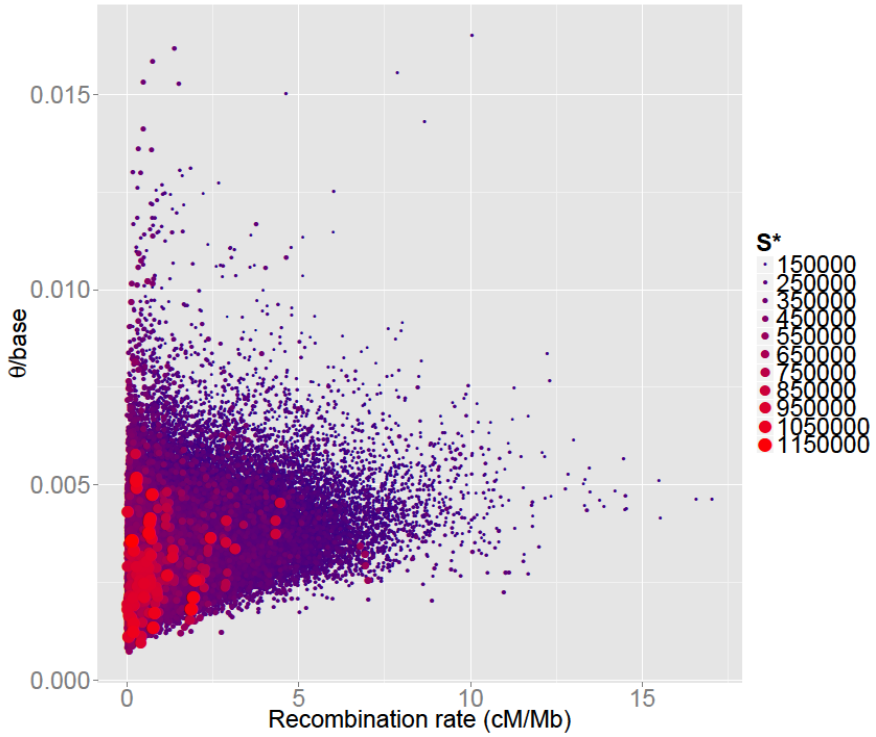
**(A)**




**(B)**

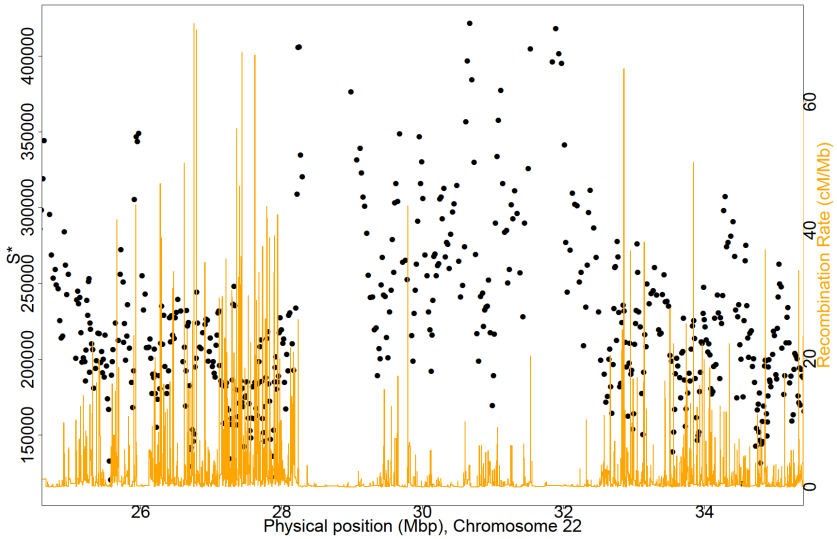



**Figure S1. Dependence of *S** statistic on local heterozygosity and recombination rate**. (A) Each point represents a window of 200 SNVs. The size of each point represents its *S** value in log-scale. *S** is negatively correlated with both local heterozygosity (θ/base, Pearson's correlation: -0.41, $p<2.2x10^{-16}$) and recombination rate (cM/Mb, Pearson's correlation: -0.40, $p<2.2x10^{-16}$). (B) An example of the negative correlation between S* and local recombination rate at locus Chr22:25000000-35000000. Each point is a window of 200 SNVs and the orange distribution represents the recombination variation across this genomic region.

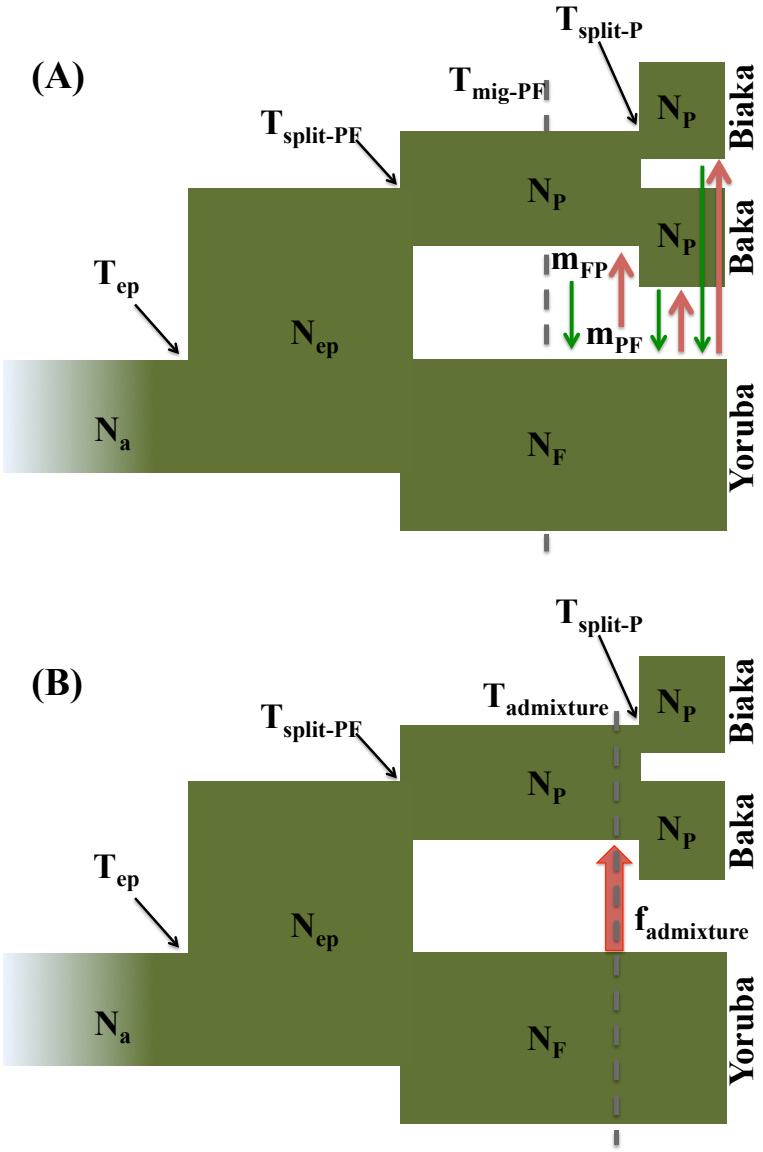



**Figure S2. Schematic of the two best-fit demographic null models without archaic admixture for African farmer (Yoruba) and Pygmy (Baka and Biaka) populations from Hsieh et al. (in review).** (A) The continuous asymmetric gene flow model (Model-1) with the 10 free parameters labeled. (B) The single-pulse admixture model (Model-2) with the 9 free parameters labeled. The corresponding demographic parameters are in Table S1.

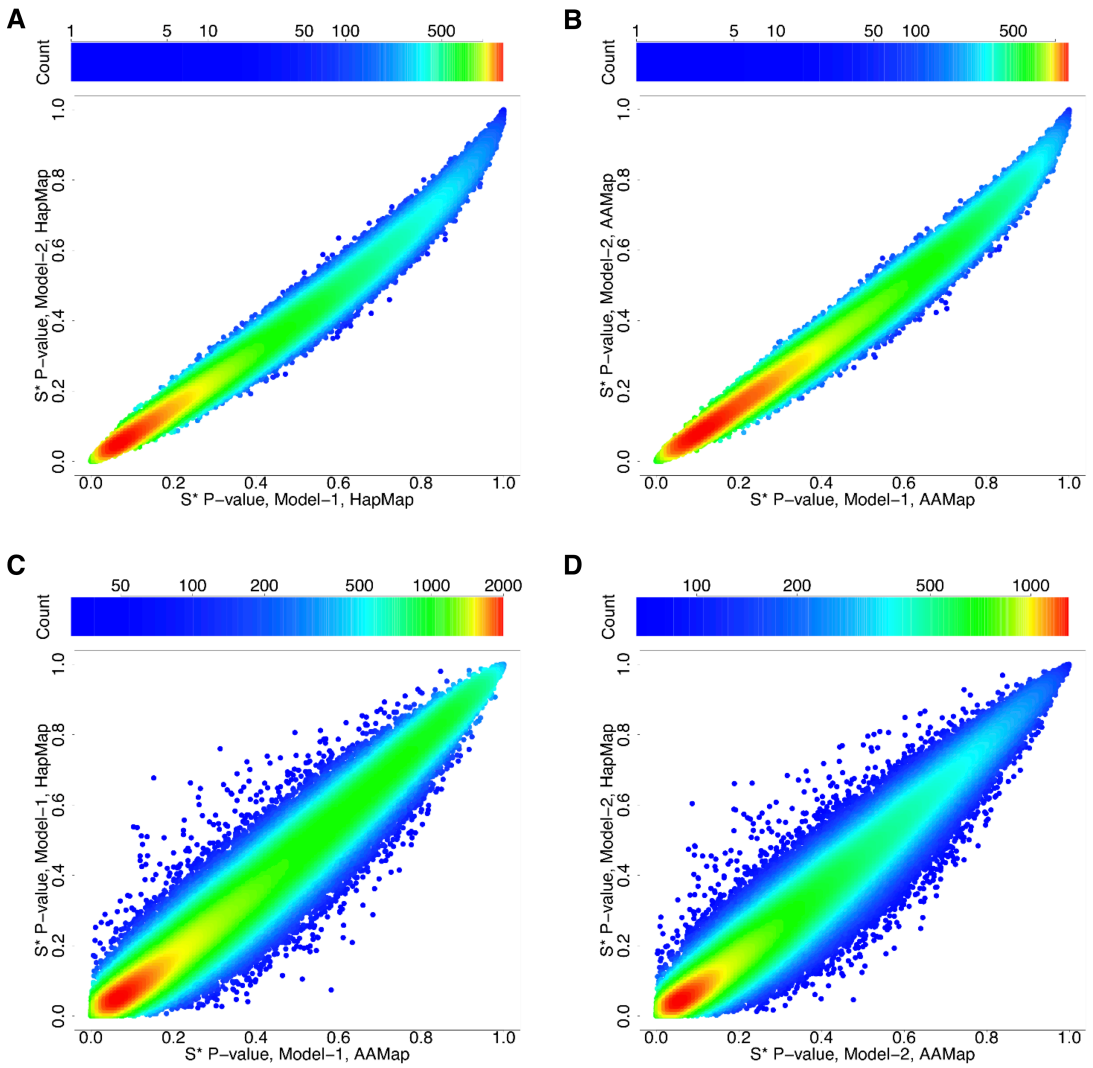



**Figure S3. Strong correlation of *S** p-values between (A, B) the two demographic null models and (C, D) the two recombination maps.** Each point is a window of 200 SNVs and color represents the density of the points. (A) Model-1 vs. Model-2, using HapMap Yoruba map ($\rho$=0.990, $p<2.2\times10^{-16}$). (B) Model-1 vs. Model-2, using African American map ($\rho$=0.990, $p<2.2\times10^{-16}$). (C) HapMap Yoruba map vs. African American map, using Model-1 ($\rho$=0.973, $p<2.2\times10^{-16}$). (D) HapMap Yoruba map vs. African American map, using Model-2 ($\rho$=0.965, $p<2.2\times10^{-16}$).

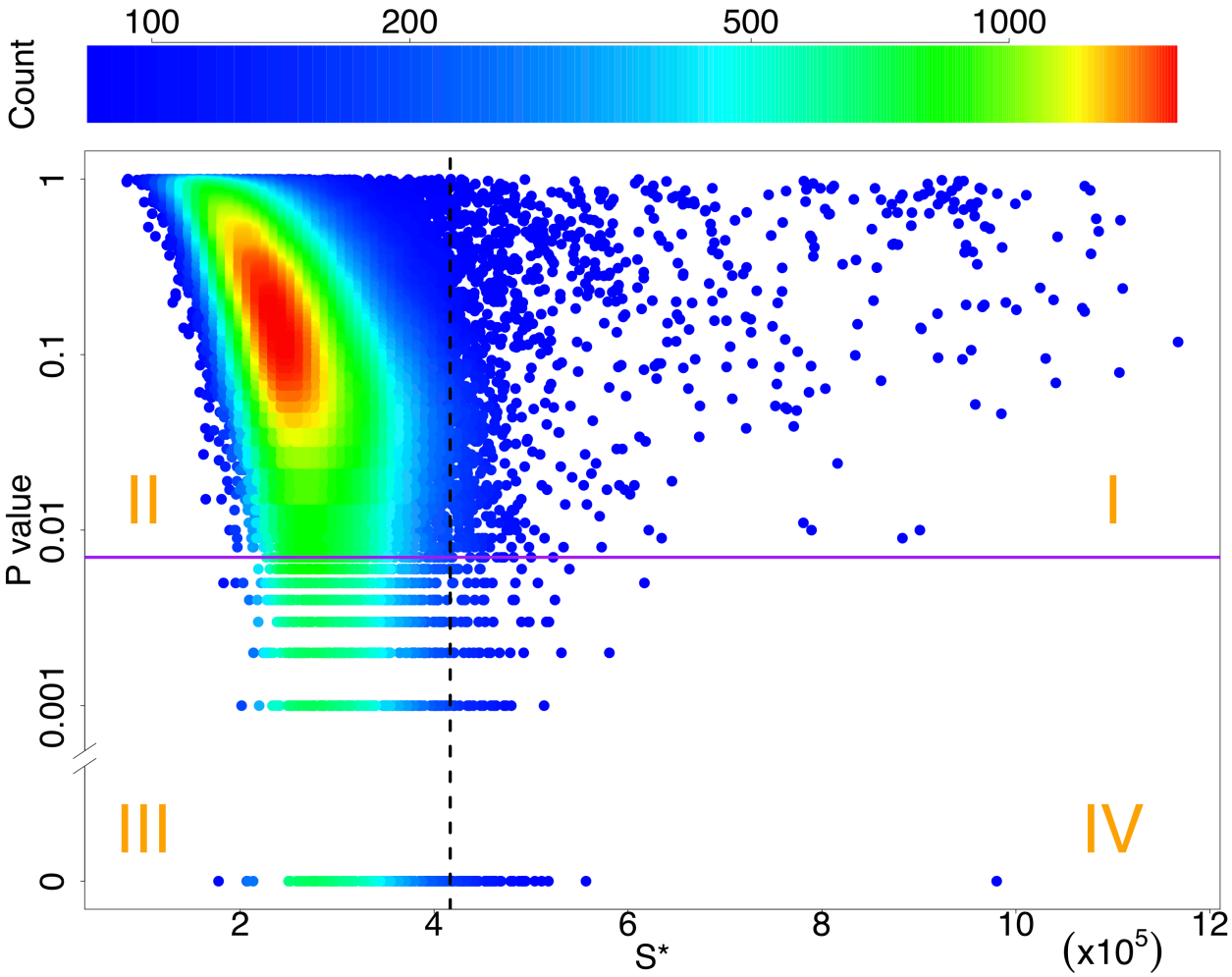



**Figure S4. Importance of using p-values to prioritizing candidates in the *S** analysis.** Each point is a window of 200 SNPs, and color represents the density of points. The vertical black dashed line and the horizontal purple solid line are the top 1% significance cutoffs for the *S** and *S** p-value distributions, respectively. Windows in Quadrant I are outliers in the *S** distribution but are not statistically significant when the effects of demography and genome architecture are controlled for. In Quadrant III are the many windows that are statistically significant even though their *S** values are modest.

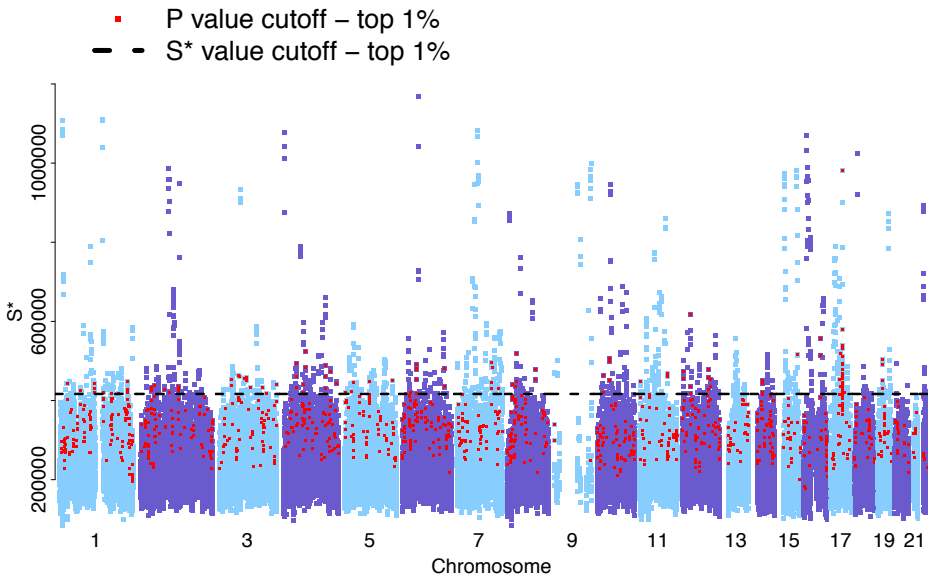



**Figure S5. Genome-wide Manhattan plot of *S** statistic**. Each point is a window of 200 SNPs. The dash line is the top 1% outlier cutoff in the empirical distribution of *S** statistic. Red points are the candidate introgressive loci from the top 1% *S** p-value distribution using all the four alternative simulation sets. Most of chromosome 9 was excluded by our quality control filters.

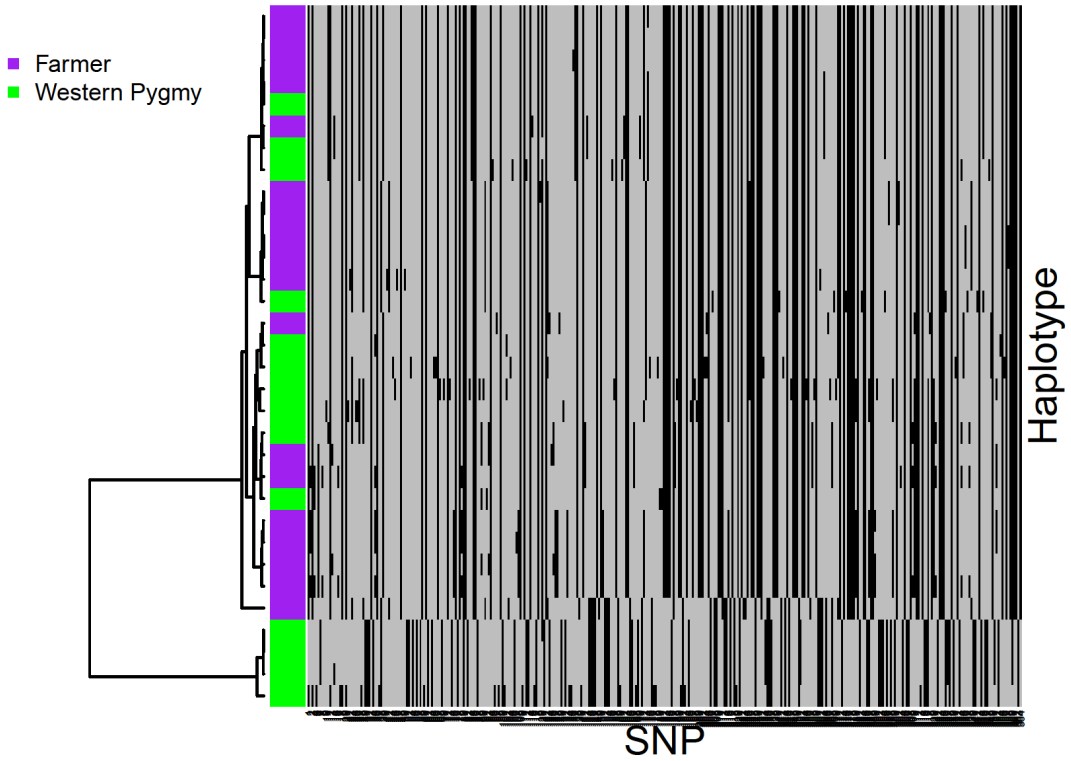



**Figure S6. Hierarchical clustering of the haplotypes for the candidate introgressive locus Chr16:8702222-8747116.** Columns are SNPs, with grey and black for ancestral and derived alleles, respectively, while rows show individual haplotypes. The four haplotypes at the bottom cluster are the four haplotypes on the basal branch in the haplotype network plot (Fig. 2).

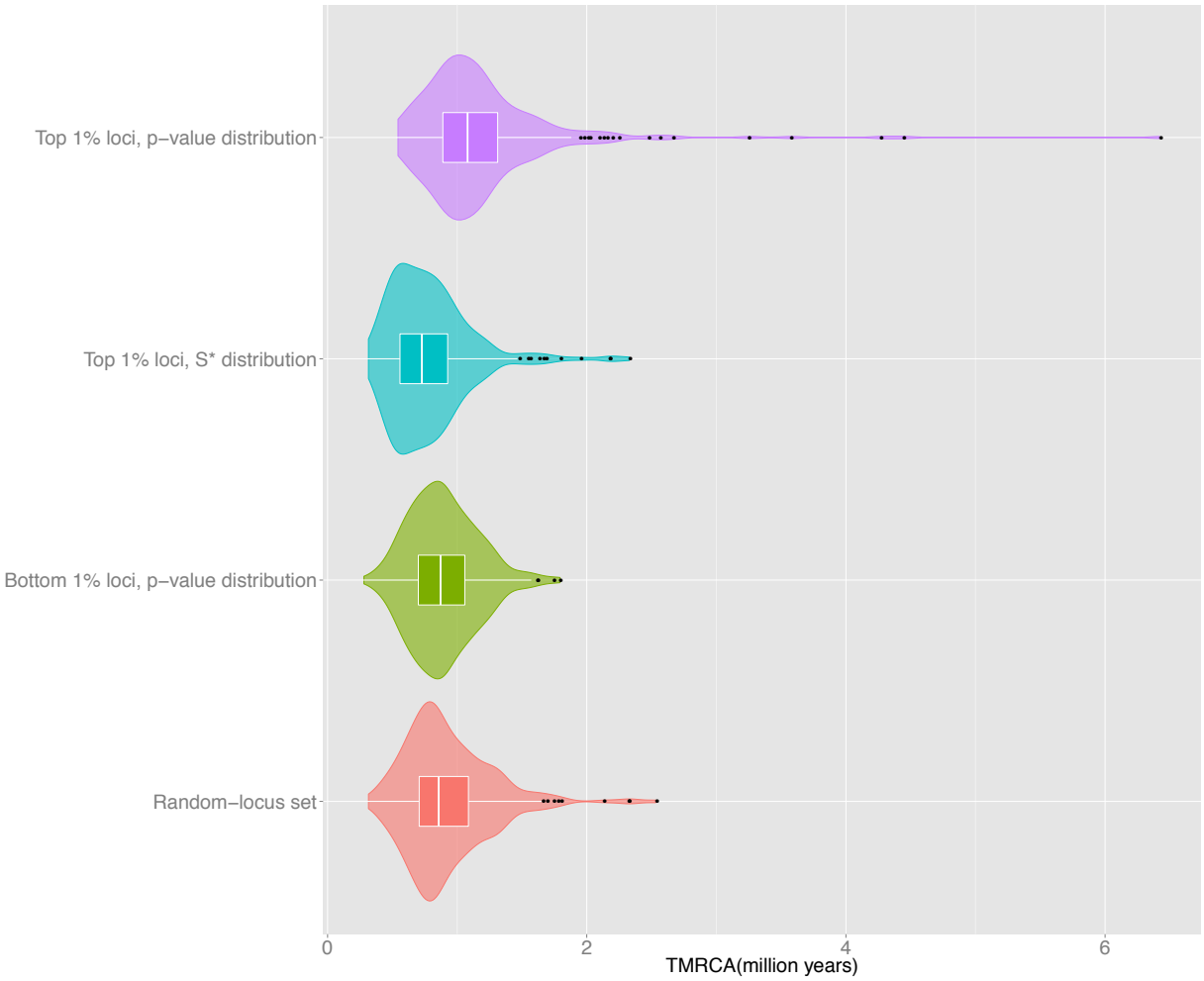



**Figure S7. Substantially older TMRCA (million years) of the top 1% *S** p-value candidate introgressive regions.** TMRCA were calculated for loci from four data sets: the top 1% loci in the *S** p-value (1st row, purple) and the *S** empirical distribution (2nd row, blue), the bottom 1% loci in the *S** p-value distribution (3rd row, green), and a set of random loci with the same amount of sequences as the top 1% loci in the *S** p-value distribution (4th row, pink). For each data set, a violin plot is drawn to show the density of TMRCA with a corresponding boxplot embedded within.

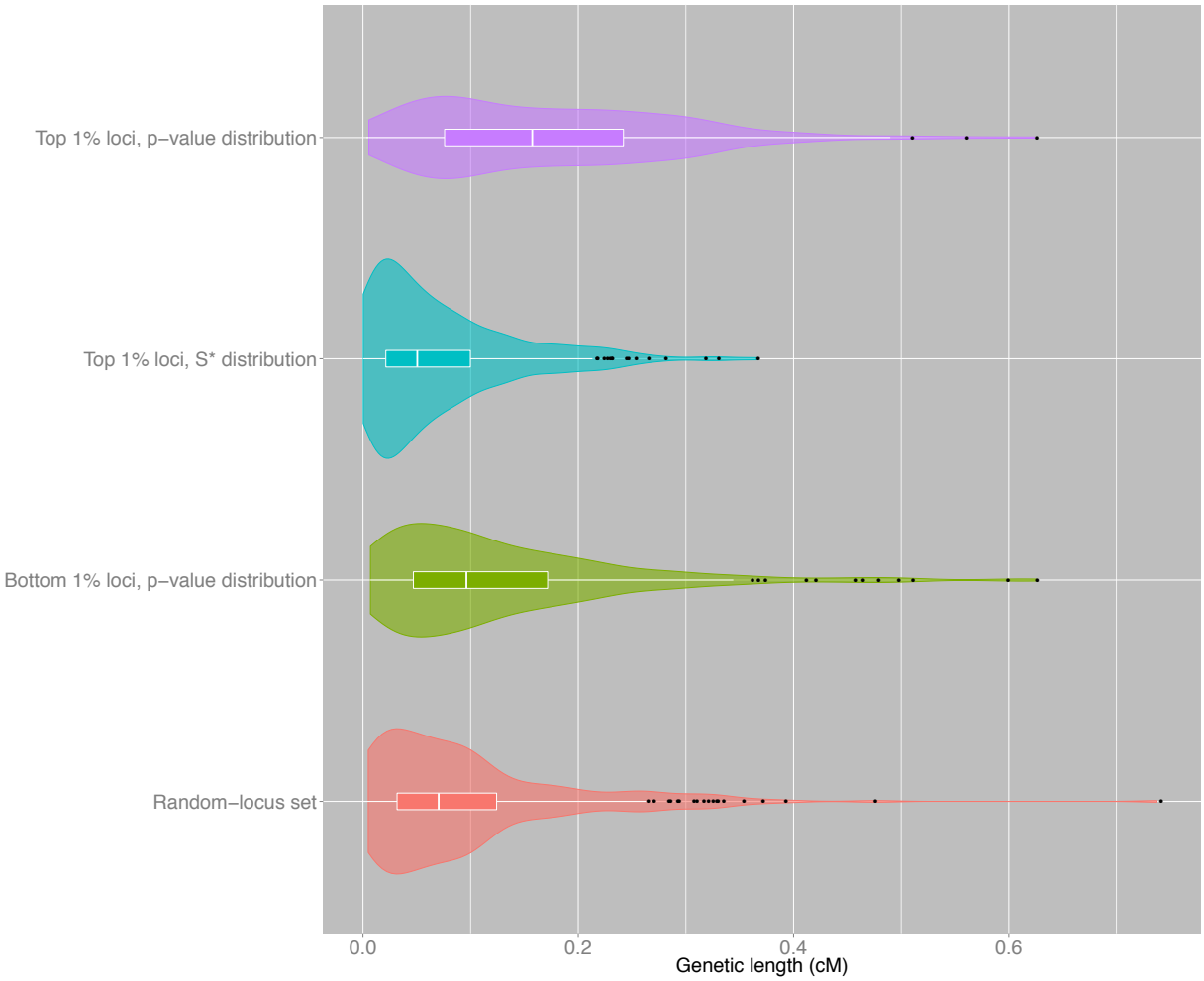



**Figure S8. Significantly longer genetic length (cM) of the top 1% *S** p-value candidate introgressive regions.** As in Figure S7, but for the genetic length of each locus.

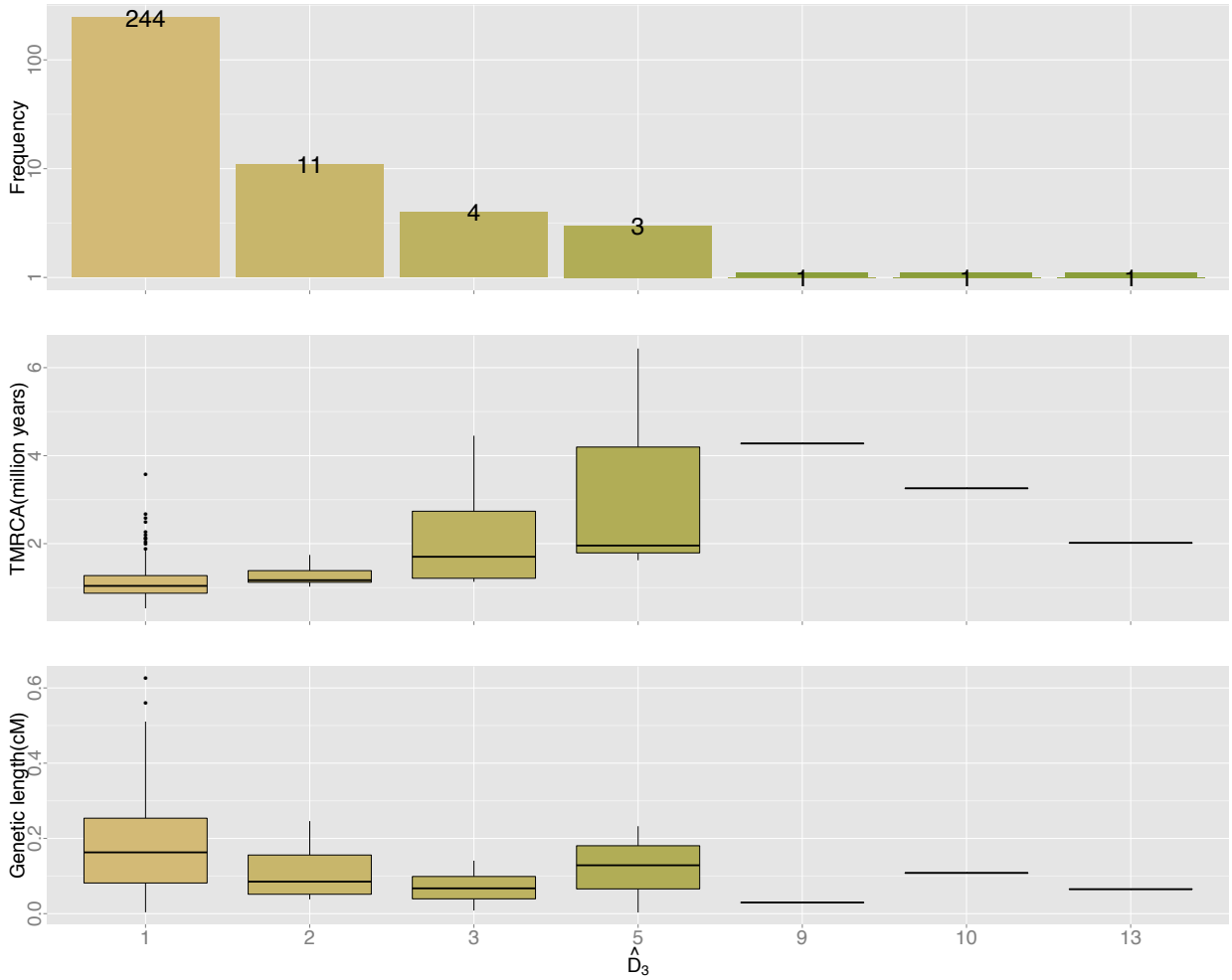



**Figure S9. Distributions of TMRCA (million years) and genetic length (cM) for the top 1% $S^*$ p-value candidate introgressive regions stratified by $\widehat{D_3}$,** which estimates the minimum of the sizes of the two most basal lineages for a given locus, and thus is sensitive to the strength of admixture. Top, middle, and bottom panels are the locus count, the distribution of TMRCA, and the distribution of genetic length, respectively, for each class of $\widehat{D_3}$.

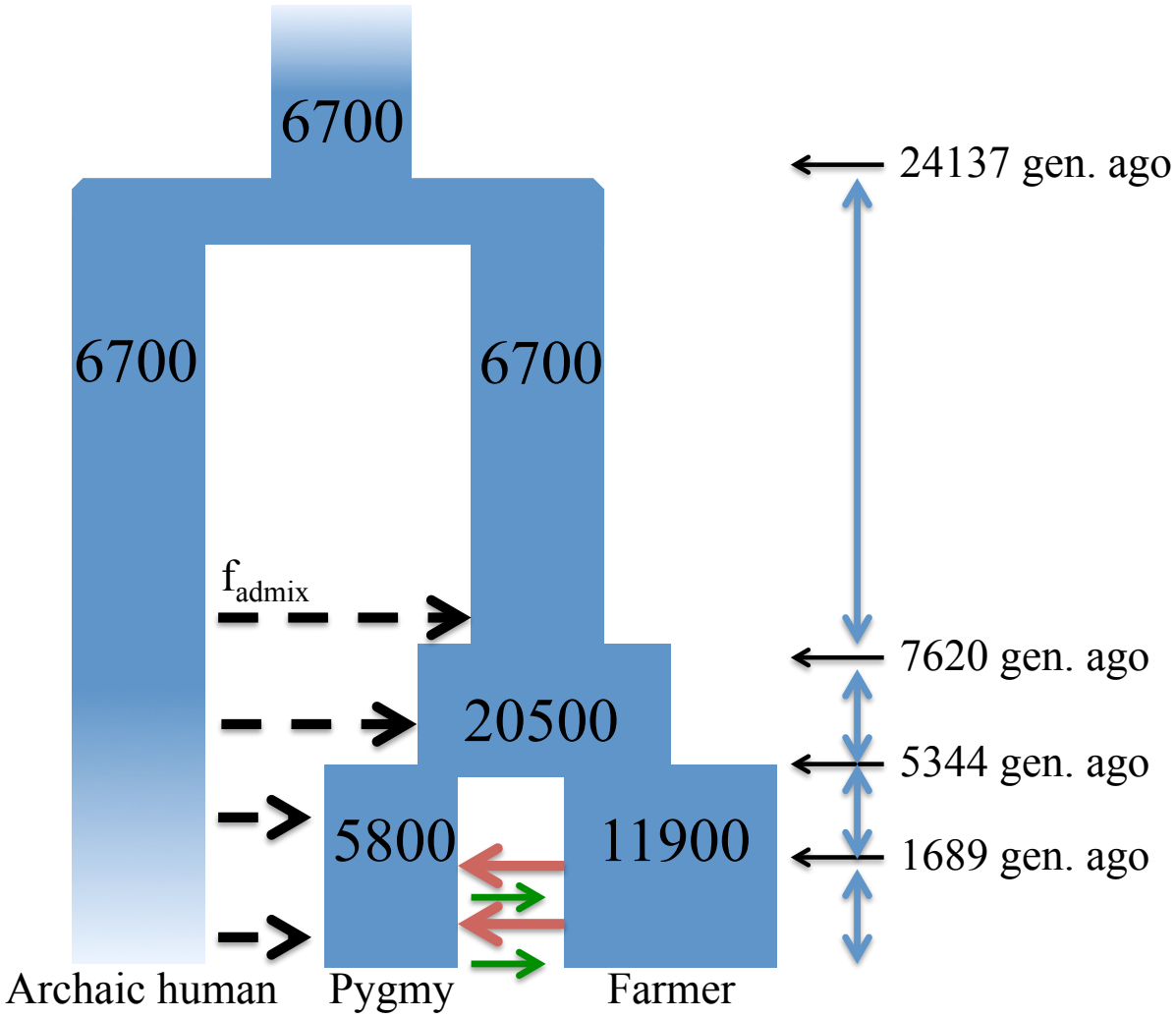



**Figure S10. Demographic schematic for whole-genome simulations with archaic introgression.** Fixed model parameters, including the effective population sizes for ancestors of modern human, Pygmy, and Farmer populations (numbers inside the tree), times of divergence between populations, as well as the asymmetric gene-flow between Pygmy and Farmer populations (red/green arrows), were chosen based on Hammer et al. (2011) and Hsieh et al. (2015). Note that for simplicity, we assumed that both the common ancestor of archaic/modern humans and the archaic human population have the same effective population size (6,700 individuals). For simulations with single-wave archaic admixture, the time of archaic admixture into modern human lineages could occur at 1200, 2700, 5500, or 7800 generations ago, which corresponds to each of the four time intervals illustrated on the right. For the two-wave archaic admixture simulations, we set the two events to be 7800 and 300 generations ago in order to investigate extreme cases. The admixture proportion ($f_{admix}$) was set to be 2%, 5%, or 10% in each individual simulation (**Table S2-5**).

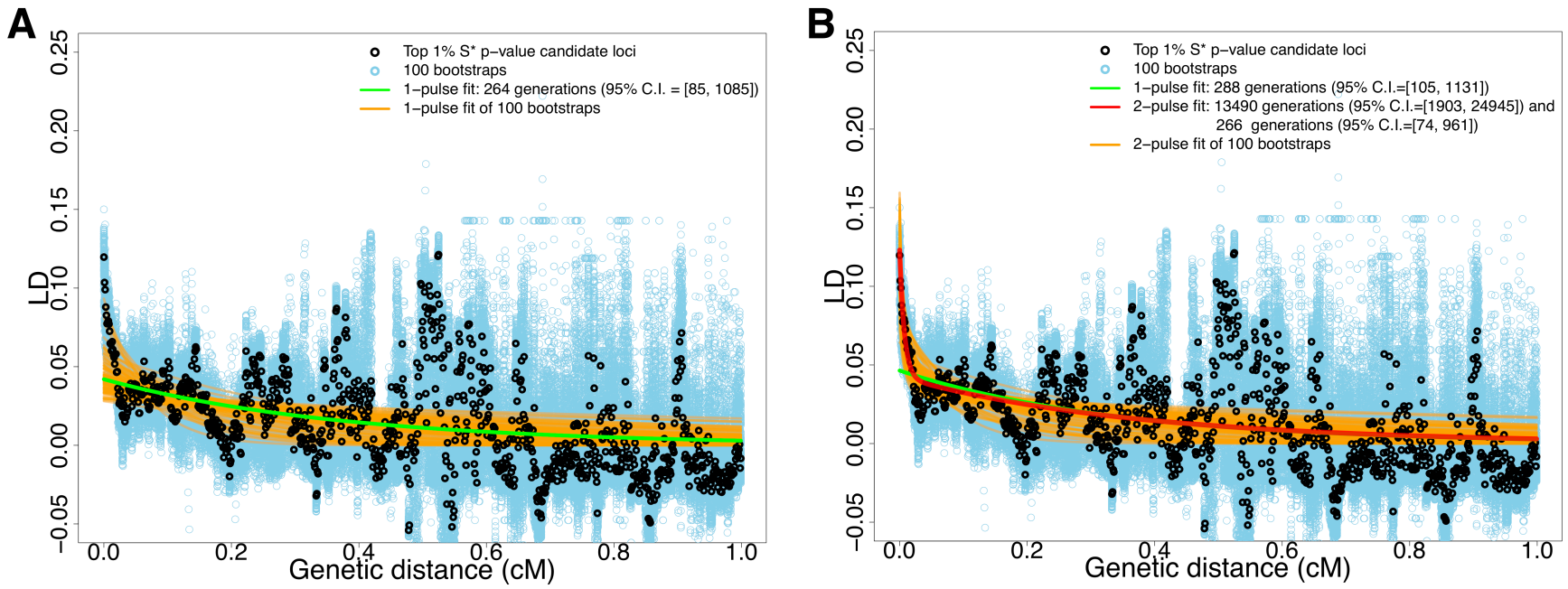



**Figure S11. Decay of pairwise LD with respect to genetic distance for SNPs ascertained from the top 1% candidate introgressive loci.** As in Figure 4, but for the case that genetic distance is calculated using the African American recombination map. **(A)** Fitting LD decay curve with genetic distance 0.02 – 1 cM. **(B)** Fitting LD decay curve with genetic distance 0.002 – 1 cM.

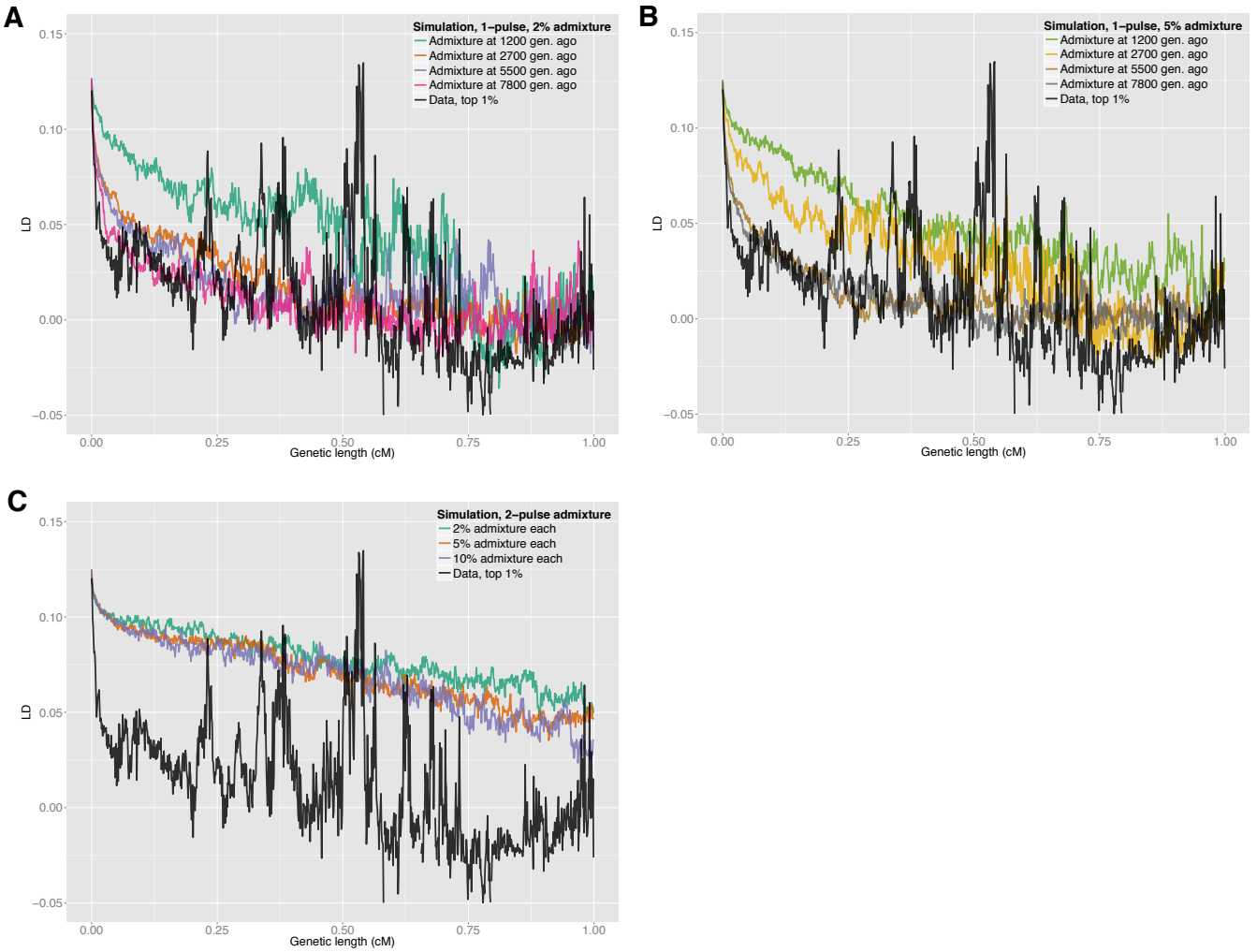



**Figure S12. Comparisons of the patterns of linkage disequilibrium (LD) between data and whole-genome archaic admixture simulations (Figure S10)**. LD estimates were estimated as described in Materials and Methods for genetic distance range 0 – 1 cM. LD of the data's top 1% *S** p-value candidate loci was compared with that from (A) single-wave, 2% archaic admixture, (B) single-wave, 5% archaic admixture, and (C) two-wave archaic admixture with 2%, 5%, or 10% admixture in each pulse.

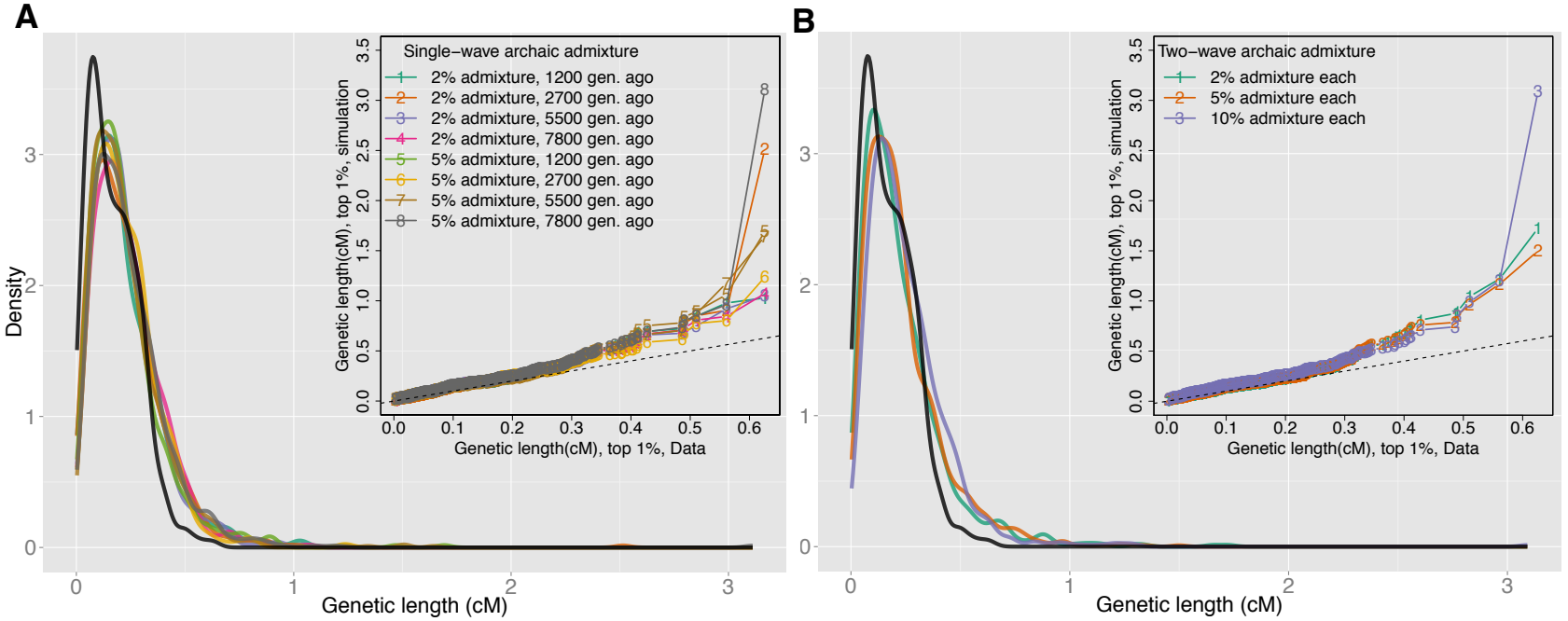



**Figure S13. Comparisons of the genetic length distributions of the top 1% *S* p*-value candidate loci from data and whole-genome archaic admixture simulations (Figure S10)**. (A) single-wave, 2% and 5% archaic admixture, (B) two-wave archaic admixture with 2%, 5%, or 10% each.

**Table S1. Parameter estimates and confidence intervals for two best-fit demographic models from Hsieh et al. (2015). Model-1: continuous asymmetric gene flow. Model-2: single-pulse gene flow.** Estimates and confidence intervals are shown for effective population sizes (N), times (T) of population divergence and gene flow onset, levels of gene flow (m) between farmer (F) and Pygmy (P) populations. $T_{admixture}$ and $f_{admixture}$ refer to the timing and strength of the single-pulse gene flow from the farmers (F) to Pygmies (P) in Model-2.

| | Model-1 (Asymmetric gene flow) | | Model-2 (Single-pulse gene flow) | |
|---|---|---|---|---|
| **Demographic parameters** | **Estimates** | **[+]95% C.I.** | **Estimates** | **[+]95% C.I.** |
| $N_a$: $N_e$[*] ancestral population | 6,727 | 6,676 – 6,819 | 6,735 | 6,671- 6,826 |
| $N_{ep}$: $N_e$ ancestral population after expansion | 20,473 | 15,560 – 27,561 | 15,236 | 14,436 – 15,894 |
| $N_F$: $N_e$ contemporary Farmer (F) | 11,900 | 11,714 – 12,138 | 13,854 | 13,721 – 14,055 |
| $N_P$: $N_e$ contemporary Pygmy (P) | 5,831 | 5,631 – 5,986 | 5,373 | 5,217 – 5,530 |
| $T_{ep}$: Time[†] of ancestral expansion | 221,118 | 210,513 – 236,634 | 232,629 | 223,172 – 244,327 |
| $T_{split\text{-}PF}$: Time of P-F split | 155,671 | 139,661 – 164,280 | 89,645 | 85,503– 91,725 |
| $T_{mig\text{-}PF}$: Time of onset of gene flow between P and F | 39,337 | 36,565 – 43,550 | – | – |
| $T_{admixture}$: Time of admixture from F to P | – | – | 7,136 | 6,887 – 7,656 |
| $T_{split\text{-}P}$: Time of split between the two P populations | 5,139 | 4,762 – 5,630 | 4,049 | 3,803 - 4,396 |
| $m_{PF}$: Gene flow[‡] (P ← F) | $9.0\times10^{-4}$ | $8.4\times10^{-4}$ – $9.4\times10^{-4}$ | – | – |
| $m_{FP}$: Gene flow (F ← P) | $9.1\times10^{-5}$ | $8.2\times10^{-5}$- $1\times10^{-4}$ | – | – |
| $f_{admixture}$: Strength of admixture (P ← F) | – | – | 0.6799 | 0.6789 – 0.6818 |

[*]Effective population size in individuals. [†]Time in years, assuming 25 years per generation and mutation rate $2.35\times10^{-8}$ per base per generation (Gutenkunst et al. 2009). [‡]Fraction of the population each generation that are new migrants. [+]Confidence intervals estimated using100 conventional bootstraps

**Table S2. Hypothesis testing using linkage disequilibrium (LD) decay information on whole genome single-wave archaic admixture simulations: fitting distance range 0.02 – 1 cM**. This LD approach was applied to Top 1% *S** p-value candidate loci in whole genome simulations based on **Figure S10**, with different admixture proportions and times specified in the table. Dates were all in units of generations. All analyses for *S**, p-values, and LD decay curve fitting were performed using the same pipeline as for the data described in the main text. Note that n.a. refers to not able to obtain stable model fit.

| Proportion of archaic admixture | Simulated time of admixture | Inference for single-wave archaic admixture | Inference for the two-wave archaic admixture | | Falsely reject single-wave admixture? (p-value) |
|---|---|---|---|---|---|
| | | Inferred date of the event (95% C.I.) | Inferred date of the 1st event (95% C.I.) | Inferred date of the 2nd event(95% C.I.) | |
| 2% | 1200 | 212 (91, 315) | n.a. | n.a. | NO (n.a.) |
| | 2700 | 424 (271, 591) | 23121 (657, 20173) | 424 (98, 516) | NO (0.9876) |
| | 5500 | 364 (86, 669) | 885 (633, 19774) | 117 (4, 626) | YES ($<2.2\times10^{-16}$) |
| | 7800 | 534 (268, 778) | 25063 (1204, 41661) | 513 (158, 750) | YES ($2\times10^{-4}$) |
| 5% | 1200 | 176 (91, 315) | 736 (404, 6212) | 165 (11, 243) | NO (0.1476) |
| | 2700 | 284 (138, 424) | 38363 (828, 29254) | 283 (101, 393) | NO (0.9088) |
| | 5500 | 631 (404, 847) | n.a. | n.a. | NO (n.a.) |
| | 7800 | 558 (375, 793) | 8315 (789, 27542) | 519 (17, 749) | YES ($<1.45\times10^{-7}$) |

**Table S3. Hypothesis testing using linkage disequilibrium (LD) decay information on whole genome single-wave archaic admixture simulations: fitting distance range 0.002 – 1 cM**. Same as Table S3, but fitting LD decay curves using LD data from genetic distance range 0.002 – 1 cM in order to explore the efficacy of this LD method for the inference of older admixture events.

| Proportion of archaic admixture | Simulated time of admixture | Inference for single-wave archaic admixture | Inference for the two-wave archaic admixture | | Falsely reject single-wave admixture? (p-value) |
|---|---|---|---|---|---|
| | | Inferred date of the event (95% C.I.) | Inferred date of the 1st event (95% C.I.) | Inferred date of the 2nd event (95% C.I.) | |
| 2% | 1200 | 214 (96, 316) | 15048 (1529, 96348) | 213 (81, 307) | YES ($4.1\times10^{-2}$) |
| | 2700 | 440 (289, 613) | 12046 (1594, 30998) | 423 (167, 588) | YES ($5.5\times10^{-14}$) |
| | 5500 | 419 (119, 711) | 1004 (694, 19463) | 128 (8, 664) | YES ($<2.2\times10^{-16}$) |
| | 7800 | 626 (358, 908) | 6982 (1584, 12479) | 501 (165, 759) | YES ($<2.2\times10^{-16}$) |
| 5% | 1200 | 178 (96, 313) | 895 (383, 44811) | 165 (7, 255) | YES ($1.3\times10^{-4}$) |
| | 2700 | 289 (148, 426) | 10214 (1026, 55904) | 283 (122, 414) | YES ($1.5\times10^{-3}$) |
| | 5500 | 702 (471, 920) | 6723 (1006, 21364) | 593 (58, 850) | YES ($<2.2\times10^{-16}$) |
| | 7800 | 629 (446, 870) | 6535 (996, 18060) | 518 (77, 788) | YES ($<2.2\times10^{-16}$) |

**Table S4. Hypothesis testing using linkage disequilibrium (LD) decay information on whole genome two-wave archaic admixture simulations: fitting distance range 0.02 – 1 cM**. Similar to Table S3, but the whole genome simulations here incorporated two waves of archaic admixture at 7800 and 300 generations ago, with the same admixture proportion in each simulation.

| Proportion of each archaic admixture | Inference for single-wave archaic admixture | Inference for the two-wave archaic admixture | | Correctly rejected single-wave admixture? (p-value) |
|---|---|---|---|---|
| | Inferred date of the event (95% C.I.) | Inferred date of the 1st event (95% C.I.) | Inferred date of the 2nd event (95% C.I.) | |
| 2% | 58 (26, 87) | 31689 (190, 23234) | 58 (17, 59) | NO (0.7446) |
| 5% | 81 (53, 122) | 22025 (595, 16473) | 81 (40, 96) | NO (0.6829) |
| 10% | 89 (56, 131) | 27981 (1311, 22063) | 89 (50, 111) | NO (0.9949) |

**Table S5. Hypothesis testing using linkage disequilibrium (LD) decay information on whole genome two-wave archaic admixture simulations: fitting distance range 0.002 – 1 cM**. Same as Table S5, but fitting LD decay curves using LD data from genetic distance range 0.002 – 1 cM in order to explore the efficacy of this LD method for the inference of older admixture events.

| Proportion of each archaic admixture | Inference for single-wave archaic admixture | Inference for the two-wave archaic admixture | | Correctly rejected single-wave admixture? (p-value) |
|---|---|---|---|---|
| | Inferred date of the event (95% C.I.) | Inferred date of the 1st event (95% C.I.) | Inferred date of the 2nd event (95% C.I.) | |
| 2% | 59 (28, 87) | 13312 (263, 37151) | 58 (12, 83) | YES ($4.0 \times 10^{-7}$) |
| 5% | 82 (54, 122) | 11538 (2926, 42250) | 81 (49, 124) | YES ($7.2 \times 10^{-10}$) |
| 10% | 90 (57, 131) | 13198 (2098, 56576) | 89 (53, 125) | YES ($1.7 \times 10^{-3}$) |

**Table S6. Estimation of the false discovery rate (FDR) for the top 1% *S** p-value candidate introgressive loci.** FDRs were estimated based on each of the four alternative simulation sets separately using the method of Williamson *et al*. (32)

| Simulation set | FDR estimate for the top 1% *S** *P*-value candidate introgressed loci |
|---|---|
| Model-1, HapMap | 0.42 |
| Model-2, HapMap | 0.19 |
| Model-1, AAMap | 0.68 |
| Model-2, AAMap | 0.31 |

**Table S7.** Maximum likelihood estimates for the parameters of the LD decay analysis based on the top 1% *S** p-*value* candidate loci. Note that n.a. refers to unable to obtain stable model fit.

| Genetic distance range | Recombination map | Inference for single-wave archaic admixture | | Inference for the two-wave archaic admixture | | | |
|---|---|---|---|---|---|---|---|
| | | Amplitude (95% C.I.) | Decay rate (95% C.I.) | Amplitude 1 (95% C.I.) | Decay rate 1 (95% C.I.) | Amplitude 2 (95% C.I.) | Decay rate 2 (95% C.I.) |
| 0.02 – 1 cM | HapMap (Yoruba) | 0.04 (0.03, 0.07) | 3.12 (0.45, 9.75) | n.a. | n.a. | n.a. | n.a. |
| | African American | 0.04 (0.03, 0.08) | 2.64 (0.85, 10.85) | n.a. | n.a. | n.a. | n.a. |
| 0.002 – 1 cM | HapMap (Yoruba) | 0.05 (0.04, 0.09) | 3.27 (1.06, 13.76) | 0.09 (0.06, 0.13) | 193.44 (24.38, 447.72) | 0.05 (0.02, 0.06) | 3.12 (0.81, 9.34) |
| | African American | 0.04 (0.03, 0.08) | 2.88 (1.05, 11.31) | 0.08 (0.04, 0.10) | 134.90 (19.03, 249.45) | 0.04 (0.01, 0.07) | 2.66 (0.74, 9.61) |

## References


Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics* **5**(10): e1000695.

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD. 2011. Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences of the United States of America* **108**(37): 15123-15128.

Hsieh P, Veeramah KR, Lachance J, Tishkoff SA, Wall JD, Hammer MF, Gutenkunst RN. 2015. Whole genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. *bioRxiv*: 022194.