Supplementary information for: Signals of recent positive selection in a worldwide sample of human populations

Joseph K. Pickrell[1,†], Graham Coop[1,12,†], John Novembre[1,2],
Sridhar Kudaravalli[1], Jun Li[10], Devin Absher[4], Balaji S. Srinivasan[6,7,8,9],
Gregory S. Barsh[3], Richard M. Myers[4], Marcus W. Feldman[5],
and Jonathan K. Pritchard[1,11,†]

[1] Department of Human Genetics, The University of Chicago.

[2] Department of Ecology and Evolutionary Biology, UCLA.

[3] Department of Genetics, Stanford University.

[4] HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, AL.

[5] Department of Biological Sciences, Stanford University.

[6] Stanford Genome Technology Center, Stanford University

[7] Program in Biomedical Informatics, Stanford University

[8] Dept. of Computer Science, Stanford University

[9] Dept. of Statistics, Stanford University

[10] Dept. of Human Genetics, University of Michigan

[11] Howard Hughes Medical Institute.

[12] Current Address: Department of Ecology and Evolution, University of California, Davis

† To whom correspondence should be addressed: pickrell@uchicago.edu,
gmcoop@ucdavis.edu, pritch@uchicago.edu

January 12, 2009

Power simulations. To estimate our power to detect selection in the ascertained Illumina data, we performed simulations under the "cosi" model of human demography [3] with a slight modification– for computational efficiency, the recent increases in population size were dropped. This had little effect on the fit of the model to the HapMap data (results not shown). Other relevant parameters are described in the main text. The selection model is one where a single selected site arises and sweeps up in frequency. Implicitly, this assumes a given selected locus in the human genome has experienced only a single selective sweep in the time from over which our statistics are applicable (approximately 10-80ky); we find this reasonable, and simulations of this kind are a standard method for testing the power of a test for selection [2, 4–7]

To approximate the ascertainment of the Illumina panel, we performed a two-stage ascertainment on all of the simulations; this is described in the main text. In Supplementary Figure 1, we show that this procedure provides a good approximation to the frequency spectrum of the Illumina chip. For calculations of power, in all cases the selected site was excluded from analysis, making the calculations conservative. In Supplementary Figure 2, we show power estimates as a function of derived allele frequency in different demographies, and in Supplementary Figure 3, we show the effect of sample size on power.

About 200 of the simulations have a selected allele that arises on the branch common to both the "European" and "Asian" populations; this allows us to get an upper bound on the expected amount of overlap between the populations for XP-EHH (with a selection coefficient of 1% almost all selected alleles that arose during that branch would have gone to fixation and are not likely to be detected by iHS). In these simulations of truly shared selective sweeps, 76% are detected in both populations.

Overlap of haplotype-based selection signals between populations. We calculated iHS and XP-EHH in each individual population, and converted scores to empirical p-values as described in the text. For a signal to be considered overlapping between populations, we required that the iHS or XP-EHH score be in the 1% tail of one population and the 5% tail of the other. In Supplementary Figures 4 and 5, we show the fraction of selection signals that overlap between all pairs of populations. The striking difference between African and non-African populations in XP-EHH is an artifact of the way XP-EHH is calculated–for non-African populations, we used the group of Bantu speakers as a reference, and for African populations we used a group of European chromosomes as the reference. We performed the test this way to most closely follow what was most powerful in the simulations, but this means that the non-overlap between African and non-African populations by XP-EHH is not meaningful.

Calculation of the CLR statistic and overlap with other signals of selection The CLR test used in the main text as an alternative to XP-EHH for detecting high-frequency sweeps was originally designed to be applied to data for which SNP ascertainment is constant across the genome and, ideally, can be properly modeled. This is not the case for the HGDP data–tag SNPs were identified in the HapMap database, which has complicated and often unknown ascertainment [1].

Because of this, we have used this statistic not as a formal test for significance, but rather to rank regions by the deviance of the allele frequency spectra from the genome-wide average. We then used the CLR statistics as test statistics like XP-EHH and iHS and convert them to empirical p-values after controlling for SNP density, as described in the main text.

As XP-EHH and CLR both detect local deviations in the allele frequency spectrum (XP-EHH detects regions of extended homozygosity), we expect them to largely overlap. This is indeed the case for non-African populations–defining overlap as above, we find that the CLR and XP-EHH statistics overlap 63%, 58%, 67%, 83%, 88%, and 71% for Europe, the Middle East, S. Asian, E. Asia, Oceania, and the Americas, respectively. For the Bantu, however, the overlap is only 28%. We interpret this to mean that many signals of fixed sweeps in the Bantu are less robust than those outside of Africa. This is consistent with the observation that there appear to be few reliable signals of recent fixations in African populations[2, 4].

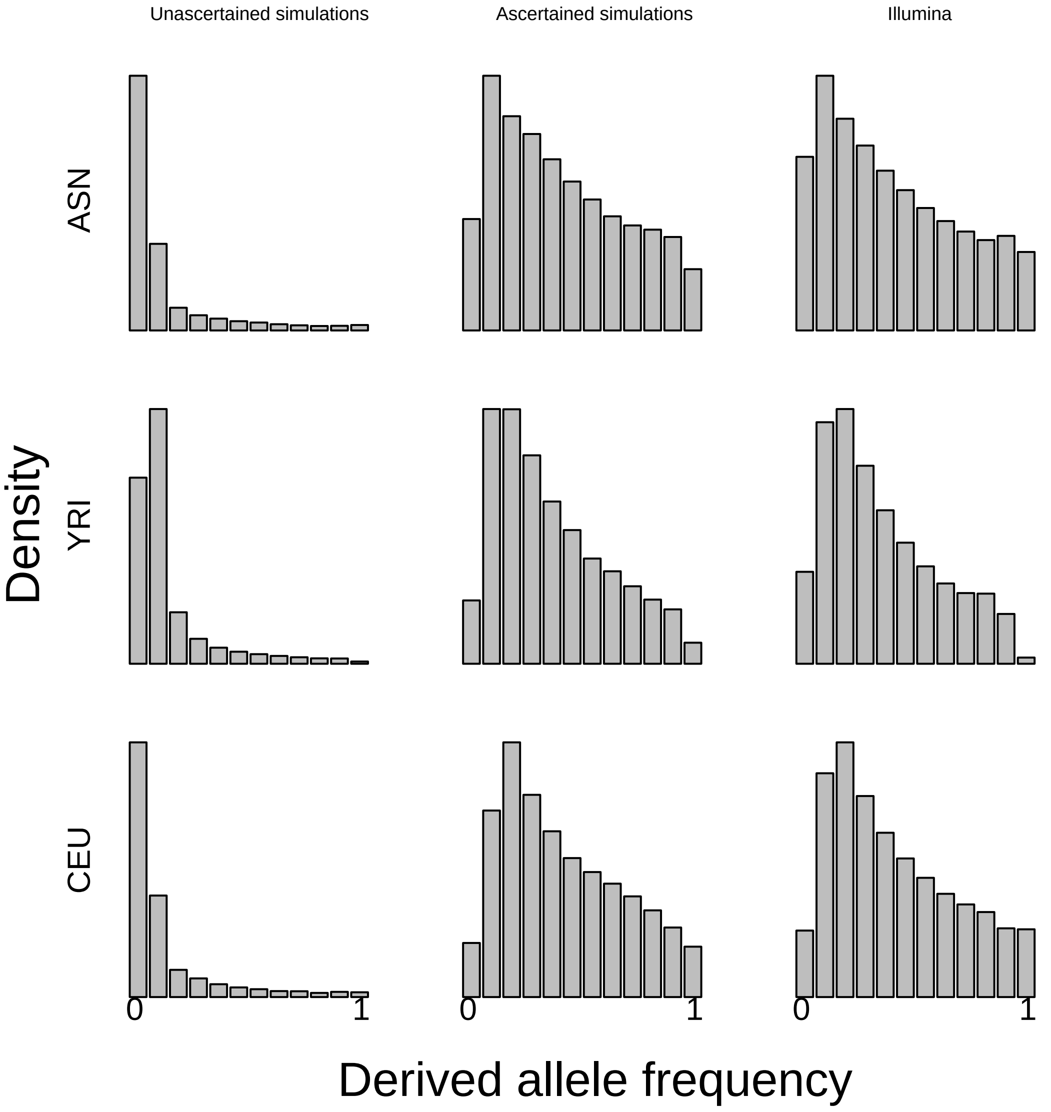

Figure 1: Marginal frequency spectra in the simulations before and after the application of the ascertainment procedure, as well as in the real data. In the first column are the marginal allele frequency spectra in the simulations before application of the ascertainment procedure, in the second column are the allele frequency spectra in the simulations after the application of the ascertainment procedure, and in the third column are the marginal allele frequency spectra in the HapMap populations at the SNPs typed on the Illumina panel.

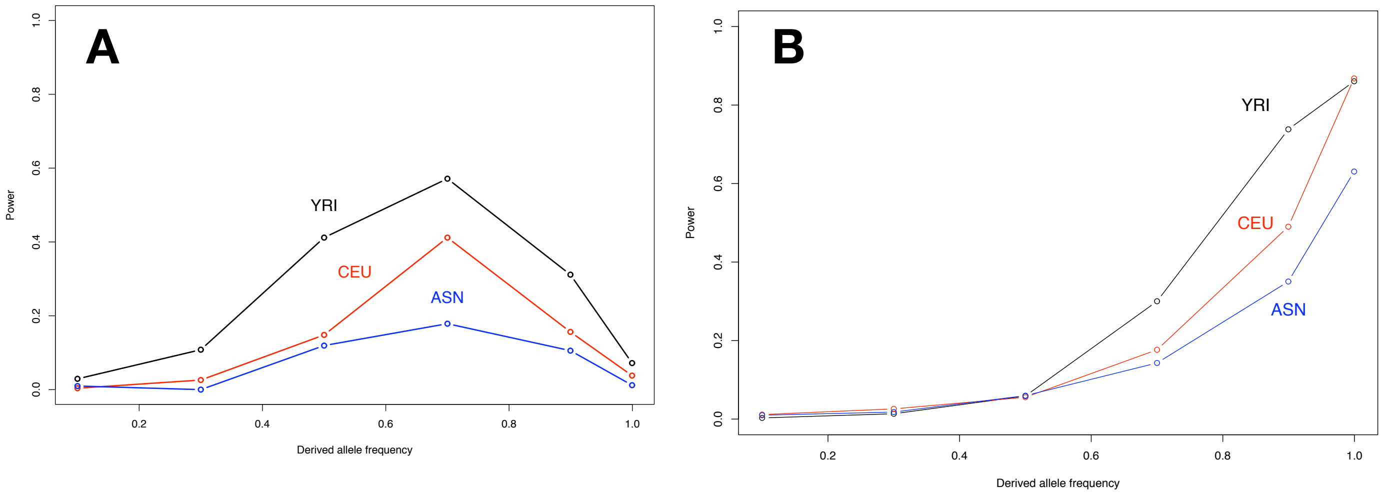

Figure 2: **Power of iHS (A) and XP-EHH (B) in three demographies based on the HapMap.** Selected alleles were introduced at a random time with a selection coefficient of 1% and simulated forwards in time. In all simulations, the selected site was excluded from analyses. Simulations were binned into six bins according the the final frequency of the selected allele. The type I error was set to 1%, with the threshold determined by neutral simulations.

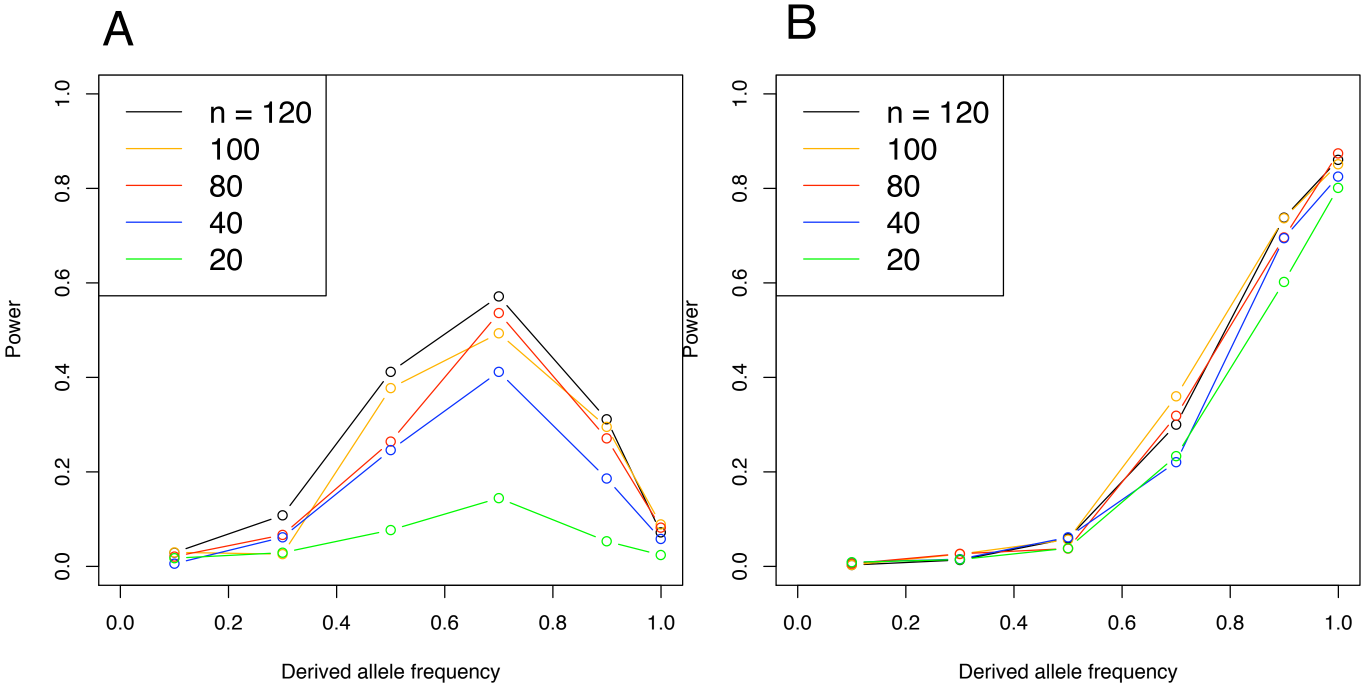

Figure 3: **Power of iHS (A) and XP-EHH (B) in the YRI demography for different sample sizes (in number of chromosomes)**. Selected alleles were introduced at a random time with a selection coefficient of 1% and simulated forwards in time. Simulations were binned into six bins according the the final frequency of the selected allele. The type I error was set to 1%, with the threshold determined by neutral simulations.

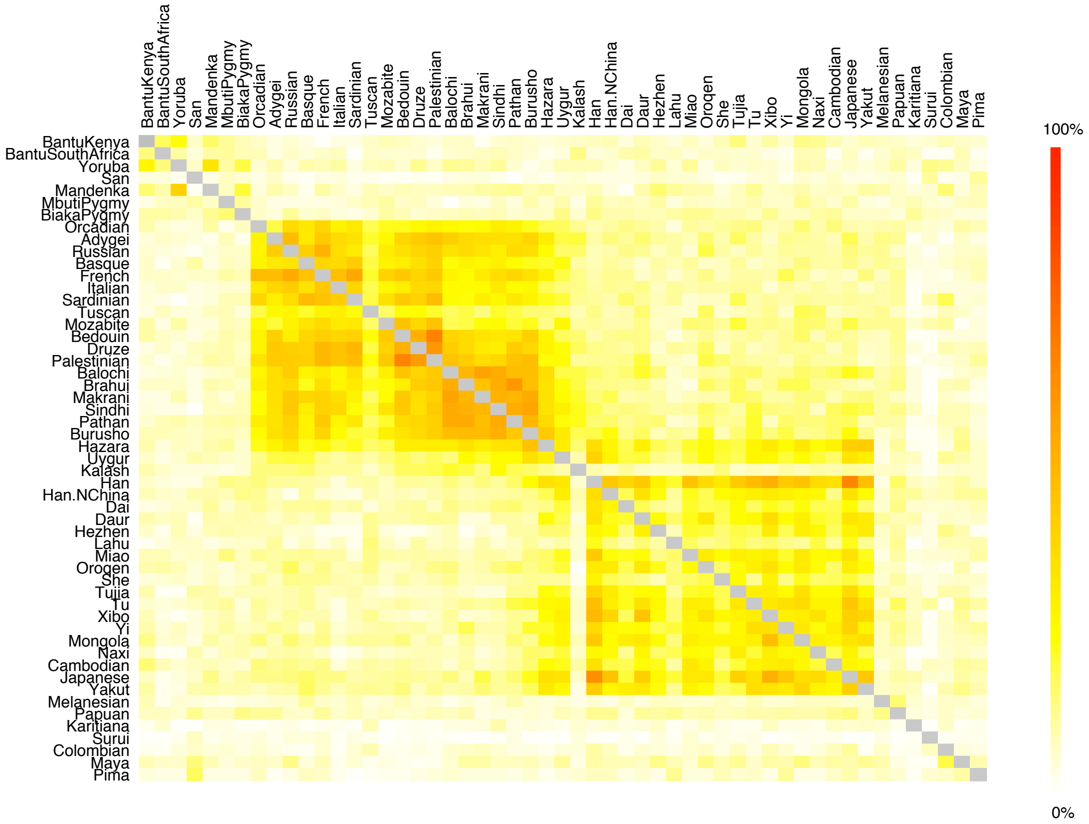

Figure 4: Overlap of iHS signals across populations. In each cell (i, j) the fraction of windows with scores in the 1% tail of iHS in population i and in the 5% tail of iHS in population j is colored according the legend on the right.

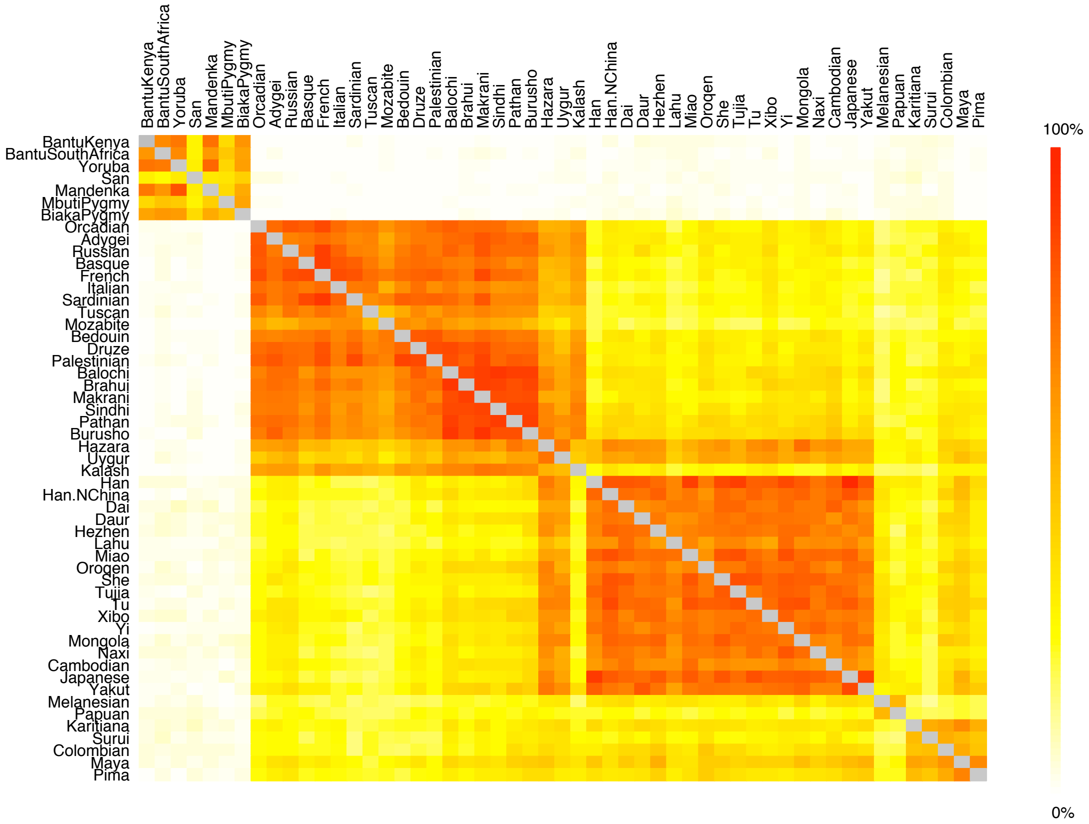

Figure 5: Overlap of XP-EHH signals across populations. In each cell (i, j) the fraction of windows with scores in the 1% tail of XP-EHH in population i and in the 5% tail of XP-EHH in population j is colored according the legend on the right.

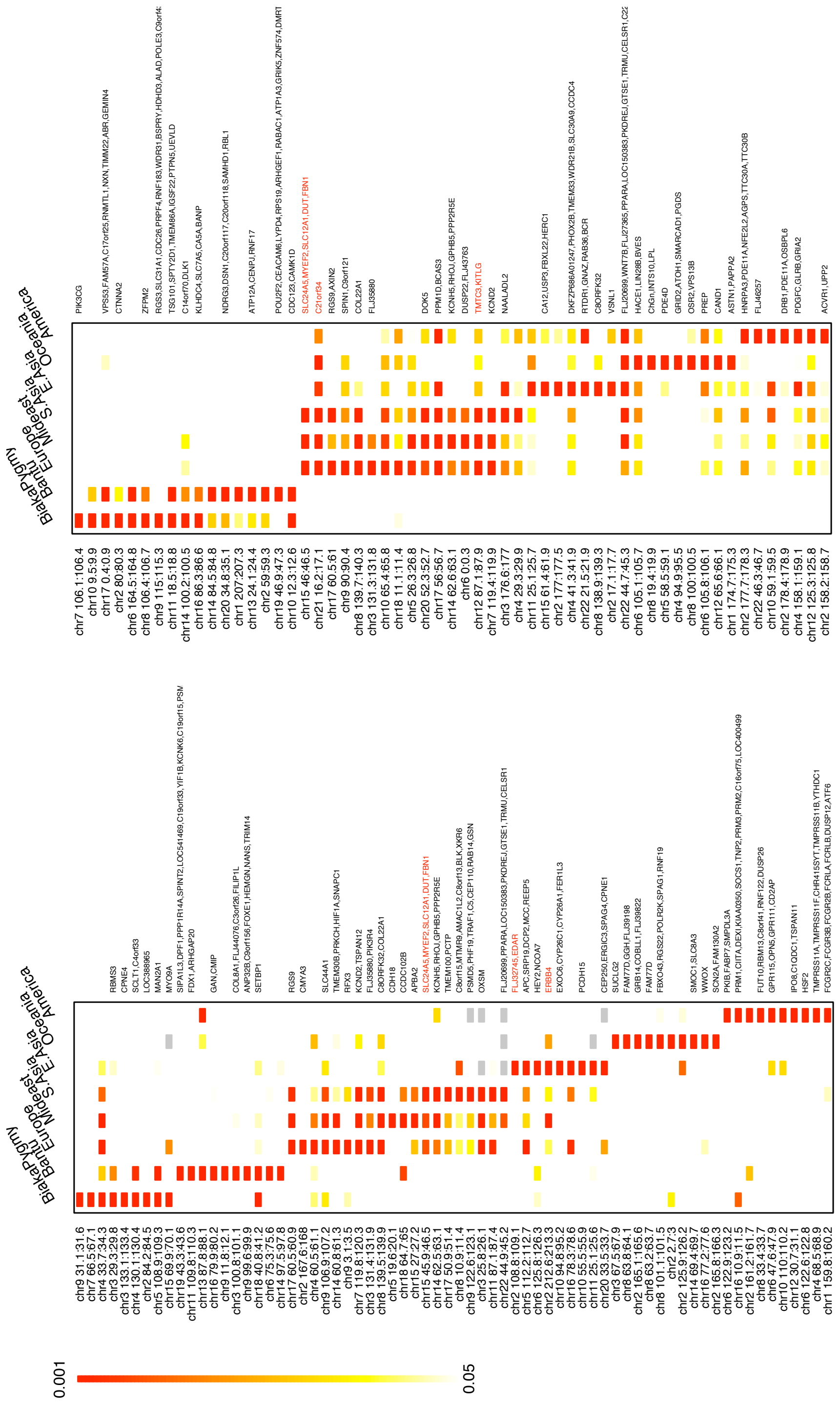

Figure 6: **Top 10 iHS (A) and XP-EHH (B) signals by population cluster**. This is analogous to Figure 1 in the main text, but uses a sliding window of 200kb that slides in steps of 100kb.

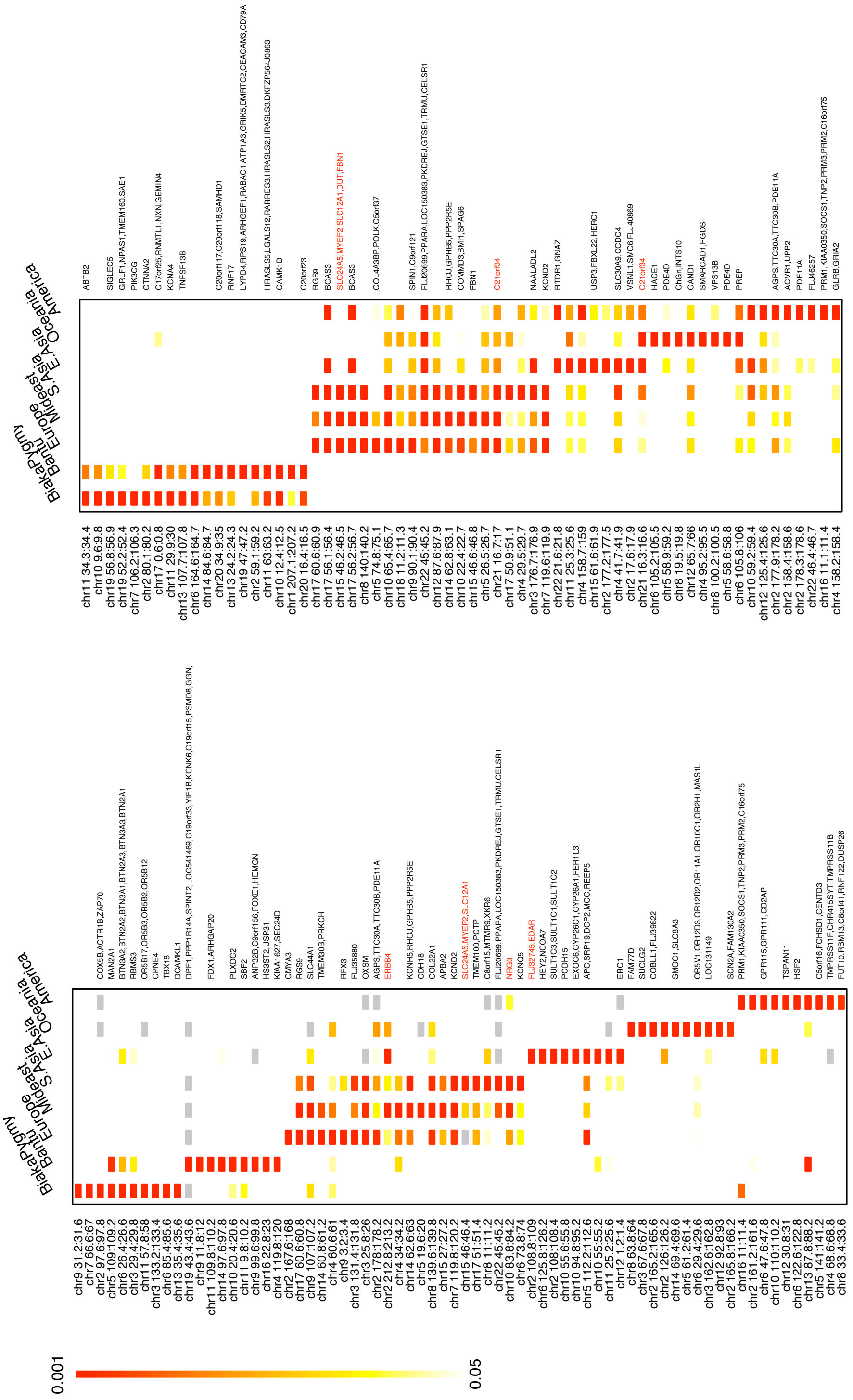

Figure 7: **Top 10 iHS (A) and XP-EHH (B) signals by population cluster**. This is analogous to Figure 1 in the main text, but uses a window size of 100kb.

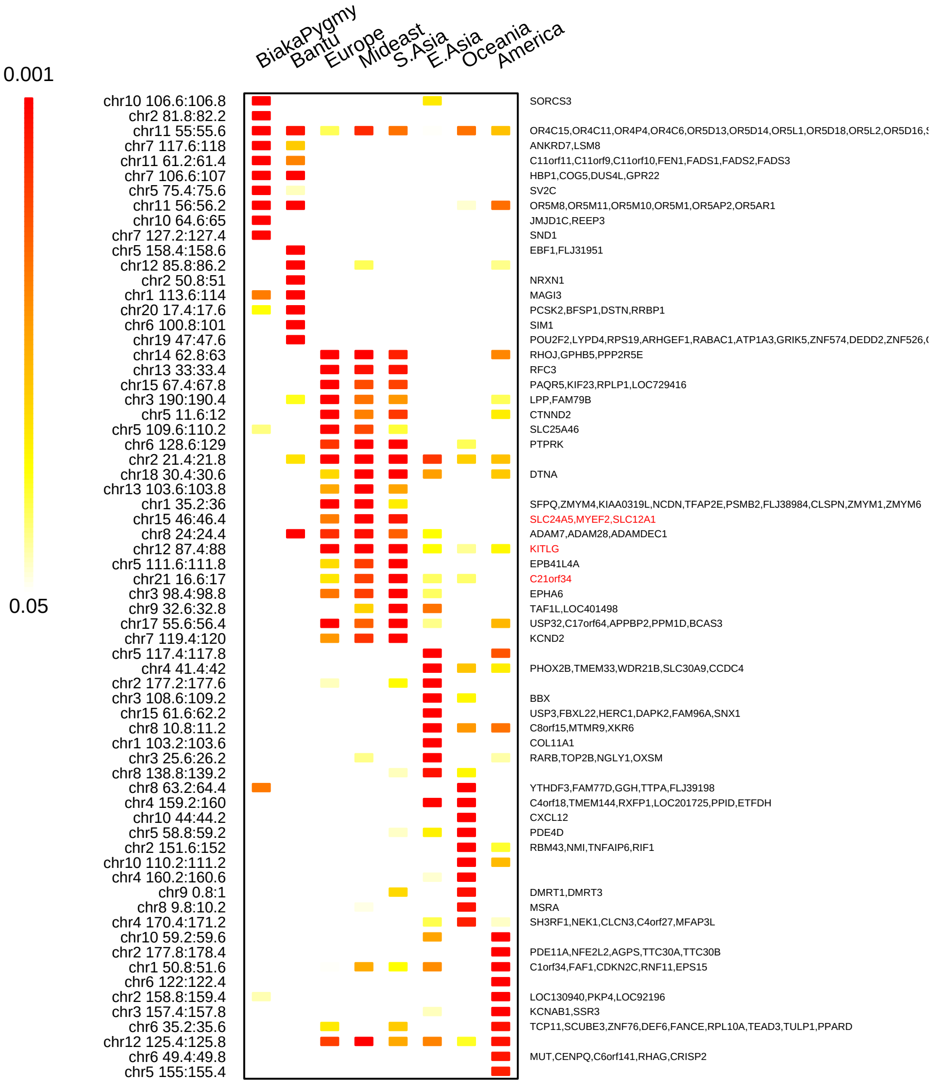

Figure 8: **Top 10 CLR signals by population cluster**. This is analogous to Figure 1 in the main text.

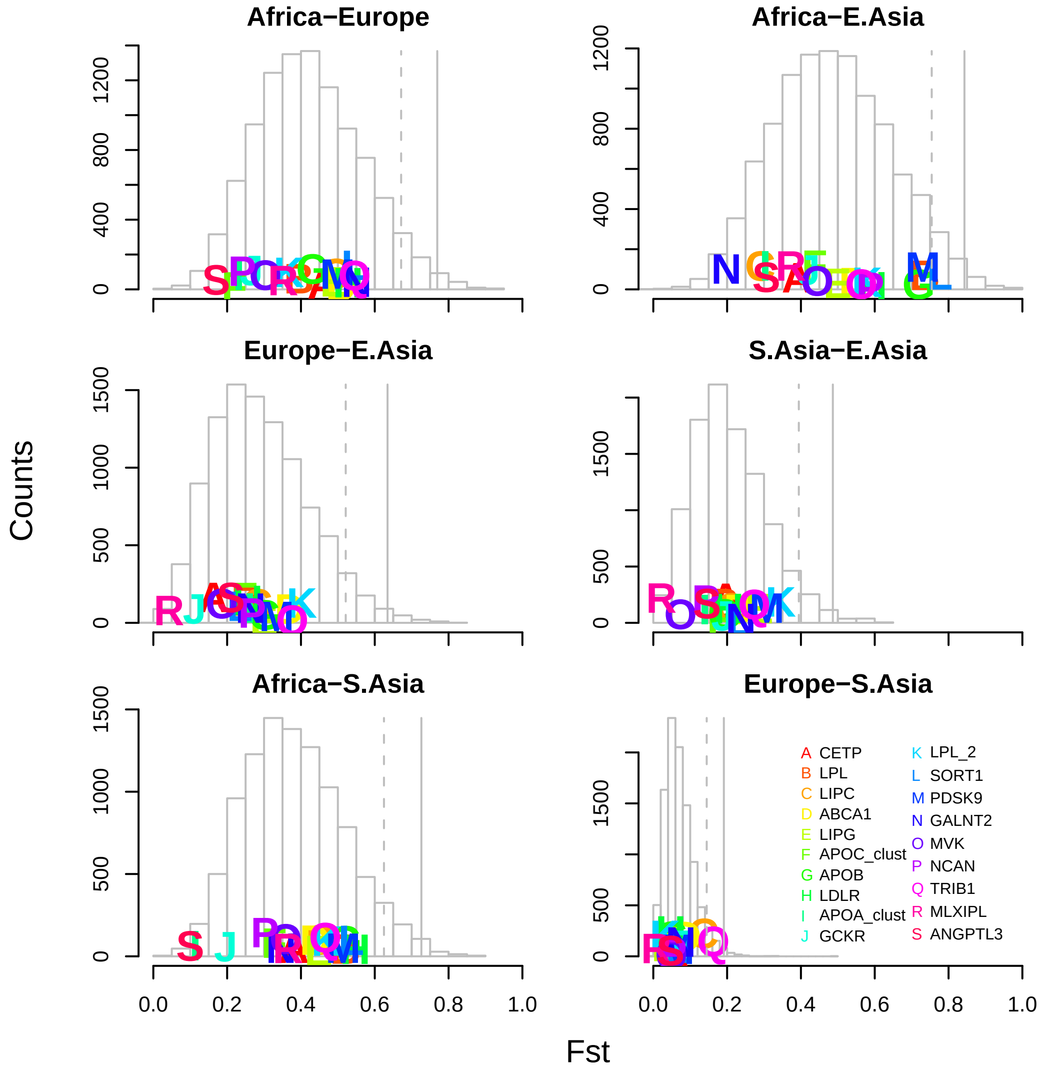

Figure 9: **F_{ST} surrounding loci involved in natural variation in lipid levels**. This figure was generated as in Figures 3 and 4 in the main text.

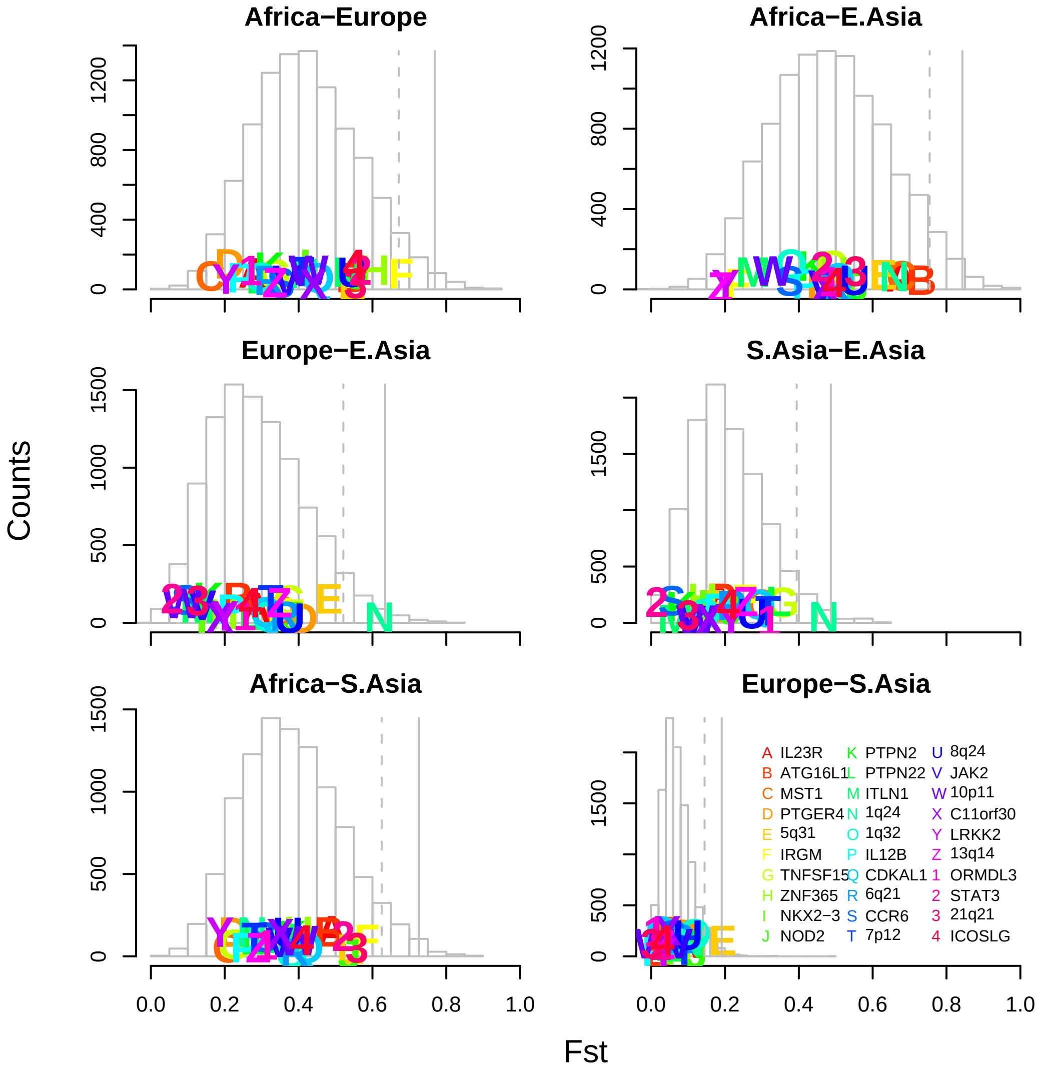

Figure 10: **$\mathbf{F}_{\mathrm{ST}}$ surrounding loci involved in natural variation in susceptibility to Crohn's disease**. This figure was generated as in Figures 3 and 4 in the main text.

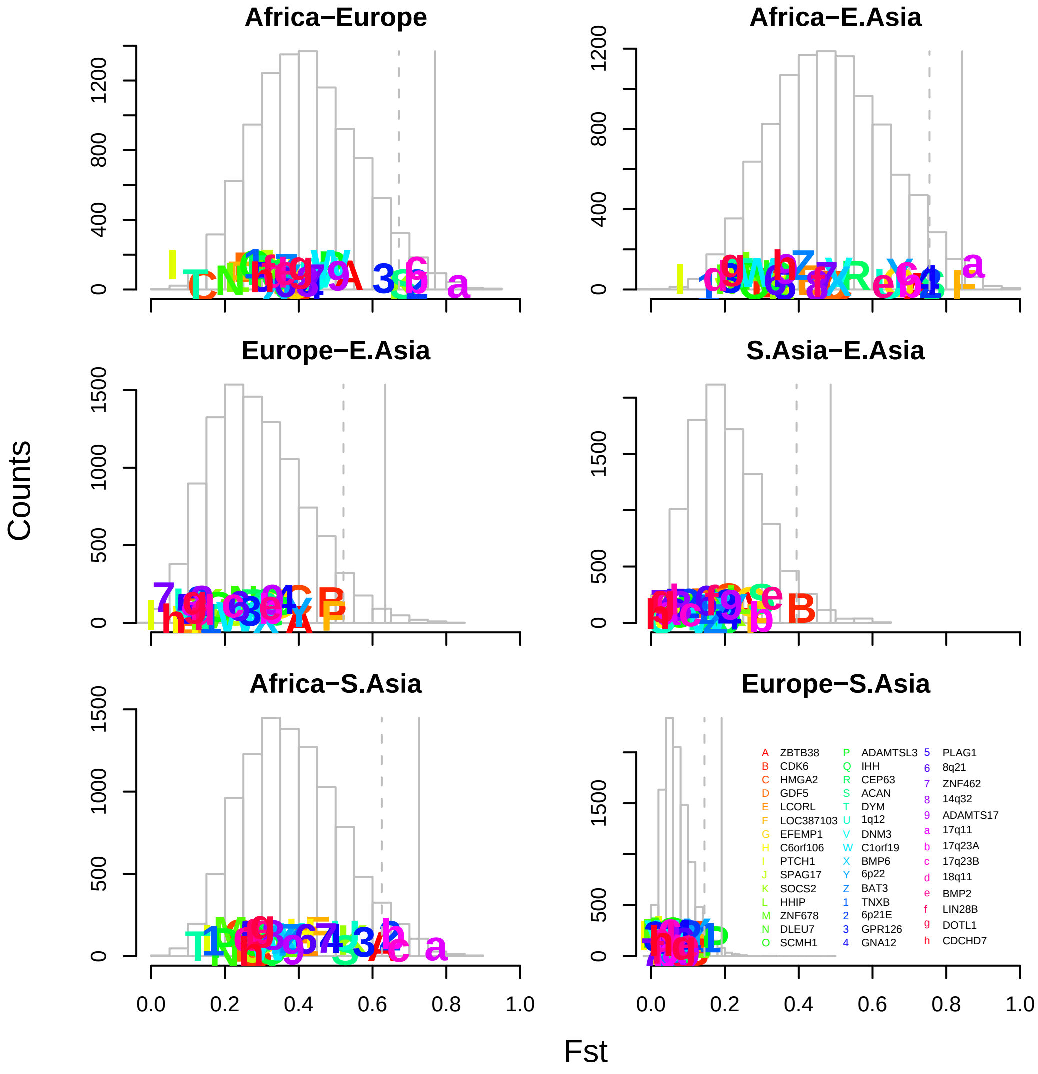

Figure 11: **F_{ST} surrounding loci involved in natural variation in height**. This figure was generated as in Figures 3 and 4 in the main text.

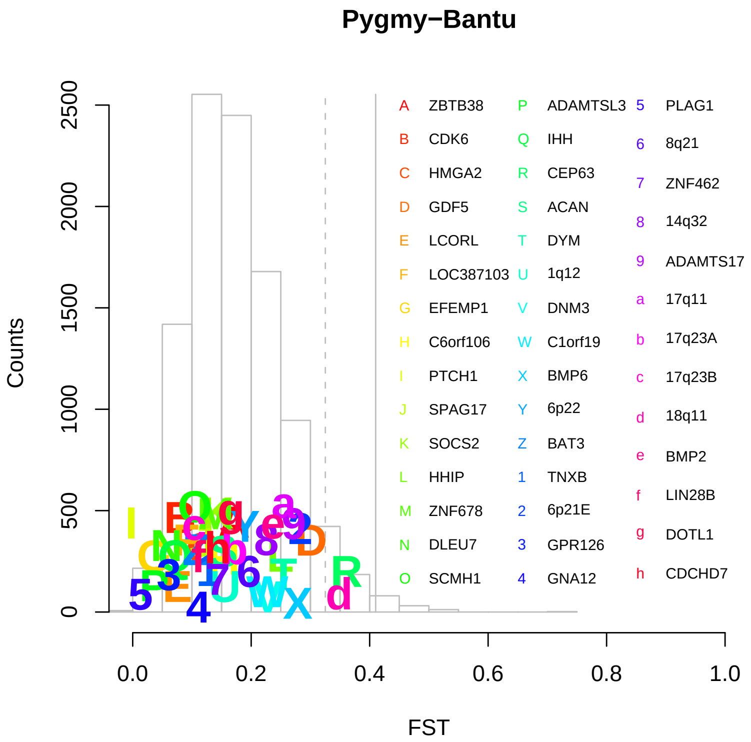

Figure 12: $\mathbf{F_{ST}}$ **between Bantu and Pygmy populations surrounding loci involved in natural variation in height**. This figure was generated as in Figures 3 and 4 in the main text.

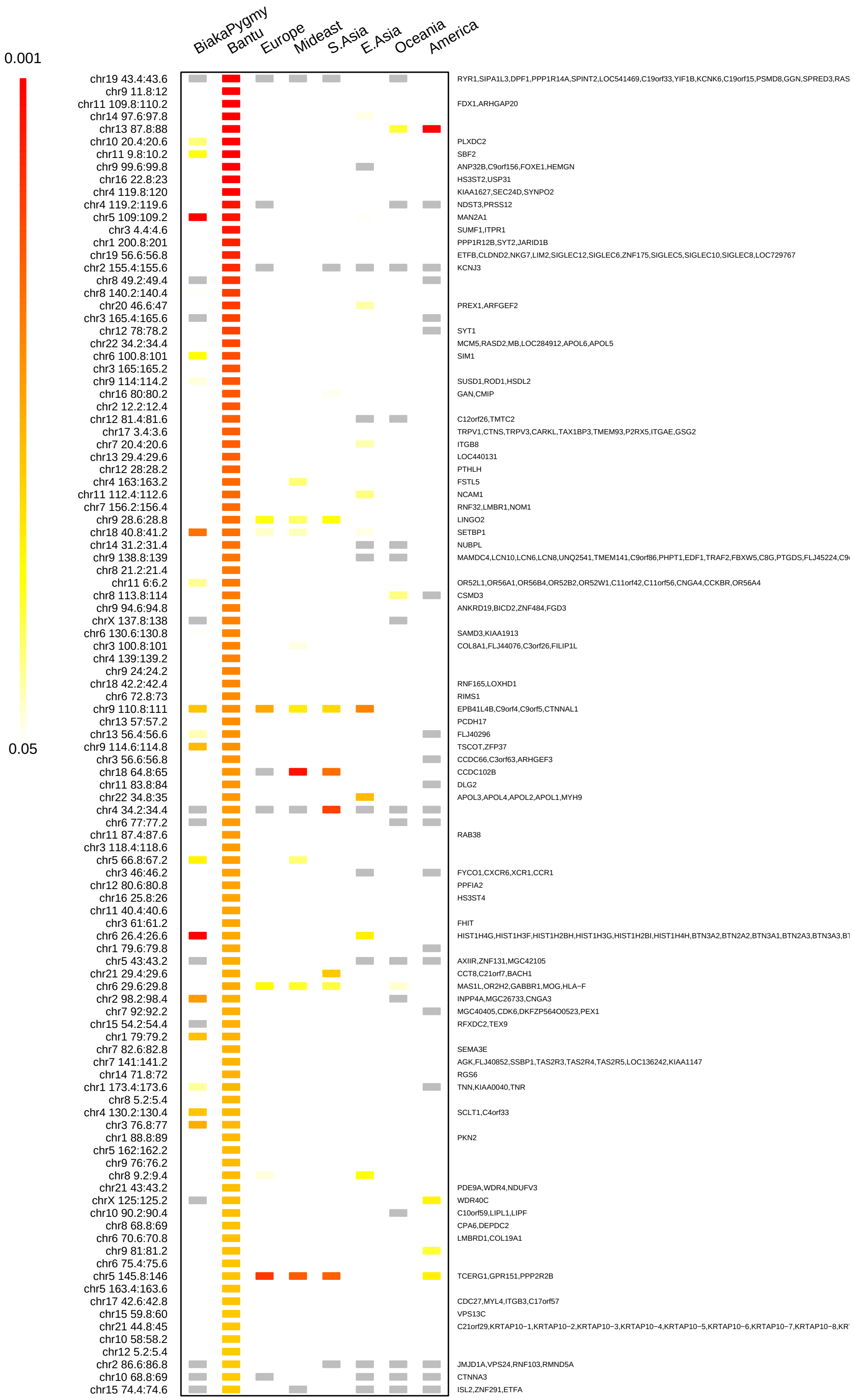

Figure 13: The top 1% of iHS signals in the Bantu, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

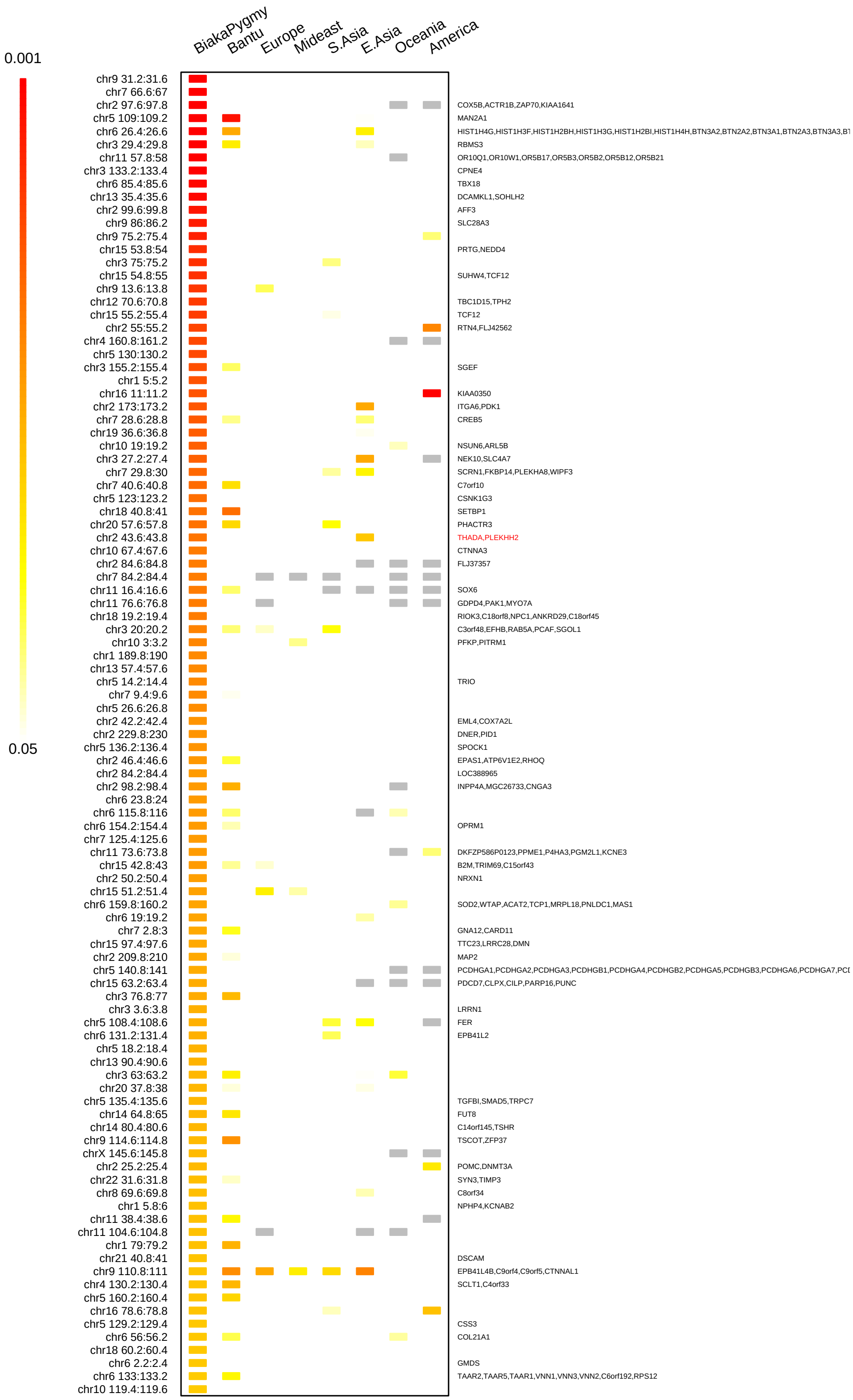

Figure 14: The top 1% of iHS signals in the Biaka Pygmies, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

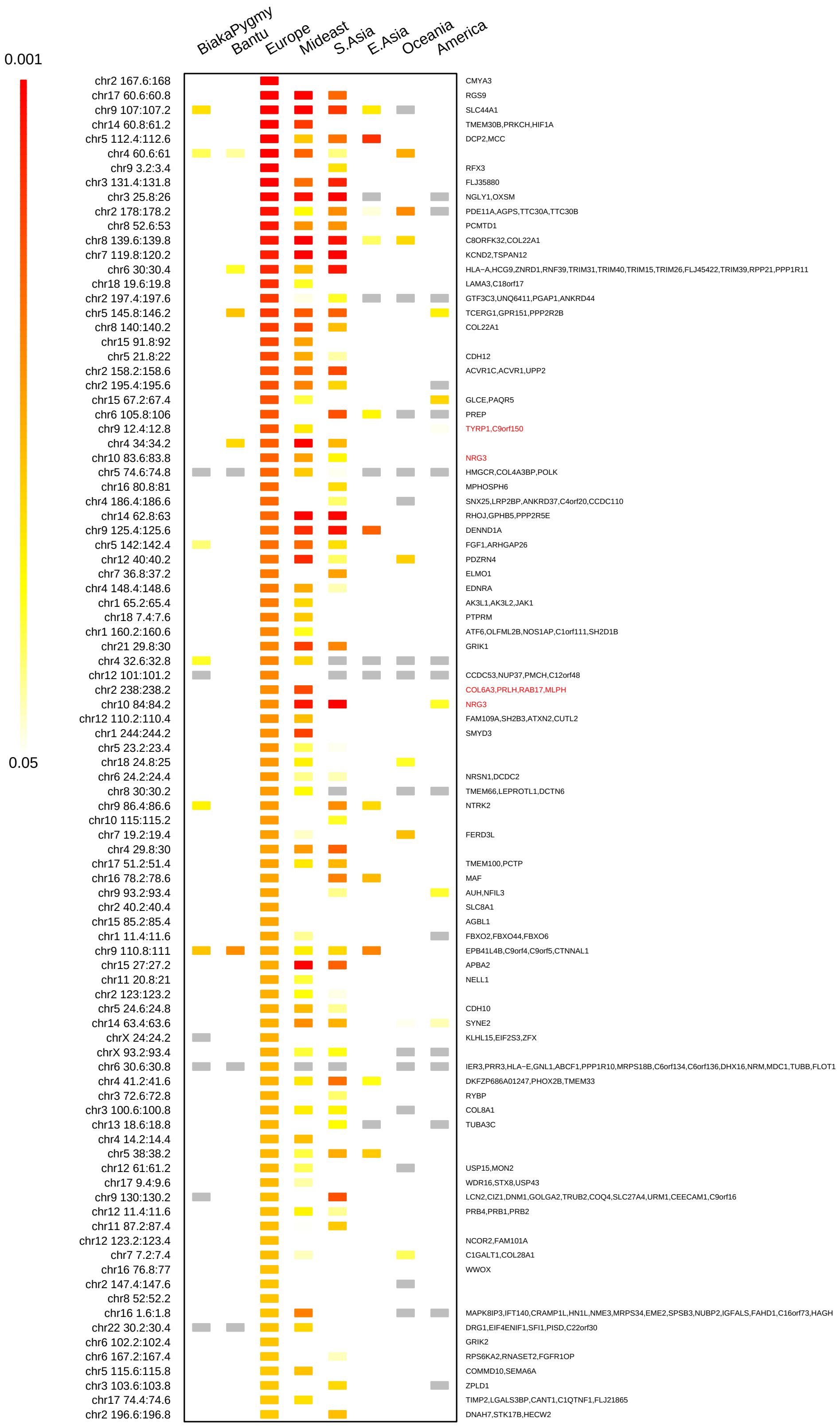

Figure 15: The top 1% of iHS signals in the Europeans, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

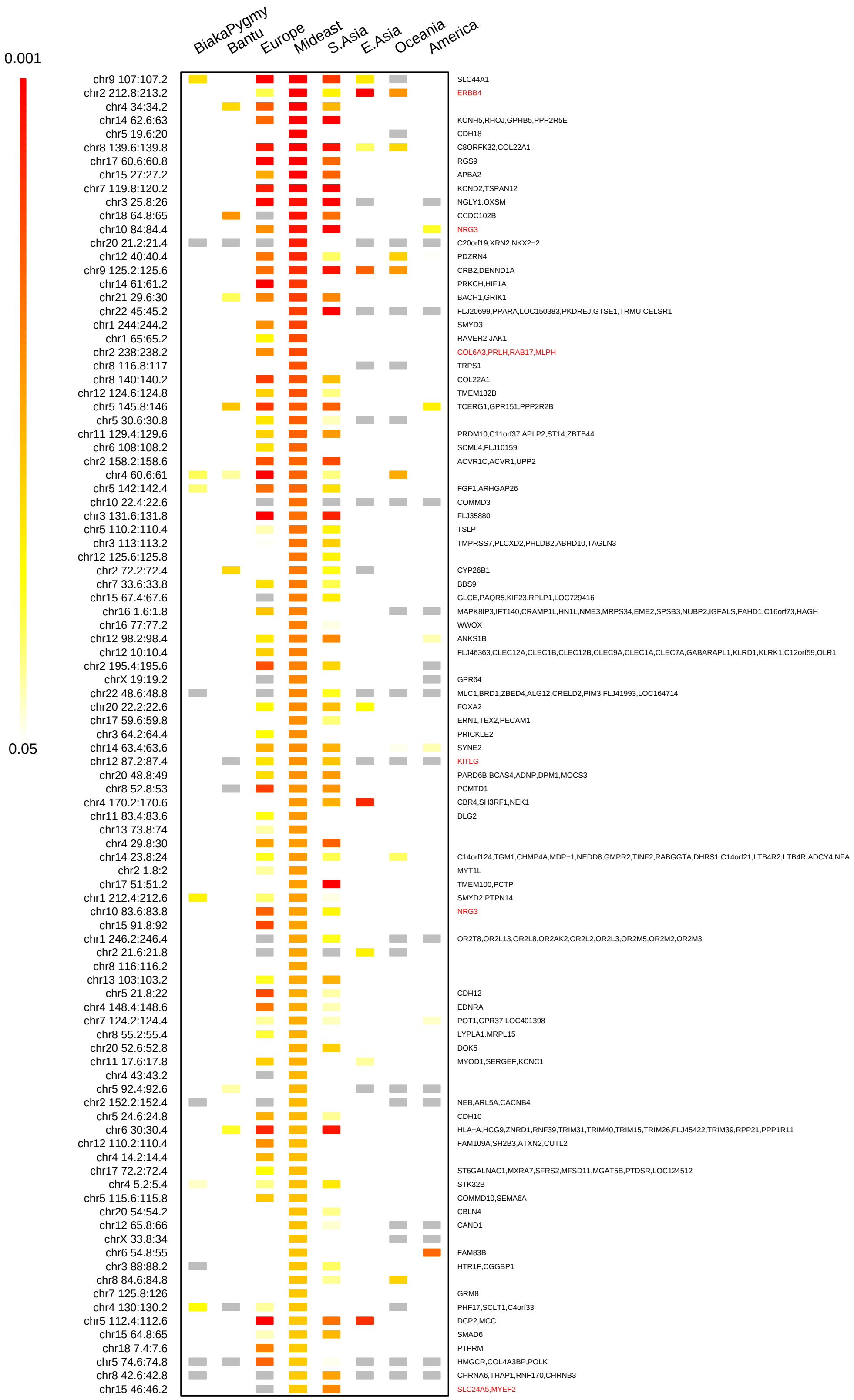

Figure 16: The top 1% of iHS signals in the Middle East, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

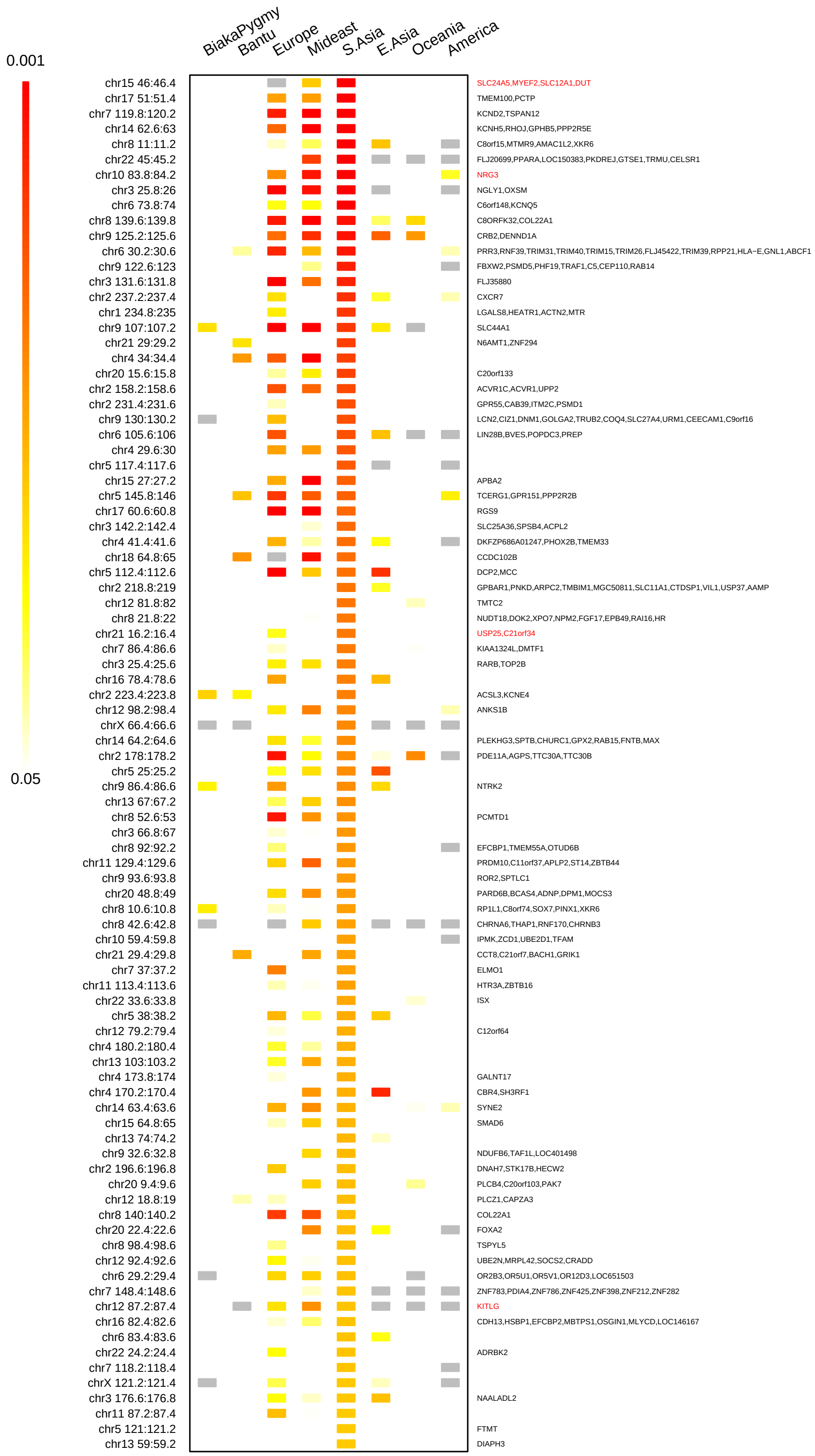

Figure 17: The top 1% of iHS signals in South Asia, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

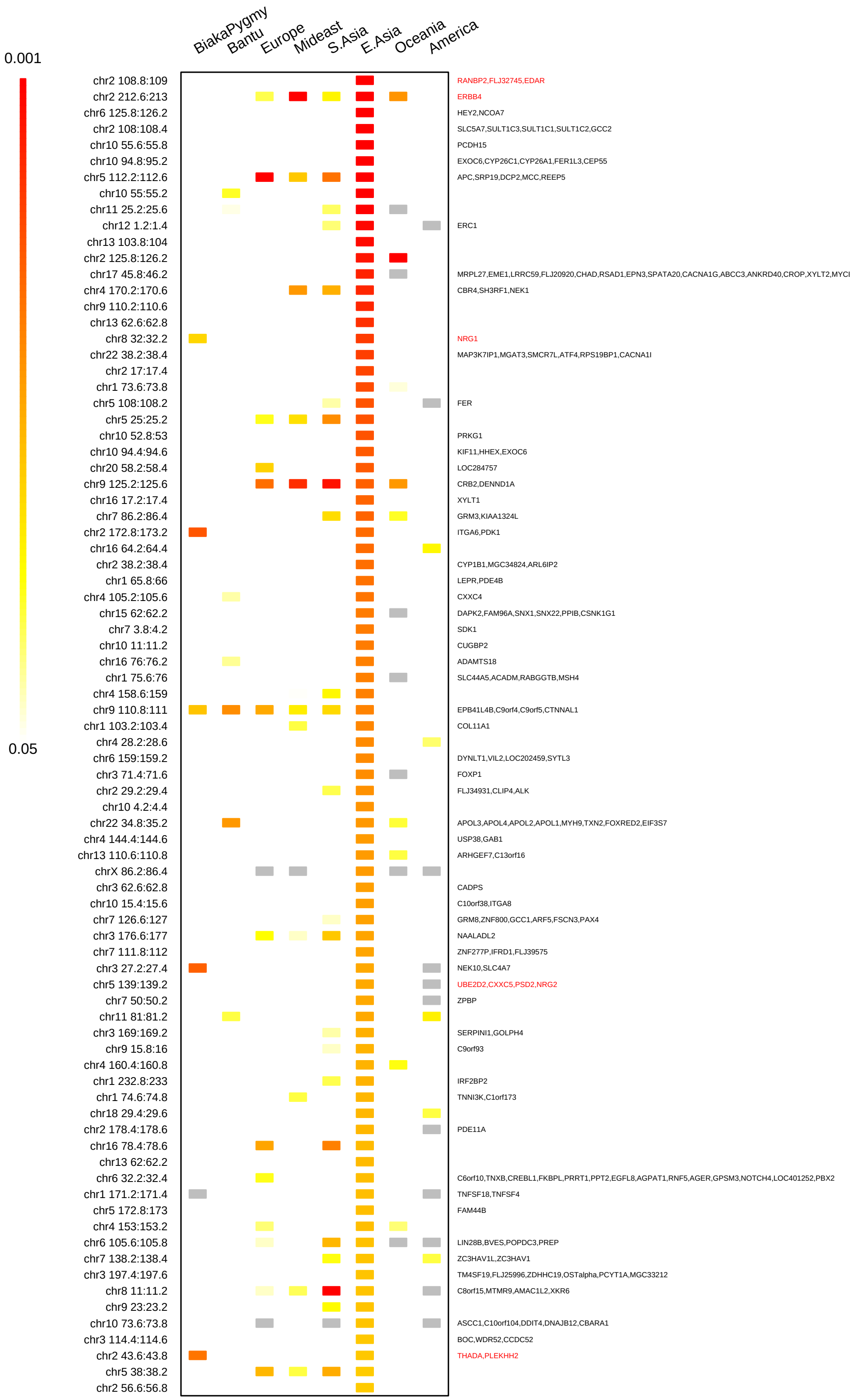

Figure 18: The top 1% of iHS signals in East Asia, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

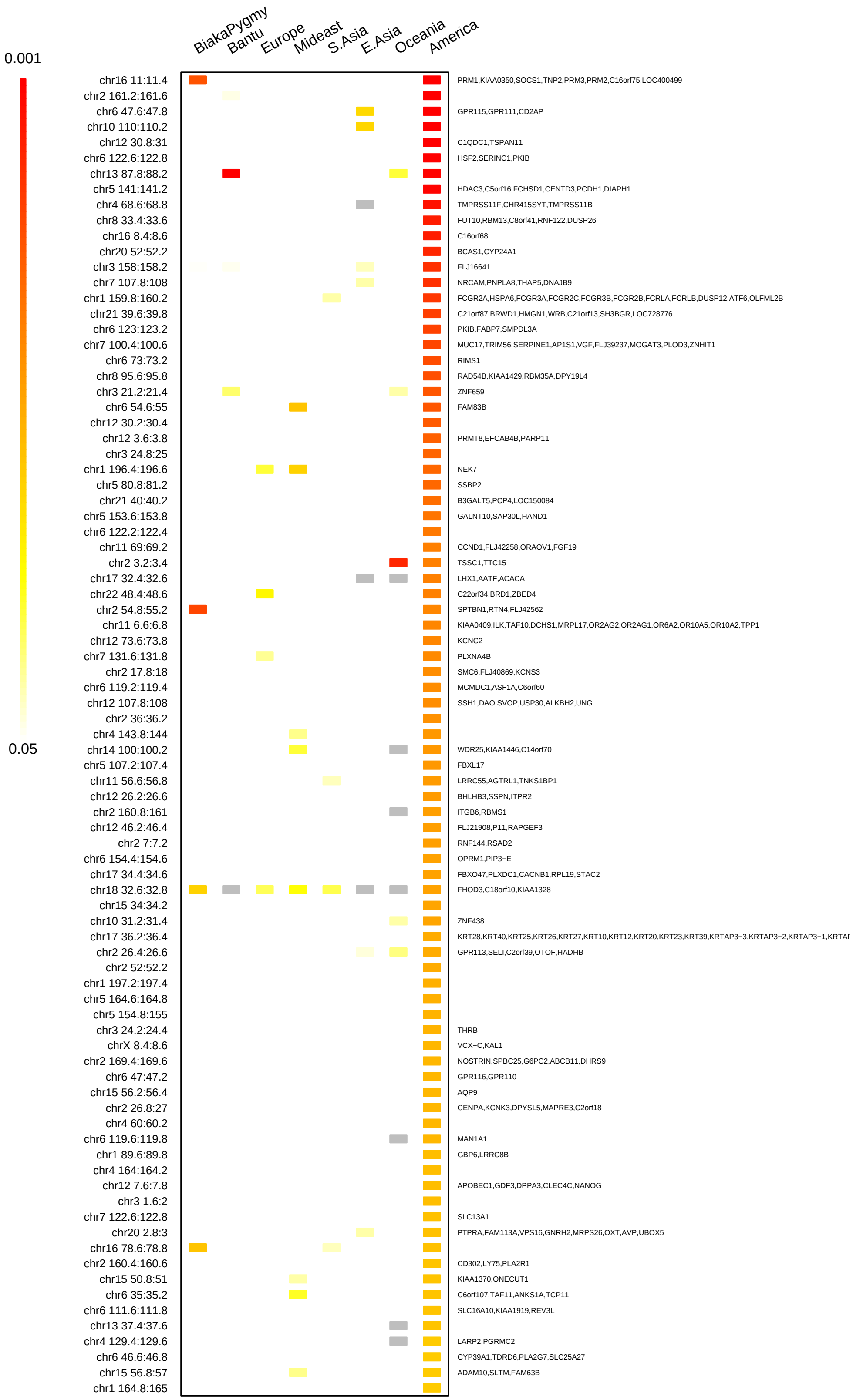

Figure 19: The top 1% of iHS signals in the Americas, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

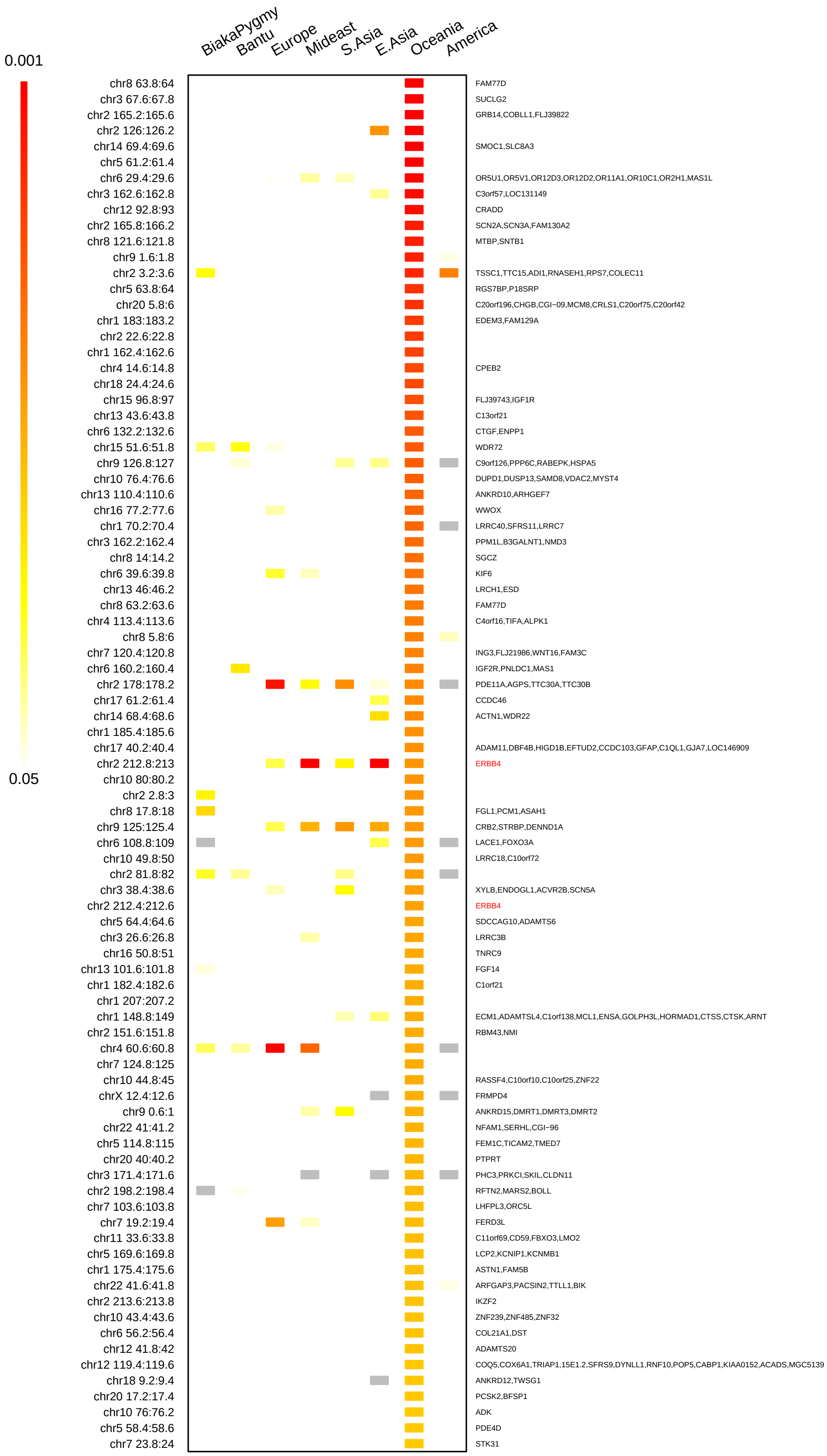

Figure 20: The top 1% of iHS signals in Oceania, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

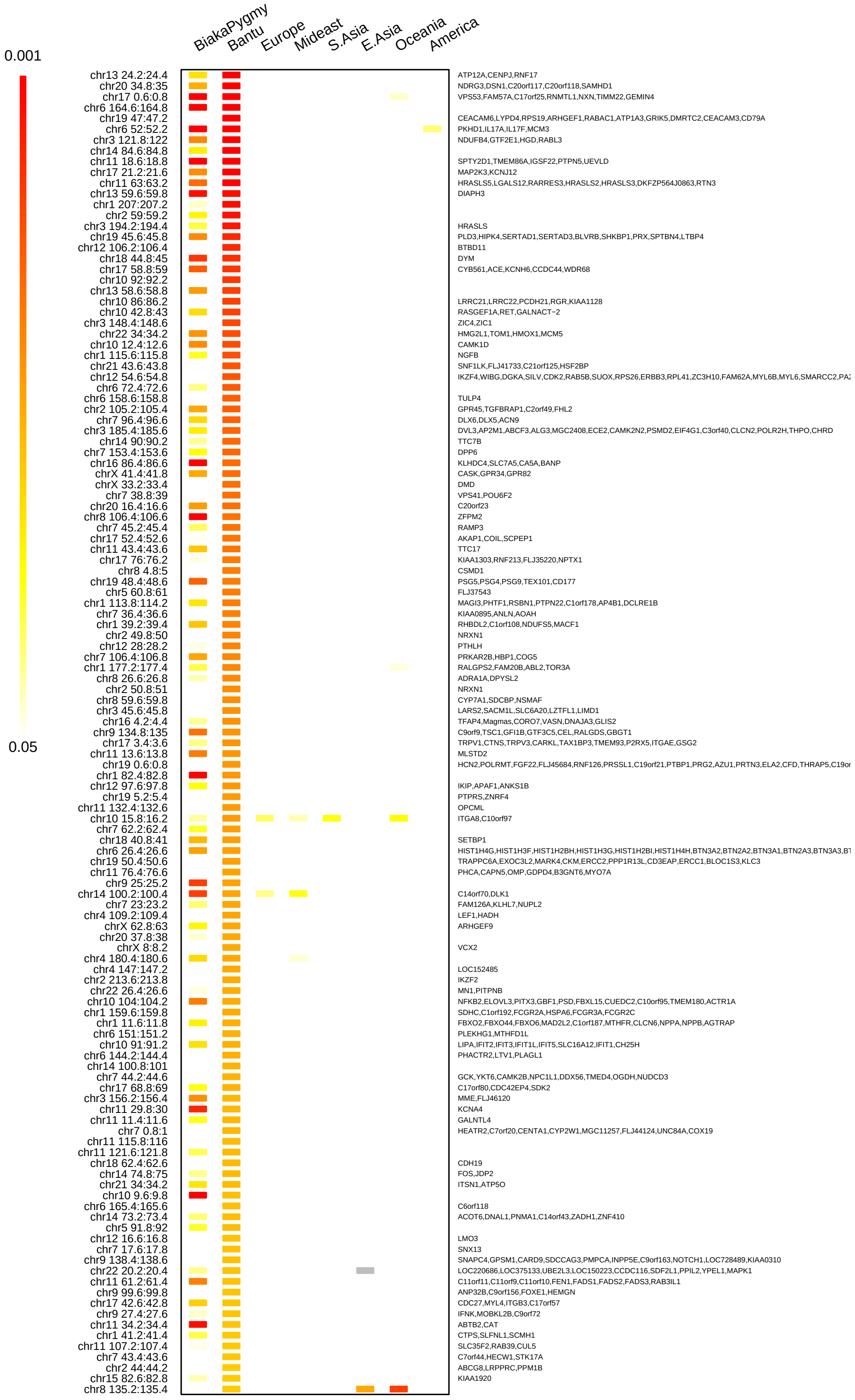

Figure 21: The top 1% of XP-EHH signals in the Bantu, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

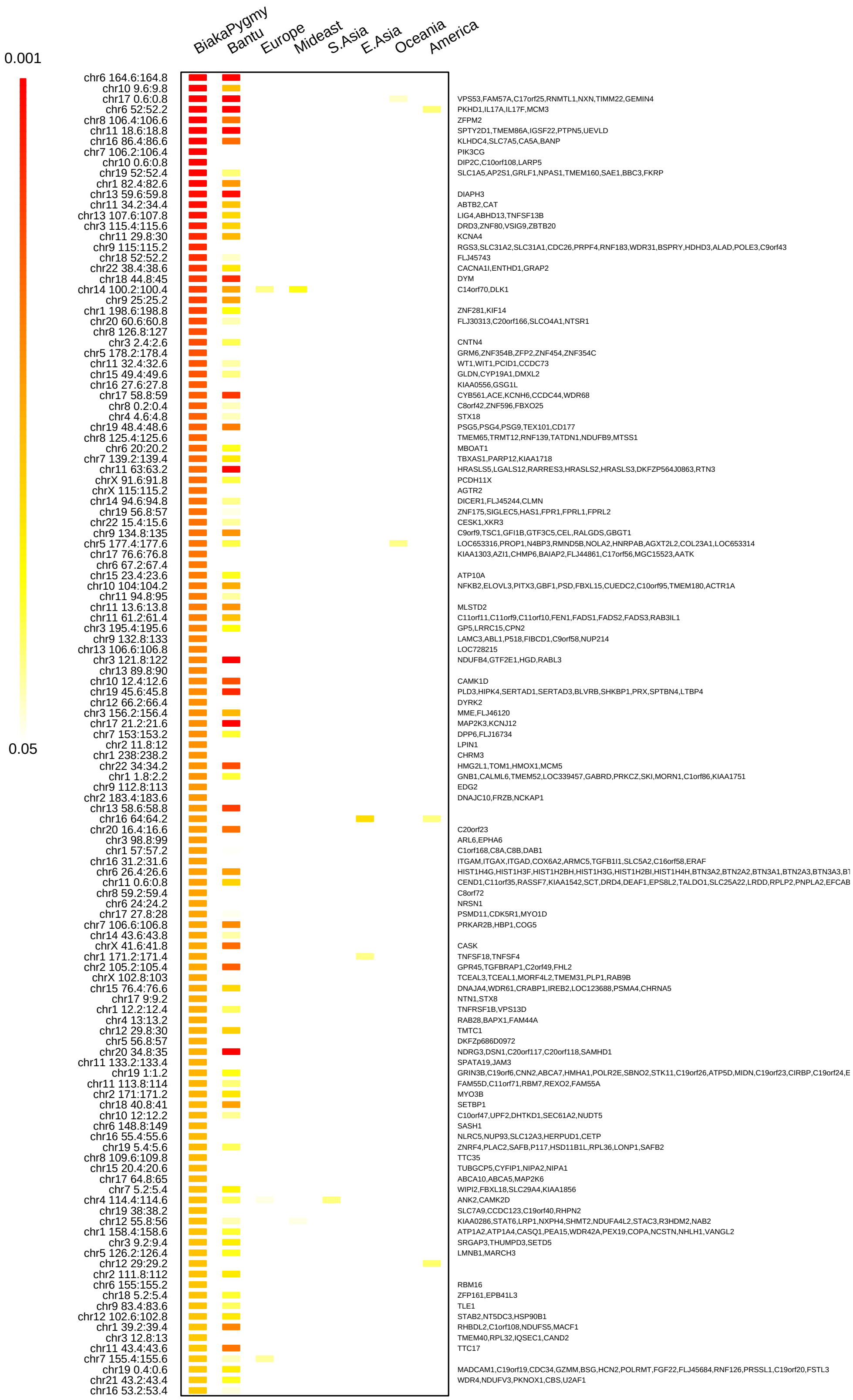

Figure 22: The top 1% of XP-EHH signals in the Biaka Pygmies, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

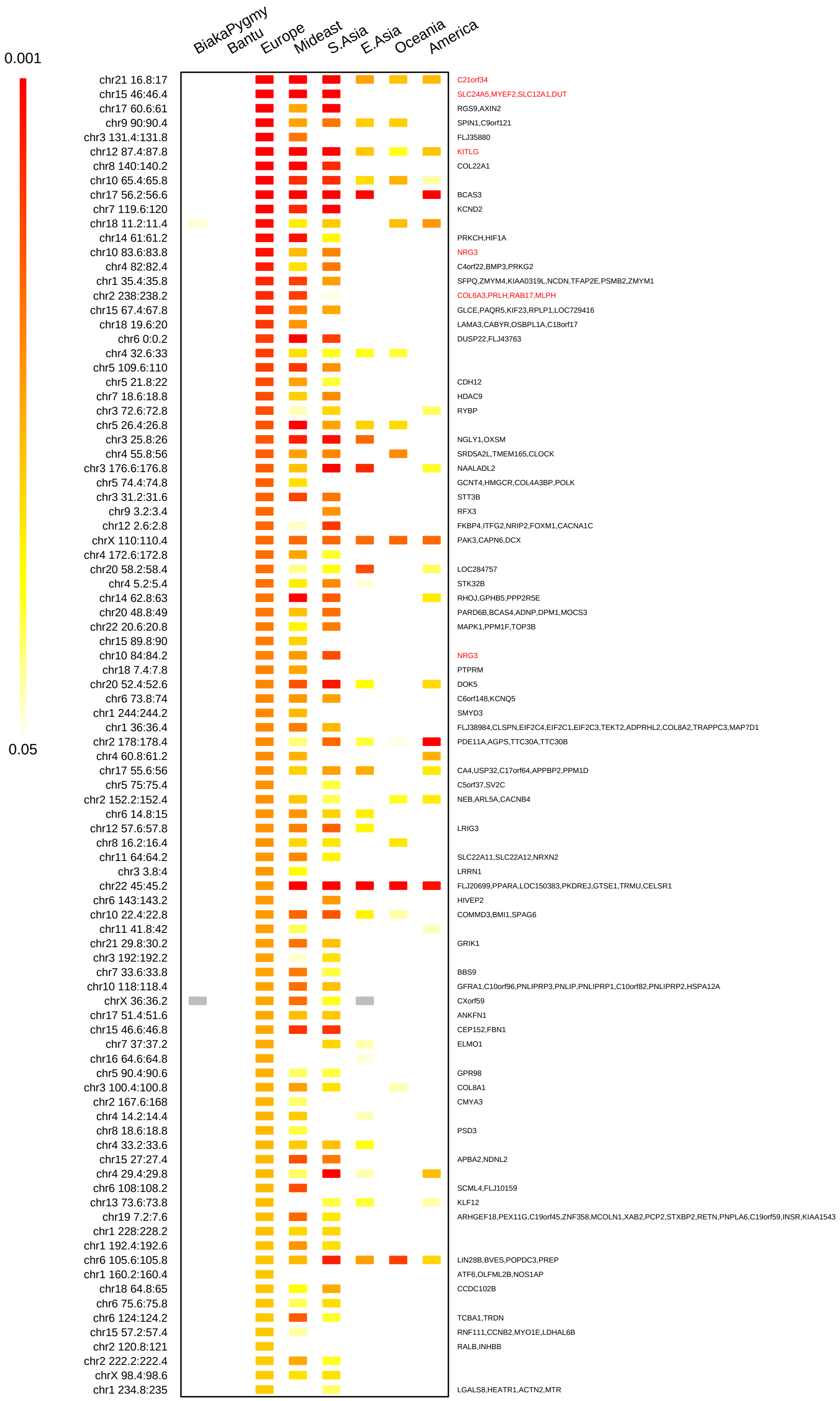

Figure 23: The top 1% of XP-EHH signals in Europe, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

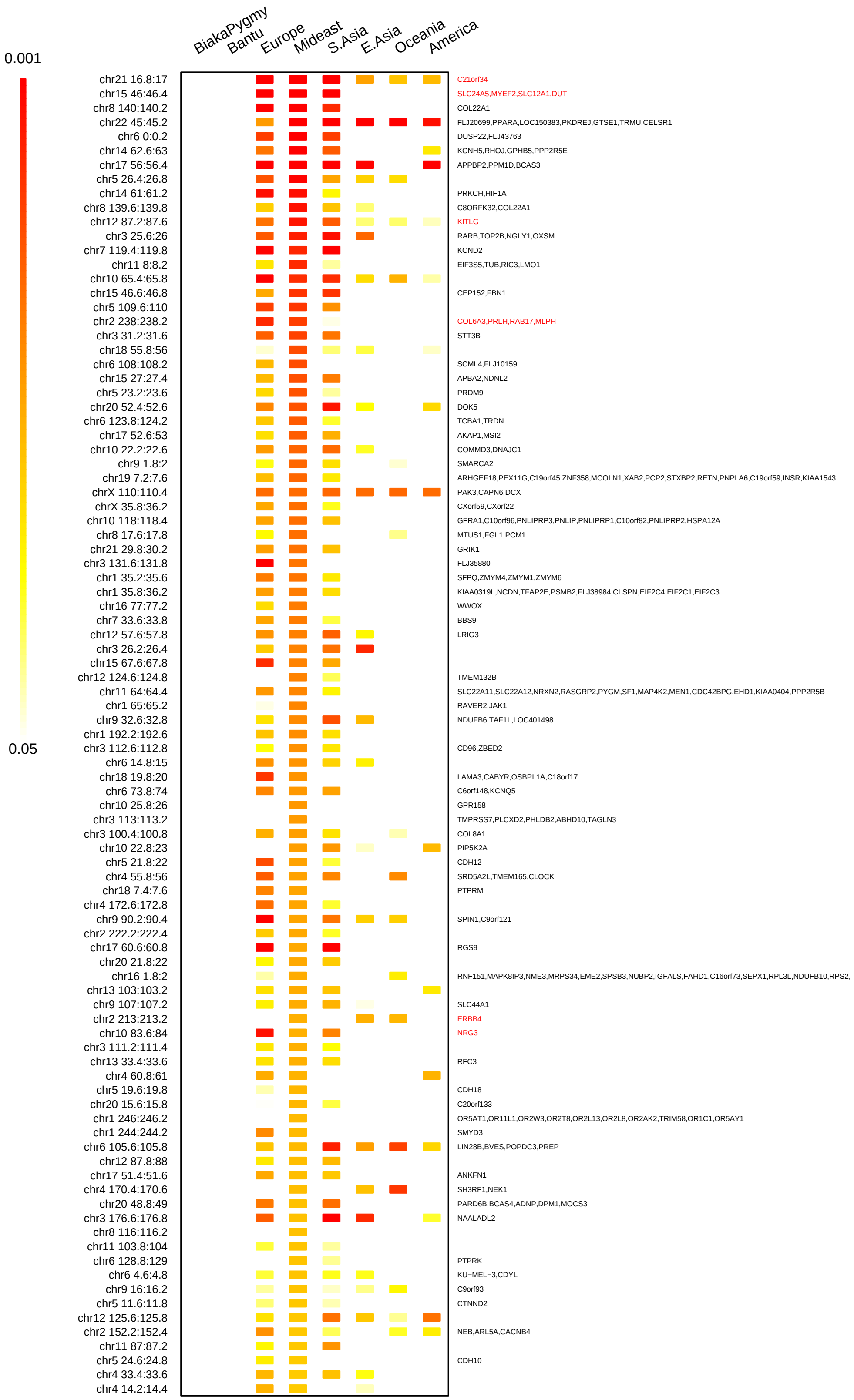

Figure 24: The top 1% of XP-EHH signals in the Middle East, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

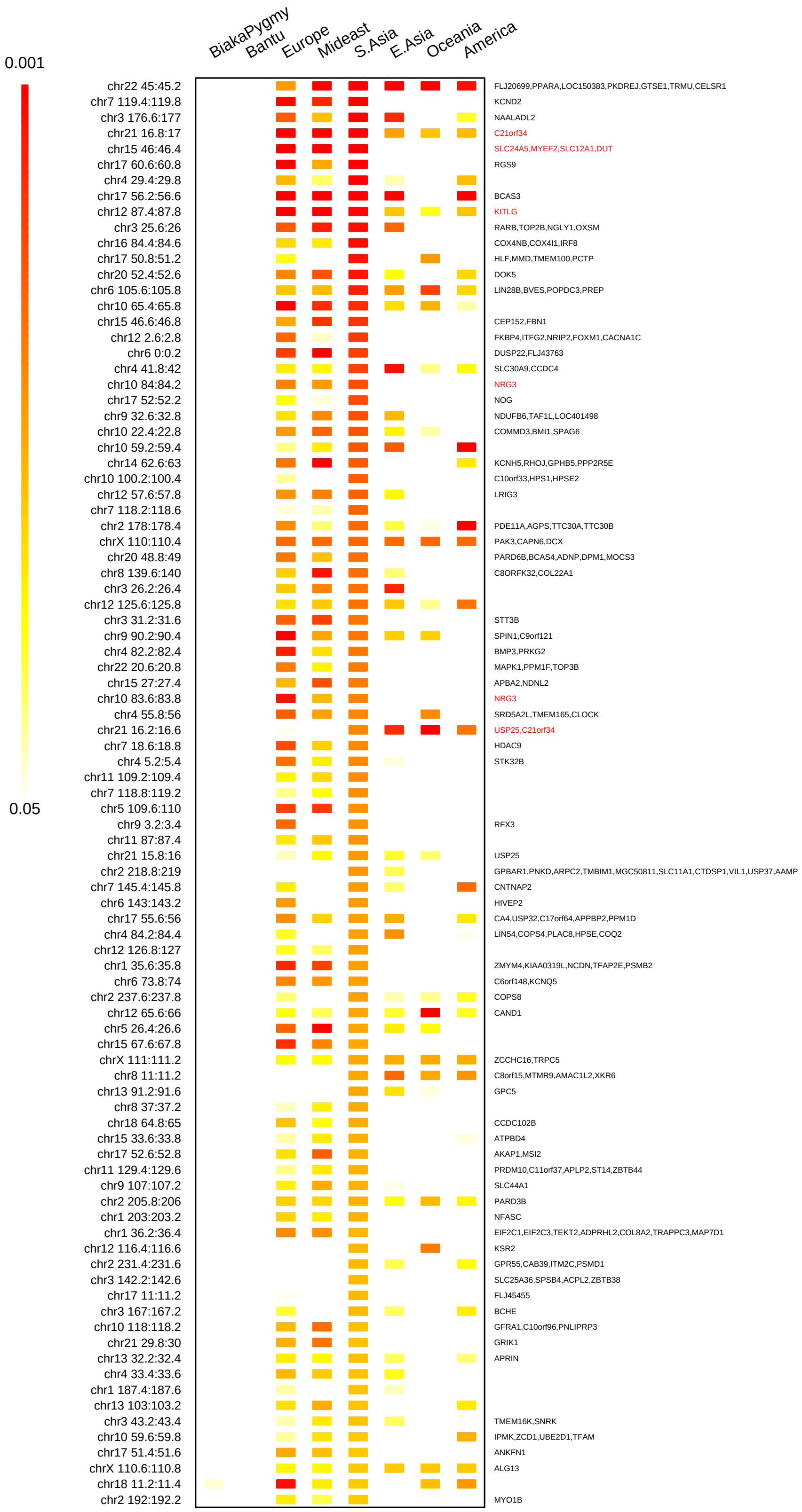

Figure 25: The top 1% of XP-EHH signals in South Asia, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

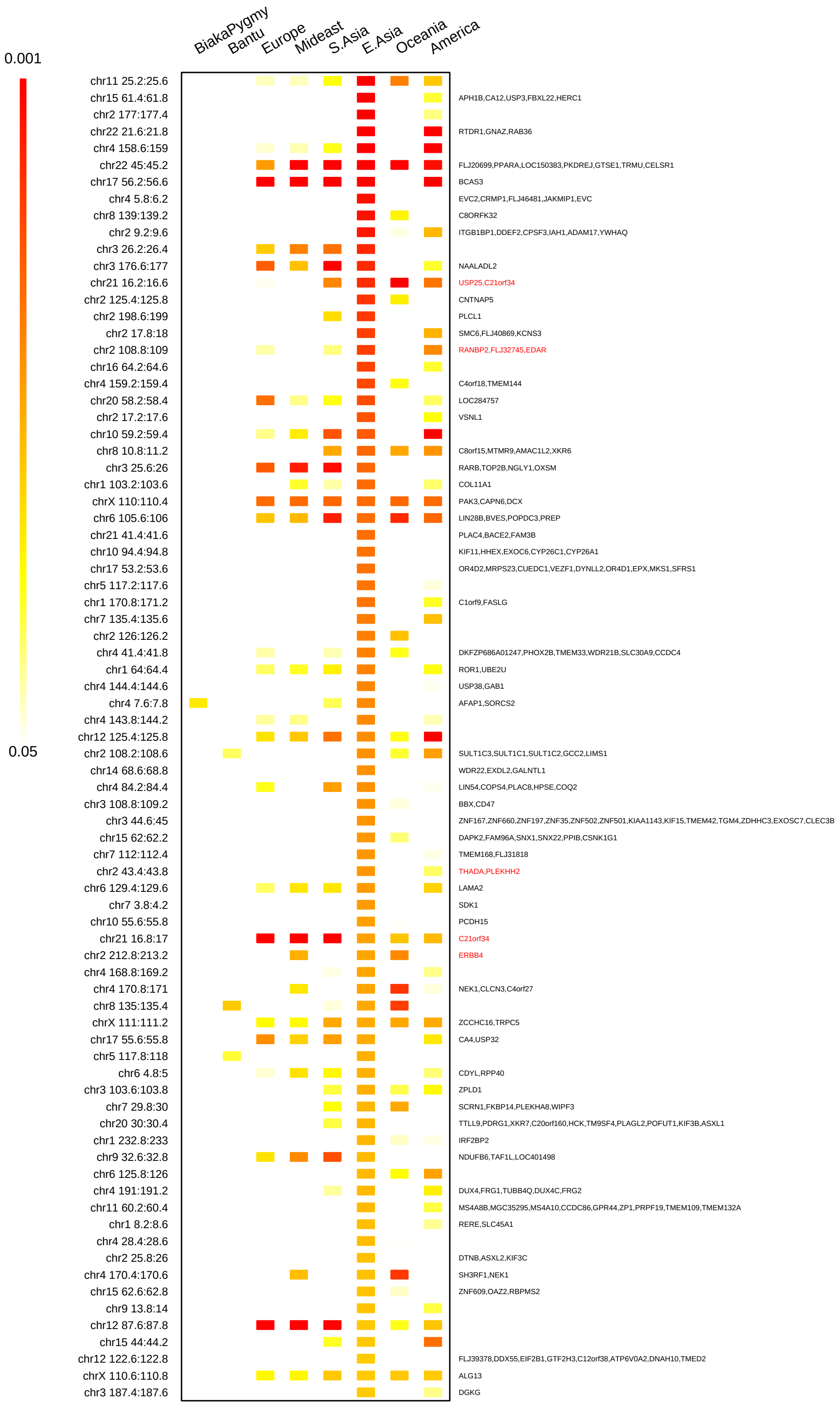

Figure 26: The top 1% of XP-EHH signals in East Asia, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

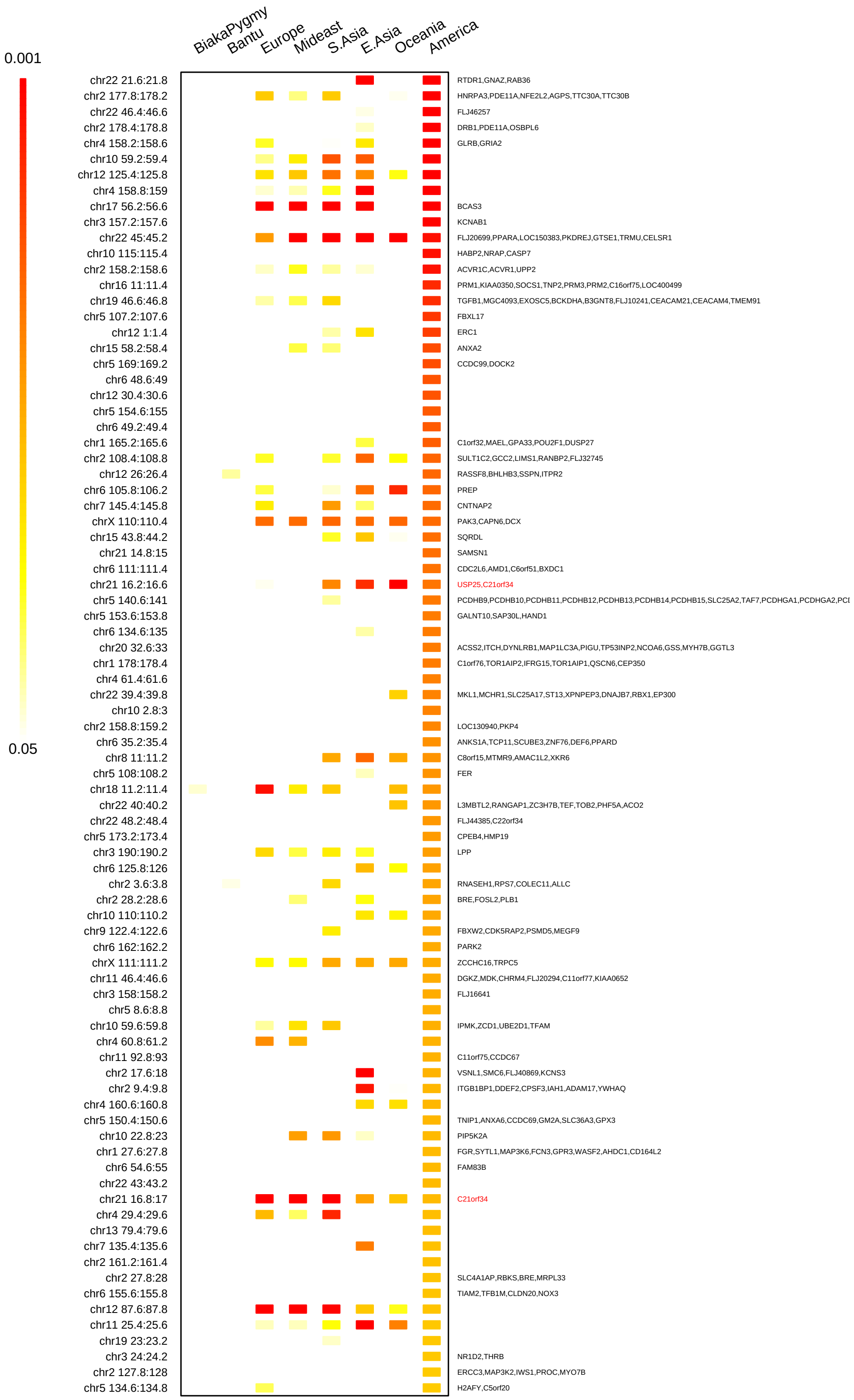

Figure 27: The top 1% of XP-EHH signals in the Americas, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

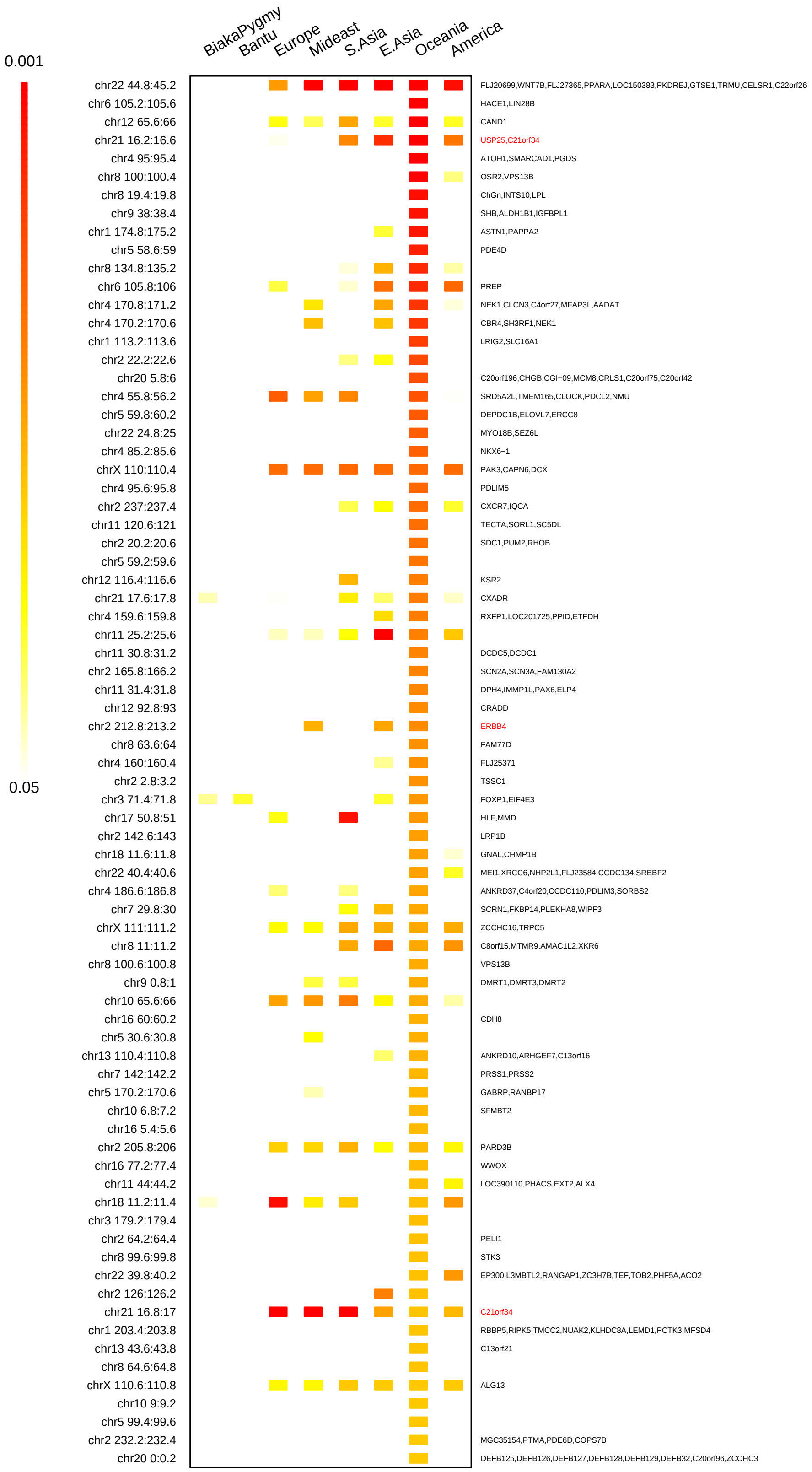

Figure 28: The top 1% of XP-EHH signals in Oceania, ordered from top to bottom in order of significance. See the caption of Figure 1 in the main text for details.

References

[1] Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15: 1496–1502.

[2] Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.

[3] Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15: 1576–1583.

[4] Tang K, Thornton K, Stoneking M (2007) A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. PLoS Biol 5: e171.

[5] Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? Genome Res 16: 702–712.

[6] Thornton KR, Jensen JD (2007) Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175: 737–750.

[7] Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.