Supplementary Online Information

**A Chronological Atlas of Natural Selection in the Human Genome during the Past Half-million Years**

Hang Zhou1,2,10, Sile Hu1,2,10, Rostislav Matveev3,4, Qianhui Yu1,2, Jing Li1, Philipp Khaitovich1, Li Jin1,5, Michael Lachmann3,6, Mark Stoneking3, Qiaomei Fu7,8,9 & Kun Tang1

May 04, 2015

1Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Science, Shanghai 200031, China.

2University of Chinese Academy of Sceiences, Beijing 100049, China.

3Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany.

4Max Planck Institute for Mathematics in the Sciences, D-04103 Leipzig, Germany.

5MOE Key Laboratory of Contemporary Anthropology, Fudan University, Shanghai 200433, China.

6Santa Fe Institute, Santa Fe, New Mexico 87501, USA.

7Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

8Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA

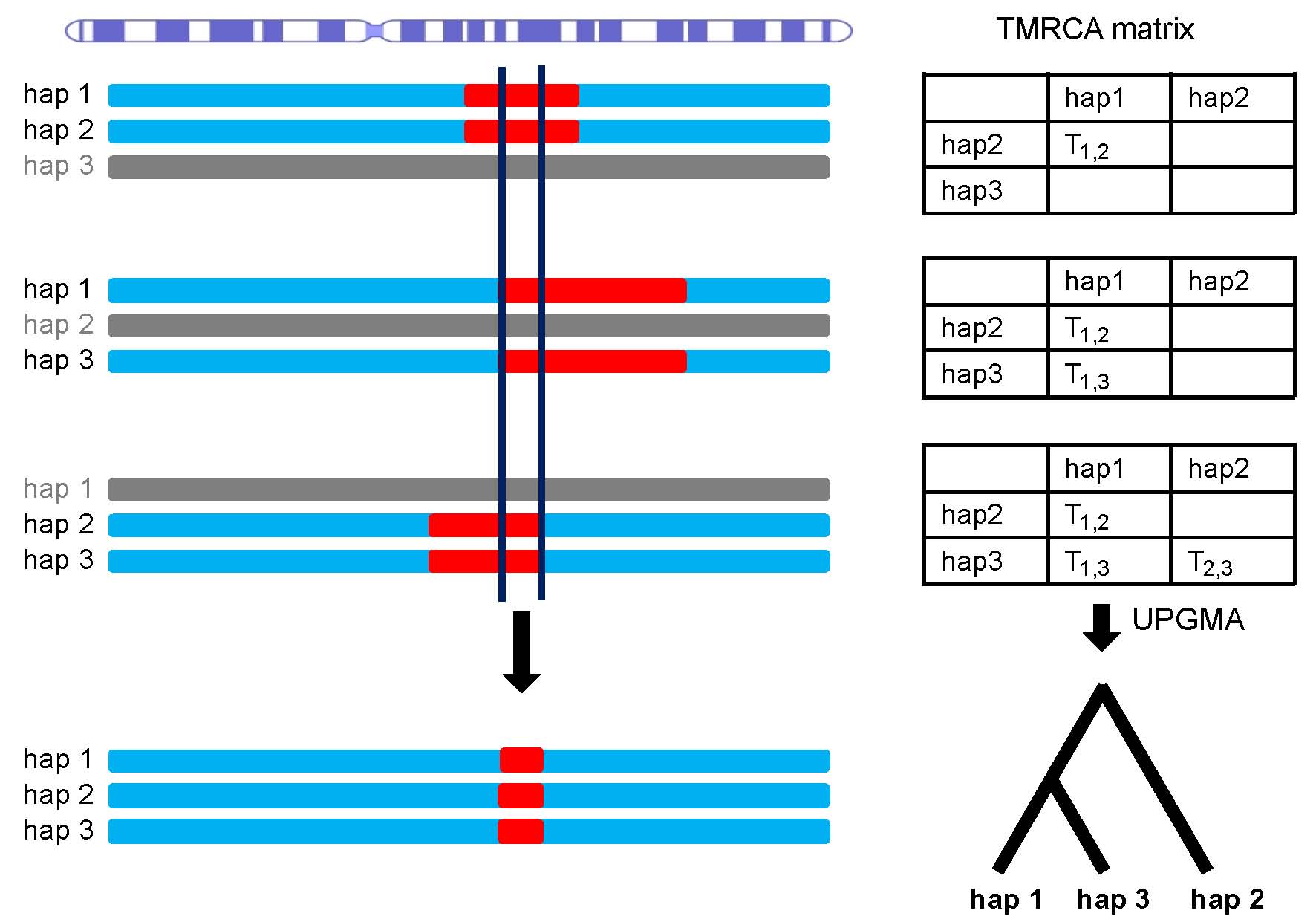
9Key Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences, IVPP, CAS, Beijing 100044, China.

10These authors contributed equally to this work.

Correspondence could be addressed to K.T. (tangkun@picb.ac.cn)

Tables and Figures 2-50

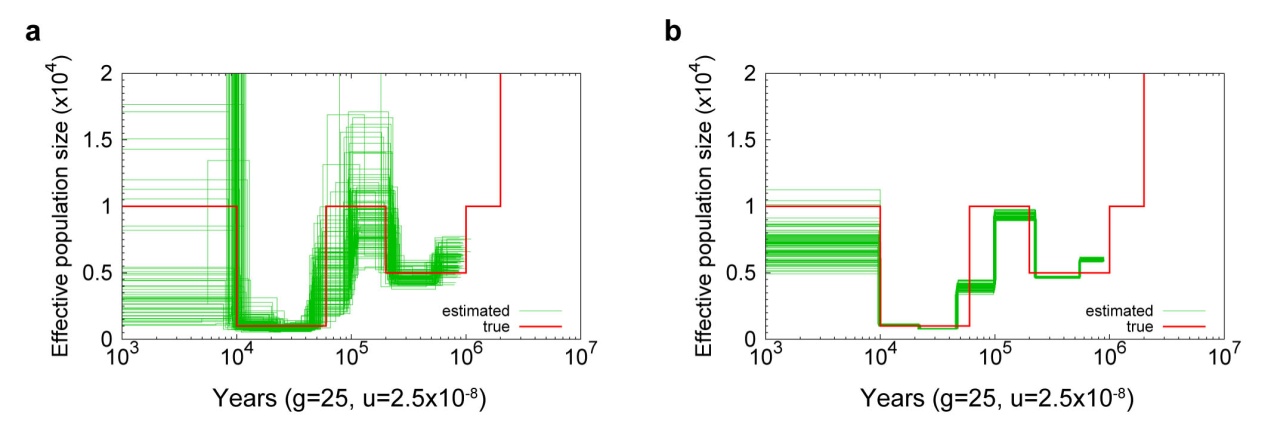
Supplementary Note 51-63



**Supplementary Figure 1**

**Illustration of reconstruction of raw coalescent tree.**

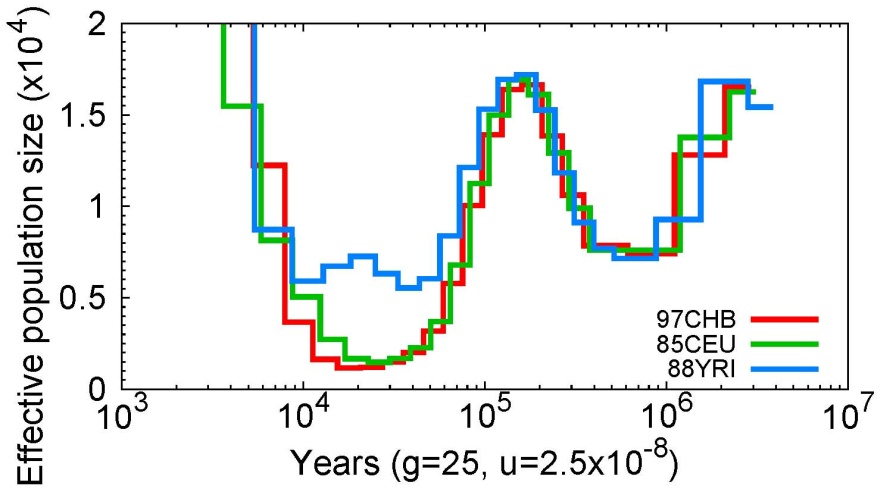
Here shows the example of 3 haplotypes. For all pair of haplotypes, the coalescent segments (red) and corresponding TMRCAs are inferred by PSMC. Coalescent tree of consensus segments (red at the bottom) was reconstructed based on the TMRCA matrix by using UPGMA.



**Supplementary Figure 2**

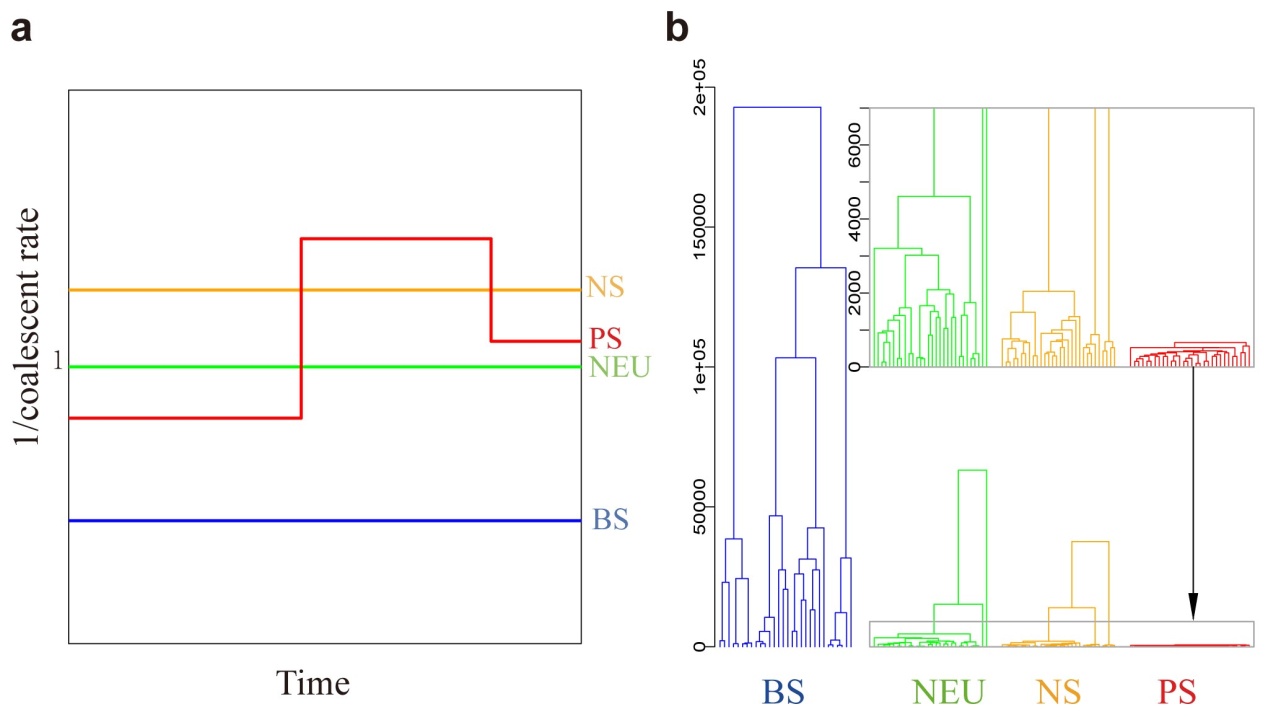
**Performance of mega-sequences.**

*N*e estimated from (**a**) single sequences and (**b**) mega-sequences. The demography is the ‘sim-1’ in Li and Durbin[1](#_ENREF_1). The red line denotes the true demography used in the simulation, the green lines are 100 repeats of the estimation.

****

**Supplementary Figure 3**

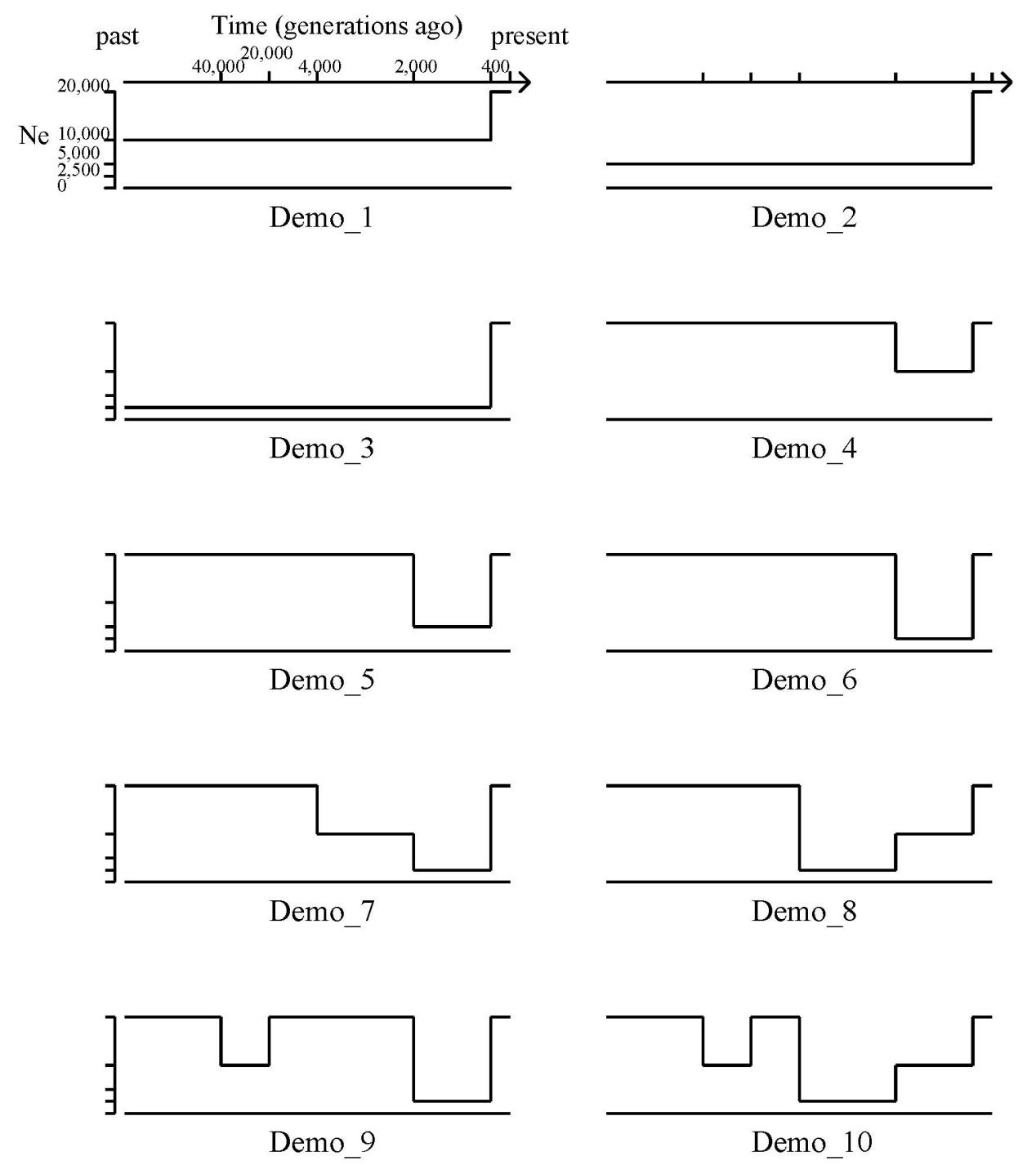
**Population sizes estimated from mega-genomes.**



**Supplementary Figure 4**

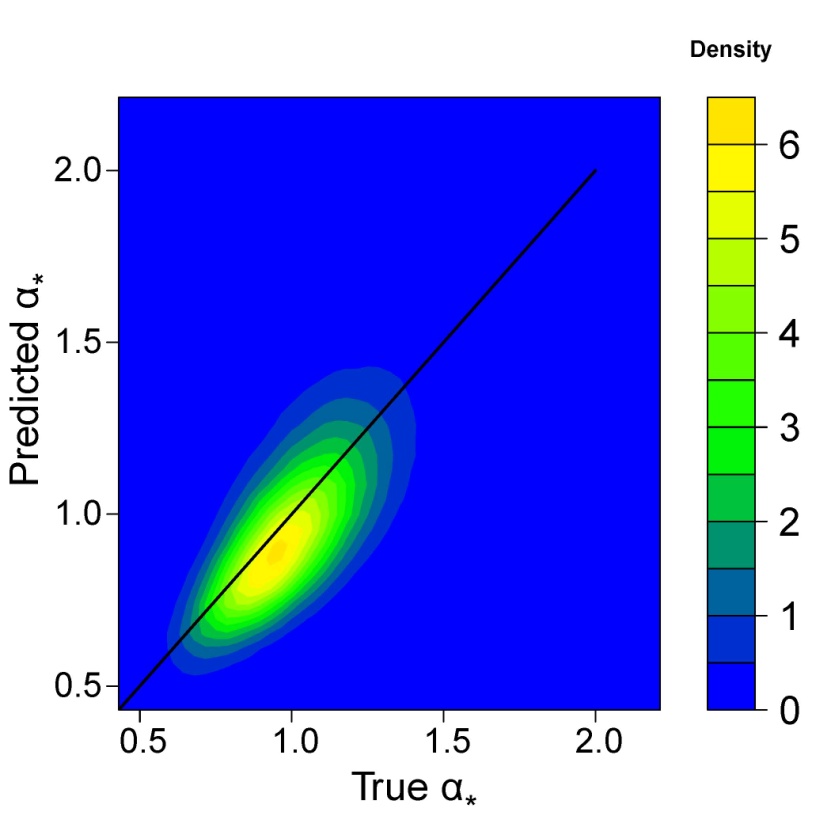
**Coalescent rate trajectories and tree patterns for different types of selection.**

(**a**) Coalescent rate trajectories for the neutral case (green), BS (blue), NS (orange) and PS (red). (**b**) Tree patterns for neutral case(green), BS (blue), NS (orange) and PS (red), with the inset showing the tree patterns in recent time.

****

**Supplementary Figure 5**

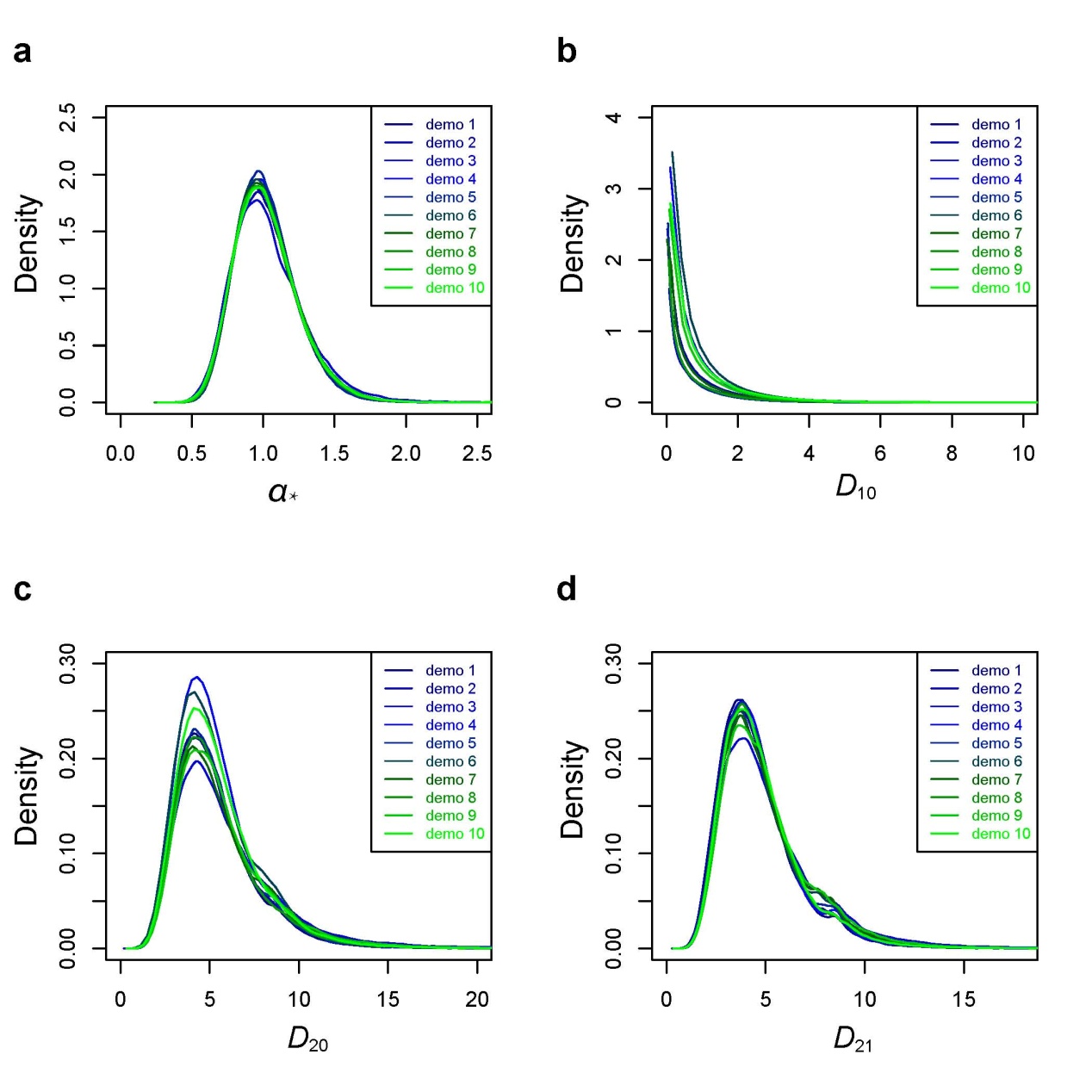
**Details of 10 models of population size change over time.**

****

**Supplementary Figure 6**

**Correlation of *α*\*****between the true and reconstructed trees.**

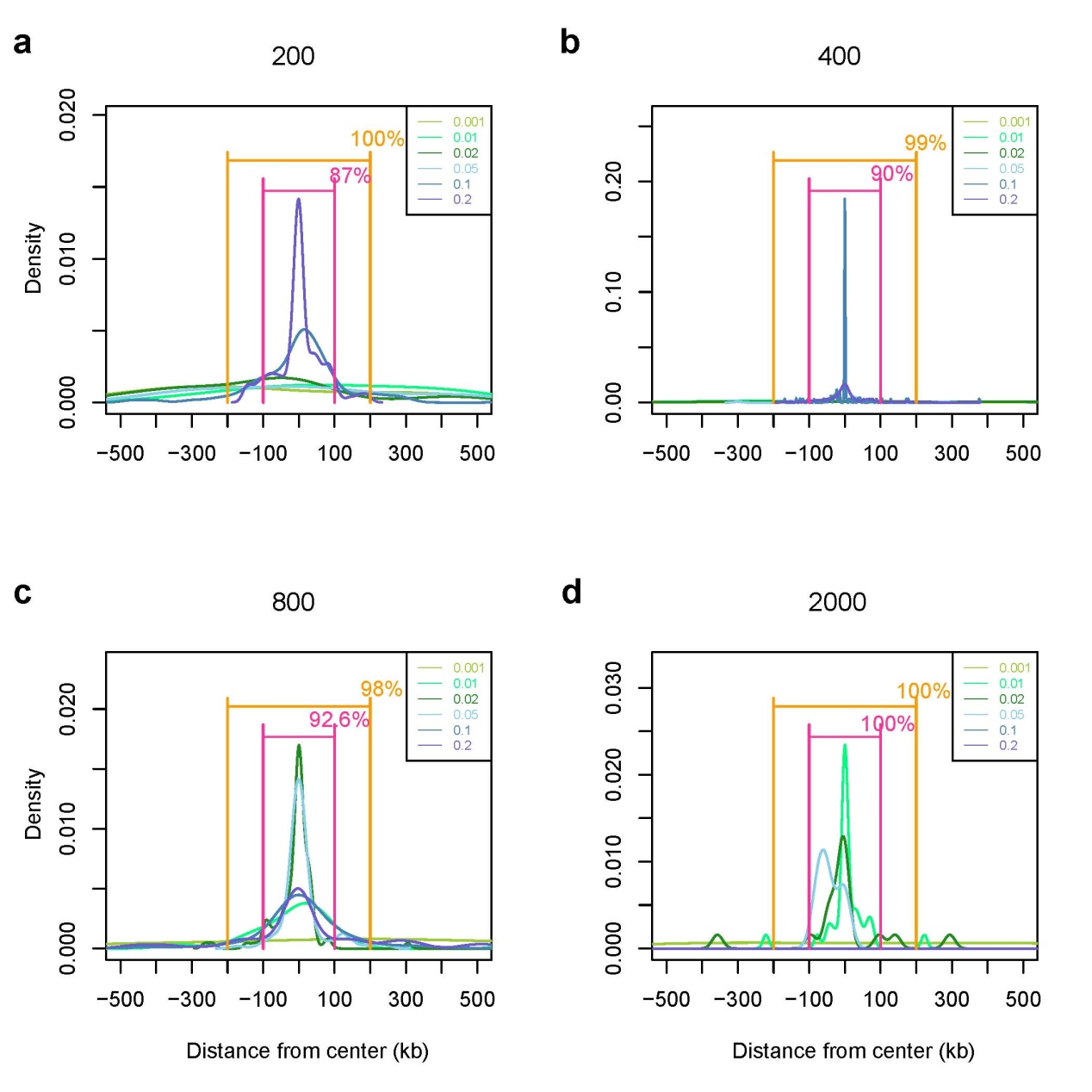
The correlation of *α*\* between true trees and reconstructed trees is 0.7, and the corresponding correlation of *D*10 is 0.5. The red line indicates x = y.

****

**Supplementary Figure 7**

**Distribution of four statistics for the 10 demographic models.**

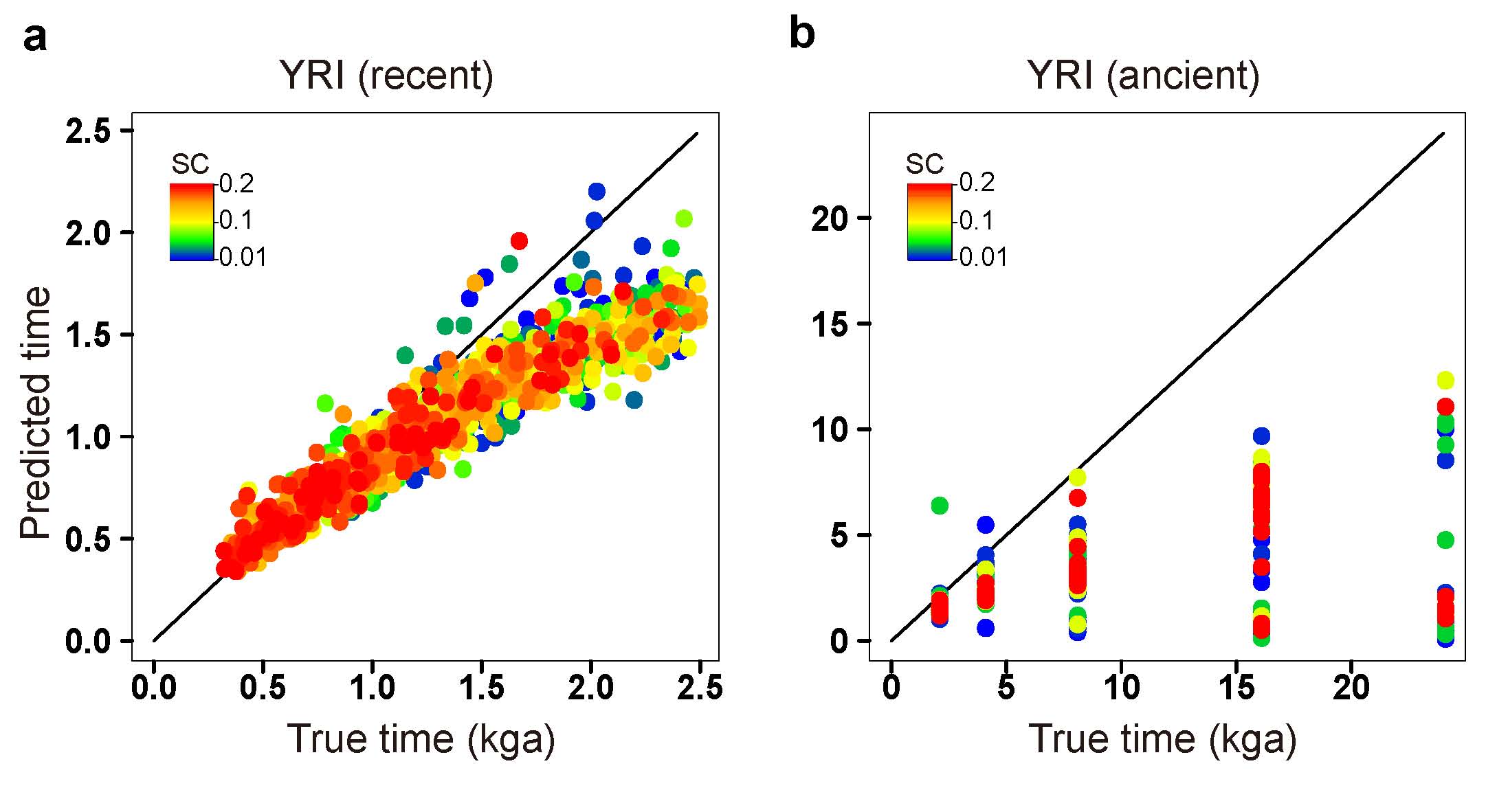
The density curves are highly consistent among different scenarios. (**a**) Density function of *α*. (**b**) Density function of *D*10. (**c**) Density function of *D*20. (**d**) Density function of *D*21.

****

**Supplementary Figure 8**

**Selection center estimation.**

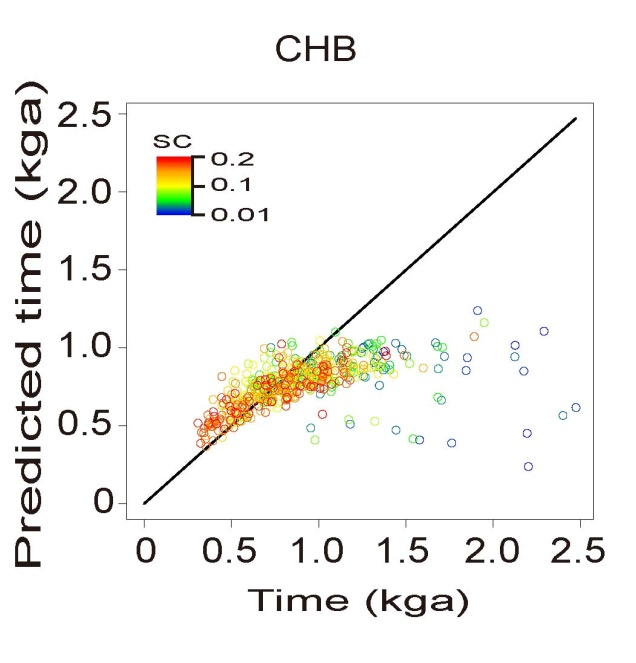
(**a**) Selection starting 200 ga. (**b**) Selection starting 400 ga. (**c**) Selection starting 800 ga. (**d**) Selection starting 2,000 ga. When the selection event starts at a very recent time (200 and 400 ga), the probability that the maximum *D*10 falls into the central 100-kb interval is 88%–90% for strong selection. But for selection occurring at more ancient time (800 ga), only selection with relatively strong selection coefficients (> 0.02) hit the central 100-kb interval with high probability. For even more ancient selection (2,000 ga), only intermediate and weak selection can be detected within the central 100-kb interval.

****

**Supplementary Figure 9**

**Time estimation without correction.**

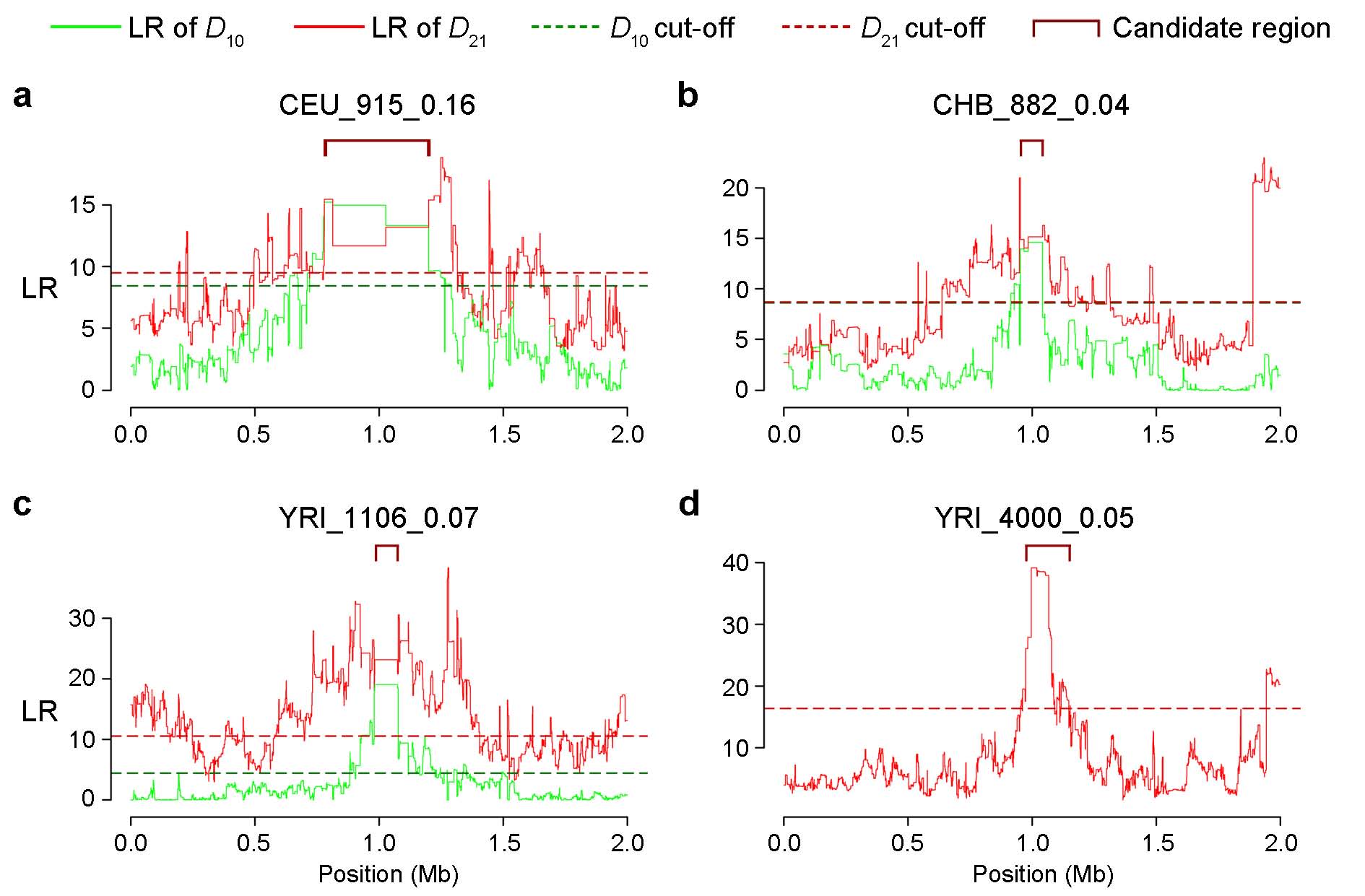
(**a**) YRI (recent). (**b**) YRI (ancient). Older selection times are clearly underestimated in both of these figures.



**Supplementary Figure 10**

**Estimation of selection starting time for CHB.**

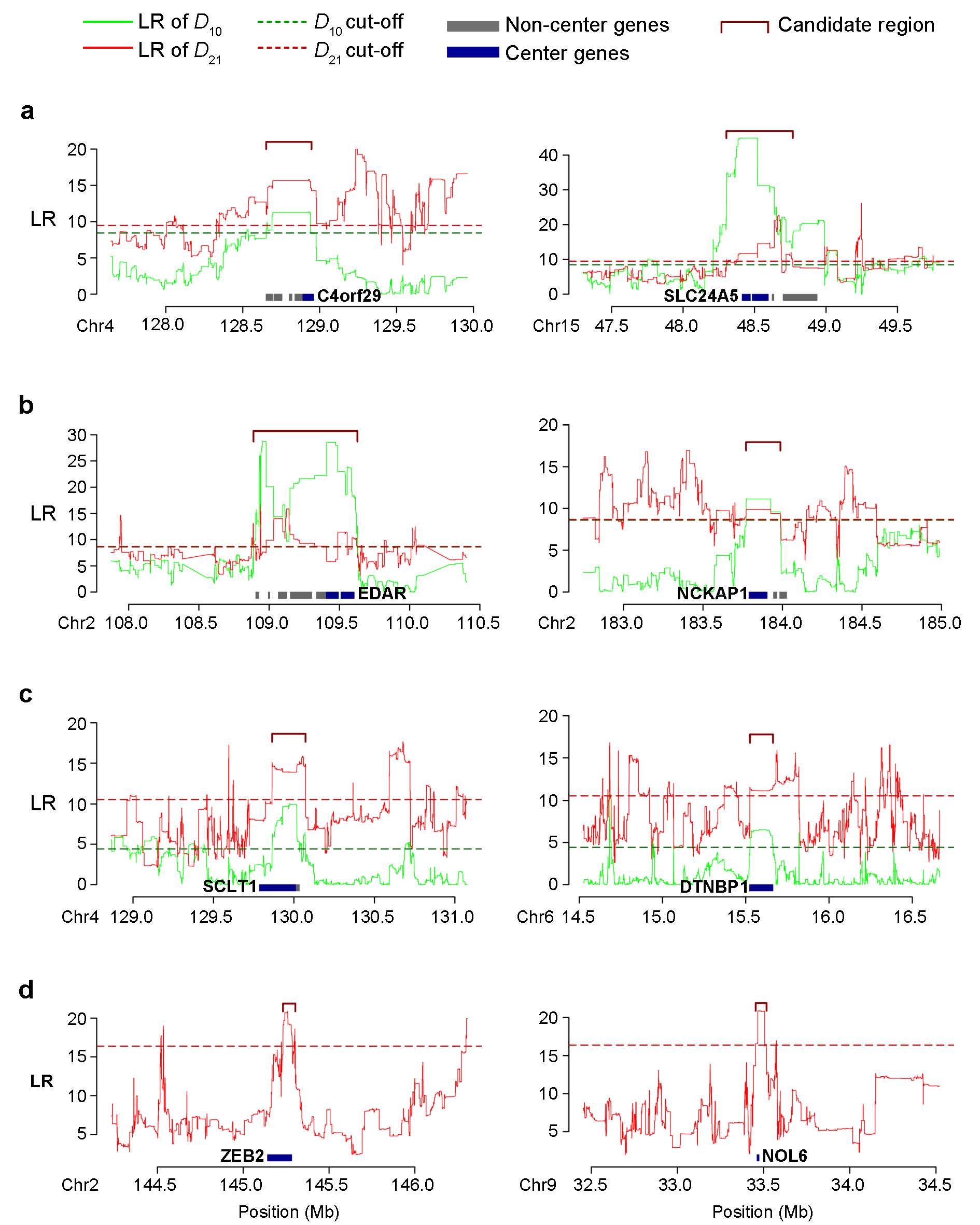
There is an obvious under-estimation of selection time in the more ancient time intervals in CHB, most likely due to the severe bottlenecks that eliminated the ancient coalescent information.

****

**Supplementary Figure 11**

**Signal of PS in simulations.**

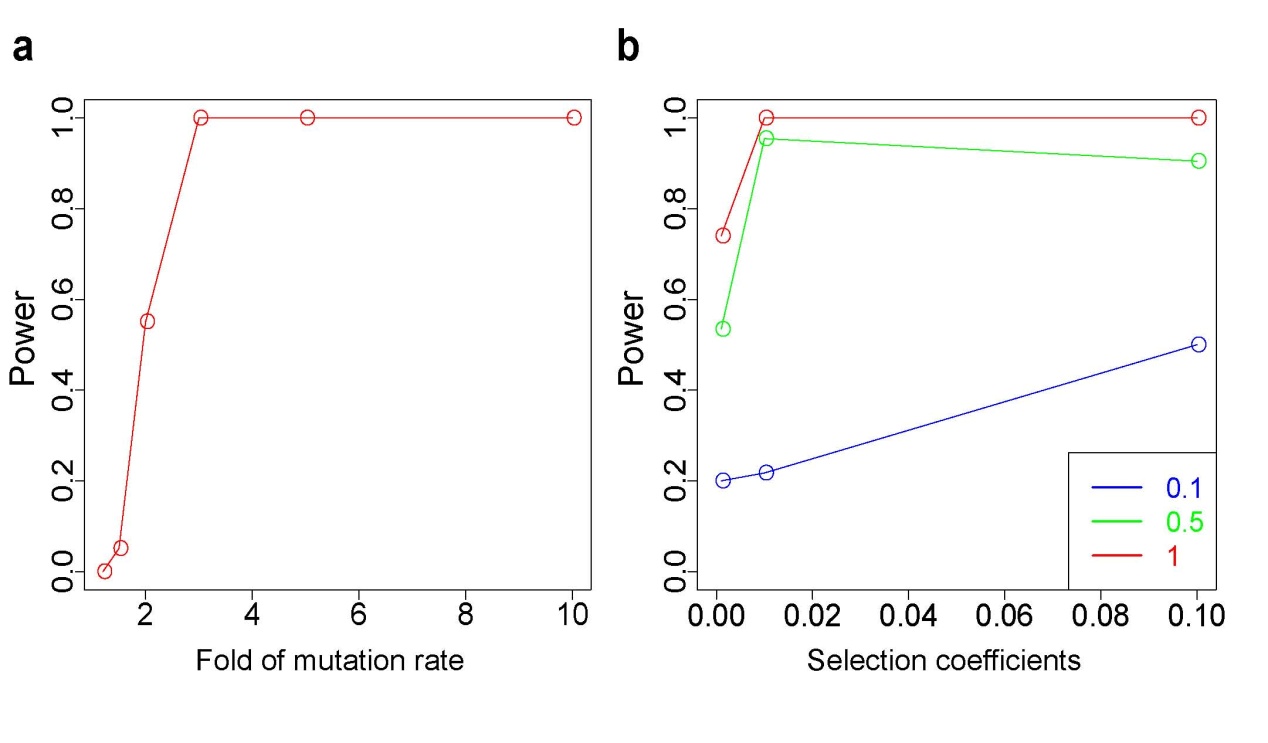
(**a**) A PS event happening 915 ga with coefficient of 0.16 under CEU demography. (**b**) A PS occuring 882 ga with coefficient of 0.04 under CHB demography. (**c**) A PS occuring 1,106 ga with coefficient of 0.07 under YRI demography. (**d**) A PS occuring 4,000 ga with coefficient of 0.07 under YRI demography. All PS occurred at the center of the segments. For recent selection, both *D*10 and *D*21 were used, for ancient selection, only *D*21 was plotted.

****

**Supplementary Figure 12**

**Signal of PS in 1000G data.**

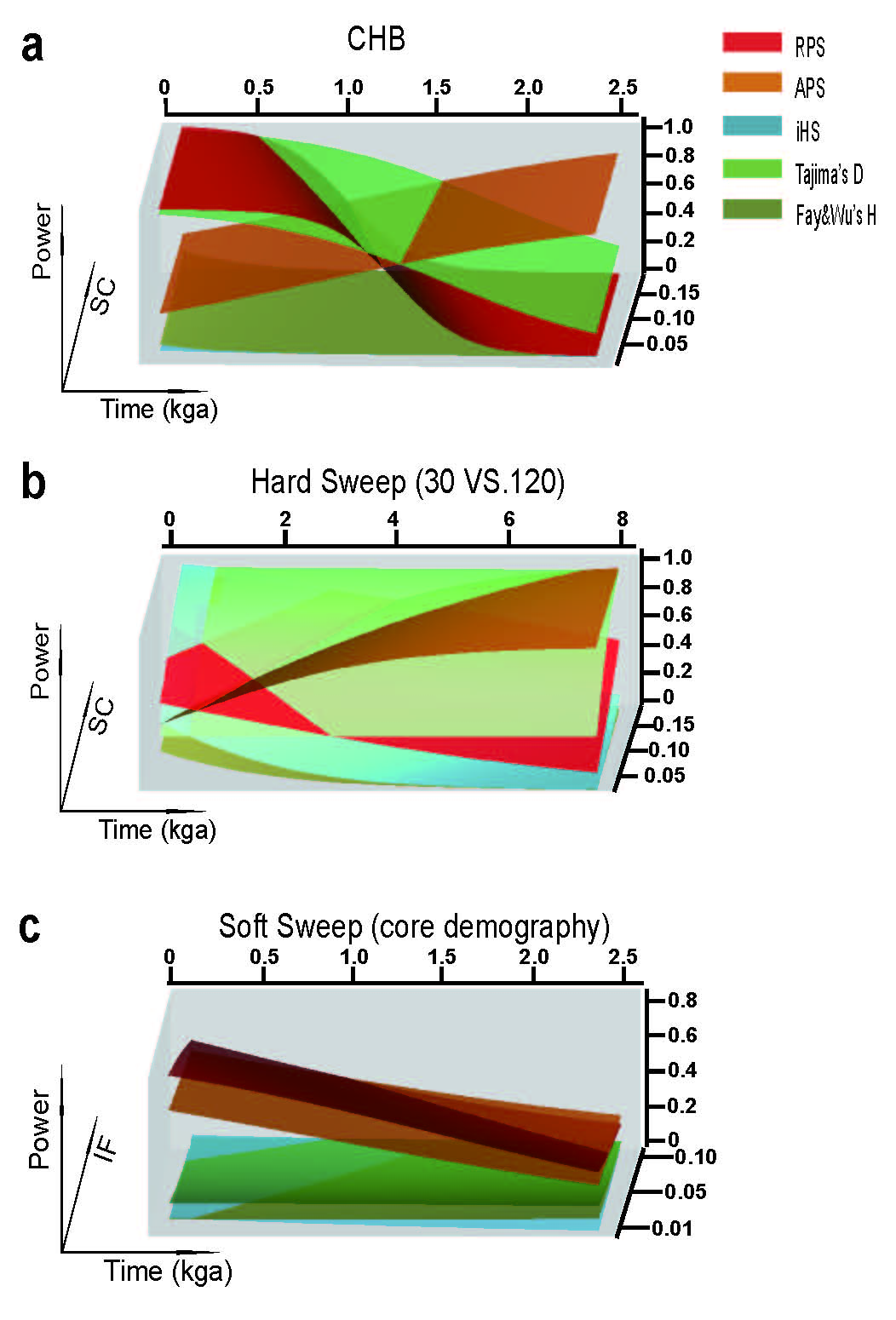
Candidate regions of recent selection in (**a**) CEU, (**b**) CHB, (**c**) YRI, and (**d**) ancient selection in YRI.

****

**Supplementary Figure 13**

**Power tests for BS and NS.**

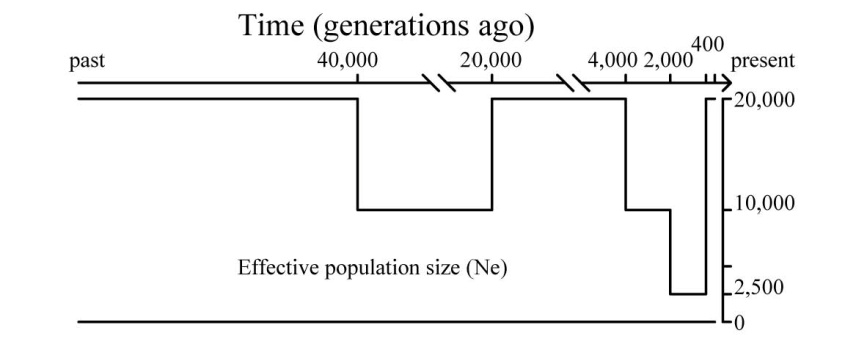
**(a**) For simulations of BS, it can be seen how test power varies with mutation rate changes. The power of the BS test is almost close to 100% for strong BS, with higher fold of mutation rate (2 ≤ fold of mutation rate ≤ 3), and has 11% to 57% power for weaker BS (1.5 ≤ fold of mutation rate ≤ 2. (**b**) The power curves for detecting NS using *D*10 for three NS scenarios with different proportions of non-synonymous mutations. As the proportion of non-synonymous mutations decreases, so does the power of detecting NS signals.

****

**Supplementary Figure 14**

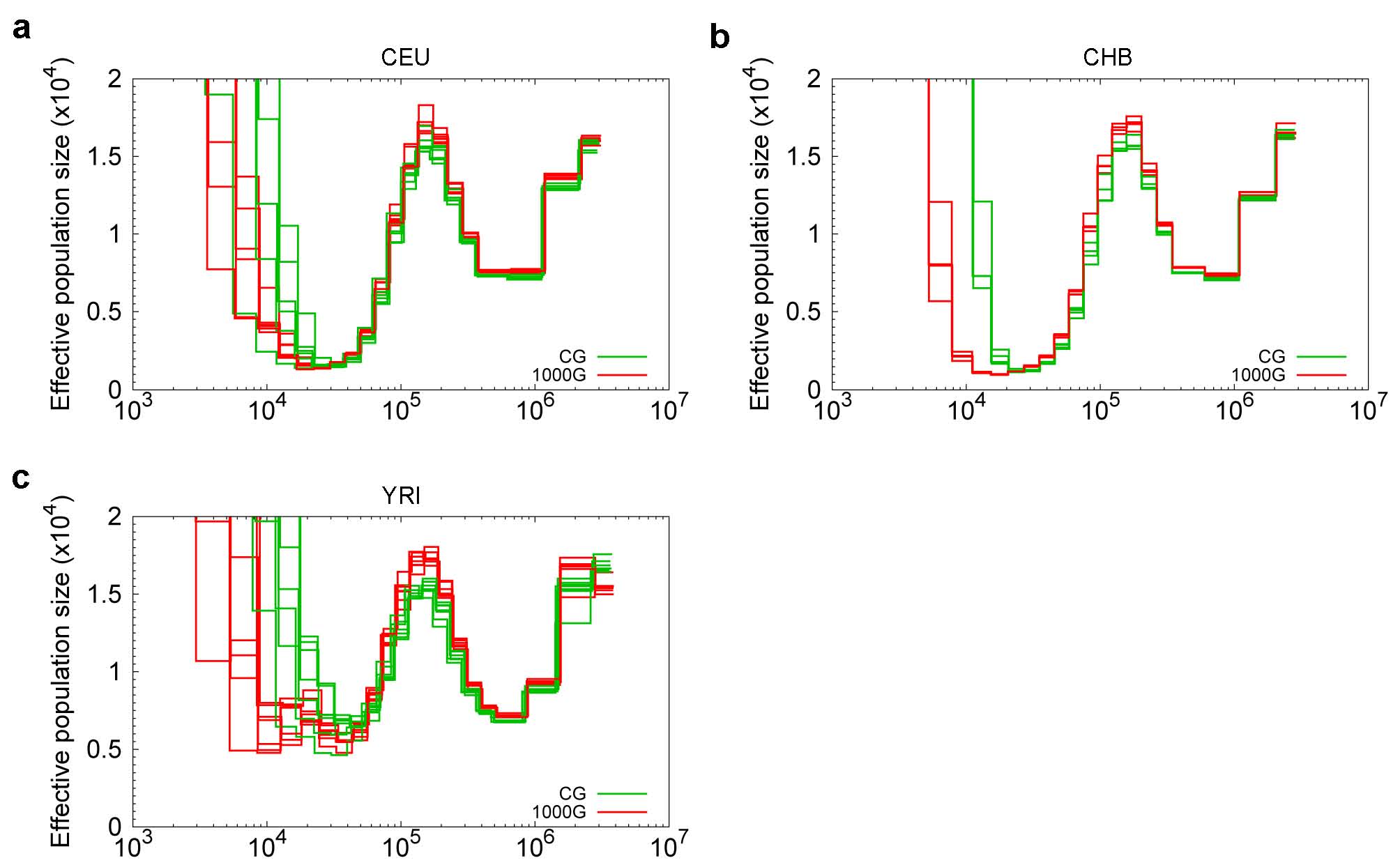
**Power of RPS and APS tests compared to other popular tests**.

Recent positive selection (RPS) (red), and ancient positive selection (APS) (orange) were compared to iHS (light blue), Tajima’s D (light green) and Fay and Wu’s H (dark green). Time is in units of thousand generations ago (kga). (**a**) Hard Sweep for CHB demography. (**b**) Hard sweep for constant size model with *Ne* of 10,000. Note that 30 haplotypes were sampled for RPS and APS due to the computational cost, whereas 120 haplotypes were sampled for iHS, Tajima’s D and Fay and Wu’s H. (**c**) Soft sweep for constant size model with *Ne* of 10,000.

****

**Supplementary Figure 15**

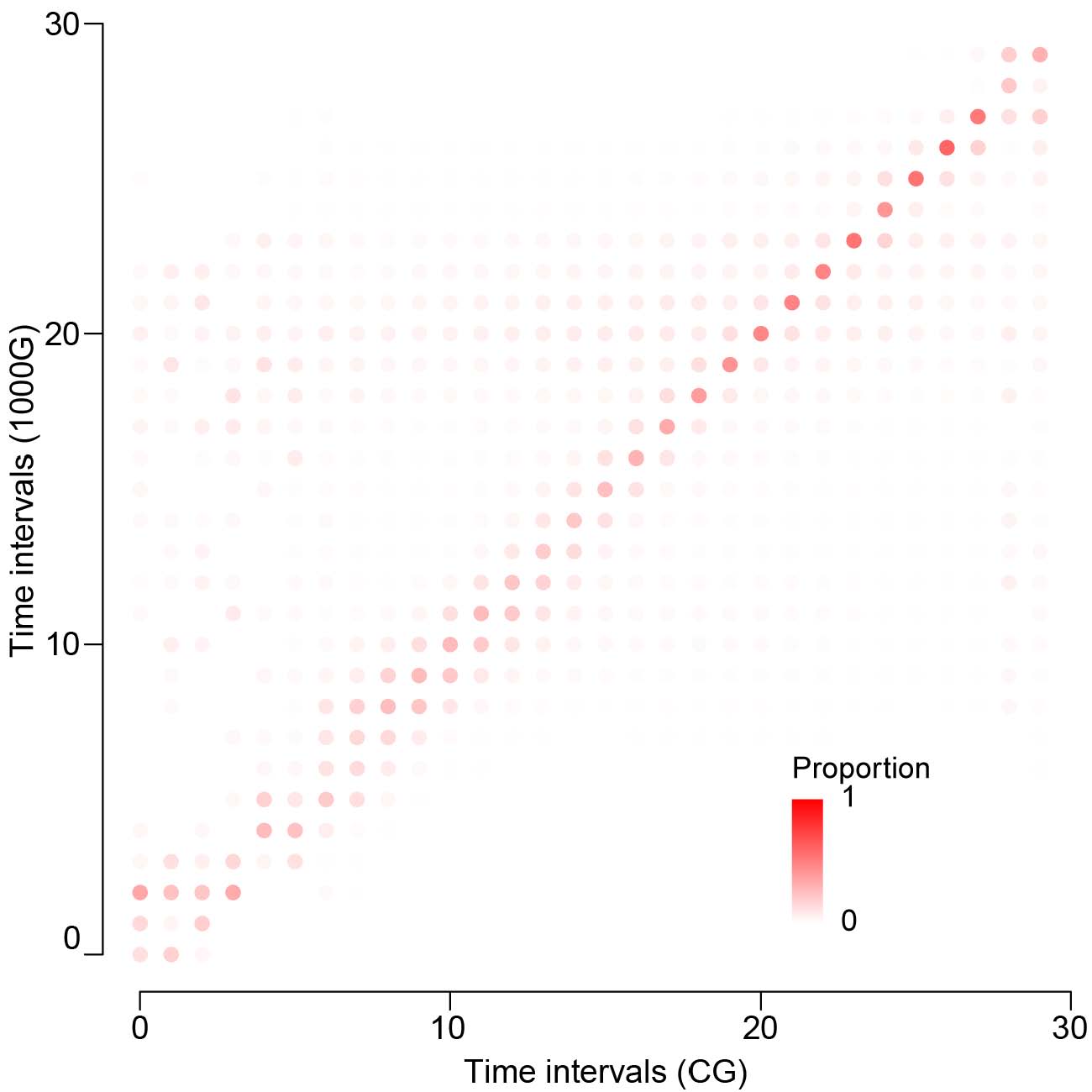
**Demography of the core model.**

****

**Supplementary Figure 16**

**Comparison of *N*e estimation between high coverage (CG) and imputed low coverage (1000G) data.**

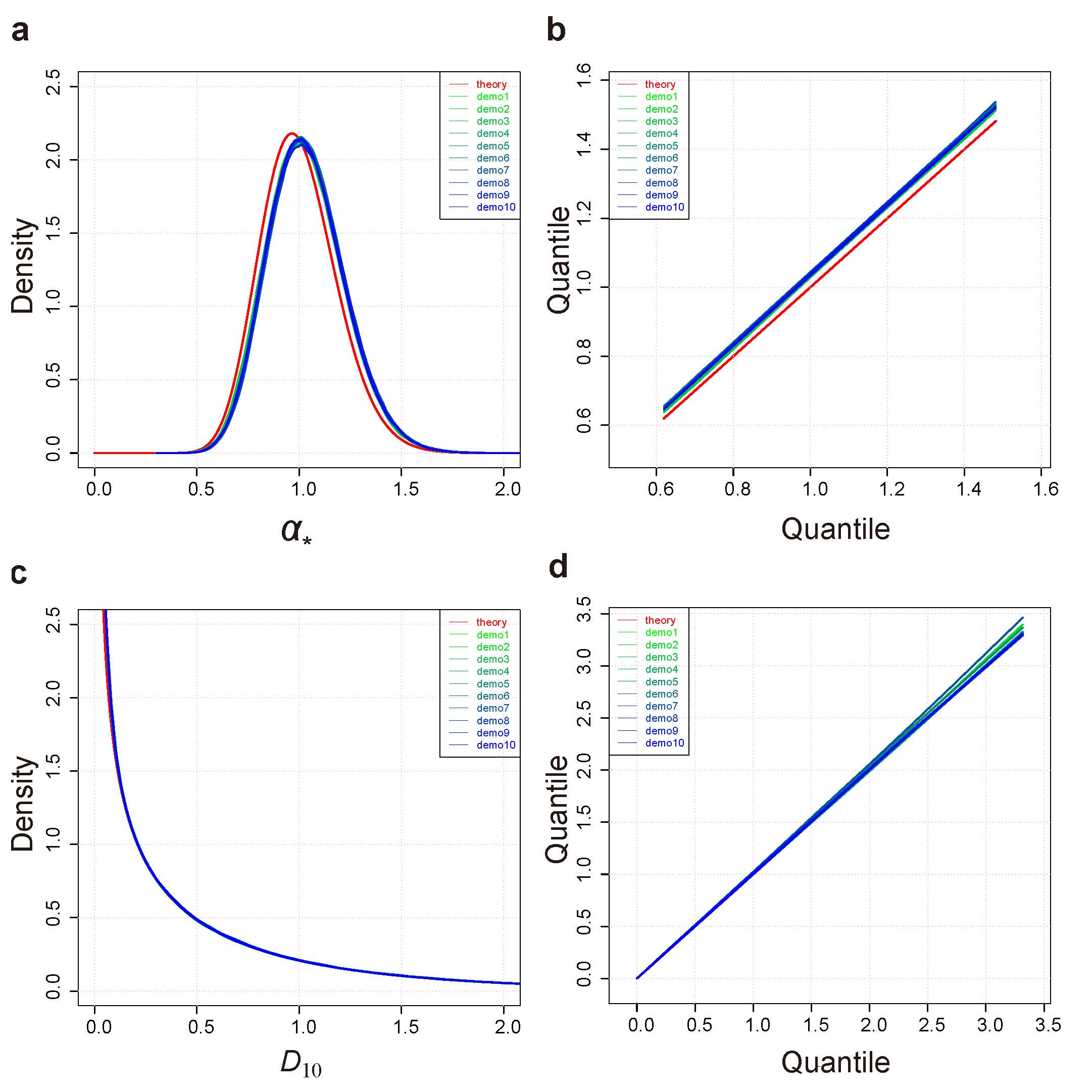
Green curves are estimated trajectories from CG individual genomes, red curves are from 1000G individual genomes. The individuals are from (**a**) CEU, (**b**) CHB and (**c**) YRI.



**Supplementary Figure 17**

**Correlation of TMRCAs estimated from CG and 1000G data.**

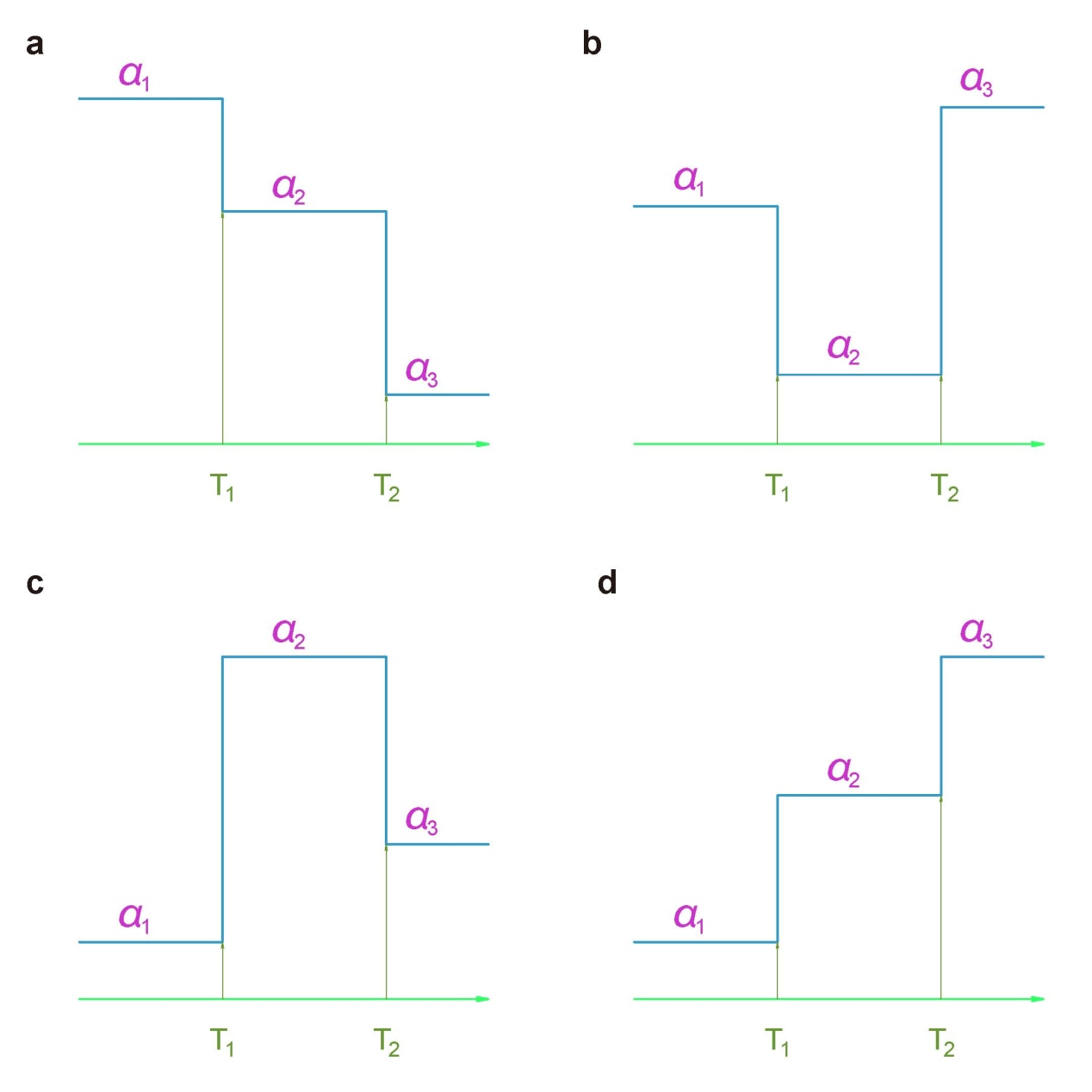
TMRCAs were estimated for the CG and 1000G data for the same individuals. The proportions of overlapping between the two different estimations were plotted for all the 30 possible time intervals.

****

**Supplementary Figure 18**

**Fitting of simulation results and theoretical curve.**

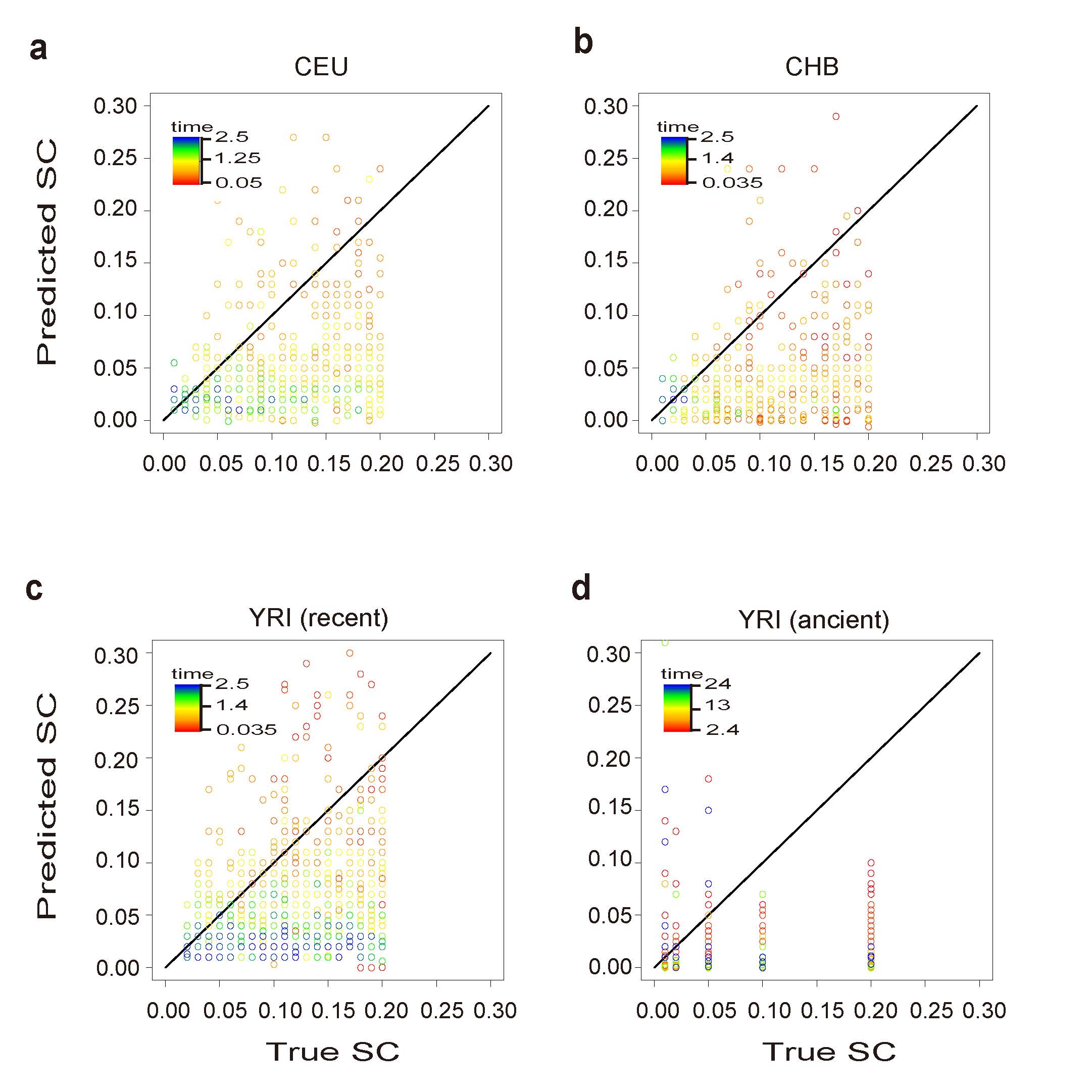
The lines in red are from theoretical distributions. The others specify the simulated data from 10 different demographic scenarios. **(a)** density curves of *α*\*. (**b**) quantile-quantile plot between the simulated and theoretical distributions of *α*\*. (**c**) density curves of *D*10. (**d**) quantile-quantile plot between the simulated and theoretical distributions of *D*10.



**Supplementary Figure 19**

**Four cases of the dynamics of *αi* (*i* = 1, 2, 3) for time estimation.**

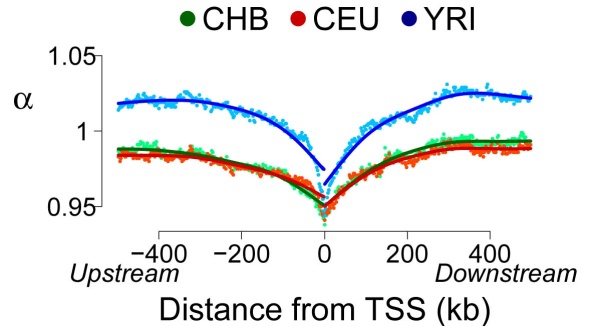
In different cases, we use separate strategies to infer the selection starting time and selection coefficients. (**a**) *α*1 > *α*2 > *α*3. (**b**) *α*1 > *α*2 < *α*3. (**c**) *α*1 < *α*2 > *α*3. (**d**) *α*1 < *α*2 < *α*3.



**Supplementary Figure 20**

**Performance of selection coefficient estimations.**

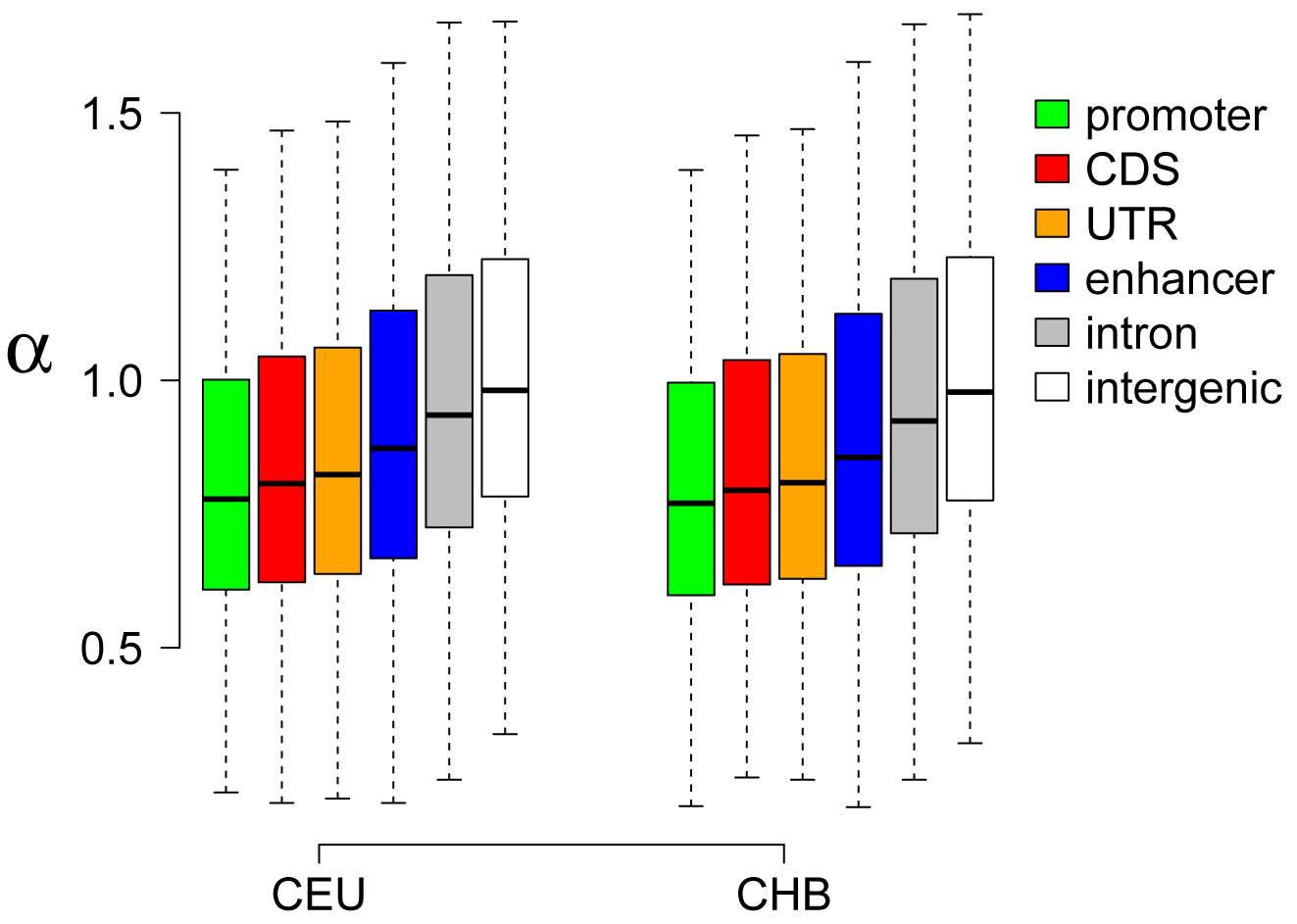
**(a**–**d**) Comparison of estimated vs. true selection coefficients in data simulated under demographic histories for CEU, CHB, YRI (recent) and YRI (ancient). The correlations of the estimated selection coefficients with the true values for CEU, CHB, YRI recent and YRI ancient are 0.18 (*P* value = 3.23 × 10-6), 0.105 (*P* value = 0.026), 0.19 (*P* value = 2.14 × 10-9), and 0.09 (*P* value = 0.0021), respectively. Nevertheless, when focusing on the events that occurred in the more recent time, CEU (about 48.3–700 generations), CHB (about 350–700 generations) and YRI (about 352–1,600 generations) have increased correlations of 0.21 (*P* value = 4.24 × 10-8), 0.13 (*P* value = 0.006), and 0.22 (*P* value = 2.76 × 10-12), respectively.

****

**Supplementary Figure 21**

***α* scores around the transcript start sites.**

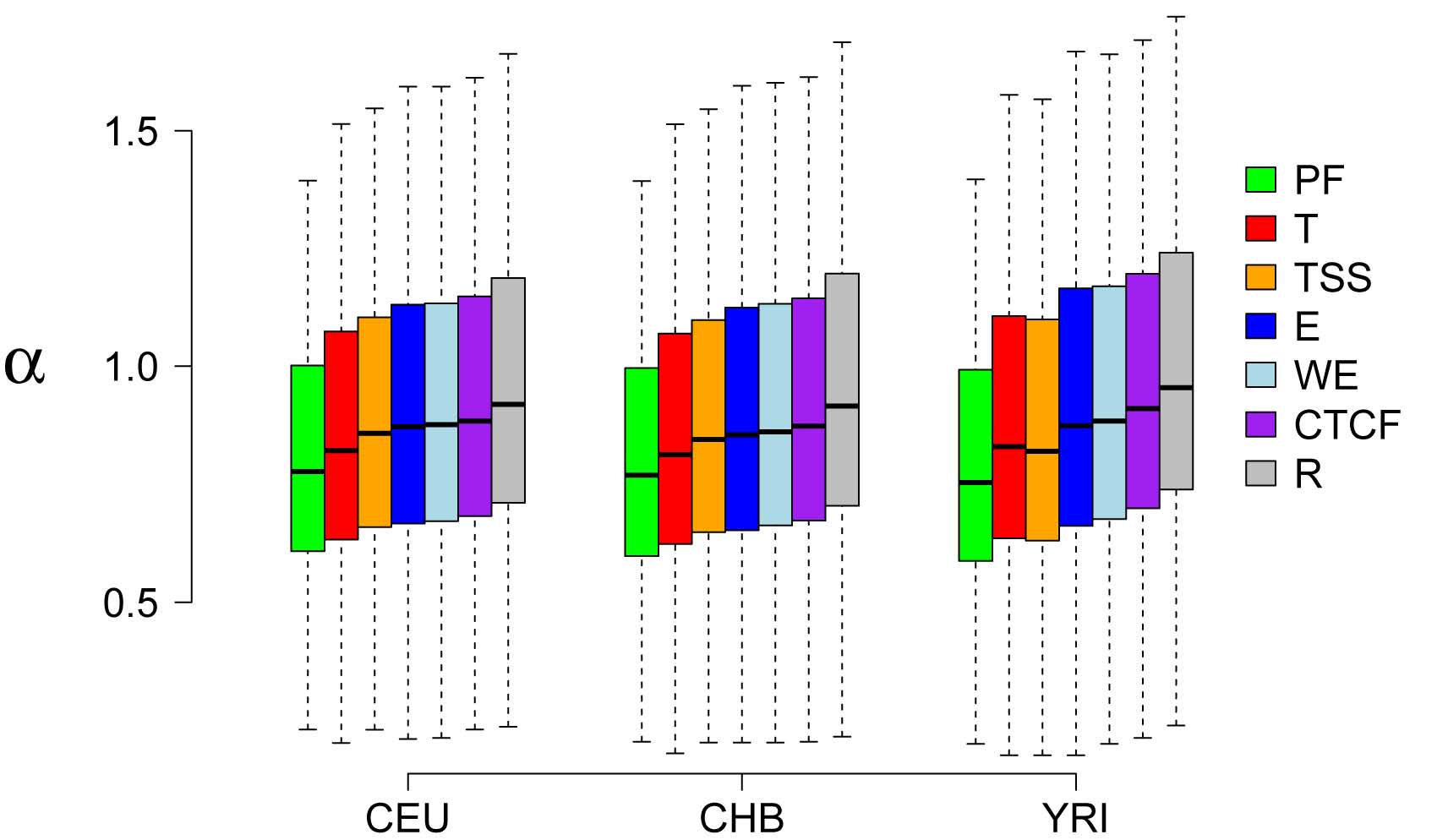
Genome average distribution of coalescent scaling coefficient *α* scores was plotted centered on the transcript start sites.

****

**Supplementary Figure 22**

***α* scores in different functional elements in CEU and CHB.**

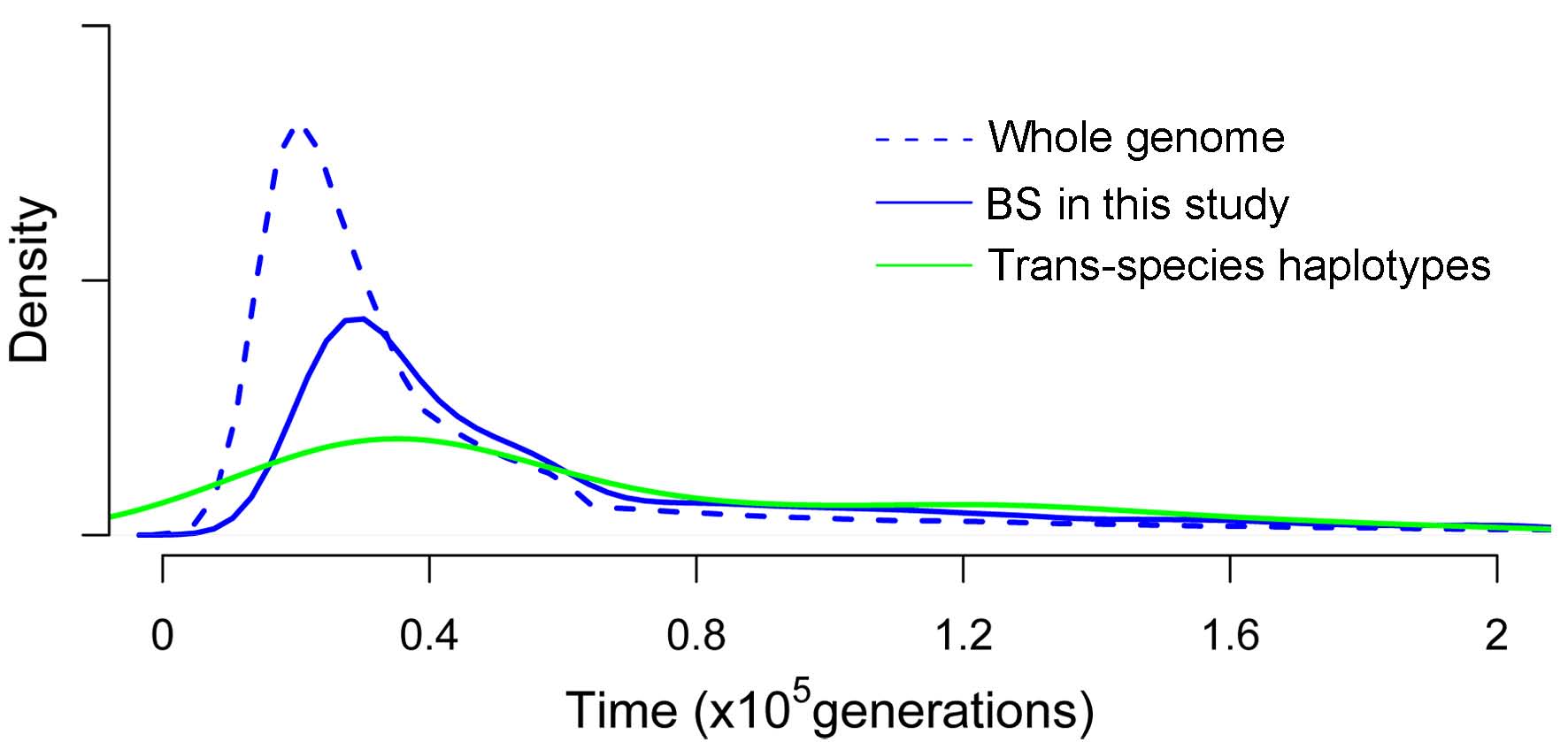
Distributions of *α* scores in different functional elements for CEU and CHB.



**Supplementary Figure 23**

***α* scores in 7 states from ENCODE.**

PF: promoter; T: transcribed regions; TSS: transcript start sites; E: enhancer; WE: weaker enhancer; CTCF: CTCF active regions; R: repressed regions.



**Supplementary Figure 24**

**Distribution of TMRCAs in YRI.**

The dashed blue line is the TMRCAs of whole genome trees, while the solid blue line is the TMRCAs of BS regions in this study. The green line is the TMRCAs distribution of trans-species haplotypes in a previous study[2](#_ENREF_2).

**Supplementary Table 1.** **Haploid genomes used in reconstruction of coalescent tree.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CEU** | | **CHB** | | **YRI** | |
| **samples** | **haplotype** | **samples** | **haplotype** | **samples** | **haplotype** |
| NA06984 | 2 | NA18528 | 2 | NA18499 | 1 |
| NA07048 | 1 | NA18530 | 2 | NA18501 | 2 |
| NA07347 | 2 | NA18538 | 1 | NA18507 | 1 |
| NA07357 | 2 | NA18541 | 2 | NA18508 | 2 |
| NA11843 | 2 | NA18542 | 1 | NA18856 | 2 |
| NA11892 | 2 | NA18543 | 2 | NA18867 | 1 |
| NA11893 | 2 | NA18545 | 1 | NA18868 | 2 |
| NA11930 | 2 | NA18546 | 1 | NA18871 | 1 |
| NA11933 | 1 | NA18547 | 2 | NA18873 | 2 |
| NA11992 | 2 | NA18549 | 1 | NA18874 | 1 |
| NA12043 | 1 | NA18550 | 2 | NA18909 | 1 |
| NA12044 | 2 | NA18552 | 1 | NA18923 | 1 |
| NA12144 | 1 | NA18553 | 1 | NA18924 | 2 |
| NA12275 | 2 | NA18558 | 1 | NA19093 | 2 |
| NA12283 | 2 | NA18562 | 2 | NA19098 | 1 |
| NA12286 | 1 | NA18564 | 1 | NA19102 | 2 |
| NA12287 | 2 | NA18571 | 1 | NA19107 | 2 |
| NA12341 | 2 | NA18573 | 1 | NA19116 | 1 |
| NA12342 | 2 | NA18579 | 2 | NA19121 | 1 |
| NA12347 | 2 | NA18582 | 1 | NA19130 | 1 |
| NA12413 | 1 | NA18596 | 2 | NA19131 | 1 |
| NA12489 | 2 | NA18603 | 2 | NA19146 | 2 |
| NA12716 | 2 | NA18611 | 1 | NA19149 | 2 |
| NA12751 | 1 | NA18613 | 2 | NA19160 | 1 |
| NA12775 | 2 | NA18614 | 1 | NA19171 | 2 |
| NA12777 | 1 | NA18624 | 2 | NA19175 | 2 |
| NA12814 | 1 | NA18628 | 2 | NA19200 | 1 |
| NA12815 | 2 | NA18634 | 2 | NA19213 | 1 |
| NA12830 | 2 | NA18639 | 1 | NA19235 | 1 |
| NA12874 | 1 | NA18747 | 1 | NA19236 | 1 |

**Supplementary Table 2.** **Thresholds of models in different populations.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | ***Dpn* 10** | ***Drp* 21** | ***Dap* 21** | ***Db* 10** |
| **threshods** | 0.99 | 0.8 | 0.99 | 0.999 |
| **CEU** | 8.434554 | 9.484881 | NA | 17.19737 |
| **CHB** | 8.601544 | 8.685523 | NA | 19.68071 |
| **YRI** | 4.425076 | 10.54162 | 16.37563 | 16.19524 |

**Supplementary Table 9. Functional enrichment of PS genes in CEU.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Term** | **Fold Enrichment** | ***P* value** | **Benjamini** |
| UNIGENE\_EST\_QUARTILE | pituitary gland\_normal\_3rd | 1.58 | 1.09E-05 | 8.30E-04 |
| GOTERM\_CC\_FAT | GO:0043232~intracellular non-membrane-bounded organelle | 1.65 | 7.31E-06 | 0.001 |
| GO:0043228~non-membrane-bounded organelle | 1.65 | 7.31E-06 | 0.001 |
| GO:0005856~cytoskeleton | 1.99 | 7.15E-06 | 0.002 |
| GO:0000786~nucleosome | 8.15 | 4.92E-05 | 0.005 |
| GO:0005929~cilium | 4.35 | 4.72E-04 | 0.026 |
| GO:0032993~protein-DNA complex | 5.84 | 4.14E-04 | 0.028 |
| SP\_PIR\_KEYWORDS | Cilium | 7.56 | 6.26E-06 | 0.002 |
| nucleosome core | 9.23 | 9.66E-05 | 0.016 |
| UP\_TISSUE | Hair root | 9.06 | 2.44E-05 | 0.004 |
| GOTERM\_BP\_FAT | GO:0031497~chromatin assembly | 7.32 | 7.81E-06 | 0.012 |
| GO:0006323~DNA packaging | 5.24 | 1.17E-04 | 0.043 |
| GO:0034728~nucleosome organization | 6.14 | 1.00E-04 | 0.049 |

**Supplementary Table 10. Functional enrichment of PS genes in CHB.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Term** | **Fold Enrichment** | ***P* value** | **Benjamini** |
| UNIGENE\_EST\_QUARTILE | testis\_normal\_3rd | 1.30 | 6.79E-06 | 5.16E-04 |
| laryngeal cancer\_disease\_3rd | 1.30 | 1.57E-03 | 0.039 |
| GNF\_U133A\_QUARTILE | Appendix\_3rd | 1.32 | 9.43E-05 | 0.007 |
| GOTERM\_BP\_FAT | GO:0006069~ethanol oxidation | 30.83 | 5.26E-06 | 0.010 |
| GO:0006067~ethanol metabolic process | 30.83 | 5.26E-06 | 0.010 |
| GO:0034308~monohydric alcohol metabolic process | 30.83 | 5.26E-06 | 0.010 |
| SP\_PIR\_KEYWORDS | alcohol metabolism | 22.47 | 3.25E-05 | 0.013 |
| UP\_TISSUE | Brain | 1.18 | 8.37E-05 | 0.017 |
| KEGG\_PATHWAY | hsa00071:Fatty acid metabolism | 6.61 | 1.54E-04 | 0.020 |
| GOTERM\_MF\_FAT | GO:0004022~alcohol dehydrogenase (NAD) activity | 21.81 | 3.62E-05 | 0.023 |
| GO:0004024~alcohol dehydrogenase activity, zinc-dependent | 30.54 | 1.35E-04 | 0.042 |

**Supplementary Table 11. Functional enrichment of PS genes in YRI.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Term** | **Fold Enrichment** | ***P* value** | **Benjamini** |
| GNF\_U133A\_QUARTILE | fetalliver\_3rd | 1.32 | 1.34E-05 | 5.28E-04 |
| Appendix\_3rd | 1.28 | 1.03E-05 | 8.11E-04 |
| Cerebellum Peduncles\_3rd | 1.48 | 3.22E-05 | 8.48E-04 |
| dorsal root ganglia\_3rd | 1.43 | 1.09E-04 | 0.002 |
| Olfactory Bulb\_3rd | 1.33 | 2.86E-04 | 0.005 |
| Occipital Lobe\_3rd | 1.27 | 1.57E-03 | 0.020 |
| Kidney\_3rd | 1.14 | 3.40E-03 | 0.038 |
| Ciliary Ganglion\_3rd | 1.18 | 6.10E-03 | 0.047 |
| Colorectal Adenocarcinoma\_3rd | 1.21 | 5.13E-03 | 0.049 |
| UP\_TISSUE | Brain | 1.16 | 1.58E-05 | 0.004 |
| Peripheral blood leukocyte | 3.59 | 2.23E-04 | 0.027 |
| GOTERM\_CC\_FAT | GO:0044456~synapse part | 2.40 | 8.25E-05 | 0.034 |
| GO:0045202~synapse | 2.04 | 2.17E-04 | 0.045 |
| UNIGENE\_EST\_QUARTILE | pituitary gland\_normal\_3rd | 1.24 | 1.87E-03 | 0.035 |
| thyroid\_normal\_3rd | 1.21 | 1.82E-03 | 0.045 |
| SP\_PIR\_KEYWORDS | Phosphoprotein | 1.16 | 9.16E-05 | 0.044 |

**Supplementary Table 12.** **Functional enrichment of BS genes in CEU.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Term** | **Fold Enrichment** | ***P* value** | **Benjamini** |
| SP\_PIR\_KEYWORDS | glycoprotein | 1.89 | 5.14E-09 | 1.31E-06 |
| polymorphism | 1.29 | 2.26E-07 | 2.89E-05 |
| disulfide bond | 2.02 | 4.90E-07 | 4.16E-05 |
| signal | 1.89 | 1.55E-06 | 9.91E-05 |
| membrane | 1.52 | 4.06E-06 | 2.07E-04 |
| transmembrane | 1.60 | 1.12E-05 | 4.75E-04 |
| cell membrane | 1.97 | 5.92E-05 | 0.002 |
| mhc ii | 44.61 | 7.67E-05 | 0.002 |
| transplantation antigen | 100.38 | 2.91E-04 | 0.008 |
| Immunoglobulin domain | 3.15 | 8.75E-04 | 0.022 |
| Signal-anchor | 3.14 | 0.002 | 0.036 |
| heterodimer | 6.77 | 0.002 | 0.037 |
| transmembrane protein | 2.61 | 0.002 | 0.038 |
| UNIGENE\_EST\_QUARTILE | normal\_disease\_3rd | 1.40 | 5.83E-05 | 0.004 |
| brain\_normal\_3rd | 1.39 | 0.001 | 0.041 |
| KEGG\_PATHWAY | hsa05416:Viral myocarditis | 12.08 | 1.03E-04 | 0.007 |
| hsa04514:Cell adhesion molecules (CAMs) | 6.99 | 3.60E-04 | 0.008 |
| hsa05320:Autoimmune thyroid disease | 14.97 | 2.77E-04 | 0.009 |
| hsa05332:Graft-versus-host disease | 16.69 | 0.002 | 0.021 |
| hsa05330:Allograft rejection | 17.30 | 0.001 | 0.023 |
| hsa04940:Type I diabetes mellitus | 14.16 | 0.002 | 0.028 |
| GOTERM\_BP\_FAT | GO:0022610~biological adhesion | 3.00 | 1.81E-05 | 0.011 |
| GO:0007155~cell adhesion | 3.00 | 1.77E-05 | 0.021 |
| GOTERM\_CC\_FAT | GO:0042611~MHC protein complex | 14.30 | 3.70E-04 | 0.013 |
| GO:0031224~intrinsic to membrane | 1.36 | 3.48E-04 | 0.015 |
| GO:0044459~plasma membrane part | 1.68 | 5.88E-04 | 0.018 |
| GO:0005901~caveola | 9.76 | 3.45E-04 | 0.018 |
| GO:0031226~intrinsic to plasma membrane | 2.06 | 3.40E-04 | 0.024 |
| GO:0005886~plasma membrane | 1.50 | 2.50E-04 | 0.026 |
| GO:0016021~integral to membrane | 1.32 | 0.002 | 0.047 |
| GO:0005887~integral to plasma membrane | 2.11 | 2.33E-04 | 0.048 |
| PANTHER\_MF\_ALL | MF00175:Major histocompatibility complex antigen | 17.63 | 1.61E-04 | 0.017 |
| MF00258:CAM family adhesion molecule | 8.53 | 6.47E-04 | 0.034 |

**Supplementary Table 13. Functional enrichment of BS genes in CHB.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Term** | **Fold Enrichment** | ***P* value** | **Benjamini** |
| SP\_PIR\_KEYWORDS | glycoprotein | 2.12 | 3.72E-09 | 8.14E-07 |
| signal | 2.04 | 5.97E-06 | 6.53E-04 |
| disulfide bond | 2.08 | 1.67E-05 | 0.001 |
| mhc ii | 66.32 | 2.33E-05 | 0.001 |
| transplantation antigen | 149.23 | 1.31E-04 | 0.006 |
| heterodimer | 10.06 | 3.15E-04 | 0.011 |
| transmembrane protein | 3.36 | 4.50E-04 | 0.014 |
| transmembrane | 1.56 | 7.25E-04 | 0.020 |
| mhc i | 49.74 | 0.002 | 0.036 |
| duplication | 5.17 | 0.002 | 0.049 |
| PANTHER\_MF\_ALL | MF00175:Major histocompatibility complex antigen | 27.29 | 2.88E-05 | 0.003 |
| KEGG\_PATHWAY | hsa04940:Type I diabetes mellitus | 19.66 | 9.26E-05 | 0.003 |
| hsa05416:Viral myocarditis | 13.43 | 6.04E-05 | 0.004 |
| hsa05332:Graft-versus-host disease | 18.54 | 0.001 | 0.017 |
| hsa04514:Cell adhesion molecules (CAMs) | 6.66 | 0.002 | 0.020 |
| hsa05330:Allograft rejection | 19.23 | 9.96E-04 | 0.020 |
| hsa05320:Autoimmune thyroid disease | 13.31 | 0.003 | 0.030 |
| UP\_TISSUE | Peripheral blood | 9.46 | 9.25E-05 | 0.009 |
| GOTERM\_CC\_FAT | GO:0042611~MHC protein complex | 21.44 | 7.63E-05 | 0.012 |
| GO:0005576~extracellular region | 1.96 | 4.74E-04 | 0.038 |

**Supplementary Table 14.** **Functional enrichment of BS genes in YRI.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Term** | **Fold Enrichment** | ***P* value** | **Benjamini** |
| SP\_PIR\_KEYWORDS | glycoprotein | 1.64 | 3.36E-14 | 1.48E-11 |
| signal | 1.68 | 2.59E-11 | 5.69E-09 |
| polymorphism | 1.21 | 4.23E-11 | 6.20E-09 |
| disulfide bond | 1.72 | 6.80E-11 | 7.48E-09 |
| Secreted | 1.81 | 8.29E-08 | 7.30E-06 |
| collagen | 5.38 | 1.71E-06 | 1.08E-04 |
| calcium | 2.13 | 1.48E-06 | 1.09E-04 |
| cell adhesion | 2.60 | 5.18E-06 | 2.85E-04 |
| extracellular matrix | 3.13 | 8.14E-06 | 3.98E-04 |
| alternative splicing | 1.22 | 1.39E-05 | 6.09E-04 |
| membrane | 1.26 | 3.25E-05 | 0.001 |
| cell membrane | 1.52 | 4.86E-05 | 0.002 |
| transmembrane | 1.25 | 6.00E-04 | 0.019 |
| guanine-nucleotide releasing factor | 3.53 | 5.85E-04 | 0.020 |
| Signal-anchor | 2.13 | 7.68E-04 | 0.022 |
| zymogen | 2.69 | 9.31E-04 | 0.025 |
| Serine protease | 3.09 | 0.002 | 0.043 |
| disease mutation | 1.47 | 0.002 | 0.044 |
| mhc ii | 14.52 | 0.002 | 0.047 |
| UNIGENE\_EST\_QUARTILE | normal\_disease\_3rd | 1.37 | 1.41E-10 | 1.07E-08 |
| brain\_normal\_3rd | 1.26 | 1.17E-04 | 0.004 |
| eye\_normal\_3rd | 1.25 | 2.64E-04 | 0.007 |
| embryo\_development\_3rd | 1.19 | 4.36E-04 | 0.008 |
| thyroid\_normal\_3rd | 1.26 | 0.003 | 0.044 |
| GOTERM\_CC\_FAT | GO:0005576~extracellular region | 1.67 | 6.75E-08 | 2.07E-05 |
| GO:0031012~extracellular matrix | 2.51 | 1.83E-05 | 0.003 |
| GO:0005578~proteinaceous extracellular matrix | 2.51 | 3.97E-05 | 0.0048 |
| GO:0044421~extracellular region part | 1.74 | 8.53E-05 | 0.005 |
| GO:0044459~plasma membrane part | 1.44 | 7.34E-05 | 0.006 |
| GO:0005581~collagen | 6.19 | 7.75E-04 | 0.039 |
| GO:0005886~plasma membrane | 1.25 | 9.19E-04 | 0.039 |
| GNF\_U133A\_QUARTILE | Pituitary\_3rd | 1.51 | 1.67E-05 | 4.40E-04 |
| leukemiachronicmyelogenous(k562)\_3rd | 1.68 | 1.44E-05 | 5.70E-04 |
| Medulla Oblongata\_3rd | 1.35 | 3.75E-05 | 5.93E-04 |
| Testis\_3rd | 1.54 | 3.51E-05 | 6.93E-04 |
| dorsal root ganglia\_3rd | 1.68 | 1.11E-05 | 8.73E-04 |
| Prostate\_3rd | 1.14 | 2.08E-04 | 0.003 |
| Cerebellum Peduncles\_3rd | 1.57 | 2.69E-04 | 0.003 |
| Colorectal Adenocarcinoma\_3rd | 1.36 | 5.58E-04 | 0.005 |
| Olfactory Bulb\_3rd | 1.41 | 9.92E-04 | 0.009 |
| Thymus\_3rd | 1.35 | 0.001 | 0.010 |
| Smooth Muscle\_3rd | 1.27 | 0.002 | 0.011 |
| Ovary\_3rd | 1.47 | 0.002 | 0.012 |
| Occipital Lobe\_3rd | 1.33 | 0.004 | 0.020 |
| Lung\_3rd | 1.30 | 0.003 | 0.021 |
| Ciliary Ganglion\_3rd | 1.26 | 0.004 | 0.021 |
| Hypothalamus\_3rd | 1.27 | 0.005 | 0.024 |
| PB-BDCA4+Dentritic\_Cells\_3rd | 1.86 | 0.005 | 0.024 |
| Prefrontal Cortex\_3rd | 1.33 | 0.006 | 0.026 |
| Skeletal Muscle\_3rd | 1.15 | 0.007 | 0.030 |
| globuspallidus\_3rd | 1.29 | 0.011 | 0.043 |
| Cerebellum\_3rd | 1.08 | 0.013 | 0.046 |
| GOTERM\_MF\_FAT | GO:0005201~extracellular matrix structural constituent | 5.09 | 7.69E-06 | 0.005 |
| GO:0005509~calcium ion binding | 1.80 | 7.15E-05 | 0.021 |
| PANTHER\_MF\_ALL | MF00101:Guanyl-nucleotide exchange factor | 3.31 | 2.97E-04 | 0.016 |
| MF00178:Extracellular matrix | 2.35 | 1.22E-04 | 0.019 |
| MF00258:CAM family adhesion molecule | 4.66 | 2.57E-04 | 0.020 |
| MF00100:G-protein modulator | 2.16 | 7.36E-04 | 0.029 |
| MF00179:Extracellular matrix structural protein | 3.90 | 9.72E-04 | 0.031 |
| MF00173:Defense/immunity protein | 2.16 | 0.002 | 0.047 |
| GOTERM\_BP\_FAT | GO:0022610~biological adhesion | 2.08 | 1.12E-05 | 0.011 |
| GO:0007155~cell adhesion | 2.08 | 1.06E-05 | 0.022 |

**Supplementary Table 15.** **GO terms of biological processes enriched of BS genes in YRI.** Terms labeled in bold font were the most specific term within the related significant GO hierarchies and were shown in Figure 2e.

|  |  |  |
| --- | --- | --- |
| **Biological Process** | **Count** | ***P* value** |
| antigen processing and presentation | 19 | 2.38E-10 |
| response to stimulus | 101 | 1.72E-09 |
| system process | 91 | 6.95E-09 |
| **antigen processing and presentation of peptide or polysaccharide antigen via MHC class II** | **14** | **3.67E-08** |
| immune system process | 99 | 5.73E-08 |
| **cellular defense response** | **36** | **3.85E-07** |
| multicellular organismal process | 97 | 8.67E-07 |
| single-multicellular organism process | 97 | 8.67E-07 |
| **cell-cell adhesion** | **39** | **2.10E-05** |
| cell communication | 144 | 3.32E-05 |
| cell adhesion | 56 | 3.95E-05 |
| biological adhesion | 56 | 3.95E-05 |
| **neurological system process** | **67** | **1.27E-04** |
| **regulation of liquid surface tension** | **10** | **1.52E-04** |
| homeostatic process | 16 | 7.23E-04 |
| **cell-cell signaling** | **48** | **1.72E-03** |
| cellular process | 224 | 2.09E-03 |
| vesicle-mediated transport | 50 | 8.33E-03 |
| **blood circulation** | **15** | **1.15E-02** |
| anatomical structure morphogenesis | 40 | 1.20E-02 |
| **cellular component morphogenesis** | **38** | **1.20E-02** |
| transport | 109 | 1.47E-02 |
| **induction of apoptosis** | **19** | **1.47E-02** |
| developmental process | 115 | 2.08E-02 |
| localization | 110 | 2.81E-02 |
| **endocytosis** | **26** | **4.10E-02** |

**Supplementary Table 16. Functional classification of PS genes in CEU.** Knowledge/text mining means the annotation was curated from literatures.

|  |  |  |
| --- | --- | --- |
| **Categories** | **GO terms** | **Genes** |
| **Brain** | GO:0007420~brain development | TBR1,NCOR1,STK36,MDK,SPTBN2,PRKDC,SLC6A4,UBB,PAFAH1B3,CEP290,FOXP2,CXCR4,ATP6V0D1,RAB3GAP1,ALX1,DPCD,RAB18 |
| **Metabolism** | GO:0006600~creatine metabolic process | GATM |
| GO:0005975~carbohydrate metabolic process | PC,GALK2,PGM3,LCT,GSK3A,SORD,PPP1R3A,PPIP5K1,B3GNT9,GALNT5,POFUT1 |
| GO:0006629~lipid metabolic process | ACSL6,PLCD4,CYP26B1,LIPE,NCOR1,GPX5,CYP3A4,PC,CYP3A5,PITPNB,HSD11B2,DGKZ,PAFAH1B3,ACAD11,NPHP3,SPTSSB,SLC44A1,ORMDL1,PIGL |
| GO:0017144~drug metabolic process | CYP3A4,CYP3A5,ADAL |
| GO:0055123~digestive system development | IL8,NPHP3,MEGF8 |
| GO:0048565~digestive tract development | IL8,NPHP3,MEGF8 |
| GO:0007586~digestion | LCT |
| **Bone** | GO:0060349~bone morphogenesis | CYP26B1 |
| **Pigmentation** | GO:0048070~regulation of developmental pigmentation | KITLG |
| **Knowledge/text mining** | **SLC24A5** |
| **Hair** | GO:0022405~hair cycle process | FGF7 |
| **Knowledge/text mining** | **KRTs** |
| **Sensory** | GO:0019233~sensory perception of pain | DLG2 |
| GO:0050953~sensory perception of light stimulus | AOC2 |
| GO:0050954~sensory perception of mechanical stimulus | SLC26A5,MAP1A,GJC3 |
| **Reproduction** | GO:0000003~reproduction | FANCC,VMP1,ACVR2A,PLCD4,PGM3,CYP26B1,PRKDC,SLC6A4,UBB,PAFAH1B3,SPAG6,KITLG,PMS1,CXCR4,HSD11B2,ADIG,NR5A1,NR6A1,RQCD1,DPCD,SPIN1 |
| **Knowledge/text mining** | **SPAG17** |
| **Stress** | GO:0001666~response to hypoxia | ACSL6,PAK1,SLC6A4,BECN1,UBB,CXCR4,HSD11B2 |
| GO:0002209~behavioral defense response | MDK |
| GO:0006954~inflammatory response | SAAL1,IL8,CXCR4,AOC3,THEMIS2 |
| GO:0098542~defense response to other organism | FGR,BECN1 |
| GO:0006979~response to oxidative stress | PXDNL,FANCC,GPX6,GATM,GPX5 |
| GO:0009409~response to cold | CDH8 |
| GO:0009611~response to wounding | MDK,SAAL1,GATM,KIF3B,FGR,IL8,FGF7,CXCR4,PAK1,DGKZ,AOC3,THEMIS2,ZFPM2 |
| GO:0042594~response to starvation | ATG13,BECN1,FNIP1 |
| **Immune** | GO:0001776~leukocyte homeostasis | PDE4B,PPP3CB |
| GO:0002200~somatic diversification of immune receptors | PRKDC,XRCC4,POLL |
| GO:0002252~immune effector process | FGR,PAK1,XRCC4,BECN1 |
| GO:0002253~activation of immune response | FGR,UBB,MAPK10,PDE4B,PAK1,THEMIS2 |
| GO:0002262~myeloid cell homeostasis | FANCC,NCOR1,PDE4B,SOX6,EPB42 |
| GO:0002440~production of molecular mediator of immune response | PRKDC,XRCC4,POLL |
| GO:0002520~immune system development | NCOR1,PGM3,SOX6,PRKDC,CBFB,PPP3CB,EPB42,KITLG,BMI1,XRCC4,FNIP1,POLL |
| GO:0006955~immune response | IL8,FGR,IFI6,UBB,PPP3CB,KITLG,FGF7,ADRBK1,GSK3A,MAPK10,PRKDC,PAK1,XRCC4,AGO3,AGO4,ACKR4,AGO1 |
| GO:0019882~antigen processing and presentation | SPTBN2,KIF3B,UBB,PSMC2,PSME3,BLMH |
| GO:0031294~lymphocyte costimulation | PAK1 |
| GO:0045321~leukocyte activation | NCOR1,IL8,PRKDC,CBFB,PPP3CB,CXCR4,XRCC4,FNIP1 |
| GO:0050900~leukocyte migration | IL8,PDE4B,CXCR4 |

**Supplementary Table 17. Functional classification of PS genes in CHB.**

|  |  |  |
| --- | --- | --- |
| **Categories** | **GO terms** | **Genes** |
| **Brain** | GO:0007420~brain development | STK36,ARF4,FYN,GNAQ,SLC6A4,HESX1,NF1,FOXP2,SEPT4,KDM1A,PFDN1,FGF2,NME5,ISL1,ATP6V0D1,CNTN2,PYGO2,LHX8,ASPM,AHI1,RAPGEF2,NME7,ROBO1 |
| **Metabolism** | GO:0006067~ethanol metabolic process | ADH7,ADH5,ADH1B,ADH1A,ADH4,ADH6 |
| GO:0005975~carbohydrate metabolic process | ACADM,RBKS,PFKFB4,SYNJ1,TIPARP,GCKR,PHKG2,PPIP5K2,RPIA,RANBP2,KHK,GPC5,PPP1R3A,GALNT16,ST6GALNAC3,NUP37,NUP35,PREB |
| GO:0006629~lipid metabolic process | ACADM,ABHD5,CYP27A1,SYNJ1,PLCD4,ADH7,TIPARP,ADH5,CYP26B1,HADHA,PLCL1,CDIPT,ACOXL,HSD3B7,CYP7B1,GPX1,ADH4,CYP3A4,FGF2,FDPS,RBP2,HADHB,MTTP,GPC5,HSD11B2,SNX17,SLC44A5,ST6GALNAC3,SLC44A1,OXSM,MTMR4,SLC27A6 |
| GO:0017144~drug metabolic process | ADH1A,CYP3A4,CAD |
| GO:0022600~digestive system process | SLC26A7,UCN |
| GO:0055123~digestive system development | RB1,AHI1,NCKAP1 |
| GO:0048565~digestive tract development | RB1,AHI1,NCKAP1 |
| GO:0007586~digestion | SLC26A7,ACE,UCN,UCN2 |
| **Bone** | GO:0060349~bone morphogenesis | RARA,CYP26B1,RAB23 |
| **Pigmentation** | GO:0043473~pigmentation | GNAQ,NF1,AP3B1,EDAR |
| **Hair** | GO:0022405~hair cycle process | RUNX1,EDAR |
| GO:0035315~hair cell differentiation | CUX1,GFI1 |
| **Sensory** | GO:0019233~sensory perception of pain | FYN,DLG2 |
| GO:0050953~sensory perception of light stimulus | RABGGTB,TRPM1,ZNF513,BBS10,RORB,HMCN1 |
| GO:0050954~sensory perception of mechanical stimulus | FYN,GPX1,SLC12A2,TIMM9 |
| GO:0021988 ~olfactory lobe development | ROBO1 |
| **Reproduction** | GO:0000003~reproduction | KDM3A,PLCD4,ADAD1,ZPBP,TIPARP,CYP26B1,SLC6A4,RARA,RAD51C,MSH4,SEPT4,DHH,SPAG9,SPAG6,CYP7B1,ZNF35,PMCH,CAD,UCN,NME5,HSD11B2,LHX8,SPATA31D1,SIAH1,TEX14,DND1,ASPM,BOLL,SPATA5,RQCD1,PVRL3,EIF2B4,RUVBL1 |
| **Stress** | GO:0001666~response to hypoxia | ATP1B1,RHOA,SLC6A4,NF1,BACH1,UBE2D2,HSD11B2,MGARP,UCN2 |
| GO:0003299~muscle hypertrophy in response to stress | HEY2 |
| GO:0006954~inflammatory response | C5,UCN,HNRNPA0,PYCARD |
| GO:0098542~defense response to other organism | DEFB112,DEFB110,PYCARD |
| GO:0006979~response to oxidative stress | MPV17,GPX1,UCN,GPX7 |
| GO:0009408~response to heat | EIF2B4 |
| GO:0009409~response to cold | ACADM,CDH8 |
| GO:0009611~response to wounding | ARF4,AP3B1,GNA13,NF1,KIF3C,RAD51C,KDM1A,C5,ATP1B1,F13B,FYN,GPX1,FGF2,ITGAM,SHC1,VEGFC,GNAQ,UCN,ISL1,RHOA,DST,HNRNPA0,MAP3K1,KIF22,TRIM72,PARD3,HBEGF,KIF15,PYCARD,KCNMB2 |
| GO:0042594~response to starvation | ACADM,RAB23 |
| GO:0051409~response to nitrosative stress | ADH5 |
| **Immune** | GO:0002252~immune effector process | WIPF1,C5,FYN,POU2F2,WIPF2,CYFIP2,CFHR5,PYCARD,NCKAP1 |
| GO:0002253~activation of immune response | PTPN22,WIPF1,C5,FYN,ITGAM,ZAP70,MAPK10,UBE2D2,ITK,MAP3K1,WIPF2,CYFIP2,CFHR5,PYCARD,NCKAP1 |
| GO:0002262~myeloid cell homeostasis | ARID4A,RB1 |
| GO:0002440~production of molecular mediator of immune response | TRAF3IP2,POU2F2 |
| GO:0002520~immune system development | ARID4A,TIPARP,ITK,PTPN22,KDM1A,ARIH2,RB1,POU2F2,ITGAM,ACE,BMI1,ZAP70,IL2,RUNX1,GON4L,SNRK |
| GO:0006955~immune response | CCL28,WIPF1,TRAF3IP2,C5,FYN,FGF2,POU2F2,ITGAM,SHC1,ZAP70,MAPK10,IL2,UBE2D2,ITK,IK,MAP3K1,PKHD1L1,RNF135,WIPF2,PYDC1,CYFIP2,HBEGF,CFHR5,AGO3,AGO1,PYCARD,NCKAP1 |
| GO:0019882~antigen processing and presentation | AP3B1,KIF3C,DYNC1LI2,PSMA3,ACTR1B,BLMH,KIF22,DYNLL2,KIF15 |
| GO:0031294~lymphocyte costimulation | FYN |
| GO:0045058~T cell selection | ZAP70 |
| GO:0045321~leukocyte activation | FYN,ITK,PTPN22,PFDN1,SKAP2,POU2F2,ITGAM,IMPDH2,ZAP70,IL2,GON4L,PYCARD |
| GO:0050900~leukocyte migration | C5,ATP1B1,FYN,ITGAM,SHC1,ZAP70 |

**Supplementary Table 18. Functional classification of PS genes in YRI.**

|  |  |  |
| --- | --- | --- |
| **Categories** | **GO terms** | **Genes** |
| **Brain** | GO:0007420~brain development | MACROD2,PAX6,MET,SEMA4C,CTTNBP2,ZEB2,CEP120,CADM1,MEF2A,DIXDC1,PAFAH1B3,CACNA1A,CEP290,SEPT4,RBFOX2,GRID2,PFDN1,SLIT1,NR2F1,C5AR1,WNT3A,NRXN1,RAB3GAP1,NTRK2,PYGO2,AHI1,DYNC2H1,FEZF2,ULK4,SMO,CNTNAP2 |
| **Knowledge/text mining** | **SPON1,SORL1,MAPT,SNCA,CADPS,ELAVL4,KCTD7,RNF180,SHC3,SLTM,AUTS2** |
| **Metabolism** | GO:0006067~ethanol metabolic process | ALDH2,ACSS2 |
| GO:0005975~carbohydrate metabolic process | ALDOC,CS,PC,MGAT4C,SYNJ1,SPAM1,HYAL4,HYAL2,PARP3,HYAL1,HYAL3,GBE1,LCT,PHKB,ST6GAL2,CACNA1A,PPIP5K2,TRAK2,LARGE,ALDH2,DLAT,PYGM,SDS,OGN,SLC5A2,VEGFB,GSK3A,CRTC2,KIAA1161,GALNT12,GPC2,GALNT16,POGLUT1,ST6GALNAC3,NANP,PTPMT1,GAL3ST4,SDF2,OMD,FUT8,POFUT1,ALG9,PIGV,B3GNT2,NISCH |
| GO:0006629~lipid metabolic process | PIGV,ABHD5,SYNJ1,NR1H3,CFTR,SNCA,APOF,PIGS,ACSS2,PROCA1,LIPE,HDAC3,LARGE,FSHB,PLA2G4B,PC,ALAS1,CD36,LIPA,CNBP,ARF1,PITPNA,MTMR3,PAFAH1B3,PMVK,HSD17B12,ARSK,ANKRD23,SGMS1,GPC2,ST6GALNAC3,PLBD2,SGMS2,PTPMT1,PIGC,SORL1,CPNE1,MECR,MTMR4,ORMDL1,GGT7,PLAGL2,SLC27A6 |
| GO:0017144~drug metabolic process | GUK1 |
| GO:0022600~digestive system process | TRPC1,TLR9 |
| GO:0055123~digestive system development | FOXE1,MEGF8,AHI1,SMO,CCDC39 |
| GO:0048565~digestive tract development | FOXE1,MEGF8,AHI1,SMO,CCDC39 |
| GO:0007586~digestion | LCT,TRPC1,TLR9 |
| **Knowledge/text mining** | **MAPKAP1** |
| **Bone** | GO:0060349~bone morphogenesis | IFT80,MEF2A,COL13A1 |
| **Knowledge/text mining** | **BMPR2,CILP2,ZSWIM6** |
| **Pigmentation** | GO:0048070~regulation of developmental pigmentation | KITLG |
| GO:0043473~pigmentation | MITF,MEF2A,ASIP,HPS3,DTNBP1,C10orf11 |
| **Hair** | GO:0022405~hair cycle process | FOXE1,RUNX1,KRT27,SMO |
| Knowledge/text mining | KRTs (except KRT27) |
| **Sensory** | GO:0007606~sensory perception of chemical stimulus | GNAT3,GNAT1,GRM8,C5AR1,B3GNT2 |
| GO:0050953~sensory perception of light stimulus | GNAT1,GRM8,PDE6D,PAX6,RP1,PITPNA,UNC119,CNNM4,RIMS1,BBS10,KRT12 |
| GO:0050954~sensory perception of mechanical stimulus | GNAT1,GRM8,PDE6D,PAX6,RP1,PITPNA,UNC119,CNNM4,RIMS1,BBS10,KRT |
| GO:0021988 ~olfactory lobe development | POU2F1 |
| **Reproduction** | GO:0000003~reproduction | DDR1,KCNU1,FANCC,CREB3L4,ACVR2A,SMC4,SPAM1,ASZ1,CELF1,MLH3,RPA1,RAD51C,DHODH,PAFAH1B3,RAD51B,SEPT4,CLDN11,PARN,FSHB,PLA2G4B,CFTR,ZNF35,LIF,KITLG,RPL29,EIF2B2,RAB13,PMS1,SPESP1,HORMAD1,DZIP1,TEX14,NPAS3,SPAG16,HORMAD2,SPATA5,ANKRD7,SCMH1,SMO,SPATA9,TSSK6,CADM1,SPAG4,STAG3,CLIC4 |
| **Knowledge/text mining** | **KIF3B,SPAG5,SPAG17** |
| **Stress** | GO:0001666~response to hypoxia | SCFD1,ALDOC,EGR1,ARNT,UBE2D2 |
| GO:0003299~muscle hypertrophy in response to stress | MEF2A,HEY2 |
| GO:0006954~inflammatory response | CCR2,HYAL3,TPST1,TUSC2,HP,PLA2G4B,C5AR1,CCR1,LIPA,CCR3,CCR5,HYAL1,HNRNPA0,ABCF1,IL23A,TLR9,STAB1,HDAC9,SGMS1 |
| GO:0098542~defense response to other organism | HYAL2,CD36,IFIT3,HP,IFIT2,IFIT1,C5AR1,STAT2,ROMO1,DNAJC3,IL23A,TLR9,STAB1 |
| GO:0006979~response to oxidative stress | FANCC,NFE2L2,SNCA,CA3,CD36,GPX2,HP,ARNT,GSS,ROMO1,HYAL1,HYAL2,MAPK7,ADAM9 |
| GO:0009408~response to heat | OSM,EIF2B2,STAC |
| GO:0009409~response to cold | CDH8,IMMT |
| GO:0009611~response to wounding | DDR1,GNAT3,CHURC1-FNTB,DTNBP1,CPQ,CCR2,FERMT3,MAX,STXBP3,KIF3B,RAD51B,RAD51C,HYAL3,TPST1,TUSC2,HP,GNAI2,PLA2G4B,GP9,CD36,ATP2B1,C5AR1,PAX6,SHC1,CCR1,LIPA,CAPZA2,FNTB,VEGFB,CCR3,CCR5,WNT3A,BNC1,HYAL1,HNRNPA0,ADAM15,RASGRP2,ABCF1,KIF18A,SCG3,HBEGF,KLC2,TRPC7,IL23A,TLR9,KIF15,STAB1,HDAC9,SGMS1 |
| GO:0042594~response to starvation | SLC39A5,MAX,DHODH,BMPR2 |
| GO:0055093~response to hyperoxia | EGR1 |
| **Immune** | GO:0001776~leukocyte homeostasis | SKIL,PDE4B,MEF2A,RAG1,PPP3CB,ZC3H8 |
| GO:0002200~somatic diversification of immune receptors | RAG1,RAG2 |
| GO:0002252~immune effector process | HYAL2,STXBP3,WASL,IFIT3,POU2F2,IFIT2,IFIT1,GNL1,STAT2,DNAJC3,CFHR5,IL23A |
| GO:0002253~activation of immune response | MAPKAPK3,MEF2A,WASL,CD36,C5AR1,CTSS,CTSK,UBE2D2,PDE4B,PDE4D,MAPK7,DUSP7,CFHR5,TLR9,TRAF6 |
| GO:0002262~myeloid cell homeostasis | FANCC,PDE4B,SOX6,SPI1 |
| GO:0002440~production of molecular mediator of immune response | POU2F2,RAG2 |
| GO:0002520~immune system development | STK3,SATB1,HYAL2,MITF,SOX6,RAG1,RAG2,TSHR,SIX4,MEF2A,AHSP,RPA1,FOXE1,TUSC2,POU2F2,PPP3CB,SPI1,EGR1,KITLG,WNT3A,RUNX1,SIX1,SNRK,HDAC9,GAB2,TRAF6 |
| GO:0006955~immune response | OSM,IFIT1,CCR2,MEF2A,STXBP3,WASL,IFIT3,IGSF6,POU2F2,IFIT2,LIF,RAG1,PPP3CB,CD36,EGR1,KITLG,C5AR1,CTSS,SHC1,CEACAM8,CCR1,CD80,GNL1,SNCA,CTSK,CAPZA2,GSK3A,CCR5,STAT2,UBE2D2,MAPK7,MAPKAPK3,DUSP7,IFIT1B,HBEGF,MAPKAP1,CFHR5,IL23A,TLR9,GAB2,TRAF6 |
| GO:0019882~antigen processing and presentation | KIF3B,CD36,PSMC3,ITGB5,CTSS,PSMA5,PSMA6,ARF1,DYNC2H1,KIF18A,KLC2,KIF15,TRAF6 |
| GO:0031294~lymphocyte costimulation | CD80,MAPKAP1 |
| GO:0050900~leukocyte migration | EPS8,PDE4D,PDE4B,PODXL,C5AR1,SHC1,CCR1,CCR2,CCR3,CCR5 |
| **Knowledge/text mining** | **LARGE** |

**Supplementary Table 19.** **Labeled functional groups in CEU.** Number in bracket denotes the distance to the center segment.

|  |  |  |
| --- | --- | --- |
| **Candidate regions** | **Genes** | **Functions** |
| chr1:27890701-28200000 | FGR(0) | stress |
| chr1:36172301-36612900 | AGO3(0) | immune |
| chr1:66473201-66645600 | PDE4B(0) | immune |
| chr1:118370501-118513400 | SPAG17(0) | reproduction |
| chr2:135295301-137805400 | LCT(154.6) | metabolism |
| chr2:148508701-148925400 | ACVR2A(0) | reproduction |
| chr2:158166001-158281300 | GALNT5(0) | metabolism |
| chr2:162272001-162573200 | TBR1(0) | brain |
| chr2:219291801-219587300 | STK36(0) | brain |
| chr3:132116101-132451100 | NPHP3(0) | metabolism |
| chr4:74535601-74698200 | IL8(15.5) | immune |
| chr4:87187001-87400900 | MAPK10(0) | immune |
| chr5:82290501-82637400 | XRCC4(0) | immune |
| chr5:130884801-131353100 | ACSL6(0) | metabolism |
| chr7:99017401-99521900 | CYP3A4(0) | metabolism |
| chr7:102886201-103064100 | SLC26A5(0) | sensory |
| chr7:113713801-113837100 | FOXP2(0) | brain |
| chr8:52525801-52667400 | PXDNL(0) | stress |
| chr8:106311101-106496200 | ZFPM2(0) | stress |
| chr9:90950701-91143000 | SPIN1(11.7) | reproduction |
| chr9:98002201-98104500 | FANCC(0) | stress |
| chr9:108003701-108201300 | SLC44A1(0) | metabolism |
| chr9:127256601-127442800 | NR6A1(0) | reproduction |
| chr10:27758601-27867300 | RAB18(1.1) | brain |
| chr10:75106501-75402700 | PPP3CB(0) | immune |
| chr11:16009101-16210400 | SOX6(0) | immune |
| chr11:17954401-18094800 | SAAL1(0) | stress |
| chr11:46331901-46707100 | DGKZ(0) | stress |
| chr11:66847201-67072800 | ADRBK1(0) | immune |
| chr11:77014001-77189800 | PAK1(0) | stress |
| chr11:83338801-83463100 | DLG2(0) | sensory |
| chr12:85584501-86133400 | ALX1(231.8) | brain |
| chr12:88453201-89381300 | KITLG(136.0) | pigmentation |
| chr15:45245401-45354200 | SORD(0) | metabolism |
| chr15:45491901-45742300 | GATM(23.9) | metabolism |
| chr15:48303201-48767900 | SLC24A5(0) | pigmentation |
| chr15:49601101-49838400 | GALK2(16.8) | metabolism |
| chr16:61780801-61991200 | CDH8(0) | stress |
| chr16:67031601-67637400 | CBFB(0) | immune |
| chr17:40923401-41043600 | BECN1(0) | stress |
| chr17:57816901-57932500 | VMP1(0) | reproduction |
| chr19:42718701-42927600 | GSK3A(0) | metabolism |
| chr20:30818001-31011400 | POFUT1(0) | metabolism |
| chr21:31941601-32162100 | KRTAP21-3(2.4) | hair |
| chr22:28215801-28316300 | PITPNB(0) | metabolism |

**Supplementary Table 20.** **Labeled functional groups in CHB.**

|  |  |  |
| --- | --- | --- |
| **Candidate regions** | **Genes** | **Functions** |
| chr1:36344801-36529800 | AGO3(0) | immune |
| chr1:75278701-75815100 | SLC44A5(2.3) | metabolism |
| chr1:75919701-76724100 | ST6GALNAC3(16.4) | metabolism |
| chr1:92512501-93005100 | GFI1(0) | hair |
| chr1:154919701-155020900 | PYGO2(0) | brain |
| chr1:155286401-155802000 | FDPS(0) | metabolism |
| chr1:169093501-169428200 | NME7(0) | brain |
| chr1:185697601-185975000 | HMCN1(0) | sensory |
| chr1:196946101-197112400 | CFHR5(0) | immune |
| chr1:205023501-205175800 | CNTN2(0) | brain |
| chr2:26167201-26518700 | KIF3C(0) | stress |
| chr2:86597201-86818400 | KDM3A(0) | reproduction |
| chr2:88958601-89090900 | RPIA(8.4) | metabolism |
| chr2:108888501-109628300 | EDAR(0) | hair |
| chr2:111473401-111576100 | ACOXL(0) | metabolism |
| chr2:175368001-175580000 | WIPF1(22.7) | immune |
| chr2:183771001-183986900 | NCKAP1(0) | metabolism |
| chr2:219271701-219684100 | CYP27A1(0) | metabolism |
| chr3:44496501-44954600 | KIF15(0) | stress |
| chr3:79198101-79403100 | ROBO1(0) | brain |
| chr3:110526801-110879500 | PVRL3(0.7) | reproduction |
| chr3:177725101-178057700 | KCNMB2(185.9) | stress |
| chr4:87075901-87630600 | MAPK10(0) | immune |
| chr4:99708601-100505500 | ADHs | metabolism |
| chr4:123070001-123502300 | ADAD1(0) | reproduction |
| chr4:123686201-123928000 | FGF2(0) | stress |
| chr4:160241001-160465800 | RAPGEF2(119.3) | brain |
| chr4:177613701-177869300 | VEGFC(0) | stress |
| chr5:50499801-50791500 | ISL1(62.1) | stress |
| chr5:56068201-56250900 | MAP3K1(40.0) | stress |
| chr5:77332401-77492400 | AP3B1(0) | pigmentation |
| chr5:102459901-102684200 | PPIP5K2(0) | metabolism |
| chr5:127346301-127603500 | SLC12A2(9.7) | sensory |
| chr5:127910501-128013900 | SLC27A6(0) | metabolism |
| chr5:137030101-137533200 | NME5(7.3) | brain |
| chr5:138889901-139110900 | UBE2D2(0) | stress |
| chr5:156673401-156920300 | CYFIP2(0) | immune |
| chr6:49958401-50078600 | DEFB110(0) | stress |
| chr6:56484901-56616600 | DST(0) | stress |
| chr6:112107201-112230300 | FYN(0) | stress |
| chr6:135592501-135813300 | AHI1(0) | brain |
| chr7:26919601-27087000 | SKAP2(17.2) | immune |
| chr7:49879701-50154900 | ZPBP(0) | reproduction |
| chr7:99365901-99471000 | CYP3A4(0) | metabolism |
| chr7:101562501-101744000 | CUX1(0) | hair |
| chr7:113422401-113909400 | PPP1R3A(0) | metabolism |
| chr8:65500901-65762600 | CYP7B1(0) | metabolism |
| chr8:92399001-92560700 | SLC26A7(69.5) | metabolism |
| chr9:77021501-77136400 | RORB(35.7) | sensory |
| chr9:80497201-80721500 | GNAQ(0) | stress |
| chr9:84498201-84977600 | SPATA31D1(74.2) | reproduction |
| chr9:108070301-108202000 | SLC44A1(0) | metabolism |
| chr9:123730801-123832800 | C5(0) | stress |
| chr10:22539301-22768300 | SPAG6(0) | reproduction |
| chr10:34530501-35074100 | PARD3(0) | stress |
| chr11:84843601-85406900 | DLG2(0) | sensory |
| chr12:76722301-76910700 | BBS10(15.5) | sensory |
| chr13:92629301-92784900 | GPC5(0) | metabolism |
| chr14:58625001-59067200 | TIMM9(0) | sensory |
| chr15:31159301-31360800 | TRPM1(0) | sensory |
| chr16:29765201-29875700 | KIF22(0) | stress |
| chr16:31180001-31330200 | ITGAM(0) | immune |
| chr16:61843801-61991200 | CDH8(0) | stress |
| chr16:67452801-67679400 | HSD11B2(0) | metabolism |
| chr17:27925501-28663400 | SLC6A4(0) | brain |
| chr17:49036501-49208400 | SPAG9(0) | reproduction |
| chr17:56023701-56181800 | DYNLL2(0) | immune |
| chr17:62936601-63063900 | GNA13(0) | stress |
| chr19:42543101-42668100 | POU2F2(0) | immune |
| chr21:30616301-30819600 | BACH1(0) | stress |
| chr21:37076601-37238000 | RUNX1(0) | immune |

**Supplementary Table 21.** **Labeled functional groups in YRI.**

|  |  |  |
| --- | --- | --- |
| **Candidate regions** | **Genes** | **Functions** |
| chr1:24949601-25176200 | CLIC4(0) | reproduction |
| chr1:41615501-41748800 | SCMH1(0.8) | reproduction |
| chr1:50564001-50668700 | ELAVL4(0) | brain |
| chr1:66279301-66496000 | PDE4B(0) | immune |
| chr1:76783901-76896500 | ST6GALNAC3(0) | metabolism |
| chr1:109268601-109494900 | STXBP3(0) | stress |
| chr1:109845401-110042100 | PSMA5(16.5) | immune |
| chr1:150544901-150919100 | CTSK(1.8) | immune |
| chr1:167166701-167281200 | POU2F1(0) | sensory |
| chr2:50884501-51376600 | NRXN1(0) | brain |
| chr2:107263301-107424500 | ST6GAL2(132.0) | metabolism |
| chr2:112958501-113097800 | ZC3H8(0) | immune |
| chr2:135778101-135934100 | RAB3GAP1(0) | brain |
| chr2:145231701-145303700 | ZEB2(0) | brain |
| chr2:148513201-148743000 | ACVR2A(0) | reproduction |
| chr2:177991801-178119000 | NFE2L2(0) | stress |
| chr2:203334601-203441100 | BMPR2(0) | bone |
| chr2:214171801-214344300 | SPAG16(0) | reproduction |
| chr3:36483001-36606600 | STAC(0) | stress |
| chr3:41461501-41529800 | ULK4(0) | brain |
| chr3:44202601-45016600 | KIF15(0) | stress |
| chr3:46197101-46419200 | CCR5(0) | immune |
| chr3:62355001-62433800 | CADPS(0) | brain |
| chr3:69835601-70124500 | MITF(105.7) | pigmentation |
| chr3:81568901-82208100 | GBE1(0) | metabolism |
| chr3:119114701-119264500 | POGLUT1(0) | metabolism |
| chr3:124471501-124498900 | ITGB5(0) | immune |
| chr3:148755701-148872100 | HPS3(0) | pigmentation |
| chr4:90637801-90759100 | SNCA(0) | brain |
| chr4:93364901-93625900 | GRID2(0) | brain |
| chr4:93911201-94205800 | GRID2(0) | brain |
| chr4:108756901-108783500 | SGMS2(0) | metabolism |
| chr4:123861301-123995400 | SPATA5(0) | reproduction |
| chr5:60732301-60843700 | ZSWIM6(0) | bone |
| chr5:63280101-63492200 | RNF180(153.7) | brain |
| chr5:92807201-92976400 | NR2F1(7.2) | brain |
| chr5:102454601-102560700 | PPIP5K2(0) | metabolism |
| chr5:135581801-135681900 | TRPC7(0) | stress |
| chr5:140887401-141013200 | DIAPH1(0) | sensory |
| chr6:15525501-15663900 | DTNBP1(0) | pigmentation |
| chr6:30798601-30921000 | DDR1(0) | stress |
| chr6:72929401-73086100 | RIMS1(0) | sensory |
| chr6:125933101-126419900 | HEY2(11.0) | stress |
| chr6:135625001-135930100 | AHI1(0) | brain |
| chr7:18467101-18600100 | HDAC9(0) | stress |
| chr7:65658001-65830300 | TPST1(0) | stress |
| chr7:65974201-66114700 | KCTD7(15.7) | brain |
| chr7:69197101-69343700 | AUTS2(0) | brain |
| chr7:80123301-80216300 | CD36(0) | stress |
| chr7:116345201-116578000 | MET(0) | brain |
| chr7:116780501-117388800 | CFTR(0) | metabolism |
| chr7:126740601-126843400 | GRM8(0) | sensory |
| chr7:146060401-146239900 | CNTNAP2(0) | brain |
| chr8:36311001-36705800 | KCNU1(0) | reproduction |
| chr8:38851401-39058000 | ADAM9(0) | stress |
| chr8:55533601-55733300 | RP1(55.5) | sensory |
| chr8:86052401-86357800 | CA3(0) | stress |
| chr8:97687501-97717400 | CPQ(0) | stress |
| chr8:99754201-99869200 | STK3(0) | immune |
| chr9:87207101-87390400 | NTRK2(0) | brain |
| chr9:91690601-91839100 | SHC3(0) | brain |
| chr9:97982801-98114000 | FANCC(0) | stress |
| chr9:128236201-128453200 | MAPKAP1(0) | immune |
| chr10:52172101-52425100 | SGMS1(0) | stress |
| chr10:71671601-71704000 | COL13A1(0) | bone |
| chr10:75163101-75269400 | PPP3CB(0) | immune |
| chr10:77285801-77612100 | C10orf11(0) | pigmentation |
| chr10:91064801-91168400 | LIPA(0) | metabolism |
| chr10:98747001-98773200 | SLIT1(8.2) | brain |
| chr11:14149101-14185700 | SPON1(0) | brain |
| chr11:16395601-16604200 | SOX6(0) | immune |
| chr11:36498101-36686700 | TRAF6(0) | immune |
| chr11:43329101-43695000 | HSD17B12(0) | metabolism |
| chr11:47572101-47729900 | PTPMT1 | metabolism |
| chr11:63864001-64004100 | FERMT3(0) | stress |
| chr11:65836101-66063700 | KLC2(0.4) | stress |
| chr11:66650001-66802900 | PC(0) | metabolism |
| chr11:78016101-78119900 | GAB2(0) | immune |
| chr11:102921501-103127200 | DYNC2H1(0) | immune |
| chr11:111553001-111719200 | SIK2(0) | metabolism |
| chr11:115179001-115392600 | CADM1(5.0) | reproduction |
| chr11:121388101-121553600 | SORL1(0) | brain |
| chr12:15872701-16020500 | EPS8(0) | immune |
| chr12:56627401-56754500 | CS(0) | metabolism |
| chr12:87059901-87161400 | MGAT4C(0) | metabolism |
| chr12:88849801-89357700 | KITLG(165.6) | pigmentation |
| chr12:89944401-90083600 | ATP2B1(0) | stress |
| chr12:112182301-112294500 | ALDH2(0) | metabolism |
| chr12:113528201-113890300 | SDS(0) | metabolism |
| chr13:96251701-96409500 | DZIP1(0) | reproduction |
| chr14:31020101-31202800 | SCFD1(0) | stress |
| chr14:34060501-34092000 | NPAS3(0) | reproduction |
| chr14:65379801-65562400 | CHURC1-FNTB(0) | stress |
| chr14:65977201-66196000 | FUT8(0) | metabolism |
| chr14:68515801-68727200 | RAD51B(0) | reproduction |
| chr14:75440201-75560000 | MLH3(0) | reproduction |
| chr15:59160201-59225000 | SLTM(0) | brain |
| chr15:69185801-69321200 | SPESP1(0) | reproduction |
| chr15:83318201-84063400 | BNC1(24.6) | stress |
| chr15:100104501-100219400 | MEF2A(0) | stress |
| chr16:14602101-14719500 | PARN(0) | reproduction |
| chr16:47472901-48244400 | PHKB(48.6) | metabolism |
| chr16:61798101-61963400 | CDH8(0) | stress |
| chr17:1422401-1443500 | PITPNA(0) | metabolism |
| chr17:1738501-1761000 | RPA1(0) | immune |
| chr17:26871601-27088000 | UNC119(0) | sensory |
| chr17:38916601-39048500 | KRT26(0) | hair |
| chr17:44039201-44079000 | MAPT(0) | brain |
| chr17:53272101-53357100 | HLF(51.1) | immune |
| chr19:13577701-13614900 | CACNA1A(0) | brain |
| chr19:19550001-19753800 | TSSK6 | reproduction |
| chr19:47799901-47847200 | C5AR1(0) | immune |
| chr20:14202201-14262900 | MACROD2(0) | brain |
| chr20:14395001-14463600 | MACROD2(0) | brain |
| chr20:25540901-25689900 | NANP(18.6) | metabolism |
| chr20:30700001-31028100 | KIF3B(14.7) | stress |
| chr20:32783501-32898500 | ASIP(0) | pigmentation |
| chr20:33430001-33557900 | GGT7(0) | metabolism |
| chr21:31643701-31664500 | KRTAP24-1(0) | hair |
| chr21:36425001-36450900 | RUNX1(0) | immune |
| chr22:30098401-30797600 | HORMAD2(0) | reproduction |
| chr22:34151101-34175800 | LARGE(0) | immune |
| chr22:36267201-36433000 | RBFOX2(0) | brain |

**Supplementary Table 22.** **aFM-like selection regions in CEU.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Regions** | **Time (g)** | **Frequency** | | ***DaHG,MHG*** | ***DaFM,MHG*** | **Center genes** |
| chr1:49402201-50047800 | 1561 | | 0.535 | 37 | 4 | AGBL4(0) |
| chr1:79829201-79938400 | 395 | | 0.318 | 54 | 9 |  |
| chr 1:215351201-215562400 | 1055 | | 0.582 | 32 | 19 | KCNK2(71.3) |
| chr 3:132116101-132451100 | 1109 | | 0.606 | 44 | 3 | ACAD11(0),NPHP3(0) |
| chr 4:128652401-128947100 | 1032 | | 0.512 | 33 | 6 | C4orf29(0) |
| chr 7:69708401-69817300 | 842 | | 0.7 | 47 | 5 | AUTS2(0) |
| chr 7:102886201-103064100 | 185 | | 0.571 | 45 | 20 | PSMC2(0),SLC26A5(0) |
| chr 7:118450501-118568600 | 885 | | 0.453 | 21 | 6 |  |
| chr 8:52525801-52667400 | 1065 | | 0.524 | 86 | 9 | PXDNL(0) |
| chr 8:86018101-86241000 | 1085 | | 0.459 | 92 | 12 | E2F5(4.4) |
| chr 9:98002201-98104500 | 296 | | 0.288 | 33 | 20 | FANCC(0) |
| chr 10:27758601-27867300 | 345 | | 0.312 | 123 | 12 | RAB18(1.1) |
| chr 10:103336801-103475500 | 1132 | | 0.306 | 36 | 12 | FBXW4(0) |
| chr 11:83338801-83463100 | 1087 | | 0.565 | 92 | 10 | DLG2(0) |
| chr 15:45491901-45742300 | 1020 | | 0.529 | 39 | 6 | C15orf48(3.5), **GATM**(23.9) |
| chr 15:48303201-48767900 | 564 | | 0.994 | 28 | 2 | **SLC24A5**(0),MYEF2(0),CTXN2(0),SLC12A1(0) |
| chr 15:49601101-49838400 | 623 | | 0.3 | 49 | 5 | FAM227B(0), **GALK2**(16.8) |

**Supplementary Table 23.** **aHG-like selection regions in CEU.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Regions** | **Time (g)** | **Frequency** | ***DaHG,MHG*** | ***DaFM,MHG*** | **Center genes** |
| chr 2:72361801-73121400 | 925 | 0.682 | 6 | 22 | EXOC6B(0) |
| chr 2:148508701-148925400 | 970 | 0.382 | 5 | 80 | ACVR2A(0) |
| chr 2:162272001-162573200 | 914 | 0.347 | 10 | 44 | TBR1(0) |
| chr 2:196705301-197023100 | 990 | 0.588 | 6 | 31 | DNAH7(0) |
| chr 2:208818901-208925800 | 835 | 0.153 | 8 | 46 | PLEKHM3(0) |
| chr 2:219291801-219587300 | 1010 | 0.471 | 4 | 25 | STK36(0) |
| chr 3:93857301-93986700 | 1435 | 0.365 | 3 | 24 |  |
| chr 4:33850301-34445900 | 1083 | 0.806 | 20 | 31 |  |
| chr 4:41902001-42193400 | 820 | 0.682 | 5 | 61 | TMEM33(5.8) |
| chr 4:87187001-87400900 | 527 | 0.2 | 5 | 85 | MAPK10(0) |
| chr 5:86583901-86699700 | 1283 | 0.512 | 3 | 39 | RASA1(0),CCNH(0) |
| chr 5:87022801-87337000 | 1165 | 0.441 | 7 | 102 |  |
| chr 6:39399301-39541400 | 281 | 0.494 | 5 | 29 | KIF6(0) |
| chr 6:147425001-147531100 | 941 | 0.365 | 17 | 23 | STXBP5(47.4) |
| chr 8:53032501-53162300 | 1005 | 0.476 | 20 | 96 | ST18(0) |
| chr 8:92974301-93165600 | 977 | 0.659 | 6 | 34 | RUNX1T1(16.6) |
| chr 9:127256601-127442800 | 791 | 0.465 | 8 | 50 | NR6A1(0) |
| chr 10:111655201-111860600 | 1220 | 0.518 | 17 | 23 | ADD3(0) |
| chr 12:85584501-86133400 | 863 | 0.594 | 5 | 41 | ALX1(231.8) |
| chr 14:67084001-67888300 | 987 | 0.453 | 4 | 27 | FAM71D(0),MPP5(0),ATP6V1D(0) |
| chr 15:35837401-35944300 | 753 | 0.635 | 4 | 41 | DPH6(77.3) |
| chr 15:43427301-43831200 | 783 | 0.653 | 3 | 59 | TP53BP1(0) |
| chr 15:45245401-45354200 | 829 | 0.682 | 4 | 70 | **SORD**(0) |
| chr 16:67031601-67637400 | 759 | 0.788 | 13 | 39 | CBFB(0),C16orf70(0),B3GNT9(0),TRADD(0),FBXL8(0),HSF4(0),NOL3(0),KIAA0895L(0),EXOC3L1(0),E2F4(0),ELMO3(0),LRRC29(0),TMEM208(0),FHOD1(0),SLC9A5(0),PLEKHG4(0),KCTD19(0),LRRC36(0),TPPP3(0),ZDHHC1(0),HSD11B2(0),ATP6V0D1(0),AGRP(0),FAM65A(0) |
| chr 17:50457701-50610500 | 724 | 0.594 | 19 | 80 |  |
| chr 18:50776401-50879800 | 750 | 0.359 | 5 | 99 | DCC(0) |

**Supplementary Table 24.** **aFM-aHG common selection regions in CEU.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Regions** | **Time (g)** | **Frequency** | ***DaHG,MHG*** | ***DaFM,MHG*** | **Center genes** |
| chr1:27890701-28200000 | 848 | 0.771 | 10 | 11 | AHDC1(0),FGR(0) |
| chr 1:36172301-36612900 | 1097 | 0.782 | 4 | 10 | AGO3(0) |
| chr 1:118370501-118513400 | 992 | 0.476 | 3 | 7 | GDAP2(0),WDR3(0),SPAG17(0) |
| chr 1:178252301-178460800 | 882 | 0.788 | 6 | 7 | RASAL2(0) |
| chr 1:198173101-198541700 | 720 | 0.576 | 3 | 9 | ATP6V1G3(39.0) |
| chr 2:158166001-158281300 | 920 | 0.782 | 6 | 8 | **GALNT5**(0) |
| chr 2:190526301-190798100 | 758 | 0.624 | 9 | 8 | C2orf88(0) |
| chr 2:194007901-194138000 | 971 | 0.606 | 12 | 12 |  |
| chr 3:40628201-40885900 | 993 | 0.5 | 6 | 9 |  |
| chr 3:96936201-97161800 | 1028 | 0.853 | 7 | 9 | EPHA6(0) |
| chr 3:127877401-127994800 | 793 | 0.459 | 5 | 16 | EEFSEC(0) |
| chr 3:188646901-188817800 | 1048 | 0.735 | 7 | 10 | TPRG1(0) |
| chr 4:33067701-33220500 | 1195 | 0.635 | 4 | 10 |  |
| chr 4:74535601-74698200 | 905 | 0.8 | 4 | 6 | IL8(15.5) |
| chr 4:85720501-85988100 | 1489 | 0.571 | 20 | 19 | WDFY3(0) |
| chr 5:82290501-82637400 | 1049 | 0.729 | 4 | 8 | XRCC4(0) |
| chr 5:130884801-131353100 | 722 | 0.647 | 11 | 15 | FNIP1(0),**ACSL6**(0) |
| chr 6:75884001-75992500 | 1068 | 0.382 | 9 | 15 | COL12A1(0),COX7A2(0),TMEM30A(0) |
| chr 7:99017401-99521900 | 857 | 0.835 | 5 | 10 | ZNF655(0),ZSCAN25(0),**CYP3A5(0),CYP3A7(0),CYP3A4**(0) |
| chr 7:113713801-113837100 | 940 | 0.624 | 7 | 8 | FOXP2(0) |
| chr 8:35540601-36280100 | 1020 | 0.771 | 6 | 12 | UNC5D(0) |
| chr 8:48632601-48808600 | 1033 | 0.753 | 6 | 6 | SPIDR(0) |
| chr 8:60344201-60563900 | 1331 | 0.382 | 12 | 17 |  |
| chr 8:85469601-85715500 | 152 | 0.682 | 3 | 5 | RALYL(0) |
| chr 9:90950701-91143000 | 883 | 0.647 | 5 | 7 | SPIN1(11.7) |
| chr 9:108003701-108201300 | 874 | 0.335 | 6 | 18 | **SLC44A1**(0) |
| chr 10:22570201-22671300 | 938 | 0.9 | 4 | 12 | COMMD3(18.4) |
| chr 10:60790301-61050700 | 268 | 0.465 | 7 | 7 | PHYHIPL(103.7) |
| chr 10:75106501-75402700 | 886 | 0.8 | 7 | 11 | PPP3CB(0),USP54(0) |
| chr 11:16009101-16210400 | 1247 | 0.371 | 12 | 14 | SOX6(0) |
| chr 11:46331901-46707100 | 1117 | 0.735 | 5 | 6 | **DGKZ**(0),MDK(0),CHRM4(0),AMBRA1(0) |
| chr 11:66431401-66616700 | 446 | 0.594 | 13 | 14 | C11orf80(0) |
| chr 11:66847201-67072800 | 992 | 0.841 | 6 | 16 | KDM2A(0),ADRBK1(0),ANKRD13D(0) |
| chr 12:79713601-80140500 | 972 | 0.588 | 5 | 12 | PAWR(32.3) |
| chr 12:88453201-89381300 | 352 | 0.871 | 4 | 4 | KITLG(136.0) |
| chr 14:45875601-46028500 | 982 | 0.5 | 6 | 6 |  |
| chr 14:76180001-76301100 | 735 | 0.741 | 6 | 6 | TTLL5(0) |
| chr 16:61780801-61991200 | 1177 | 0.394 | 18 | 9 | **CDH8**(0) |
| chr 17:40923401-41043600 | 968 | 0.7 | 7 | 4 | VPS25(0),WNK4(0),COA3(0),CNTD1(0),**BECN1**(0),PSME3(0),AOC2(0),AOC3(0) |
| chr 17:53924001-54065300 | 788 | 0.818 | 7 | 6 |  |
| chr 17:59121901-59238200 | 1044 | 0.594 | 11 | 7 | BCAS3(0) |
| chr 18:30406801-30649400 | 940 | 0.947 | 8 | 6 | CCDC178(0) |
| chr 19:42718701-42927600 | 1104 | 0.529 | 4 | 7 | **GSK3A**(0),ERF(0) |
| chr 20:33915001-34157500 | 955 | 0.559 | 5 | 11 | CEP250(0) |
| chr 20:37085001-37225400 | 1402 | 0.276 | 20 | 7 | RALGAPB(1.2) |
| chr 21:31941601-32162100 | 913 | 0.371 | 5 | 13 | **KRTAP21-3**(2.4) |
| chr 22:28215801-28316300 | 758 | 0.594 | 4 | 13 | **PITPNB**(0) |

**Supplementary Table 25.** **Classification of ancient selections (≥ 1,900 generations) in YRI.** aEA: Ust’-Ishim; Nean: Neanderthal.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Regions** | **Time (g)** | **Freq** | ***DaEA,MHG*** | ***DNean,MHG*** | **Groups** | **Genes** |
| chr1:20504201-20536600 | 2373 | 0.216 | 36 | 48 | aEA-like | UBXN10 |
| chr1:25541901-25585000 | 2489 | 0.148 | 34 | 97 | aEA-like | SYF2,C1orf63 |
| chr1:50564001-50668700 | 2556 | 0.136 | 80 | 61 | aYRI-rest | ELAVL4 |
| chr1:61536901-61638000 | 15643 | 0.438 | 5 | 49 | Nean-like | NFIA |
| chr1:68752401-68776900 | 3250 | 0.17 | 47 | 87 | aEA-like |  |
| chr1:118370201-118729400 | 2390 | 0.148 | 18 | 92 | aEA-like | GDAP2,WDR3,SPAG17 |
| chr1:218638201-218704900 | 2196 | 0.17 | 23 | 100 | aEA-like | C1orf143 |
| chr2:13689001-13716000 | 2666 | 0.307 | 21 | 203 | aEA-like |  |
| chr2:133117001-133151400 | 7167 | 0.21 | 76 | 67 | aYRI-rest |  |
| chr2:145231701-145303700 | 2780 | 0.097 | 41 | 43 | aEA-like | ZEB2 |
| chr2:203334601-203441100 | 2492 | 0.159 | 24 | 40 | aEA-like | BMPR2 |
| chr2:214171801-214344300 | 2345 | 0.295 | 37 | 110 | aEA-like | SPAG16 |
| chr3:43304401-43782100 | 1934 | 0.125 | 32 | 118 | aEA-like | SNRK,ANO10,ABHD5 |
| chr3:62355001-62433800 | 2490 | 0.034 | 63 | 71 | aYRI-rest | FEZF2,CADPS |
| chr3:94759501-95086100 | 2040 | 0.097 | 173 | 142 | aYRI-rest |  |
| chr4:26087901-26130100 | 2161 | 0.119 | 11 | 29 | aEA-like |  |
| chr4:73949801-74204200 | 2880 | 0.375 | 27 | 42 | aEA-like | ANKRD17 |
| chr4:90637801-90759100 | 2234 | 0.136 | 169 | 126 | aYRI-rest | SNCA |
| chr4:126686301-126722500 | 3212 | 0.199 | 84 | 58 | aYRI-rest |  |
| chr4:150440401-150634100 | 14970 | 0.386 | 5 | 64 | aEA-like |  |
| chr4:153145601-153277900 | 1975 | 0.403 | 38 | 72 | aEA-like | FBXW7 |
| chr5:60732301-60843700 | 8801 | 0.165 | 98 | 60 | aYRI-rest | ZSWIM6 |
| chr5:63280101-63492200 | 16601 | 0.159 | 61 | 43 | Nean-like | RNF180 |
| chr5:87769901-87901600 | 9128 | 0.33 | 31 | 116 | aEA-like |  |
| chr5:88437801-88743300 | 7199 | 0.091 | 19 | 27 | Nean-like |  |
| chr5:102454601-102560700 | 10552 | 0.222 | 3 | 114 | aEA-like | GIN1,PPIP5K2 |
| chr5:138866901-139053500 | 2013 | 0.409 | 37 | 60 | aEA-like | UBE2D2,CXXC5 |
| chr5:139657601-139801800 | 11246 | 0.097 | 75 | 59 | aYRI-rest | CYSTM1,PFDN1,HBEGF,SLC4A9,ANKHD1,ANKHD1-EIF4EBP3 |
| chr6:294901-376600 | 2490 | 0.227 | 8 | 24 | aEA-like | DUSP22 |
| chr6:18982901-19095900 | 7965 | 0.085 | 54 | 124 | aYRI-rest |  |
| chr6:126727901-126993900 | 13775 | 0.443 | 21 | 33 | Nean-like |  |
| chr7:65658001-65830300 | 5918 | 0.193 | 102 | 32 | Nean-like | TPST1 |
| chr7:65974201-66114700 | 4440 | 0.153 | 157 | 110 | aYRI-rest | KCTD7 |
| chr7:69197101-69343700 | 18027 | 0.21 | 33 | 39 | Nean-like | AUTS2 |
| chr7:73350901-73401900 | 2810 | 0.08 | 38 | 77 | aEA-like |  |
| chr7:85642201-85666800 | 2370 | 0.142 | 35 | 368 | aEA-like |  |
| chr7:124459101-124569400 | 2086 | 0.085 | 189 | 49 | aYRI-rest | POT1 |
| chr7:133828801-133859000 | 2694 | 0.477 | 446 | 450 | aYRI-rest | LRGUK |
| chr7:157369601-157404500 | 2181 | 0.256 | 60 | 118 | aYRI-rest | PTPRN2 |
| chr8:55533601-55733300 | 23219 | 0.17 | 37 | 41 | Nean-like | RP1 |
| chr9:91690601-91839100 | 7586 | 0.165 | 76 | 42 | Nean-like | SHC3 |
| chr9:128236201-128453200 | 4814 | 0.182 | 71 | 69 | aYRI-rest | MAPKAP1 |
| chr10:71671601-71704000 | 11090 | 0.131 | 78 | 130 | aYRI-rest | COL13A1 |
| chr10:75163101-75269400 | 3038 | 0.267 | 21 | 55 | aEA-like | ANXA7,MSS51,PPP3CB,USP54 |
| chr11:14149101-14185700 | 6354 | 0.091 | 21 | 157 | aEA-like | SPON1 |
| chr11:47242801-47462300 | 2988 | 0.136 | 108 | 119 | aYRI-rest | DDB2,ACP2,NR1H3,MADD,MYBPC3,SPI1,SLC39A13,PSMC3,RAPSN |
| chr11:47572101-47729900 | 7341 | 0.159 | 48 | 89 | aEA-like | CELF1,NDUFS3,PTPMT1,KBTBD4,FAM180B,C1QTNF4,MTCH2,AGBL2 |
| chr11:72620601-72857300 | 2520 | 0.085 | 41 | 26 | aEA-like | FCHSD2 |
| chr11:111553001-111719200 | 2381 | 0.136 | 43 | 124 | aEA-like | SIK2,PPP2R1B,ALG9 |
| chr11:121388101-121553600 | 3285 | 0.08 | 31 | 50 | aEA-like | SORL1 |
| chr12:76721501-76969400 | 2463 | 0.199 | 46 | 41 | aEA-like | BBS10,OSBPL8 |
| chr12:79058701-79225200 | 4364 | 0.21 | 7 | 139 | aEA-like |  |
| chr13:55839501-55871300 | 2218 | 0.131 | 558 | 65 | aYRI-rest |  |
| chr13:63622701-63775200 | 7876 | 0.142 | 9 | 34 | Nean-like |  |
| chr14:48737301-48772600 | 1931 | 0.273 | 302 | 68 | aYRI-rest |  |
| chr15:59160201-59225000 | 17355 | 0.074 | 15 | 40 | Nean-like | RNF111,SLTM |
| chr15:100104501-100219400 | 7522 | 0.216 | 31 | 37 | Nean-like | MEF2A |
| chr16:14602101-14719500 | 27154 | 0.511 | 16 | 40 | Nean-like | PARN |
| chr17:26871601-27088000 | 2051 | 0.216 | 19 | 39 | aEA-like | UNC119,PIGS,ALDOC,SPAG5,KIAA0100,SDF2,SUPT6H,PROCA1,RAB34,RPL23A,TLCD1,NEK8,TRAF4,FAM222B |
| chr17:44039201-44079000 | 3078 | 0.08 | 120 | 73 | aYRI-rest | MAPT,STH |
| chr17:47657301-47828200 | 1911 | 0.097 | 40 | 36 | aEA-like | NXPH3,SPOP,SLC35B1,FAM117A |
| chr17:53272101-53357100 | 7270 | 0.068 | 106 | 54 | aYRI-rest | HLF |
| chr19:9576201-9606400 | 2616 | 0.102 | 22 | 355 | aEA-like | ZNF560 |
| chr19:19550001-19753800 | 19289 | 0.432 | 46 | 71 | aEA-like | GATAD2A,TSSK6,NDUFA13,YJEFN3,CILP2,PBX4,LPAR2,GMIP |
| chr19:37794701-38145500 | 2138 | 0.301 | 66 | 81 | aYRI-rest | HKR1,ZNF527,ZNF569,ZNF570,ZNF793,ZNF540,ZNF571,ZFP30 |
| chr20:30700001-31028100 | 18469 | 0.085 | 142 | 112 | aYRI-rest | TM9SF4,PLAGL2,POFUT1,KIF3B,ASXL1 |
| chr20:33430001-33557900 | 2665 | 0.119 | 102 | 58 | aYRI-rest | GGT7,ACSS2,GSS |
| chr20:43108101-43245000 | 13792 | 0.142 | 42 | 42 | Nean-like | TTPAL,SERINC3,PKIG |
| chr21:34943601-35034500 | 2965 | 0.108 | 58 | 59 | aYRI-rest | SON,DONSON,CRYZL1,ITSN1 |

**Supplementary Table 26.** **Human specific expressed genes of ancient selection in YRI.** PFC: prefrontal cortex; VC: visual cortex; CBC: cerebellum cortext; K: kidney; M: muscle.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Regions** | **Time (g)** | **Genes** | **Tissues** | **H v.s. C** | **H v.s. R** | **C v.s. R** |
| **chr16:14602101-14719500** | **27154** | **PARN** | **PFC** | **0.0423** | **0.0198** | **0.4369** |
| chr19:19550001-19753800 | 19289 | LPAR2 | VC | 0.0145 | 0.0037 | 0.9481 |
|  |  | GATAD2A | VC | 0.0168 | 0.0082 | 0.5731 |
| chr20:30700001-31028100 | 18469 | TM9SF4 | VC | 0.0217 | 0.0047 | 0.4874 |
| **chr7:69197101-69343700** | **18027** | **AUTS2** | **CBC** | **0.0421** | **0.0003** | **0.0752** |
| chr15:59160201-59225000 | 17355 | RNF111 | CBC | 0.0038 | 0.0342 | 0.8904 |
|  |  |  | VC | 0.0334 | 0.0023 | 0.0836 |
| chr5:139657601-139801800 | 11246 | CYSTM1 | CBC | 0.0422 | 0.0027 | 0.0596 |
| chr11:47572101-47729900 | 7341 | KBTBD4 | CBC | 0.0402 | 0.0020 | 1 |
| chr9:128236201-128453200 | 4814 | MAPKAP1 | CBC | 0.0016 | 0.0003 | 0.8667 |
|  |  |  | PFC | 0.0056 | 0.0009 | 0.1269 |
|  |  |  | VC | 0.0049 | 0.0005 | 0.1625 |
| **chr11:121388101-121553600** | **3285** | **SORL1** | **CBC** | **0.0083** | **0.0452** | **0.5842** |
| chr10:75163101-75269400 | 3038 | ANXA7 | K | 0.0380 | 0.0230 | 0.1321 |
| chr11:47242801-47462300 | 2988 | MADD | CBC | 0.0128 | 0.0046 | 0.6297 |
|  |  | ACP2 | VC | 0.0175 | 0.0140 | 0.2871 |
| chr20:33430001-33557900 | 2665 | GSS | PFC | 0.0072 | 0.0057 | 0.0686 |
|  |  |  | VC | 0.0415 | 0.0056 | 0.1806 |
| chr1:118370201-118729400 | 2390 | GDAP2 | K | 0.0384 | 0.0014 | 0.0586 |
| chr2:214171801-214344300 | 2345 | SPAG16 | VC | 0.0162 | 0.0021 | 0.0909 |
| **chr4:90637801-90759100** | **2234** | **SNCA** | **CBC** | **0.0330** | **0.0419** | **0.7082** |
| chr17:26871601-27088000 | 2051 | SPAG5 | VC | 0.0303 | 0.0035 | 0.2968 |
|  |  | PIGS | CBC | 0.0038 | 0.0061 | 0.7585 |
|  |  | TRAF4 | PFC | 0.0403 | 0.0066 | 0.1223 |
| chr17:47657301-47828200 | 1911 | NXPH3 | CBC | 0.0141 | 0.0288 | 0.3773 |

**Supplementary Note**

# 1. Empirical data processing

## 1.1 Time interval division

To increase the resolution for more recent history, and considering the computational cost, we allocated 30 time intervals as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

## 1.2 Comparisons of 1000G phase I data with the high-quality genome sequencing data

To evaluate the inferences of PSMC on the imputed genomes, we compared the *N*e and TMRCAs estimated from individuals who were sequenced by both 1000G and Complete Genomics (CG). Six CEU (NA06994, NA07357, NA10851, NA12004, NA12889, and NA12890), 4 CHB (NA18526, NA18537, NA18555, and NA18558), and 7 YRI individuals (NA18501, NA18502, NA18504, NA18505, NA18508, NA18517, and NA19129) were analyzed. We merged the high-quality variants (ftp://ftp2.completegenomics.com/Multigenome\_summaries/Complete\_Public\_Genomes\_69genomes\_VQHIGH\_testvariants.tsv.bz2) and reference genome into diploid genomes, as was done with the 1000G data. Supplementary Figure 16 shows that the *N*e estimated by the two datasets were similar along the entire history, with a small underestimation by 1000G in recent history because less rare variants were called, while a minor overestimation in ancient history was observed because more common variants are imputed in 1000G than that in CG.

The accuracy of local TMRCA is another factor for the reconstruction of coalescent trees. We compared the TMRCAs estimated by 1000G and CG data in all individuals. Supplementary Figure 17 shows the proportion of TMRCAs in CG estimated into other TMRCAs in 1000G. The correlation of all TMRCAs was 0.703. These analyses demonstrate that the overall coalescent information is present in the haploid genome sequences of 1000G, and indicate that it is feasible to reconstruct coalescent trees from these data.

# 2. Coalescent-based model

## 2.1 Hypothesis *H*0: Neutral case

Following the time scaling described in equation (1) and (2), Griffiths and Tavaré[3](#_ENREF_3),[4](#_ENREF_4) showed the number of ancestors *n*(*t*) at time *t* is a Markovian random variable, and there is a non-homogeneous Markovian process of *n*(*t*) with transition probability from state of *i* ancestors to *j* ancestors. It is convenient to define *τ*(*s*) = *Ʌ*(*t* + *s*) – *Ʌ*(*t*) and define *δτ* = *τ*(*δt*), where *τ* can be considered as rescaled coalescent time; and then the transition probability is:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

and with density function

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

Here, we use *τi* (*i* = *n*, *n* – 1,…, 2) to represent the rescaled waiting time to go from state *i* to state *i*–1, then the joint density function for (*τn, τn*–1,…, *τ*2) is equation (3). This defines the tree pattern for neutral evolution. Notably, the trees defined by this function are neither too compressed nor too enlarged on average. Mathematically, the coalescent rate is constant  in the *i*th coalescent interval, *i* = *n, n* – 1,…, 2.

## 2.2 Hypothesis *H*1

Selection events change the neutral tree pattern. For example, a tree that becomes extremely compressed indicates either PS or NS. Under strong PS or NS, any two individuals would find their common ancestor faster than usual, and the haplotype heterozygosity drops. In contrast, a tree that strongly elongates may indicate BS, as any pair of individuals took a longer time to find their common ancestor and the local heterozygosity increases.

As mentioned in Materials and Methods, the rescaled coalescent rate from *k* to *k*-1 can be written as a function of *α*: . Therefore the density function of *τi*, *i* = *n*, *n* – 1,…, 2, can be written as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

and the joint density function of (*τn*, *τn*–1,…, *τ*2) is given by equation (4) to define the *H*1 model with *α* ≠ 1.

To test whether the tree state is neutral, we can construct log-likelihood ratio test, with *H*0 as the null hypothesis and *H*1 as the alternative hypothesis. Then, we rewrite equations (3) and (4) as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

and

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

Combining equations (S5) and (S6), we compute for the log-likelihood ratio by using equation (5). We can then calculate the derivative of equation (5) and solve for it being equal to zero. It is straightforward to calculate the optimal solution that maximizes *D*10. The value of the optimized solution of *α* is as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

and the corresponding log-likelihood ratio is:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

By using equation (S7), the expectation of *α*\* should be 1, which indicates that the rate of the coalescent process is neutral on average. It is easy to verify this conclusion. Because *τi* (*i* = *n*, *n* – 1,…, 2) are exponential random variables with parameters , *i* = *n*, *n* – 1,…, 2, then the cumulative distribution of *α*\* is the convolution of such exponential random variables , *i* = *n*, *n* – 1,…, 2, with parameters 1, i.e.,

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

According to the central limit theorem, as *n* grows, the distribution of *α*\* will be asymptotically close to the normal distribution . In addition, the median of *α*\* is also approximately 1 based on the properties of the Gamma distributions.

Next, to validate the robustness of our statistic *α*\* under variable demographic histories, we conducted simulations on 10 demographic models (Supplementary Fig. 5 and Supplementary Note 3.1). In these simulations, we assume that the exact demographic trajectory and coalescent trees are all known. Supplementary Figure 18a,b show the fitting of theoretical and simulated density curves and the quantile-quantile plot. Similar to the test statistic *D*10, by Wilks’ theorem, the distribution of 2*D*10 can be asymptotically close to *χ*2(1). Simulation results demonstrate that the probability density of *D*10 is asymptotically close to *Γ*(0.5, 1), as shown in Supplementary Figure 18c,d.

Under this framework, if *α*\* is substantially smaller or larger than 1, the corresponding *D*10 will be associated with a significant *P* value. To simplify the classification, when *α*\* is significantly greater than 1, we assign the tree to the mode of BS; when *α*\* is significantly smaller than 1, we assign the tree to a wobble mode of NS/PS.

## 2.3 Hypothesis *H*2: Identify PS

*H*1 can be used to distinguish whether the tree is compressed or elongated, and BS can be identified as elongated trees. For the case of PS, the situation is more complicated. It is easy to understand that recent strong PS can also induce highly compressed coalescent trees, thus very small *α* as NS does; therefore, *H*1alone cannot distinguish PS from NS. Here, we propose a third hypothesis, *H*2, which is sensitive to the coalescent patterns induced by PS. The original idea of *H*2 is based on the unique PS property of strong time dependence compared to other modes of selection. For PS, we are looking at a state change in the population in which a newly arisen mutation goes to fixation (hard sweep), or through an environmental change in which pre-existing neutral mutation becomes positively selected (soft sweep). In the case of NS, the state of the population stays the same. Newly arising deleterious mutations are constantly eliminated from the population. Because we are looking at the surviving lineages and average over many elimination events, there will not be a time signature in NS.

In view of this distinction between PS and NS, a straightforward strategy is to propose a hypothesis that features more parameters that reflect such time-dependent changes at a coalescent rate. We simplify the PS process by introducing three consecutive time intervals, within each of which the coalescent rates remain constant. Therefore, we divided the whole tree into three parts by two discrete time parameters, *τi*1 and *τi*2, where 2 < *i*1< *i*2 < *n* with respect to the time of ancestor states, *n – i*1 and *n* – *i*2. The total coalescent rate then can be change rate *α* wa divided into three parts: *α*1, *α*2, and *α*3. Analogous to *H*1, it is natural to get the following results:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

and equations (6–7).

Similarly, given *τi*1 and *τi*2, the best estimators *α*\**i* (*i* = 1, 2, 3) with corresponding partition triplet [*u*, *n* – (*u* + *v*) – 1, *v*] (*u*, *v* ≥ 2) can be written as:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |
|  |  | (S ) |
|  |  | (S ) |

and given the partition and estimators of *αi* (*i* *=* 1, 2, 3), the formula of the *D*20 (equation (6)) and *D*21 (equation (7)) should be as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

And

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

Therefore, *D*21 and *D*20 signal the local coalescent rate changes, which can be used to detect PS.

**2.4 Estimating the selection starting time**

Apart from detecting signals of PS, we may use further information generated from the model to test *H*2 to infer the selection starting time and selection coefficients.

The essential parameters for inferring the selection time are two time parameters,  and , 2 < *i*1< *i*2 < *n*, indicating at which time point the local coalescent rate significantly changed. For convenience, we define *T*0 = 0 and  as the current and most ancient time points, respectively. Three parameters, *α*1*, α*2,and *α*3, measure the local average coalescent rate change in three disjoint time intervals, [*T*0, *T*1], [*T*1, *T*2], and [*T*2,*Te*].

To clearly elucidate the dynamics of coalescent rate change, we considered four cases by comparing the values of *α*1, *α*2,and *α*3. Because we can only obtain the estimators of *αi*, we then use these estimators *α*\**i* instead of *αi*.

All four cases are shown in Supplementary Figure 19. A further discussion for each case is as presented below.

At the beginning, we use *ai* (*i* = *n*, *n* – 1,…, 2) to measure the coalescent rate at the relative coalescent time interval *τi*:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

Analogous to our previous discussion, the value of *ai* (*i* = *n*, *n* – 1,…, 2) indicates whether the coalescent rate accelerated (*ai* < 1) or decelerated (*ai* > 1). *α*\* should be considered as the average of *ai* (*i* = *n*, *n* – 1,…, 2) and *αi* (*i* = 1, 2, 3), should be considered as the local average of *ai* (*i* = *n*, *n* – 1,…, 2).

In general, we must locate a coalescent time region [*rs*, *re*] where  and  in which we can search for the selection starting time point. After that, we will seek the starting time point step by step from the leftmost bound of the region. Once there is an interval *τi*\* satisfying the condition that the corresponding *ai*\*< *M*and *ai*\*> *M*, where *M* is a restriction to *ai* (*i* = *s*, *s* – 1,…, *e*) or *ai*\* = *ae*, which indicates that the seeking process arrived at the rightmost point of the searching region, then we will consider the result time  as the selection starting time. We commonly use *M* = 1.

In practice, for recent selection, we need to add another restriction to the absolute time interval to avoid the case of a single interval with extreme large absolute time value. The criteria we use here is that *Ti*\*– *Ti*\*+1>1,000 then we go back from *i*\*+ 1 to *n* to find the first interval *τi*\*\*<300, *i*\* ≤ *i*\*\*≤ *n*, then the resulting selection starting time  will be the ultimate selection time.

Case 1: *α*\*1 > *α*\*2 > *α*\*3

In this case, the selection starting time should be more ancient than *T*2. We then choose the time region that ranges from *T*2 to the most ancient time *T*e as our searching region. As a result, we can get a time estimation within the time interval [*T*2, *Te*] by using the above algorithm.

Case 2: *α*\*1 > *α*\*2 < *α*\*3

In this case, it is natural to consider that the selection starting time is located within the time interval [*T*1, *T*2]. There may be an exception that the difference between *α*\*2and *α*\*3 is small.

The reason for this phenomenon is that the selection event occurs in the consecutive time interval [*T*1, *Te*], but *H*2 is forced to divide [*T*1, *Te*] into two separate intervals, [*T*1, *T*2] and [*T*2, *Te*].

To determine the right interval for estimating time, we set a criterion by introducing a bias related to the average *α* value, as well as the number of coalescent events in the time interval. Because the *α* value reflects the intensity of coalescent rate, the more coalescent events occur in a specific interval; therefore, it is more likely that such an interval contains the starting time. In terms of limitation of the total coalescent events, the bias should be neither too large nor too small. Here, we define the bias as , where *Ni* (*i* = 2, 3) is the number of coalescent events in [*T*1, *T*2] and [*T*2, *Te*].

The algorithm for searching time intervals is as follows: if *α*\*3 > 1, then we just consider *T*2 as the selection starting time; otherwise, we will check whether bias is < 1. If the *bias* < 1, then we search the selection starting time at [*T*2, *Te*]; otherwise, we still consider *T*2 as the selection starting time.

Case 3: *α*\*1 < *α*\*2 > *α*\*3

In this case, if *α*\**i* (*i* = 1, 2, 3) are all < 1, then we consider the selection starting time located in [*T*2, *T*3]. But if *α*\*2 > 1, then the case is more complex because we must judge whether the selection starting time is within [0, *T*1] corresponds to *α*\*1,or [*T*2, *Te*] that represents *α*\*3.

If only one of *α*\*1 and *α*\*3 is < 1, then the selection starting time must be located within one of the relative intervals. However, when both of these two values are > 1, we should again construct a bias to decide which time interval is what we want. Analogous to case 2, the bias used here is , where *Ni* (*i* = 1, 3) is the number of coalescent events in [0, *T*1] and [*T*2, *Te*]. If *bias* > 1, then we search for the selection starting time in [0, *T*1], and if *bias* < 1, then we search for the selection starting time in [*T*2, *Te*].

Case 4: *α*\*1 < *α*\*2 < *α*\*3

For the last case, if *α*\*2 > 1, then we just need to focus on the time interval [0, *T*1]. However, if *α*\*2 < 1 and *α*\*3 < 1, then selection starting time may fall in interval [*T*1, *T*2] or [*T*2, *Te*], and then we can use the same strategy as in Case 2 to determine the interval. If *α*\*2 < 1 and *α*\*3 > 1, then we just need to search the selection starting time within interval [*T*1, *T*2].

Going through all these four possible cases, we can then determine a time interval, within which the selection starting time can be inferred using the algorithm introduced at the beginning of this section. In addition, selection coefficient is also estimated within this time interval, following section 2.5.

## 2.5 Estimate selection coefficient

Slatkin and Rannala[5](#_ENREF_5) described the change in the mean number of ancestors *k* under the rescaled coalescent time τ, ignoring the variance of ancestor number by using the following equation:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

As mentioned above, we use the time interval where the selection starting time is inferred, to also estimate the selection coefficient. Analogous to equation (S17) and considering that the population size change rate *λ* is independent from the selection, and the change of ancestor number during selection only happens in individuals carrying the advantage allele. The allele frequency of the advantageous allele, say *x*, is a function of selection coefficient and coalescent time as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

Note that *t* is the absolute time scale and expressed in unit of generations.

Equation (S18) provides an approximate equation that connect to our *H*2 model. Supposing , then . We can get the discretization form , (*k* = *n*, *n* – 1,…, 2) as mentioned in section 2.4, then

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

For the *k*th time interval from one coalescent event to the next, *δk* = 1 and *δa* = *ak*, *k* = *n*, *n* – 1,…, 2, then we use the following equation:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

where *x*(*s*, *tk*) is the frequency of the selected allele in the *k*th coalescent time interval. We also assume that during the selection phase, every coalescent event occurs in the haplotype carrying the advantageous allele, thus the coalescent rate of the tree is equivalent to the coalescent rate of the selected allele.

To infer the selection coefficient, we introduced two models. First we applied a deterministic model. Supposing there are two alleles, A and B, with allele frequencies *x* and 1 – *x*. Then following the *Hardy-Weinberg* law, the genotypes AA, AB, and BB have frequencies *x*2, 2*x*(1 – *x*), and (1 – *x*)2. We assume A is the advantageous allele, and the fitness of AA, AB, and BB are 1 + 2s, 1 + s, and 1, respectively. Then under such model, *x* has the following expression:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

where *t*mid is the half-time when the allele frequency of A achieves 0.5.

The second model is an exponential model. Campbell[6](#_ENREF_6) pointed out that when selection is strong, the differential form of the equation governing the coalescent will be as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

So we simply assumed that

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

For the logistic model, we can get the following approximation of *s* by using equations (S20) and (S21):

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

And for an exponential model, we can generate the following approximation by using equations (S20) and (S23):

|  |  |  |
| --- | --- | --- |
|  |  | (S ) |

Equations (S24) and (S25) both show a simple relationship between *s* and *a*. Based on these findings, we can use linear regression to fit a straight line that has a slope equal to *s*. We found in most cases, the two models gave very similar estimates. However, in extreme cases, the estimates may differ by a magnitude. Simulations showed that the smaller value is always a better estimate. Therefore, we applied both models to data and choose the smaller value as the final estimate. In addition, to reduce the noise and to improve the accuracy of the prediction of the selection coefficient, we only used the consective subintervals which compose the biggest slope (). Supplementary Figure 20 shows the performance of selection coefficient estimation in simulations on variable scenarios.

The pseudo code of the algorithm to allocate candidate time intervals for our linear square estimation is as follows:

*len ← length\_of\_selection\_time\_interval*

*la ← log(a)*

*do j from 2 to len-2*

*range ← time\_beg:(time\_end-j)*

*do i in range*

*dla ← la[i+j]-la[i]*

*dt ← t[i+j]-t[i]*

*abs\_slope[i] ← dla/dt*

*end do*

*s[j] ← index of max(abs\_slope[i]);*

*end do*

*s\_range[1] ← max(s[j])*

*s\_range[2] ← s\_range[1]+index of max(abs\_slope[i])*

The final step is using the linear square method to fit the logarithm log(*a*) (for the exponential curve) or (for the logistic curve) and the absolute time *t*. The slope from the linear square is the estimate of the selection coefficient *s*.

# 3. Simulations

## 3.1 Neutral simulation

To define the thresholds of the models, we performed neutral simulations based on the demography inferred from mega-genomes (Supplementary Fig. 3) and the core model (Supplementary Fig. 15). The command line for simulations based on CEU demography is:

msms -N 10000 -ms 30 100 -t 100000 -r 52000 100000000 -eN 0 31992.4263670592 -eN \ 0.000217705365 937032.677745933 -eN 0.00087928974 30375.3116123553 -eN \ 0.002009452275 9.12237206427 -eN 0.003650887215 1.54691976123 -eN 0.005868165195 \ 0.81437755104 -eN 0.008750203155 0.50508914922 -eN 0.012417674085 0.27262145511 \

-eN 0.01703041677 0.16742636388 -eN 0.022799785365 0.14780535912 -eN 0.03000382173\ 0.16885326906 -eN 0.0390105432 0.22700777058 -eN 0.05030652303 0.36921771369 -eN \ 0.06454099602 0.67861126332 -eN 0.082584430785 1.12489314777 -eN 0.10561285971 \ 1.4983845201 -eN 0.135226082025 1.68978740742 -eN 0.173618440785 1.61094701493 \

-eN 0.223824755835 1.29181141485 -eN 0.290077050105 0.99000332703 -eN \ 0.378329230125 0.76061102994 -eN 0.65828033442 0.75980795472 -eN 1.186401441015 \ 1.37679060465 -eN 2.22525580545 1.62494861022 -eN 4.35856514373 2.83136165541 \

-oFP .00000000

The command line for simulations based on CHB demography is:

msms -N 10000 -ms 30 100 -t 100000 -r 52000 100000000 -eN 0 20537.7602903039 -eN \ 0.00019734936 411299.837605472 -eN 0.0007969878 2908.02698340927 -eN \ 0.001822002665 165.49303401968 -eN 0.003310978285 9.38693812165 -eN \ 0.005323688745 1.22404359215 -eN 0.007941730415 0.36658592415 -eN 0.011275796045\ 0.16308837255 -eN 0.015473265125 0.11556006835 -eN 0.02072927316 0.12029898311 \

-eN 0.027300311065 0.14845605606 -eN 0.035526996245 0.2014102026 -eN \ 0.045860638855 0.31748135519 -eN 0.058904039985 0.57772696827 -eN 0.07546778683 \ 1.00424004968 -eN 0.09664868641 1.39055015682 -eN 0.123942672175 1.63869736859 \

-eN 0.159407364215 1.6647430564 -eN 0.2058948403 1.38610347092 -eN 0.26739320208 \ 1.0610184726 -eN 0.34952912017 0.78468762162 -eN 0.61124346791 0.74231000921 -eN \ 1.108006279915 1.28028183445 -eN 2.091753622655 1.65223351059 -eN 4.12663331036 \ 2.67817250192 -oFP .00000000

The command line for simulations based on YRI demography is:

msms -N 10000 -ms 30 100 -t 100000 -r 52000 100000000 -eN 0 51565.3305880467 -eN \ 0.00032229121 228347.806087211 -eN 0.00130126451 1134.76714946095 -eN \ 0.00297101897 4.64512601129 -eN 0.005392602925 0.8730901878 -eN 0.008655113945 \ 0.59044079663 -eN 0.01288284864 0.67243256043 -eN 0.018242452465 0.72654778452 \

-eN 0.02495501939 0.63147407751 -eN 0.03330929154 0.55386679413 -eN 0.04368310861 \ 0.60528159187 -eN 0.056573107055 0.83921001183 -eN 0.07262991913 1.21296111834 \

-eN 0.092714821345 1.53237150691 -eN 0.11797068253 1.69328501824 -eN \ 0.14992701097 1.71967659845 -eN 0.19064240052 1.527088351 -eN 0.242910224995 \ 1.18271854316 -eN 0.310548480265 0.912535112 -eN 0.398818872825 0.7706829808 -eN \ 0.515030703325 0.71566138143 -eN 0.876456096 0.92831418165 -eN 1.539915651155 \ 1.68208292376 -eN 2.80669995155 1.54266282623 -eN 5.325485505525 4.39060505296 \

-oFP .00000000

The command line for simulations based on the core model is:

msms -N 10000 -ms 30 100 -T -t 10000 -r 5200 10000000 -eN 0 2 -eN 0.01 0.25 -eN 0.05 1 \

-eN 0.1 2 -eN 0.5 1 -eN 1 2 -oFP .0000000

We also performed simulations under 10 different demographic scenarios of population size change over time (Supplementary Fig. 5):

demo 1: msms -N 10000 -ms 30 1 -T -t 10000 -r 5200 10000000 -eN 0 2 -eN 0.01 1

demo 2: msms -N 10000 -ms 30 1 -T -t 10000 -r 5200 10000000 -eN 0 2 -eN 0.01 0.5

demo 3: msms -N 10000 -ms 30 1 -T -t 10000 -r 5200 10000000 -eN 0 2 -eN 0.01 0.25

demo 4: msms -N 10000 -ms 30 1 -T -t 10000 -r 5200 10000000 -eN 0 2 -eN 0.01 1 -eN \

0.05 2

demo 5: msms -N 10000 -ms 30 1 -T -t 10000 -r 5200 10000000 -eN 0 2 -eN 0.01 0.5 -eN \

0.05 2

demo 6: msms -N 10000 -ms 30 1 -T -t 10000 -r 5200 10000000 -eN 0 2 -eN 0.01 0.25 -eN \

0.05 2

demo 7: msms -N 10000 -ms 30 1 -T -t 10000 -r 5200 10000000 -eN 0 2 -eN 0.01 0.25 -eN \

0.05 1 -eN 0.1 2

demo 8 msms -N 10000 -ms 30 1 -T -t 10000 -r 5200 10000000 -eN 0 2 -eN 0.01 1 -eN 0.05 \

0.25 -eN 0.1 2

demo 9: msms -N 10000 -ms 30 1 -T -t 10000 -r 5200 10000000 -eN 0 2 -eN 0.01 0.25 -eN \

0.05 2 -eN 0.5 1 -eN 1 2

demo 10: msms -N 10000 -ms 30 1 -T -t 10000 -r 5200 10000000 -eN 0 2 -eN 0.01 1 -eN \

0.05 0.25 -eN 0.1 2 -eN 0.5 1 -eN 1 2

## 3.2 Simulation of PS

The command line for hard sweep simulations is:

msms -N 10000 -ms 30 <repeat> -t 2000 -r 1040 2000000 <demography> -SAA SAA -SaA \

SAa -Sp 0.5 -Smark -oOC -oTrace -SFC

The command line for soft sweep simulations is:

msms -N 10000 -ms 30 1 -t 2000 -r 1040 2000000 <demography> -SI time 1 initial-frequency \

-SAA 1000 -SaA 500 -Sp 0.5 -Smark -oOC -oTrace -SFC

The demography includes realistic population trajectories, the constant size model and the core model.

## 3.3 Simulation of NS

For expediency, we simulated an ancestral population with effective population size of 10,000, but scaled it to 2,000 as done previously[7](#_ENREF_7). 15 diploid individuals were sampled in each simulation. The command line for the constant model is:

sfs\_code 1 100 -n 15 -N 2000 -TE 3 -t 0.001 -r 0.00052 -L 200000 -W 1 2Ns 0 prop -Z -A

The command line for the core model is:

sfs\_code 1 100 -n 15 -N 2000 -Td 0 2 -Td 1 0.5 -Td 2 2 -Td 2.8 0.5 -Td 2.9 0.25 -Td 2.98 8 \

-TE 3 -t 0.001 -r 0.00052 -L 200000 -W 1 2Ns 0 prop -Z -A

## 3.4 Simulation of BS

The command line for the constant model is:

msms -N 10000 -ms 30 100 -T -t mrate -r 104 200000

The command line for the core model is:

msms -N 10000 -ms 30 100 -T -t mrate -r 104 200000 -eN 0 2 -eN 0.01 0.25 -eN 0.05 1 -eN \

0.1 2 -eN 0.5 1 -eN 1 2

# 4. Functional annotation

## 4.1 Functional annotation of NS

To study the selection patterns in CDS, we took all trees within 500 kb of any CDS and averaged the *α* values from the centers of CDS to the distal ends using a 2-kb sliding window. Figure 2c shows mean *α* as a function of distance from CDS. The elevated *α* score from the center toward the distal ends indicates that purifying selection pressure is highest inthe vicinity of the CDS. A similar was observed for TSS (Supplementary Fig. 21).

To investigate the distribution of selection pressures in essential regions or functional elements, we mapped the tree scaling coefficient *α* along the whole genome. Based on GENCODE v7[8](#_ENREF_8), we annotated *α* in the UTRs, CDS, intron and intergenic regions. We also annotated *α* the TSS, promoters regions (PF), transcribed regions (T), enhancers regions(E), weaker enhancers regions(WE), CTCF active regions (CTCF), and repressed regions (R) according to the ENCODE[9](#_ENREF_9) data.

To have a general view of selection pressures in different functional elements, we calculated the median *α* values in the CDS, UTRs, intron, intergenic, PF and E regions. Figure 2d and Supplementary Figure 22 show similar trends of selection pressure in different categories in all populations. When we compared the selection pressure only in functional elements, the PF regions always showed the highest pressure, then the transcript body (T and TSS), followed by enhancers (E and WE) and CTCF regions, whereas R regions are the most similar to neutrality (Supplementary Fig. 23).

## 4.2 Additional analysis of BS

An increase of in intermediate frequency alleles and the existence of long-term trans-species polymorphisms are commonly used signatures of BS. Nonetheless, the BS test developed in this study is sensitive to a slightly different signature: the increased heterozygosity that is captured as the largely decelerated coalescent rate. We compared the TMRCAs of BS regions in this study and of trans-species haplotypes in a previous study[2](#_ENREF_2). The overall distribution of TMRCAs in our BS signals in YRI are clearly shifted to older ages than the whole genome distribution, but less so than long-term trans-species haplotypes (Supplementary Fig. 24).

# 5. Divergence in gene expression

To further characterize the selection events, we compared the gene expression pattern in humans with that in the chimpanzee and rhesus macaque, utilizing the published RNA-seq data of the prefrontal cortex, visual cortex, cerebellum, kidney, and muscle from six young adults of each species[10](#_ENREF_10). Reads were mapped to the human-chimpanzee-rhesus consensus genome with STAR[11](#_ENREF_11), as described previously[12](#_ENREF_12). Briefly, chained and netted alignment files of for human (hg19), chimpanzee (panTro3), and rhesus (rheMac3) were downloaded from the UCSC Genome Browser, which were earlier aligned by using BLASTZ[13](#_ENREF_13). The human-chimpanzee-rhesus consensus genome was constructed from these alignments. The sites showing discordances among the three genomes or located in a 6-bp flanking regions of insertion/deletion sites were masked as “N”s. Only uniquely mapped reads were used for gene expression estimation, which is based on GENCODE (v.14) annotation. The expression levels were normalized as the number of reads per kilobase per million mapped reads (RPKM).

Only genes whose selection time was ≥ 1,900 generations were used for further between-species comparisons. Pairwise t-test between species was conducted for each ancient gene in different tissues separately, and a *P* value < 0.05 after Benjamini-Hochberg correction was used as a cutoff to define a significant expression difference between species. The genes with significant differences between human and chimpanzee as well as between human and rhesus but without significant differences between chimpanzee and rhesus were defined to exhibit human-specific expression profiles in a particular tissue. Supplementary Table 26 lists the corrected *P* values for all ancient genes with human-specific expression based on the pairwise between-species comparisons.

**References**

1. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-6 (2011).

2. Leffler, E.M. *et al.* Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**, 1578-82 (2013).

3. Griffiths, R.C. & Tavare, S. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci* **344**, 403-10 (1994).

4. Griffiths, R.C. & Tavaré, S. The ages of mutations in gene trees. *Ann Appl Prob* **9**, 567-590 (1999).

5. Slatkin, M. & Rannala, B. Estimating allele age. *Annu Rev Genomics Hum Genet* **1**, 225-49 (2000).

6. Campbell, R.B. Coalescent size versus coalescent time with strong selection. *Bull Math Biol* **69**, 2249-59 (2007).

7. Lohmueller, K.E. *et al.* Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet* **7**, e1002326 (2011).

8. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-74 (2012).

9. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

10. Bozek, K. *et al.* Exceptional evolutionary divergence of human muscle and brain metabolomes parallels human cognitive and physical uniqueness. *PLoS Biol* **12**, e1001871 (2014).

11. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

12. He, Z., Bammann, H., Han, D., Xie, G. & Khaitovich, P. Conserved expression of lincRNA during human and macaque prefrontal cortex development and maturation. *RNA* **20**, 1103-11 (2014).

13. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res* **13**, 103-7 (2003).