F_{min}, i.e., $F_{min}=(T_X-E_0)/(E_{100}-E_0)$ and then multiplying the result by 100 to convert F_{min} to a percentage. Counts in table are rounded to nearest integer which may produce a slight difference between F_{min} calculated from table counts and F_{min} in the corresponding figure.

Supplementary Table

Table S8. RHH detection of individual outlier with African admixture (Unthinned rare-het counts, Illm550K chip, p<0.001/1500)

Mean % admixture (Y) in the entire dataset	Rare-het frequency cutpoints (C_{het})					
Y% (=Yx100)	C_{het}=0.1%	C_{het}=0.3%	C_{het}=0.5%	C_{het}=1%	C_{het}=3%	C_{het}=5%
1.0%	83; 0; 9	1520; 4; 44	2495; 10; 66	3359; 24; 97	4755; 117; 228	5711; 315; 465
0.8%	185; 0; 11	1772; 5; 43	2711; 11; 63	3498; 27; 93	4773; 120; 221	5718; 317; 455
0.6%	394; 0; 13	2199; 7; 44	2869; 12; 58	3536; 27; 85	4773; 120; 209	5735; 319; 445
0.4%	817; 0; 16	2541; 8; 42	2975; 13; 52	3597; 29; 78	4837; 125; 205	5782; 326; 441
0.3%	1056; 1; 16	2616; 8; 39	2975; 13; 48	3662; 30; 75	4837; 125; 199	5804; 328; 437
0.2%	1313; 1; 17	2745; 9; 36	3034; 14; 44	3667; 31; 71	4837; 125; 193	5808; 330; 433
0.1%	1697; 3; 17	2788; 10; 33	3056; 14; 40	3706; 33; 69	4837; 125; 188	5814; 332; 430

The 3 values separated by semi colons in each cell of the table are:
(A) E_{100}, total genome-wide rare-het counts expected in a subject with 100% genome admixture from Africa (HapMap Yorubans)
(B) E_0, total genome-wide rare-het counts expected in an unadmixed Caucasian subject with 0% non-Caucasian admixture
(C) T_X, rare-het count threshold whose probabilty is p<0.001/1500 of being reached or exceeded under the null hypothesis

Note: As described in Methods, each minimum detectable percentage of genome admixture (F_{min}) shown in Supplementary Figures S40-S43 was found by solving the equation $T_X=(1-F_{min})E_0+F_{min}E_{100}$ for F_{min}, i.e., $F_{min}=(T_X-E_0)/(E_{100}-E_0)$ and then multiplying the result by 100 to convert F_{min} to a percentage. Counts in table are rounded to nearest integer which may produce a slight difference between F_{min} calculated from table counts and F_{min} in the corresponding figure.

Supplementary Table

Table S9. RHH detection of individual outlier with Asian admixture (Unthinned rare-het counts, Illm500K chip, p<0.001/1500)

Mean % admixture (Y) in the entire dataset	Rare-het frequency cutpoints (C_{het})					
Y% (=Yx100)	C_{het}=0.1%	C_{het}=0.3%	C_{het}=0.5%	C_{het}=1%	C_{het}=3%	C_{het}=5%
1.0%	81; 3; 15	317; 9; 31	546; 13; 42	791; 31; 71	1303; 120; 190	2071; 326; 436
0.8%	103; 3; 15	353; 9; 31	546; 13; 41	836; 32; 71	1303; 120; 187	2071; 326; 432
0.6%	121; 3; 15	446; 10; 32	582; 14; 40	836; 32; 69	1303; 120; 184	2097; 330; 433
0.4%	167; 3; 15	506; 10; 31	590; 14; 39	836; 32; 66	1330; 122; 185	2097; 330; 429
0.3%	210; 3; 15	520; 10; 31	611; 14; 39	836; 32; 65	1343; 123; 185	2110; 332; 430
0.2%	261; 3; 16	535; 11; 31	626; 15; 38	854; 33; 66	1362; 125; 185	2127; 335; 430
0.1%	271; 3; 15	564; 11; 31	653; 15; 38	869; 34; 67	1373; 126; 185	2127; 335; 428

The 3 values separated by semi colons in each cell of the table are:
(A) E_{100}, total genome-wide rare-het counts expected in a subject with 100% genome admixture from Asia (HapMap Chinese+Japanese)
(B) E_0, total genome-wide rare-het counts expected in an unadmixed Caucasian subject with 0% non-Caucasian admixture
(C) T_X, rare-het count threshold whose probabilty is p<0.001/1500 of being reached or exceeded under the null hypothesis

Note: As described in Methods, each minimum detectable percentage of genome admixture (F_{min}) shown in Supplementary Figures S40-S43 was found by solving the equation $T_X=(1-F_{min})E_0+F_{min}E_{100}$ for F_{min}, i.e., $F_{min}=(T_X-E_0)/(E_{100}-E_0)$ and then multiplying the result by 100 to convert F_{min} to a percentage. Counts in table are rounded to nearest integer which may produce a slight difference between F_{min} calculated from table counts and F_{min} in the corresponding figure.

Supplementary Table S10.
Inflated type-1 error rate at rare SNPs in modestly admixed cases and controls[a]

SNP MAF[b]	P-value Inflation[c]	Case-Control Sample Size[a]	Outliers as % of dataset[a]	Additional SNP p-values ≤0.001[d]		
				Affy500K	Illm550K	HapMap
0.05	0.2-fold	5000-5000	20%	2	2	Not Done
0.03	0.5-fold	5000-5000	20%	5	2	Not Done
0.01	1.9-fold	5000-5000	20%	5	3	Not Done
0.005	4.4-fold	5000-5000	20%	22	10	124
0.001	21-fold	5000-5000	20%	74	35	243
0.0005	37-fold	5000-5000	20%	106	68	475
0.00025	52-fold	5000-5000	20%	179	106	749
0.05	0.1-fold	10000-10000	10%	1	1	Not Done
0.03	0.2-fold	10000-10000	10%	2	1	Not Done
0.01	0.9-fold	10000-10000	10%	3	2	Not Done
0.005	2.0-fold	10000-10000	10%	10	5	56
0.001	10-fold	10000-10000	10%	35	17	116
0.0005	21-fold	10000-10000	10%	60	38	270
0.00025	37-fold	10000-10000	10%	127	76	533
0.05	0.06-fld	20000-20000	5%	1	1	Not Done
0.03	0.1-fold	20000-20000	5%	1	0	Not Done
0.01	0.4-fold	20000-20000	5%	1	1	Not Done
0.005	1.0-fold	20000-20000	5%	2	1	28
0.001	4.2-fold	20000-20000	5%	15	7	49
0.0005	11-fold	20000-20000	5%	32	20	141
0.00025	25-fold	20000-20000	5%	86	52	360

[a]Disease association results assume a model dataset with an equal number of cases and controls (5000, 10000 or 20000 of each). All subjects are unadmixed except for 1000 case outliers with admixed DNA covering 5% of the subject's genome and 1000 control outliers with admixed DNA covering 1% of their genomes, implying that modest ethnic outliers represent 20%, 10% or 5% of each model dataset.

[b]Disease association results are for "neutral" (i.e. non-disease causing) SNPs with minor allele frequency (MAF) in the non-outlier population as shown (0.05 to 0.00025) but with corresponding allele frequency in the outlier population that is 0.2 or higher.

[c]Increase in type I error rate for disease-neutral SNPs with non-outlier and outlier allele frequencies as specified in footnote b assuming a 0.001 significance level.. Since the baseline (random) type I error rate is approximately 1 "false-positive" SNP per 1000 tested, "1-fold" means that at least 1 additional false-positive association (p-value≤0.001) would be observed for every 1000 SNPs tested having the non-outlier and outlier allele frequencies specified in footnote b.

[d]Lower-bound estimate for additional false-positive SNP associations (p-value≤0.001) in a GWA scan with the Affymetrix 500K or Illumina 550K array or if all HapMap SNPs were tested. Since sequencing of ENCODE regions found a SNP density 10-times higher than HapMap SNPs (5,6), all HapMap SNPs with non-outlier and outlier allele frequencies as specified in footnote b were used to roughly estimate the number of such SNPs in future GWA scans aimed at testing many more rare SNPs (now being discovered by large-scale sequencing projects such as '1000 Genomes'). Allele frequencies in ~1400 58BC controls and in HapMap Yorubans were considered as the frequencies for the non-outlier and outlier populations, respectively, in order to count SNPs having appropriate non-outlier and outlier allele frequencies as in footnote b (see SUPPLEMENTARY METHODS for more details).